Compositionality and Cognitive Control in Neural Networks

By

JACOB RUSSIN

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Randall C. O'Reilly, Chair

Charan Ranganath

Erie Boorman

Committee in Charge

2023

ABSTRACT

Compositionality and Cognitive Control in Neural Networks

JACOB RUSSIN

Compositionality, a natural property of symbolic systems, is thought to be a key principle underlying human intelligence: known concepts can be combined in novel ways according to systematic rules, allowing for the potentially infinite expressivity of human thought and language. Neural network models of cognition have long been criticized for failing to capture this important property. Despite their massive success in cognitive domains such as natural language processing, modern deep neural networks still struggle to generalize on compositional problems in the same ways that humans do, leading some to conclude that these networks must be augmented with symbolic or rule-like operations to fully account for key aspects of human cognition. Others have attempted to discover the inductive biases that would encourage compositionality to emerge in neural networks, without the need for explicit symbols or rules.

The work presented in this dissertation takes the latter approach, exploring in particular the possibility that the mechanisms in the human brain responsible for cognitive control and top-down attentional modulation may constitute just such an inductive bias. Deep neural networks are used to study compositionality, cognitive control, and their relationship in both machine learning and cognitive neuroscience settings. Methods for measuring the extent to which compositional processing has emerged in neural networks are developed, and cognition-inspired attentional mechanisms are tested on compositional generalization problems. Additionally, neural networks are used to model phenomena observed with fMRI regarding control processes in the context of cognitive map formation.

Copyright

Acknowledgements

Note: Manuscript-specific acknowledgements are found at the end of each corresponding chapter.

Individuals and Groups

First and foremost I would like to thank my advisor, Randall O'Reilly, for his guidance and mentorship throughout my time in graduate school, and for giving me the resources and freedom to pursue the ideas that interest me most. I would also like to thank the other members of my committee, Erie Boorman and Charan Ranganath, as well as the other members of the Computational Cognitive Neuroscience Lab (Maryam Zolfaghar, Seth Herd, Thomas Hazy, John Rohrlich, Kai Kruger, Ananta Nair, Kevin McKee, Will Chapman, Andrew Carlson, Riley DeHaan, April Luo), the other members of the Learning and Decision Making Lab including Seongmin Park, the members of the Analogy Group including Jonathan Cohen, Alexander Petrov, Tim Buschman, Taylor Webb, Steven Frankland, and Randy Gobbel, the members of the Memory Meeting at UC Davis, members of the Cognition in Context Lab including Yuko Munakata, Jesse Niebaum and Winnie Zhuang, and Paul Smolensky, Roland Fernandez, Brenden Lake, Yoshua Bengio, Jason Jo, Matt Jones, Cristopher Summerfield, Sebastian Musslick, Roman Feiman, Ellie Pavlick, and Sam McGrath. All have helped me develop the ideas in this dissertation through useful discussions, as well as offered me additional advice, mentorship, and/or friendship throughout graduate school.

Funding

The work in this dissertation was supported in part by: the NIMH award number T32MH112507, the John Templeton Foundation (Duke SSNAP subaward 383-001202), ONR N00014-19-1-2684 / N00014-18-1-2116, ONR N00014-14-1-0670 / N00014-16-1-2128, and ONR N00014-18-C-2067.

Table of Contents

Dedication

To my parents. Thank you for all of your support.

Chapter 1

Introduction

Deep neural networks, direct descendants of the connectionist models explored in the Parallel Distributed Processing framework (McClelland et al., 1986), have revolutionized the field of artificial intelligence (Brown et al., 2020; LeCun et al., 2015; Silver et al., 2016). These models, like their predecessors, are based on some of the principles thought to underlie the functioning of the brain, e.g., they assume that representations are distributed across populations of neurons, each of which computes a linear function of its inputs followed by a nonlinearity. Although there is much that is biologically implausible about them including the backpropagation algorithm by which they are trained (O'Reilly, 1996; Rumelhart et al., 1986a; Whittington & Bogacz, 2019), the basic principles upon which they are built, along with their success in cognitive domains such as vision, language, and reinforcement learning, have led many researchers to suggest that modern neural networks can also support a mini-revolution in cognitive and computational neuroscience (Botvinick et al., 2019a; Hassabis et al., 2017; Kietzmann et al., 2018; Marblestone et al., 2016; Richards et al., 2019; Saxe et al., 2021).

Part of this optimism stems from the fact that these networks, like the human brain, are trained on increasingly large, high-dimensional datasets of images and video, speech, and natural language data, allowing them to be compared with humans on tasks with a more realistic scale compared to previous models, which were much smaller and trained on much simpler tasks. This scale also allows for fairer comparisons to be made because these networks can be pretrained on large datasets and then instructed or prompted to perform a specific task (Webson et al., 2023), analogous to how humans have a lot of prior experience before being instructed on a particular task given in a psychological or neuroimaging experiment. This is again opposed to earlier work which usually unrealistically assumed (for practical purposes) that all learning took place exclusively within the particular task being studied. Furthermore, the high-dimensional representations that emerge in these networks through learning can be systematically tested against brain data with methods such as Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008) — see e.g., the work presented in chapters 7 and 8.

There are also reasons to be skeptical about this kind of work: although it does have roots in cognitive science, many recent advances in deep learning have not been motivated by understanding cognition but by engineering purposes. Furthermore, the rapid success of these models – especially in popular applications (e.g., chatGPT) – have led some to make extravagant claims about their cognitive capabilities or psychological plausibility.

This state of affairs motivates the need for cognitively-focused work that aims to leverage the advantages of modern neural networks to gain insights into the functioning of the human brain. This is the broad aim of much of the work of this dissertation. The remainder of this introductory section is organized as follows: Section 1.1 briefly summarizes key historical debates that began with early work in connectionism, largely focusing on the debates about compositionality and rule learning. Section 1.2 reviews recent work that revisits these debates in light of advances in neural networks for machine learning. Section 1.3 outlines the middle-ground approach to these issues that I have taken in my own work. Section 1.4 connects the two major themes of my work – compositionality and cognitive control – and situates them in the context of this middle-ground approach, and section 1.5 summarizes each of the chapters, elaborating on their place in this broader research program.

1.1 The Connectionism Debates

In this section I briefly summarize key challenges raised by critics of connectionist models during their rise in the 80s and 90s. A full review of this literature is beyond scope of the current chapter (Buckner & Garson, 2019; Medler, 1998), so I will focus on the challenges most relevant to my own work.

The connectionist models of the 80s and 90s (McClelland et al., 1986) garnered a lot of attention because they seemed to offer solutions to well-known problems with classical symbolic models that were the state-of-the-art in artificial intelligence and cognitive science at the time (Hofstadter, 1982) – e.g., robustness to uncertainty, graceful degradation in response to noise, etc. Connectionists emphasized the advantages of distributed representations that can be learned from data and do not require any built-in rules or hand-crafted programs, hypothesizing that rule-like or symbolic behavior can emerge through learning. For example, Rumelhart & McClelland (1986) famously trained a neural network model on the English past tense, and argued that the model challenged the standard assumption that knowledge of the English past tense took the form of explicit rules, thus sparking intense debate (Pinker & Prince, 1988; Seidenberg & Plaut, 2014). Some connectionists concluded from this and other findings that high-level psychological phenomena normally characterized in terms of discrete rules or symbolic operations were in fact epiphenomenal: these characterizations were best thought of as approximations that would not appear in the final, more fun-

damental account of behavior. This more radical position is sometimes labeled *eliminative connectionism* (Marcus, 1998; Place, 1992).

Many cognitive scientists pushed back against these claims and argued that neural networks lacked the very properties that made classical symbolic models so successful in the first place. The specific claims about the English past tense were hotly contested (Pinker & Prince, 1988; Seidenberg & Plaut, 2014), and some theorists challenged the connectionist approach to cognitive architecture as a whole. Perhaps the most influential of these were Fodor & Pylyshyn (1988) who argued that neural networks could never explain human productivity and systematicity because they fundamentally lacked the compositionality inherent to the classical approach. Compositionality, or the ability to compose familiar elements into novel combinations according to known rules (Szabó, 2020), is a natural property of symbolic systems and readily explains the apparent productivity and systematicity of human thought and language — when known elements can be combined into novel concepts, the number of complex combinations that are expressible grows exponentially in the number of elements that are learned.

Marcus (1998) argued along similar lines that neural networks, which are universal function approximators, can only generalize by interpolating within their training distribution, and are incapable of human-like extrapolation, i.e., generalization outside of this distribution. For example, upon learning the basic definition of a novel verb like "dax," humans are capable of systematically generalizing its usage to many different constructions (e.g., "dax twice") — thus extrapolating from their direct experience to a truly novel situation (Lake et al., 2015; Lake & Baroni, 2018; Lake et al., 2019a)

Marcus (1998) argued that neural networks could not achieve this kind of out-of-distribution generalization because they are not endowed with symbolic structures and rule-like operations. Contrary to radical eliminative connectionism, these theorists took a strong anti-reductionist position, in some cases arguing that neural networks could not in principle be relevant to higher-level cognitive theorizing: even if all of the relevant properties of human cognition did somehow emerge in a neural network, nothing would be learned at the cognitive level (Fodor & Pylyshyn, 1988).

Many connectionists took these challenges seriously, and attempted to show how the apparent compositionality of human thought and language could be accounted for in purely neural models, i.e., ones that were not augmented with specialized symbolic or rule-like operations (O'Reilly et al., 2014; Smolensky, 1987, 1988). However, further debate about compositionality ensued (Chalmers, 1990; Fodor, 1997; Fodor & McLaughlin, 1990; MacDonald & MacDonald, 1991; McLaughlin, 1993, 2014), and the field as a whole did not come to a clear resolution on these issues.

1.2 Revisiting the Debates in Light of Advances in Deep Learning

Despite these older criticisms, many of the most successful deep learning models share the core assumptions of connectionism, including an emphasis on the power of distributed representations learned from data, which do not require domain knowledge to be built in (Bengio et al., 2014; LeCun et al., 2015). A major aspect of the design philosophy of deep learning systems is a humility about the extent to which the researcher or engineer can come up with the "right" operations or features on which the system should operate. Neural networks are instead trained on raw inputs such as images or text, and the optimization procedure is trusted to guide the emergence of representations that are useful enough for the system to achieve its objective.

Indeed, one way of characterizing the history of artificial intelligence and machine learning is as a progression from intensive top-down programming of entire systems from scratch (Newell & Simon, 1956), to merely hand-designing the features on which a learning system is trained (Schölkopf & Smola, 2002), to hard-coding as little as possible other than the learning systems themselves (e.g., the architecture, objective function, optimization algorithm, etc.), allowing effective representations at multiple levels of a deep hierarchy to emerge through training on raw inputs alone (Goodfellow et al., 2016). In retrospect, it may not be so surprising that a machine trained on large amounts of data could discover better features than human engineers — consider, for example, how challenging it would be to design by hand the features that would allow a program to recognize whether or not a given image contains a certain object such as a cat.

Although deep learning systems are often criticized for their lack of interpretability (Murdoch et al., 2019), this opacity may stem from the same principle: it may be precisely because these systems are not constrained to produce clean, succinctly characterizable features (i.e., those that a human engineer or cognitive scientist could readily understand) that they are free to learn the strange, complicated, heterogeneous features (Serre, 2019) that allow them to reach superior performance (Lipton, 2017).

The success of this design philosophy in artificial intelligence applications has been taken by some to at least partially vindicate the principles emphasized by the early connectionists (Piantadosi, 2023), and has therefore reinvigorated many of the older debates in cognitive science. For example, some recent work has revisited the challenges posed by Pinker & Prince (1988) concerning the ability of neural networks to learn discrete rules such as those thought to govern the English past tense. Kirov & Cotterell (2018), as well as follow up work (Corkery et al., 2019; Wiemerslage et al., 2022) showed that modern natural language processing (NLP) systems avert prior empirical criticisms, and learn to conjugate both regular and irregular verbs from their training set, achieving high accuracy (98%-99%) on held-out regular verbs — without any built-in capacity to represent explicit rules.

This work is part of a growing body of research investigating the extent to which modern neural networks

trained on large amounts of natural language data capture the structures posited by classical linguistic theory (Baroni, 2021; Linzen & Baroni, 2021; Pavlick, 2022). For example, state-of-the-art language models (Brown et al., 2020) demonstrate surprisingly extensive syntactic abilities (Linzen & Baroni, 2021) and are capable of producing long strings of coherent text where grammatical errors are rare. A growing body of research has scrutinized the syntactic abilities of deep neural networks by developing test datasets designed to isolate competence with specific syntactic phenomena (Gulordava et al., 2018; Linzen et al., 2016; McCoy et al., 2020; Wilcox et al., 2022). While some findings have been mixed (McCoy et al., 2019), taken together they suggest that syntactic competence in these models is robust (Linzen & Baroni, 2021), challenging prior assumptions (Piantadosi, 2023; Wilcox et al., 2022). Furthermore, the internal representations of the models align to some extent with the structures posited by classical linguistic theory (Manning et al., 2020; Tenney et al., 2019), suggesting that although these representations are not symbolic or rule-like in the classical sense, the relevant structures may nonetheless emerge through learning.

The representations learned by deep neural networks for vision (Cadieu et al., 2014; Lee & DiCarlo, 2023; Nayebi et al., 2021a; O'Reilly et al., 2021b; Yamins et al., 2014), navigation (Nayebi et al., 2021b; Whittington et al., 2020) and language (Caucheteux & King, 2022; Goldstein et al., 2022; Luckett et al., 2020; Mahowald et al., 2023) have also been shown to have a surprising level of correspondence with those of the brain, suggesting that they can inform our understanding of the processes by which it operates in these settings.

Although deep neural networks have overcome some of the criticisms associated with earlier work in connectionism, other recent studies suggest that the old challenges remain relevant. Many studies have found that although some kinds of compositionality have been shown to emerge in them (Lepori et al., 2023; Lewis et al., 2023; Russin et al., 2021a), modern neural networks still struggle to generalize in the ways one would expect of a truly compositional system (Hupkes et al., 2020; Kim & Linzen, 2020; Lake & Baroni, 2018; Lake et al., 2017; Loula et al., 2018; Marcus, 2018). For example, in one influential study that has inspired some of my own work (Russin et al., 2019), Lake & Baroni (2018) found that even state-of-the-art neural networks struggle to systematically compose known elements or apply known rules in unfamiliar situations, even though humans succeed when tested on an analogous task (Lake et al., 2019a). In particular, Lake & Baroni (2018) tested models on their ability to generalize the usage of a known verb (e.g., "dax") to unseen constructions (e.g., "dax twice"), and showed that neural networks were unsuccessful in doing so. Other follow up work has shown that the failure of neural networks on compositional generalization problems such as this one is a robust phenomenon and reappears in many domains (Bahdanau et al., 2019b,a; Lake et al., 2017). This is consistent with the data inefficiency of large neural networks, which are sometimes trained on more data than humans will encounter in an entire lifetime (e.g., GPT-3 (Brown et al., 2020) trained on

something like 100 human lifetime's worth of linguistic data (Wilcox et al., 2022)).

These and other findings suggest that modern neural networks are still incapable of the kind of sample efficiency and extrapolation/out-of-distribution generalization that the inherent compositionality of symbolic or rule-based systems affords (Marcus, 2018, 2020; Marcus & Davis, 2020). This suggests that even in the cases where large language models seem to possess powerful generalization abilities, they are still merely interpolating — their training set is so large that it is difficult to even come up with out-of-distribution tests. Despite current disagreements about how to resolve these issues, the findings exploring the successes and failures of deep neural networks have demonstrated the importance of the cognitive perspective for artificial intelligence, and suggests an important role for these networks in future cognitive theorizing.

1.3 Revisionary Connectionism: A Middle Ground

It is unclear how the current iterations of these debates will be resolved. Some in the deep learning community seem to have gravitated toward an extreme position similar to the eliminative connectionism that critics accused early connectionists of, and believe that any cognitive capacity can emerge if a large enough model is trained on a large enough dataset. On the other hand, prominent critics argue that regardless of scale, neural networks have inherent limitations that they will never overcome and will need to be augmented with symbolic or rule-like processes in order to fully account for key aspects of human cognition (Marcus, 2018; Quilty-Dunn et al., 2022).

It is important here to distinguish between the goals of research in artificial intelligence (AI), which seeks to engineer machines that are capable of intelligent behavior, and the goals of cognitive science, which seeks to explain how the human mind/brain accomplishes such behaviors: there is in principle a dissociation between the question of what will work for AI in the future, and the principles underlying human cognition. For example, it might be the case that many aspects of human cognition are in fact best characterized by symbolic processes, but we won't need to build explicit symbolic operations into large neural networks because scale can compensate for such limitations. On the other hand, it could be that human symbolic behavior is merely emergent, but that the best way forward for AI is to use hybrid architectures that incorporate symbolic components — after all, symbolic programs are much better than humans at certain kinds of computation (e.g., long division). So although there is good reason to think that there will be convergence between the two fields (Hassabis et al., 2017; Marblestone et al., 2016; Zador et al., 2023), in principle the difference in their goals highlights the fact that cognitive science cannot naively treat the volatile "state-of-the-art" in AI as the current best model of the mind/brain.

In my own work, I have sought to both bring insights from cognitive science and neuroscience to bear

on machine learning models (see chapters 2, 3, 4, 5), and to use neural network models to gain insight into the brain (see chapters 6, 7, 8). Inspired by previous work (Elman et al., 1997; O'Reilly et al., 2014; Smolensky, 1990), I have sought a middle ground between the following two extreme positions: 1) an eliminative connectionism that disregards cognitive accounts of the inner workings of the mind, emphasizing instead the emergent quality of intelligent behaviors, and 2) the extreme kind of anti-reductionism that maintains that neural networks cannot inform cognitive theorizing, even in principle (Fodor & Pylyshyn, 1988).

On the one hand, a growing body of work from both earlier connectionist research and current deep learning research shows that neural networks can offer insights into the processes underlying human cognition and neural computation. Modern neural networks have shown time and again that it is easy to underestimate emergence — many phenomena that were thought to require specific representational or architectural assumptions have been shown to emerge in relatively generic neural networks at larger scales. However, it is also clear that cognitive scientists need to be cautious in drawing cognitive conclusions from models trained in wildly inhuman ways or at implausible scales, and need to think explicitly about how the brain may embody specific inductive biases sculpted by evolution to encourage symbol-like or rule-like phenomena to emerge in human behavior. The criticisms in the work of (Lake & Baroni, 2018; Marcus, 2018) contain important kernels of truth that can inform ongoing work seeking to identify the inductive biases that are missing from current systems. These inductive biases may be architectural (Elman et al., 1997; O'Reilly et al., 2021a; Russin et al., 2019; Smolensky, 1990), but could also relate to the lack of human-like objective functions or multimodal and interactive environments (Hill et al., 2020; Linzen, 2020; McClelland et al., 2020).

The important task for research at the intersection of cognitive science, neuroscience, and AI is to translate the kinds of insights from work in cognitive science, e.g., about the importance of compositionality, into inductive biases that can be implemented in purely neural systems, i.e., without hand-coding rigid symbolic structures that are biologically implausible and suffer from the same limitations present in classical symbol programs. Note that this strategy does not seek "mere implementations" (Fodor & Pylyshyn, 1988) of current cognitive theory assumed to be rigid or final. It embodies a substantial shift in perspective: rather than positing innate concepts or rule-like structures to explain behavioral phenomena, it seeks understanding of the inductive biases that would lead those phenomena to emerge, and contends that this empirical investigation may revise current cognitive theory and offer further insights into the development and functionality of the brain. This strategy is a middle ground because it rejects the notion that neural networks will never fully account for the complete set of human competences without explicit symbols or rules, and is "revisionary" rather than "eliminative" in that it rejects the assumption that high-level cognitive explanations are merely

epiphenomenal — it welcomes these explanations and hopes to constructively revise them through the process of translating them into the language of neural networks.

1.4 Cognitive Control as an Inductive Bias to Encourage Compositionality

To demonstrate how this middle-ground approach has operated in the context of my own work, I will briefly describe my approach to compositionality, and its relation to cognitive control. Although much progress has been made in showing that neural networks can succeed on tasks thought to require rule-like or systematic representations such as natural language processing (Brown et al., 2020), the problem of compositionality still seems to plague modern neural networks (Lake & Baroni, 2018; Lake et al., 2017). The problem has received much attention from the deep learning community (Hupkes et al., 2020; Jiang & Bansal, 2021; Keysers et al., 2020; Kim & Linzen, 2020; Lake, 2019; Li et al., 2019; Liu et al., 2020; Loula et al., 2018), but there is still no consensus on how it can be solved in a purely neural learning system (i.e., one that is not augmented with explicit symbols or rules).

It is for this reason a suitable test bed for the middle-ground approach outlined in the previous section. Indeed, it has been the motivation for similar middle-ground approaches in the past (O'Reilly et al., 2014; Smolensky, 1990). For this approach to succeed, the concept of compositionality, which is deeply entangled with the language of symbolic programs, must be deconstructed into more basic principles suitable for translation into an inductive bias that can be readily implemented in neural networks.

One way of doing so has been to strip away assumptions about how precise mechanisms by which compositionality operates in the context of symbolic programs, and focus instead on the more basic notion of *content-independence* or role-filler independence (Quilty-Dunn et al., 2022; Smolensky, 1990) The notion of compositionality captures the fact that symbolic programs apply the same exact operations (or structure) to any content (e.g., consider how a function written in a programming language applies the same operations to its arguments, regardless of the particular values those arguments take). Typical neural networks, however, do not embody this principle, and exhibit an extreme sensitivity to contextual information, leading to failures on the compositional generalization problems discussed earlier.

Following others (Behrens et al., 2018; Kriete et al., 2013; O'Reilly et al., 2014; Rougier et al., 2005; Summerfield et al., 2020), my work has attempted to translate this abstract property into an inductive bias suitable for implementation in neural networks (O'Reilly et al., 2021a; Russin et al., 2019, 2020b). By learning structure and content separately, an out-of-distribution problem where a model must *extrapolate* to unseen structure-content combinations can be transformed into two *interpolation* problems — a known structure must be recognized, and applied to known content (see Figure 4.3).

A brain area thought to be involved in learning abstract, content-independent structure and rule-like representations is the prefrontal cortex (Mian et al., 2014; Milner, 1963; Wallis et al., 2001; Shallice & Burgess, 1991). Its connection to compositionality, out-of-distribution generalization, and ongoing problems with current deep neural networks is outlined in chapter 2 (Russin et al., 2020b). The prefrontal cortex is known to be involved in high-level cognitive functions that are currently not captured well by deep learning models: e.g., planning (Daw et al., 2005; Smittenaar et al., 2013), reasoning (Crescentini et al., 2011; Donoso et al., 2014; Goel, 2007; Hampshire et al., 2011; Krawczyk et al., 2011), and control in novel environments (Domenech & Koechlin, 2015; Miller & Cohen, 2001). The prefrontal cortex is thought to accomplish these functions through top-down attentional modulation of other areas (Miller & Cohen, 2001; O'Reilly & Frank, 2006).

In my work I have explored how top-down control may be an important architectural inductive bias implemented in the brain that encourages content-independent processing (Russin et al., 2019). This fits with the middle ground position outlined above: rather than hand-coding symbolic operations into the brain, evolution may have encouraged compositionality to emerge through the implementation of a mechanism for cognitive control, specifically designed to overcome habitual responses in the service of accomplishing novel goals (Miller & Cohen, 2001). Although there is still work to be done in both discerning the precise computations underlying cognitive control in the prefrontal cortex and in implementing analogous mechanisms in neural networks at scale, the results of the work outlined in these chapters suggest that this middle-ground approach is promising, and can inform cognitive science, neuroscience and AI.

1.5 Chapter summaries

The work presented in the following chapters use neural networks to explore the computational mechanisms underlying compositionality, cognitive control, and their interaction. Each chapter is a self-contained paper that was published previously. The chapters can be broadly separated into two parts. The first part (chapters 2 through 5) explores compositionality in the context of deep learning research, using insights from neuroscience and cognitive science, including the use of a specialized attentional control mechanism to achieve content-independence processing in chapter 4. Chapter 2 (Russin et al., 2020b) is a short but intensive literature review that makes the case for the need for cognitive control in deep learning by highlighting the degree of alignment between the functions of the prefrontal cortex and current limitations of state-of-the-art deep neural networks. Chapter 3 (Russin et al., 2021a) explores the possibility that compositional processing emerges in transformers trained on a large dataset of math word problems, and finds evidence for a weak kind of emergent compositionality, suggesting that compositionality can to some extent be learned in the absence

of strong inductive biases. However, this weak form of compositionality does not allow these networks to generalize sufficiently outside of their training distribution: chapter 4 (Russin et al., 2020a) explores top-down attentional control as a strong inductive bias for encouraging compositionality in recurrent neural networks, and finds that when such networks are equipped with this mechanism, content-independence emerges and compositional generalization is greatly improved. Chapter 5 (Chakravarthy et al., 2022)[1] explores a similar mechanism to induce content-independence in a transformer architecture (Vaswani et al., 2017) on the same task, and shows that similar compositional generalization performance can be achieved when abstract role information is provided as input.

The second part (chapters 6 through 8) presents work done in collaboration with Erie Boorman's fMRI lab at UC Davis, and explores cognitive control in neural network models in the context of cognitive maps (Behrens et al., 2018; O'Keefe & Nadel, 1978; Park et al., 2020a), using these models to gain insight into phenomena observed with fMRI. Chapter 6 (Russin et al., 2021b) describes preliminary work exploring whether neural networks can capture the basic phenomena observed in a particular fMRI experiment investigating cognitive map formation in humans (Park et al., 2020b). Chapter 7 (Zolfaghar et al., 2022)[1] investigates direct links between control processes and the geometry of cognitive maps learned by neural networks, and chapter 8 (Russin et al., 2022) extends this work with a neural network model equipped with an explicit control mechanism, showing that it can account for empirical findings related to blocked vs. interleaved learning (Flesch et al., 2018).

[1] Co-first authors on original publication.

Chapter 2

Deep Learning Needs a Prefrontal Cortex

Jacob Russin[1], Randall C. O'Reilly[1], Yoshua Bengio[2]

The original version of this article (Russin et al., 2020b) was accepted for publication at the "Bridging AI and Cognitive Science Workshop" at the International Conference on Learning Representations (ICLR 2020). The opinions expressed here are the author's own and do not necessarily reflect the views of the conference, workshop, or publisher. The original version is available online at `https://baicsworkshop.github.io/pdf/BAICS_10.pdf`.

2.1 Abstract

Research seeking to build artificial systems capable of reproducing elements of human intelligence may benefit from a deeper consideration of the architecture and learning mechanisms of the human brain. In this brief review, we note a connection between many current challenges facing artificial intelligence and the functions of a particular brain area — the prefrontal cortex (PFC). This brain area is known to be involved in executive functions such as reasoning, rule-learning, deliberate or controlled processing, and abstract planning. Motivated by the hypothesis that these functions provide a form of out-of-distribution robustness currently not available in state-of-the-art AI systems, we elaborate on this connection and highlight some computational principles thought to be at work in PFC, with the goal of enhancing the synergy between neuroscience and machine learning.

2.2 Introduction

Deep learning, which has historically taken inspiration from the brain, has had unexpected and massive success in many applications. Though much progress has been made, new advances will be needed to meet the substantial challenges remaining on the path toward recreating the most powerful aspects of human intelligence. State-of-the-art methods remain inferior to human learners in their ability to transfer knowledge to new domains (Lake et al., 2017), to capture compositional or systematic structure (Lake & Baroni, 2018),

[1] Center for Neuroscience, University of California, Davis
[2] MILA, Université de Montréal, CIFAR Senior Fellow

to plan efficiently (Hamrick, 2019), and to reason abstractly (Bhagavatula et al., 2019; Xu et al., 2020). All of these abilities share similarities with the collection of human capacities known as executive functions, and are often associated with conscious processing, i.e., they can be reported verbally by human subjects. The human brain must embody principles lacking in current deep learning systems that allow it to perform these powerful functions. This has led some to consider the possibility of taking inspiration from the architecture of the human brain to build more flexible learning systems (Marblestone et al., 2016; Hassabis et al., 2017). Here, we observe an intriguing correspondence between some of the current open questions in deep learning research and the functions of the human prefrontal cortex (PFC), a brain area known to be involved in executive functions such as planning (Duncan, 1986), abstract reasoning (Donoso et al., 2014), rule-learning (Wallis et al., 2001), and controlled or deliberate processing (Miller & Cohen, 2001). We explore this connection, and the potential of translating what is known about this brain area into architectural assumptions or inductive biases in deep learning (Marblestone et al., 2016; Battaglia et al., 2018). First, we elaborate on some of the current challenges in deep learning research mentioned above, and then briefly survey some findings from neuroscience about the PFC, noting connections to these current challenges. We then discuss some theoretical ideas about PFC function from cognitive and computational neuroscience, with the aim of stimulating a fruitful synergy between neurocience and deep learning research.

2.3 The Need for Neural Networks with Executive Functions

Current deep learning methods excel in perceptual tasks in which complicated patterns must be recognized in high-dimensional data. However, no one yet knows how to build learning machines which fare well on tasks that require deliberate, controlled processing over multiple steps or dealing with changes in distribution (Bengio, 2017, 2019; Lake et al., 2017; Marcus, 2018). In the following, we highlight some of the aspects of human cognition that have so far proven difficult for neural networks to reproduce, and have become active areas of research in deep learning.

Reasoning Bottou (2011) offers a helpful working definition of reasoning as "algebraically manipulating previously acquired knowledge in order to answer a new question." What this definition entails is the reuse of dynamically selected computational modules, with the results of recently produced computations feeding the currently selected computation. Most of the tasks at the center of the rise of deep learning (e.g., object recognition, video-game playing, machine translation) generally do not require reasoning, i.e., algebraic manipulation of existing knowledge. Current neural nets involve composition of functions (e.g. layers) but in a fixed order. Recently, there has been growing interest and much progress on datasets and tasks that require reasoning over multiple steps (e.g. Bhagavatula et al., 2019; Johnson et al., 2017; Graves et al., 2016;

Hudson & Manning, 2019; Weston et al., 2015; Xu et al., 2020; Barrett et al., 2018), with some cases of systems surpassing human performance (e.g. Hudson & Manning, 2018; Perez et al., 2018; Yi et al., 2018). However, as we discuss in the next section, most models trained on these tasks still fail to "answer new questions", in the sense of generalizing outside of the training distribution (Barrett et al., 2018; Bahdanau et al., 2019a).

Compositionality and Systematicity It has been argued that one of the most powerful aspects of human cognition is its systematicity: concepts can be composed in novel ways, so that the number of expressible combinations grows exponentially in the number of primitive concepts learned (Fodor & Pylyshyn, 1988; Lake et al., 2017; Lake & Baroni, 2018). This topic is closely related to reasoning, because "algebraic manipulation" requires that existing knowledge be represented in a form that is systematically composable. Interest in compositionality among deep learning researchers has grown over the past few years, where experiments have revealed that standard approaches in deep learning show weak generalization for compositions of known elements which are unlikely under the training distribution (Lake & Baroni, 2018; Bahdanau et al., 2019b,a; Keysers et al., 2020, but see also Hill et al. 2020). These experiments show that standard architectures fail to capture the compositional structure or systematic rules governing the data-generation process.

Control in Novel Environments Just as standard deep networks have weaker generalization outside of their training distribution in the settings described above, they are less efficient than humans at transferring knowledge about learned environments to novel ones (Hagendorff & Wezel, 2019; Kansky et al., 2017; Lake et al., 2019b, 2017). For example, when trained on the Atari games, the generalization of standard methods in deep reinforcement learning is not robust to slight changes in the rules of the game or the layout of the inputs (Kansky et al., 2017). Generalization to novel environments has continued to be an important topic in deep learning research, where an increased focus on one-shot learning (e.g. Vinyals et al., 2017), transfer (Weiss et al., 2016), and meta-learning (e.g. Finn et al., 2017; Bengio et al., 2019) has emerged. Much progress has been made in these areas, but human-level transfer remains elusive (Lansdell & Kording, 2019; Griffiths et al., 2019).

Abstract Planning It has long been recognized that the standard planning algorithms used in model-based reinforcement learning (RL) are too computationally expensive to be useful in many real-world domains (Barto & Mahadevan, 2003), and that humans and other animals seem to possess planning strategies that avoid much of this computational cost (Botvinick, 2008; Botvinick et al., 2009). In particular, it has been suggested that humans plan using temporally abstract representations, whereas model-based algorithms usually treat each time-step independently (Botvinick et al., 2009; Botvinick & Weinstein, 2014). The most successful algorithms in deep RL are model-free (Arulkumaran et al., 2017; Hamrick, 2019), and

though model-based deep RL methods have had some recent success (Corneil et al., 2018; Finn & Levine, 2017; Nagabandi et al., 2018; Feinberg et al., 2018), most still plan each time step individually or lack the abstraction and compositionality displayed in human planning (Hamrick, 2019).

2.4 Some Functions of the Prefrontal Cortex

All of the challenges described above have been noted by others, and are active areas of research. The first of our main contributions is to draw connections between them and the functioning of the human PFC. The PFC comprises a large swath of the most anterior portion of the cerebral cortex and appears to have undergone a disproportionate amount of development over the course of human evolution (Schoenemann et al., 2005; Rilling, 2006; Semendeferi et al., 2001; Falk, 2012). It receives highly processed, multimodal information from perceptual areas, and sits at the top of the decision-making hierarchy (Fuster, 2009; Hunt & Hayden, 2017; O'Reilly et al., 2012). Much remains unknown about the PFC, and in particular there is ongoing investigation into functional differentiation between different areas within it (Hunt et al., 2018). However, it has been argued that much of the PFC retains a canonical computational role, with functional differentiation among subareas emerging due to differences in connectivity (Miller & Cohen, 2001; O'Reilly, 2010; Thompson-Schill, 2005). Here we highlight some aspects of the general functionality of the PFC.

Reasoning One of the most well-established findings about the PFC is that it is specialized for working memory, or the ability to maintain and manipulate information over short periods of time (Fuster & Alexander, 1971; Kubota & Niki, 1971; Miller & Desimone, 1994; Goldman-Rakic, 1995; Sommer & Wurtz, 2000; Lara & Wallis, 2015). Working memory can be seen as an important aspect of the capacity to reason, as it allows for 1) computation on information that is not currently observable in the environment, and 2) the integration of intermediate results in a larger reasoning process (e.g., in a serial summation of a list of numbers; Menon, 2016). Indeed, evidence of prefrontal engagement has been found in many experiments investigating the neural underpinnings of human reasoning (Donoso et al., 2014), including deductive (Goel, 2007), inductive (Crescentini et al., 2011), relational (Krawczyk et al., 2011), and analogical reasoning (Hampshire et al., 2011).

Representing Abstract Rules One domain in which humans excel at generalizing outside of the training distribution is the ability to apply known rules to novel elements (Lake et al., 2019a). The PFC has been found to be important for success on tasks that require the induction, maintenance, updating, or application of rules (Mian et al., 2014; Milner, 1963; Wallis et al., 2001; Shallice & Burgess, 1991). For example, patients with damage to PFC struggle to sort cards according to a changing rule (e.g., color or shape) (Milner, 1963; Buchsbaum et al., 2005; Berg, 1948). In a seminal electrophysiology study on rule

application (Wallis et al., 2001), monkeys were trained to either select the picture that was a 'match' to the previously presented one, or select the 'nonmatch'. Single neurons in PFC were found to respond when invoking such abstract rules, regardless of the particular pictures presented on a given trial (Wallis et al., 2001). Some computational models (Rougier et al., 2005) have attempted to capture this important property of the PFC, showing, e.g., how indirection might be implemented in a canonical PFC circuit (Kriete et al., 2013, see also Hayworth & Marblestone 2018 and Müller et al. 2016).

Control in Novel Environments Overwhelming evidence implicates the PFC in decision-making and control processes (Domenech & Koechlin, 2015; Miller & Cohen, 2001). However, it is in general not crucial for the execution of habitual responses that have been trained extensively (as would be the case, e.g., in a model that had played an Atari game for hundreds of hours) — rather, it is required for *overriding* these habitual responses in novel situations, with new rules or in the pursuit of a novel goal (Miller & Cohen, 2001; Botvinick & Cohen, 2014). This function, generally termed "cognitive control," is illustrated well in studies using the classic Stroop task (Stroop, 1935). In this task, participants are presented with color words (e.g., 'red', 'blue') written in colored ink, which may or may not match the words. Patients with damage to PFC perform reliably poorly on this task, which requires them to override habitual responses (reading text) according to the color-naming rule (Perret, 1974; Vendrell et al., 1995). In general, it is thought that the functioning of the PFC is crucially important when a novel goal is being pursued in a familiar environment where habits have become entrenched, or in novel environments when no such habits yet exist (Miller & Cohen, 2001).

Abstract Planning Humans and other mammals demonstrate evidence of both model-free and model-based RL (Momennejad et al., 2017; Daw et al., 2011, 2005), but the PFC has been implicated in model-based RL in particular (Daw et al., 2005; Smittenaar et al., 2013). Humans with damage to the PFC can exhibit deficits in routine behaviors that require planning and coordinating sequences of actions like cooking or making coffee (Miller & Cohen, 2001; Levine et al., 1998; Duncan, 1986; Shallice, 1982). Some have theorized that the planning processes in PFC are temporally abstract or hierarchical, as in, e.g., the options framework (Sutton et al., 1999; Botvinick, 2008; Botvinick et al., 2009; Botvinick & Weinstein, 2014; Frank & Badre, 2012). This idea accords well with experiments indicating that PFC represents actions at multiple timescales simultaneously (Hunt & Hayden, 2017; Botvinick et al., 2009; Sarafyazd & Jazayeri, 2019).

2.5 Computational Principles and Learning Mechanisms in PFC

The section above describing some of the functions of the PFC was structured to draw out their connection to current challenges facing deep learning. However, the structure of this section is somewhat arbitrary, as

all of these functions are related to one another. Here, we cover some theoretical ideas about the underlying computational mechanisms of PFC that can unify these various functions, with an eye toward principles that may be transferable to deep learning.

Top-down Attention and Modulation In an influential framework, Miller & Cohen (2001) argue that many of the cognitive capacities associated with the PFC, including reasoning, rule-learning, planning, and cognitive control, can be explained by its role in top-down attentional modulation of other brain areas. The PFC sends projections to much of neocortex, allowing it to modulate activity in other areas, possibly according to a current goal or in agreement with currently conscious contents. In the Stroop task, e.g., the PFC represents the instruction to name the colors rather than read the words, and modulates the activity of color features in higher-order visual areas of the brain to bias behavior toward naming them (Miller & Cohen, 2001).

Top-down attentional modulation has some analogues in deep learning research. The use of attention has become an increasingly popular approach in many tasks (e.g. Bahdanau et al., 2014; Xu et al., 2016; Hudson & Manning, 2018). One major difference with these mechanisms may be that PFC is thought to modulate activity through multiple brain areas at once, conditioned on the current goal. This kind of conditioning may be more similar to HyperNetworks (Ha et al., 2016), FiLM (Perez et al., 2018), where the mapping learned by a single feedforward network can be modulated with transformations at each layer, or RIM, which tries explicitly to model a top-down attentional modulation mechanism (Goyal et al., 2019).

Recurrence, Gating, and Seriality Recurrence is ubiquitous in the brain, but the PFC has a special role in maintaining information in working memory over longer timescales (Lara & Wallis, 2015). Work in computational neuroscience examining the detailed biological mechanisms that would allow PFC to accomplish this has emphasized its interaction with the basal ganglia and the importance of LSTM-like gating operations (O'Reilly & Frank, 2006). Although computations in the brain are massively parallelized, the amount of information that can be maintained in working memory at a given time is notoriously small (Petri et al., 2021; Feng et al., 2014; Oberauer & Kliegl, 2006). This means that seriality is also an important aspect of how the PFC operates: top-down attention must be applied serially over the course of a planning or reasoning process, and intermediate results must be integrated over time. However, the serial processing of a few elements at a time can also be an advantage, as it enables arbitrary sequences of complex computational processing at each of these steps to be combined to obtain more powerful and compositional computation. This may be an important factor for supporting Turing-machine like universal computation (Graves et al., 2014; Newell, 1990), and for generalizing outside of the training distribution (Bengio, 2017, 2019).

Learning: Dopamine and Reinforcement Recent proposals from deep learning researchers have encouraged neuroscientists to focus on the architectures, learning algorithms, and cost functions in the brain,

as opposed to the more traditional approach of characterizing the low-level biological mechanisms or tuning properties of neurons or neuronal populations (Marblestone et al., 2016; Richards et al., 2019). This approach has been emphasized in research in connectionism and parallel distributed processing for decades (Rumelhart et al., 1986b), but much remains unknown about the learning mechanisms and cost functions that might be at work in biological neurons (Richards et al., 2019). Reinforcement learning is thought to be especially important for learning in the PFC, which receives ample dopamine signals conveying reward prediction errors (O'Reilly & Frank, 2006), and is heavily involved in decision-making and planning (Rushworth & Behrens, 2008). A recent proposal shows that a number of empirical findings can be explained by a model in which the PFC implements a meta-reinforcement learning system, trained by dopamine to instantiate an RL procedure within the dynamics of its neural activity (Wang et al., 2018).

A PFC Module for Deep Learning? Much remains unknown, but an overall picture of the PFC that has emerged in cognitive and computational neuroscience is one where it selects, maintains, and manipulates learned representations in other areas of the brain through a serial process of top-down attentional modulation (Miller & Cohen, 2001; O'Reilly & Frank, 2006; Hazy et al., 2007). This serial processing may be tuned through reinforcement or meta-reinforcement learning and dopamine signals to optimize performance on tasks that require reasoning, rule-like representations, sequential and dynamic recombination of computations, cognitive control, or temporally abstract planning. This kind of system may be critical to ensuring flexibility in familiar environments and controlled decision-making in novel ones, and may allow for efficient planning on multiple timescales.

Many of the current major challenges facing deep learning research involve tasks that require an extended notion of generalization, not just to examples from the same distribution as the past observations, but also to out-of-distribution inputs (Lake & Baroni, 2018; Bahdanau et al., 2019b; Bengio, 2019). The ability to handle such non-stationarities would naturally evolve because learning agents (who change their policy and thus end up visiting different states of the environment) naturally face them, and even more so in a social multi-agent context where the environment itself changes. Some of the paradigmatic cases in which humans are able to do this involve the application of known rules to novel elements (Lake et al., 2019a) — a cognitive function that has been associated with the PFC (Miller & Cohen, 2001; Wallis et al., 2001). This systematicity is natural in symbolic systems typical of classical approaches to AI, but these lack many of the powerful advantages brought by deep learning (such as the ability to learn efficiently on a large scale, to handle uncertainty, to generalize well across symbols through distributed representations, and to ground these symbols in a complex perceptual reality). These symbolic systems utilize the notions of indirection or of variables — arrays of memory that can be manipulated by computations *that do not depend on the specific content stored there* — ensuring the kind of abstraction necessary for this kind of systematic generalization

to emerge. An analogous independence may exist between the PFC and posterior sensory and association areas: the PFC may be able to select and manipulate representational content in these areas according to learned rules that can be applied to many different elements (Russin et al., 2019; Kriete et al., 2013). This may provide the kind of abstraction and compositionality currently missing from standard architectures in deep learning (Bengio, 2017; Bengio et al., 2019).

2.6 Conclusion

We have argued that there is a striking correspondence between the tasks on which humans outperform current AI systems and the executive functions associated with the PFC. We believe that a greater focus on the principles and inductive biases at work in the PFC may inspire novel architectures that can accomplish similar functions. Much remains to be learned in making these principles more concrete and in implementing them in working systems, but we hope that we have taken a step in this direction and that this work will facilitate greater synergy between neuroscience and AI in the future.

2.7 Acknowledgments

We would like to thank the reviewers for their thorough comments and useful suggestions and references. We would also like to thank Jason Jo, and all of the members of the Computational Cognitive Neuroscience Lab at UC Davis for helpful ongoing discussion on these topics. This work was supported by ONR N00014-19-1-2684 / N00014-18-1-2116, ONR N00014-14-1-0670 / N00014-16-1-2128, and ONR N00014-18-C-2067.

Chapter 3

Compositional Processing Emerges in Neural Networks Solving Math Problems

Jacob Russin[1], Roland Fernandez[2], Hamid Palangi[2], Eric Rosen[3], Nebojsa Jojic[2], Paul Smolensky[2,3], Jianfeng Gao[2]

The original version of this article (Russin et al., 2021a) was accepted for publication in the Proceedings for the 43rd Annual Meeting of the Cognitive Science Society (CogSci 2021). The opinions expressed here are the author's own and do not necessarily reflect the views of the conference, workshop, or publisher. The original version is available online at `https://arxiv.org/abs/2105.08961`.

3.1 Abstract

A longstanding question in cognitive science concerns the learning mechanisms underlying compositionality in human cognition. Humans can infer the structured relationships (e.g., grammatical rules) implicit in their sensory observations (e.g., auditory speech), and use this knowledge to guide the composition of simpler meanings into complex wholes. Recent progress in artificial neural networks has shown that when large models are trained on enough linguistic data, grammatical structure emerges in their representations. We extend this work to the domain of mathematical reasoning, where it is possible to formulate precise hypotheses about how meanings (e.g., the quantities corresponding to numerals) should be composed according to structured rules (e.g., order of operations). Our work shows that neural networks are not only able to infer something about the structured relationships implicit in their training data, but can also deploy this knowledge to guide the composition of individual meanings into composite wholes.

3.2 Introduction

A key to the power of human cognition is the principle of compositionality (Hinzen et al., 2012): complex stimuli such as sentences are understood by 1) recognizing their relevant subcomponents, 2) extracting the meanings of these subcomponents, and 3) combining these partial meanings according to structured rules to produce a final result — the meaning of the whole stimulus (Martin & Baggio, 2020). This allows a potentially

[1] Center for Neuroscience, University of California, Davis
[2] Microsoft Research, Redmond
[3] Department of Cognitive Science, Johns Hopkins University

infinite number of stimuli to be understood as novel compositions of familiar parts. Compositionality was explicitly built into traditional symbolic artificial intelligence (AI) systems, but in the currently dominant approach — deep neural networks — encodings are continuous vectors that do not overtly possess the relevant kind of compositionality (Fodor & Pylyshyn, 1988; Lake & Baroni, 2018; Lake et al., 2017; McClelland et al., 2020).

Although these recent models are the most powerful AI systems ever created, explaining how they achieve their remarkable success is a profound mystery — a grand challenge for current science. Is it possible that deep neural networks are somehow implicitly exploiting the principle of compositionality (See Fig. 3.1A)?

State-of-the-art neural networks called transformers (Vaswani et al., 2017) have shown impressive performance on many natural language tasks (Devlin et al., 2019). Briefly, these networks learn to encode each symbol in their inputs with a high-dimensional vector that is successively refined over several layers by adding information from other symbols in the input. There is evidence that these networks reflect in their activation patterns the relevant linguistic subcomponents in their inputs (Linzen & Baroni, 2021; Manning et al., 2020; Tenney et al., 2019). However, it is unclear how, if at all, these models use such structure to extract and compose the meanings of parts to succeed at their tasks (Lake et al., 2017).

Studying this question in the natural language setting is challenging because it is extremely difficult to precisely characterize the meaning of a part of a natural language expression. We therefore study the question in a domain where the meaning of a subcomponent is clear: arithmetic expressions. The 'meaning' of a part — a sub-expression — is simply its numerical value, and the principle of compositionality holds perfectly: the value of $4 * 5 + 2 * 3$ is precisely obtained by taking the meanings of the sub-expressions $4 * 5$ and $2 * 3$ (i.e., the quantities denoted by the numerals '20' and '6') and composing them with the addition operation to derive the meaning of the entire expression, the number written '26'.

Do deep neural networks evaluate arithmetic expressions using the principle of compositionality in this way? We investigated this question by analyzing models trained on the Mathematics Dataset (Saxton et al., 2019). This dataset contains 112 million mathematical word problems separated into different modules, each of which covering a different mathematical domain such as arithmetic, algebra, calculus and probability. Models receive as input a sequence of characters (e.g., `Evaluate (12/3 + 10/2)/3`) and must output the sequence of characters exactly matching the correct answer (e.g., `3`). To an untrained model, these strings of characters have no structure or semantic content whatsoever — nothing in the characters themselves conveys their semantics (e.g., that '2' is larger than '1') or the rules governing them (e.g., the correct order of operations).

Previous work showed that transformers can achieve an impressive 77.42% accuracy on this dataset, and that when the standard transformer is augmented with explicitly-structured, tensor product representations

(Smolensky, 1990) — the TP-Transformer — this improves to 80.67% (Schlag et al., 2019). These models can produce the correct answers to problems that would be challenging even to humans — for example, problems involving simultaneous differentiation and factorization:

> **Question**: `Let r(g) be the second derivative of 2*g**3/3 - 21*g**2/2 + 10*g. Let z be r(7). Factor -z*s + 6 - 9*s**2 + 0*s + 6*s**2`
>
> **Answer**: `-(s + 3)*(3*s - 2).`

However, these models were found to exhibit poor generalization performance on problems with significant deviations from their training set (i.e., "extrapolation" problems, including problems with larger numbers, more terms, etc.). This may suggest that although the models were capable of answering unseen word problems as complicated as the one above, they fail to fully capture the systematicity of the rules governing mathematical expressions, possibly relying instead on a "mix-and-match" strategy (Lake & Baroni, 2018).

In this work, we analyzed the representations of these trained models, probing them on new problems to investigate whether they evaluated arithmetic expressions compositionally, i.e., whether representations of the partial results corresponding to each sub-expression in the overall question could be found in different parts of the network. Our results suggest that to a surprising extent, both the standard transformer and the TP-Transformer learn to solve arithmetic problems by evaluating sub-expressions separately, thus demonstrating some ability to compose the meanings of symbols according to their structured relationships.

3.3 Methods

The code used in our analyses can be found online[4].

3.3.1 Models

The models we used in our analyses were already trained on the Mathematics Dataset, and were freely available online. Details about the architectures and procedures used to train them can be found in the original publication (Schlag et al., 2019). Briefly, both the standard transformer and TP-Transformer contain an encoder that processes the question, and a decoder that generates the final answer. The encoder and decoder of both architectures had 6 transformer layers containing multi-head attention modules with 8 heads, as described in Vaswani et al. (2017). Each head in each layer generates a query (Q), key (K), and value (V) vector for every input to that layer. Attention distributions are generated by taking a softmax of the scaled dot product of the queries and keys. The final output of the attention mechanism is the average of

[4] `https://github.com/jlrussin/interpret-math-transformer`

the value vectors, weighted by the attention distribution:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.1}$$

where d_k is the dimension of the key vectors.

The TP-Transformer adapts the transformer architecture to use a role-filler binding mechanism, where roles are meant to explicitly capture structural or relational information in the inputs (Schlag et al., 2019; Smolensky, 1990). The architecture shares much of the organization of the standard transformer, but an additional role vector (R) is generated in each head, and this vector is bound to the existing output of the attention mechanism with a Hadamard product:

$$\text{TP-Attention}(Q, K, V, R) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \odot R \tag{3.2}$$

where $\odot$ denotes the Hadamard product. We included the TP-Transformer in our exploration of compositionality because it had been shown to achieve state-of-the-art performance on the Mathematics Dataset (Schlag et al., 2019), and because it was hypothesized to better capture structured relationships.

Both models were trained on all modules in the Mathematics Dataset (Saxton et al., 2019) simultaneously and were not fine-tuned on the arithmetic module. The TP-Transformer used in our analyses had about 49 million parameters and was trained for 1.7 million steps, and the standard transformer had about 44 million parameters and was trained for 700,000 steps. These differences were perhaps reflected in the results of the arithmetic module ("arithmetic mixed") of the dataset, where the TP-Transformer and standard transformer achieved about 83% and 61% accuracy, respectively, on the test set.

3.3.2 Test Datasets

To investigate the degree to which these models had learned to break arithmetic expressions into subcomponents, we tested them on problems containing well-defined sub-expressions so that we could systematically probe their internal representations. These test sets were derived from the arithmetic module of the dataset, but were highly controlled in a number of ways. We created a total of six test sets, and each contained problems that were generated from one of the following arithmetic expressions: 1) $\frac{x_1+x_2}{x_3}$, 2) $\frac{x_1*x_2}{x_3*x_4}$, 3) $x_1+x_2*x_3$, 4) $\frac{x_1+x_2*x_3}{x_4}$, 5) $\frac{x_1+x_2}{x_3}+x_4$, 6) $\frac{x_1+x_2}{x_3}+\frac{x_4+x_5}{x_6}$. These expressions were chosen in order to have a good mix of the possible operations, while retaining unambiguous sub-expressions that we could probe. Each test set consisted of roughly 200 problems that were generated by randomly sampling single-digit numbers for each of the x_i, while constraining the final answers to be positive integers between 0 and 20. This was

done so that each number was restricted to a single character, and in order to avoid complications that may have occurred due to negative numbers and arithmetic carrying. The samples were also selected such that the distributions of final answers in each test set were as uniform as possible. The dataset used to train the models did not contain any of the exact samples used for testing, but it did contain some problems with the same arithmetic forms. The models that had been trained on the entire dataset were tested on these custom test sets without fine-tuning on them.

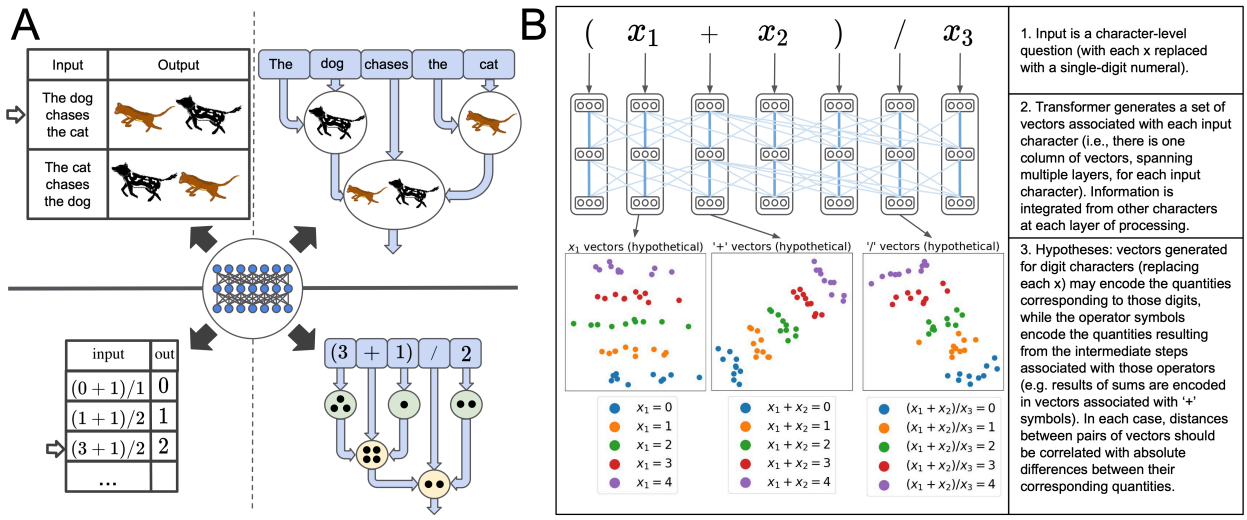

Figure 3.1: Conceptual illustrations of hypotheses. **(A)** Two extreme strategies that might be learned by a neural network trained on language tasks (top) or arithmetic problems (bottom). A large model might memorize in its weights a lookup table containing answers to each problem (i.e., a holistic pattern-matching strategy), thus bypassing their implicit compositionality altogether (left). Alternatively, the model may infer and leverage the underlying structure of the problem to decompose it into sub-expressions (right). In the latter case, the model's representations should show evidence of encoding the meaning of each symbol (e.g., quantities corresponding to numerals, shown in green circles) and the meaning of each sub-expression (e.g., quantities corresponding to results of intermediate operations, shown in yellow circles). **(B)** Hypothetical compositional representations in a transformer model trained on math problems. The vectors representing digit characters encode the values of the corresponding digits, in the sense that vectors associated with similar values are closer together. Similarly, the vectors representing operators encode the intermediate results of those operations. Distances between pairs of vectors should be highly correlated with the absolute differences between their corresponding values. Only three layers are shown, but both models had six layers.

The models generally achieved high accuracies on the probing test sets we created (numbers corresponding to expressions above): TP-Transformer 1) 100%, 2) 100%, 3) 92%, 4) 89.5%, 5) 100%, 6) 96%; standard transformer: 1) 100%, 2) 100%, 3) 38%, 4) 98.5%, 5) 100%, 6) 98%. Differences in accuracy on the test sets were not critical to our analyses, as our results were qualitatively reproduced for both models. The problems incorrectly answered by the models were not excluded from the analyses.

3.4 Results

Two extreme strategies for solving problems with implicit compositional structure are shown in Figure 3.1A. At one extreme, a large neural network might overlook the compositional structure of the problems and memorize in its weights a simple lookup table containing the answers to each problem individually. At the other extreme, a perfectly compositional learner would infer whatever structured relationships exist and use them to decompose problems into appropriate sub-components.

We hypothesized that if the models had learned to evaluate arithmetic expressions by processing their sub-components separately, then the representations corresponding to similar quantities would be closer together, as measured by Euclidean distance (see Figure 3.1B). For example, in expressions of the form $\frac{x_1+x_2}{x_3}$ (where each x_i is replaced by a particular numeral in each problem), the vectors corresponding to the partial result in the numerator $x_1 + x_2 = 1$ would be closer to those corresponding to $x_1 + x_2 = 2$ than they would be to those corresponding to $x_1 + x_2 = 3$. We therefore extracted multiple vectors (e.g., queries, keys, values, and role vectors for TP-Transformer) from each layer of both models in order to analyze them. Our analyses across these different kinds of vectors from each attention head yielded qualitatively similar results, so for simplicity we report results for queries across both models. Unless noted otherwise, we report results from vectors extracted from the highest layer (layer 6) of the encoder of each model.

Table 3.1: Spearman correlations for operators in each test set. All were highly significant ($p < 0.001$). Vectors were from the last layer of the encoder.

Expression	Operator	TP	TF
1. $(x_1 + x_2)/x_3$	'+'	.315	.453
	'/'	.565	.539
2. $(x_1 * x_2)/(x_3 * x_4)$	'$*$' (1st)	.479	.612
	'$*$' (2nd)	.662	.654
	'/'	.590	.536
3. $x_1 + x_2 * x_3$	'$*$'	.132	.251
	'+'	.502	.381
4. $(x_1 + x_2 * x_3)/x_4$	'$*$'	.141	.297
	'+'	.159	.174
	'/'	.147	.171
5. $(x_1 + x_2)/x_3 + x_4$	'+' (1st)	.424	.423
	'/'	.510	.459
	'+' (2nd)	.353	.411
6. $(x_1 + x_2)/x_3 + (x_4 + x_5)/x_6$	'+' (1st)	.410	.393
	'/' (1st)	.610	.536
	'+' (2nd)	.405	.455
	'/' (2nd)	.566	.526
	'+' (3rd)	.303	.421

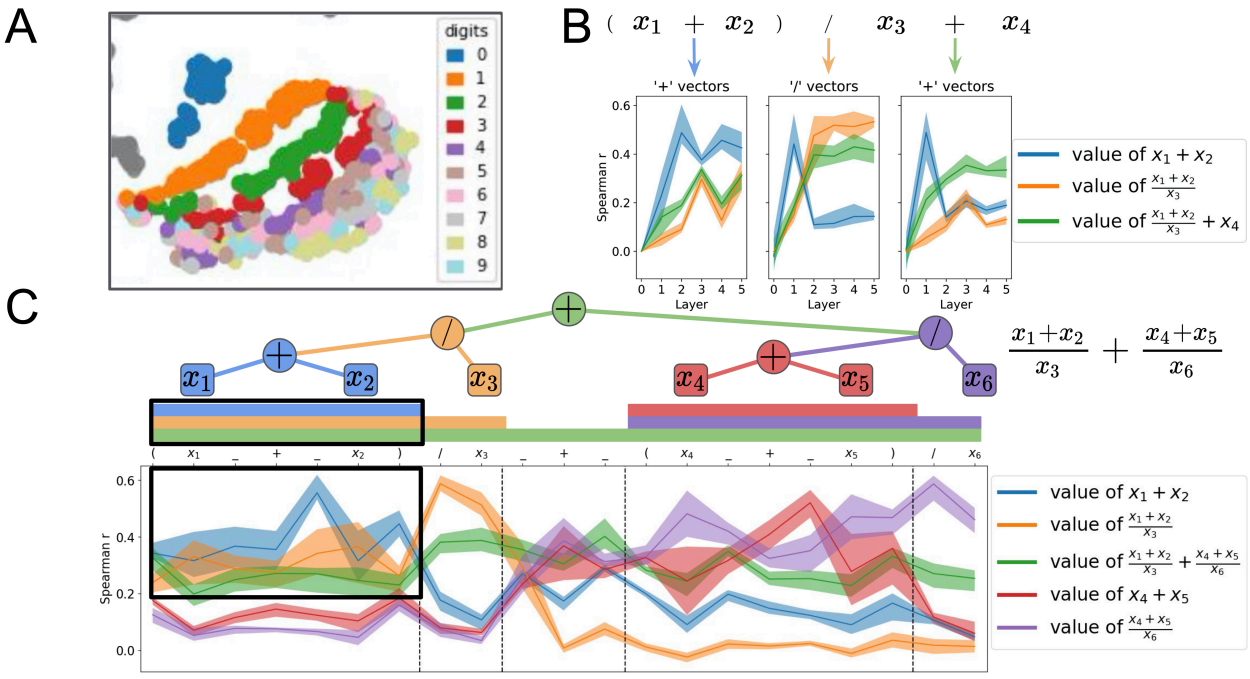

Figure 3.2: Results. **(A)** t-SNE embedding of vectors representing digits in layer 1 of the TP-Transformer on the test set of the mixed arithmetic module. Each point designates a vector associated with a particular digit character, colored by its value. The semantics of the digit characters is apparent in the model's representations, which are partially organized according to their natural order (along the top-left to bottom-right axis). **(B)** Correlations with vectors associated with operators in TP-Transformer. In each problem, each of the x_i were replaced with single-digit numerals. Vectors representing the first '+' symbol encode the results of the sum in the numerator $(x_1 + x_2)$, vectors representing the '/' symbol encode the results of the division $((x_1 + x_2)/x_3)$, and vectors representing the second '+' symbol encode the results of the final sum $((x_1 + x_2)/x_3 + x_4)$. In general, analyses focused on the vectors generated in the final layer, but correlations displayed here across all layers show how these can change with further processing (e.g., representations of the '/' symbol seem to encode the result of the sum in the first layer, but in later layers strongly encode the result of the division). Colored envelopes show the minimum and maximum correlation across heads. **(C)** Alignment of correlation measures across the entire input sequence with the parse tree of the expression, colored to match the lines in the plot for the appropriate quantities. Correlations with the values of an intermediate result peak at the vectors associated with the corresponding operator, and are elevated over the constituent containing its arguments (shown by matching colors in the parse tree, and colored bars covering the extent of the appropriate constituent). Black boxes highlight an example: the vectors representing each element in the constituent $(x_1 + x_2)$ show the highest correlations with the value of the sum of x_1 and x_2 (blue), but also show correlations with its division by x_3 (orange), and the result of the whole expression (green). Results are from vectors in the final layer of the TP-Transformer encoder. Vertical dotted lines delineate sub-expressions, and underscores indicate space characters.

As a first step, we probed the models for evidence that they were encoding the quantities associated with digit characters (akin to the dots shown on the right of Figure 3.1A). Figure 3.2A shows the vectors representing digits in layer 1 of the TP-Transformer model, visualized using t-SNE (van der Maaten & Hinton, 2008). The model's representations capture the semantics of the digit characters in that they are organized

according to their natural order (e.g., vectors corresponding to 1's are closer to 2's than to 9's, etc.). We confirmed this pattern quantitatively by measuring Spearman correlations of the distances between pairs of digit vectors with the absolute differences between the values of their corresponding digits. This correlation was significant for both TP-Transformer (TP: $r = 0.440, p < 0.001$), and the standard transformer (TF: $r = 0.436, p < 0.001$). These were compared to correlations between these same distances and the absolute differences between the values of the *other* digits in the same problems. For example, in problems of the form $\frac{x_1+x_2}{x_3}$, the distances in this "unmatched" correlation would be taken between vectors corresponding to x_1 vectors, but the absolute differences would be taken between the values of x_3 from the same problems. This further analysis revealed that the matched correlations (TP: $r = 0.440$, TF: $r = 0.436$) were much higher than the unmatched (TP: $r = 0.086$, TF: $r = 0.087$), indicating that the models were representing the quantities associated with each numeral in the appropriate position in the sequence. This relationship was found to be statistically significant in a more formal linear regression comparing "matched" and "unmatched" pairs ($p < 0.001$ for both models).

Next, we investigated whether the encoding of intermediate results of arithmetic sub-expressions could be detected in a similar way. We reasoned that if this pattern of correlation were observed for the intermediate results corresponding to arithmetic sub-expressions, this would indicate that the model had encoded these partial results. We repeated the analyses, but with distances between vectors associated with operators and the differences between their corresponding intermediate results. For example, if the vectors representing the '+' symbol in expressions with the form $\frac{x_1+x_2}{x_3}$ were encoding the sum of x_1 and x_2, we would expect those '+' vectors encoding similar values for their sums to be closer together, leading to a significant correlation between distances between the pairs of '+' vectors and the differences between their corresponding sums. This correlation would be expected to peak at the position of the operator, but also be elevated over the positions of symbols within its constituent (e.g., in $(x_1 + x_2)/x_3$, the positions over the $x_1 + x_2$ constituent for the '+' operator).

These analyses revealed a striking pattern: transformer models trained on mathematics encode the quantities of intermediate results in the vectors associated with the appropriate operators (see Table 3.1). Figure 3.2B shows how these correlations unfold over the layers of the network, with comparisons to the inappropriate operators (e.g., '+' vectors and the result of the division). When these data were aggregated across all the expressions we tested, a significant correlation was observed between corresponding operator representations and partial values for both models (TP: $r = 0.310, p < 0.001$; TF: $r = 0.314, p < 0.001$). This correlation was higher than when distances were correlated with differences between intermediate results corresponding to the other operators in the same problems (TP: $r = 0.122$, TF: $r = 0.167$). Again, a more formal linear regression revealed that this relationship was statistically significant ($p < 0.001$ for both

models), indicating that the models were representing these intermediate results in the vectors corresponding to the appropriate operators. Figure 3.2C shows the same correlation measures on representations across the entire input sequence of a more complicated expression. The correlations with partial results are seen to peak at the corresponding operators, but also to be relatively elevated over the corresponding constituents. We repeated our analyses and confirmed that the vectors representing symbols within constituents also carried information about their corresponding partial results (TP: $r = 0.201$, TF: $r = 0.189$).

3.5 Discussion

Compositionality, a hallmark of human cognition, requires knowledge of constituent structure to guide the assembly of partial meanings into coherent semantic wholes. Our analyses reveal that the compositional semantics implicit in character-level math problems can emerge to a surprising extent in neural networks even when they are instructed only on the characters comprising the problems' answers. Our results are consistent with previous work in the natural language setting (Manning et al., 2020; Tenney et al., 2019) suggesting that when these networks are trained on many observations, they can learn to represent structured relationships. However, our work shows further that these models can use this knowledge to guide a partially compositional process whereby semantic content is integrated across symbols.

Though our results were surprising, we expect that the compositionality we observed is imperfect, and that the trained models lie somewhere between the two extremes depicted in Figure 3.1A. A perfectly compositional learner would completely separate the process of evaluating each of the sub-components of an arithmetic expression (e.g., first completely evaluate the expression in the numerator, then divide the result by the denominator). Our results suggest that transformers perform these separable computations in different parts of the network: the vectors aligned with the input positions of the operators tended to encode the quantities corresponding to the results of those operations. However, our analyses did not find that these relationships were perfect (e.g., see the non-zero correlations across the sequence in Figure 3.2C). Furthermore, it is likely that these vectors are encoding more than pure quantities; the high-dimensional vectors in these models may encode the sum $(x_1 + x_2)$ while also encoding each of the constituent elements (such that a decoder could be trained to predict the constituents, as well as the sum, from the vector).

Previous work on transformers trained on the Mathematics Dataset (Schlag et al., 2019) showed that these models suffered large reductions in generalization performance on arithmetic problems with significantly different surface-level features (e.g., problems with larger numbers, more terms, etc.). A perfectly compositional agent with a true understanding of arithmetic rules would in principle be able to generalize to any problem following those rules. It is possible that although our analyses revealed a significant degree

of compositionality, its imperfection prevented the models from generalizing on these extrapolation tests.

It should be noted that although compositionality allows for flexible cognition and a powerful form of combinatorial generalization (Fodor & Pylyshyn, 1988; Lake et al., 2017), a strict form of compositionality may not always be desirable — for example, when learning the meanings of idioms or other idiosyncrasies of natural language that violate the principle of compositionality (Szabó, 2020). Human language learners must negotiate the tension between a strict compositionality assumption and the ability to learn exceptions (Rumelhart & McClelland, 1986; Pinker & Prince, 1988). Machine learning methods for natural language processing face a similar tension, and may benefit from a greater dialogue with ongoing research in cognitive science (Lake et al., 2017).

The success of modern neural networks has shown that major advances in artificial intelligence are possible when large models are trained on enough data (Brown et al., 2020; LeCun et al., 2015). Our results show that compositionality, which is often thought to be an inherent property of human cognition (Fodor & Pylyshyn, 1988), can to some extent emerge when a large neural network is trained on enough data. However, much work remains to clarify whether these approaches will continue to scale to human-level compositionality, or whether this will require learning systems that have been explicitly designed to facilitate compositional processing (Smolensky, 1990; Russin et al., 2020b).

3.6 Acknowledgments

We would like to thank Imanol Schlag, Jürgen Schmidhuber, Randall O'Reilly, the members of the Computational Cognitive Neuroscience lab at UC Davis, and reviewers for helpful comments and discussions. J.R. was supported by the NIMH, Award Number T32MH112507. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Illustrations by H. Tomkiewicz.

Chapter 4

Systematicity in a Recurrent Neural Network by Factorizing Syntax and Semantics

Jacob Russin[1], Jason Jo[2], Randall C. O'Reilly[1], Yoshua Bengio[2,3]

The original version of this article (Russin et al., 2020a) was accepted for publication in the Proceedings for the 42nd Annual Meeting of the Cognitive Science Society (CogSci 2020). The opinions expressed here are the author's own and do not necessarily reflect the views of the conference, workshop, or publisher. The original version is available online at `https://cognitivesciencesociety.org/cogsci20/papers/0027/0027.pdf`.

4.1 Abstract

Standard methods in deep learning fail to capture compositional or systematic structure in their training data, as shown by their inability to generalize outside of the training distribution. However, human learners readily generalize in this way, e.g. by applying known grammatical rules to novel words. The inductive biases that might underlie this powerful cognitive capacity remain unclear. Inspired by work in cognitive science suggesting a functional distinction between systems for syntactic and semantic processing, we implement a modification to an existing deep learning architecture, imposing an analogous separation. The resulting architecture substantially outperforms standard recurrent networks on the SCAN dataset, a compositional generalization task, without any additional supervision. Our work suggests that separating syntactic from semantic learning may be a useful heuristic for capturing compositional structure, and highlights the potential of using cognitive principles to inform inductive biases in deep learning.

4.2 Introduction

A crucial property underlying the power of human cognition is its systematicity (Lake et al., 2017; Fodor & Pylyshyn, 1988): known concepts can be combined in novel ways according to systematic rules, allowing the number of expressible combinations to grow exponentially in the number of concepts that are learned.

[1] Center for Neuroscience, University of California, Davis
[2] MILA, Université de Montréal
[3]CIFAR Senior Fellow

Recent work has shown that standard algorithms in deep learning fail to capture this important property: when tested on unseen combinations of known elements, standard models fail to generalize (Lake & Baroni, 2018; Loula et al., 2018; Bastings et al., 2018). It has been suggested that this failure represents a major deficiency of current deep learning models, especially when they are compared to human learners (Marcus, 2018; Lake et al., 2017, 2019a).

A recently published dataset called SCAN (Lake & Baroni, 2018) tests compositional generalization in a sequence-to-sequence (seq2seq) setting by systematically holding out of the training set all inputs containing a basic primitive verb ("jump"), and testing on sequences containing that verb. Success on this difficult problem requires models to generalize knowledge gained about the other primitive verbs ("walk", "run" and "look") to the novel verb "jump," without having seen "jump" in any but the most basic context ("jump" $\rightarrow$ JUMP). It is trivial for human learners to generalize in this way (e.g., if I tell you that "dax" is a verb, you can generalize its usage to all kinds of constructions, like "dax twice and then dax again", without even knowing what the word means) (Lake & Baroni, 2018; Lake et al., 2019a). However, powerful recurrent seq2seq models perform surprisingly poorly on this task (Lake & Baroni, 2018; Bastings et al., 2018).

From a statistical-learning perspective, this failure is quite natural. The neural networks trained on the SCAN task fail to generalize because they have memorized biases that do indeed exist in the training set. Because "jump" has never been seen with any adverb, it would not be irrational for a learner to assume that "jump twice" is an invalid sentence in this language. The SCAN task requires networks to make an inferential leap about the entire structure of part of the distribution that they have not seen — that is, it requires them to make an out-of-domain (o.o.d.) *extrapolation* (Marcus, 2018), rather than merely *interpolate* according to the assumption that train and test data are independent and identically distributed (i.i.d.) (see left part of Figure 4.3). Seen another way, the SCAN task and its analogues in human learning (e.g., "dax"), require models *not* to learn some of the correlations that are actually present in the training data (Kriete et al., 2013). To the extent that humans can perform well on certain kinds of o.o.d. tests, they must be utilizing inductive biases that are lacking in current deep learning models (Battaglia et al., 2018).

It has long been suggested that the human capacity for systematic generalization is linked to mechanisms for processing syntax, and their functional separation from the meanings of individual words (Chomsky, 1957; Fodor & Pylyshyn, 1988). Furthermore, recent work in cognitive and computational neuroscience suggests that human learners may factorize knowledge about structure and content, and that this may be important for their ability to generalize to novel combinations (Behrens et al., 2018; Ranganath & Ritchey, 2012). In this work, we take inspiration from these ideas and explore operationalizing a separation between structure and content as an inductive bias within a deep-learning attention mechanism (Bahdanau et al., 2014). The resulting architecture, which we call *Syntactic Attention*, separates structural learning about

the alignment of input words to target actions (which can be seen as a rough analogue of syntax in the seq2seq setting) from learning about the meanings of individual words (in terms of their corresponding actions). The modified attention mechanism achieves substantially improved compositional generalization performance over standard recurrent networks on the SCAN task.

An important contribution of this work is in showing how changes in the connectivity of a network can shape its learning in order to develop a separation between structure and content, without any direct manual imposition of this separation per se. These changes act as an inductive bias or soft constraint that only manifests itself through learning. Furthermore, our model shows that attentional modulation can provide a mechanism for structural representations to control processing in the content pathway, similar to how spatial attention in the dorsal visual pathway can generically modulate object-recognition processing in the ventral visual stream (O'Reilly et al., 2017). Thus, attentional modulation may be critical for enabling structure-sensitive processing — a natural property of symbolic models — to be realized in neural hardware. This provides a more purely neural framework for achieving systematicity, compared to hybrid approaches that combine symbolic and neural network mechanisms (Yi et al., 2018).

4.3 Model

The Syntactic Attention model improves the compositional generalization capability of an existing attention mechanism (Bahdanau et al., 2014) by implementing two separate streams of information processing for syntax and semantics (see Figure 4.1). In the seq2seq setting, we operationalize 'semantics' to mean the information in each word in the input that determines its *meaning* in terms of target outputs, and we operationalize 'syntax' to mean the information contained in the input sequence that should determine the structure of the *alignment* of input to target words. We describe the mechanisms of this separation and the other details of the model below, following the notation of (Bahdanau et al., 2014), where possible.

4.3.1 Factorizing Syntax and Semantics in Seq2seq

In the seq2seq setting, models must learn a mapping from arbitrary-length sequences of inputs $\mathbf{x} = \{x_1, ..., x_{T_x}\}$ to arbitrary-length sequences of outputs $\mathbf{y} = \{y_1, ..., y_{T_y}\}$: $p(\mathbf{y}|\mathbf{x})$. In the SCAN task, the inputs are a sequence of instructions, and the outputs are a sequence of actions. The attention mechanism of Bahdanau et al. (2014) models the conditional probability of each target action given the input sequence and previous targets: $p(y_i|y_1, y_2, ..., y_{i-1}, \mathbf{x})$. This is accomplished by processing the instructions with a recurrent neural network (RNN) in an encoder. The outputs of this RNN are used both for encoding individual words for subsequent translation, and for determining their alignment to actions during decoding.

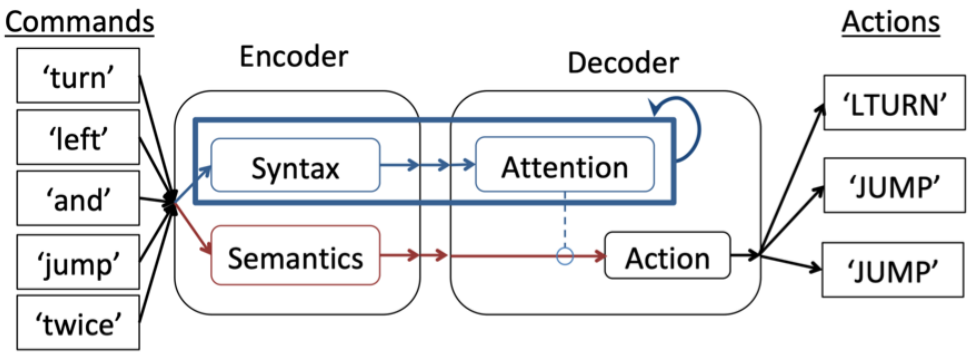

Figure 4.1: Syntactic Attention architecture. Syntactic and semantic information are maintained in separate streams. The semantic stream is used to directly produce actions, and processes words with a simple linear transformation, so that sequential information is not maintained. The syntactic stream processes inputs with a recurrent neural network, allowing it to capture temporal dependencies between words. This stream determines the attention over semantic representations at each time step during decoding.

The underlying assumption made by the Syntactic Attention architecture is that the dependence of target actions on the input sequence can be separated into two independent factors. One factor, $p(y_i|x_j)$, which we refer to as "semantics," models the conditional distribution from individual words in the input to individual actions in the target. Note that, unlike in the model of Bahdanau et al. (2014), these x_j do not contain any information about the other words in the input sequence because they are not processed with an RNN. They are "semantic" in the sense that they contain the information relevant to translating the instruction words into corresponding actions. The other factor, $p(j \rightarrow i|\mathbf{x}, y_{1:i-1})$, which we refer to as "syntax," models the conditional probability that word j in the input is relevant to word i in the action sequence, given the entire set of instructions. This is the alignment of words in the instructions to particular steps in the action sequence, and is accomplished by computing the attention weights over the instruction words at each step in the action sequence using encodings from an RNN. This factor is "syntactic" in the sense that it must capture all of the temporal dependencies in the instructions that are relevant to determining the serial order of outputs (e.g., what should be done "twice", etc.). The crucial architectural assumption, then, is that any temporal dependency between individual words in the instructions that can be captured by an RNN should largely be relevant to their alignment to words in the target sequence, and less relevant to the meanings of individual words. We argue that this can be seen as a factorization of syntax and semantics, because the grammatical rules governing the composition of instruction words' meanings (e.g., how adverbs modify verbs) must be learned in a module that does not have access to those meanings. This assumption will be made clearer in the model description below.

4.3.2 Encoder

The encoder produces two separate vector representations for each word in the input sequence. Unlike the previous attention model (Bahdanau et al., 2014), we separately extract the semantic information from each word with a linear transformation:

$$m_j = W_m x_j \tag{4.1}$$

where W_m is a learned weight matrix that multiplies the one-hot encodings $\{x_1, ..., x_{T_x}\}$. This weight matrix W_m can be thought of as extracting the information from the inputs that will be relevant to translating individual words into their corresponding actions (e.g. "jump" $\rightarrow$ JUMP).

As in the previous attention mechanism (Bahdanau et al., 2014), we use a bidirectional RNN (biRNN) to extract what we now interpret as the syntactic information from each word in the input sequence. The biRNN processes the (one-hot) input vectors $\{x_1, ..., x_{T_x}\}$ and produces a hidden-state vector for each word on the forward pass, $(\overrightarrow{h_1}, ..., \overrightarrow{h_{T_x}})$, and a hidden-state vector for each word on the backward pass, $(\overleftarrow{h_1}, ..., \overleftarrow{h_{T_x}})$. The syntactic information (or "annotations" (Bahdanau et al., 2014)) of each word x_j is determined by the two vectors $\overrightarrow{h_{j-1}}$, $\overleftarrow{h_{j+1}}$ corresponding to the words surrounding it:

$$h_j = [\overrightarrow{h_{j-1}}; \overleftarrow{h_{j+1}}] \tag{4.2}$$

In all experiments, we used a bidirectional Long Short-Term Memory (LSTM) for this purpose. These representations h_j differ from the previous model in that only the surrounding words are used to infer the relevant syntactic information about each input. Our motivation for doing this was to encourage the encoder to rely on the role each word plays in the input sentence. Note that because there is no sequence information in the semantic representations, all of the information required to parse (i.e., align) the input sequence correctly (e.g., phrase structure, modifying relationships, etc.) must be encoded by the biRNN.

4.3.3 Decoder

The decoder models the conditional probability of each target word given the input and the previous targets: $p(y_i|y_1, y_2, ..., y_{i-1}, \mathbf{x})$, where y_i is the target action and $\mathbf{x}$ is the whole instruction sequence. As in the previous model, we use an RNN to determine an attention distribution over the inputs at each time step (i.e., to align words in the input to the current action). However, our decoder diverges from this model in that the mapping from inputs to outputs is performed from a weighted average of the *semantic* representations

of the input words:

$$d_i = \sum_{j=1}^{T_x} \alpha_{ij} m_j \qquad p(y_i|y_1, y_2, ..., y_{i-1}, \mathbf{x}) = f(d_i) \tag{4.3}$$

where f is parameterized by a linear function with a softmax nonlinearity, and the α_{ij} are the weights determined by the attention model. The softmax in f produces a distribution over the possible actions. We note again that the m_j are produced directly from corresponding x_j, and do not depend on the other inputs. The attention weights are computed by a function measuring how well the syntactic information of a given word in the input sequence aligns with the current hidden state of the decoder RNN, s_i:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \qquad e_{ij} = a(s_i, h_j) \tag{4.4}$$

where e_{ij} can be thought of as measuring the importance of a given input word x_j to the current action y_i, and s_i is the current hidden state of the decoder RNN. Bahdanau et al. (2014) model the function a with a feedforward network, but we choose to use a simple dot product:

$$a(s_i, h_j) = s_i \cdot h_j, \tag{4.5}$$

relying on the end-to-end backpropagation during training to allow the model to learn to make appropriate use of this function. Finally, the hidden state of the RNN is updated with the same weighted combination of the *syntactic* representations of the inputs:

$$s_i = g(s_{i-1}, c_i) \qquad c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \tag{4.6}$$

where g is the decoder RNN, s_i is the current hidden state, and c_i can be thought of as the information in the attended words that can be used to determine what to attend to on the next time step. Again, in all experiments an LSTM was used.

4.4 Simulations

4.4.1 SCAN task

The SCAN[4] task was specifically designed to test compositional generalization (see figure 4.2). In the task, sequences of commands (e.g., "jump twice") must be mapped to sequences of actions (e.g., JUMP JUMP), and is generated from a simple finite phrase-structure grammar that includes things like adverbs

[4]The SCAN task can be downloaded at `https://github.com/brendenlake/SCAN`

and conjunctions (Lake & Baroni, 2018). The splits of the dataset include: **1) Simple split**, where training and testing data are split randomly, **2) Length split**, where training includes only shorter sequences, and **3) Add primitive split**, where a primitive command (e.g., "turn left" or "jump") is held out of the training set, except in its most basic form (e.g., "jump" → JUMP).

jump	⇒	JUMP
jump left	⇒	LTURN JUMP
jump around right	⇒	RTURN JUMP RTURN JUMP RTURN JUMP RTURN JUMP
turn left twice	⇒	LTURN LTURN
jump thrice	⇒	JUMP JUMP JUMP
jump opposite left and walk thrice	⇒	LTURN LTURN JUMP WALK WALK WALK
jump opposite left after walk around left	⇒	LTURN WALK LTURN WALK LTURN WALK LTURN WALK LTURN LTURN JUMP

Figure 4.2: Examples from the SCAN dataset. Details about the detaset can be found in (Lake & Baroni, 2018). Figure reproduced from (Lake & Baroni, 2018).

Here we focus on the most difficult problem in the SCAN dataset, the add-jump split, where "jump" is held out of the training set. The best test accuracy reported in the original paper (Lake & Baroni, 2018), using basic seq2seq models, was 1.2%. More recent work has tested other kinds of seq2seq models, including Gated Recurrent Units (GRU) augmented with attention (Bastings et al., 2018), convolutional neural networks (CNNs) (Dessì & Baroni, 2019), meta-seq2seq (Lake, 2019), and a novel architecture (Li et al., 2019). Here, we compare the Syntactic Attention model to the best previously reported results.

4.4.2 Implementation details

Train and test sets were kept as they were in the original dataset, but following Bastings et al. (2018), we used early stopping by validating on a 20% held out sample of the training set. All reported results are from runs of 200,000 iterations with a batch size of 1. Unless stated otherwise, each architecture was trained 5 times with different random seeds for initialization, to measure variability in results. All experiments were implemented in PyTorch. Our best model used LSTMs, with 2 layers and 200 hidden units in the encoder, and 1 layer and 400 hidden units in the decoder, and 120-dimensional vectors for the semantic representations, m_j. The model included a dropout rate of 0.5, and was optimized using an Adam optimizer (Kingma & Ba, 2015) with a learning rate of 0.001.

4.4.3 Results

The Syntactic Attention model improves compositional generalization performance on the standard seq2seq SCAN dataset (see table 4.1). The table shows results (mean test accuracy (%) ± standard deviation) on

Table 4.1: Compositional generalization results. The Syntactic Attention model achieves an improvement on the compositional generalization tasks of the SCAN dataset in the standard seq2seq setting, compared to the standard models (Bastings et al., 2018; Dessì & Baroni, 2019). Recent results from another novel architecture (Li et al., 2019), developed concurrently using very similar principles, are also reported. Star* indicates average of 25 runs with random initializations. Others are averages of 5 runs.

Model	**Simple**	**Length**	**Add turn left**	**Add jump**
GRU + attn (Bastings et al., 2018)	100.0 ± 0.0	18.1 ± 1.1	59.1 ± 16.8	12.5 ± 6.6
GRU + attn - dep (Bastings et al., 2018)	100.0 ± 0.0	17.8 ± 1.7	90.8 ± 3.6	0.7 ± 0.4
CNN (Dessì & Baroni, 2019)	100.0 ± 0.0	-	-	69.2 ± 8.2
Li et al. (2019)	99.9 ± 0.0	20.3 ± 1.1	99.7 ± 0.4	98.8 ± 1.4
Syntactic Attention	100.0 ± 0.0	15.2 ± 0.7	99.9 ± 0.16	78.4* ± 27.4

the test splits of the dataset. Syntactic Attention is compared to the previous best models, which were a CNN (Dessì & Baroni, 2019), and GRUs augmented with an attention mechanism (+ attn), which either included or did not include a dependency (- dep) in the decoder on the previous action (Bastings et al., 2018). Transformers (Vaswani et al., 2017) were not included in our experiments, but have been shown to suffer similar problems with compositional generalization on the SCAN dataset (Keysers et al., 2020).

The best model from the hyperparameter search showed strong compositional generalization performance, attaining a mean accuracy of 91.1% (median = 98.5%) on the test set of the add-jump split. However, as in Dessì & Baroni (2019), we found that our model showed variance across initialization seeds. For this reason, we ran the best model 25 times on the add-jump split to get a more accurate assessment of performance. These results were highly skewed, with a mean accuracy of 78.4 % but a median of 91.0 %. Overall, this represents an improvement over the best previously reported results from standard seq2seq models in this task (Bastings et al., 2018; Dessì & Baroni, 2019).

Recently, Lake (2019) showed that a meta-learning architecture using an external memory achieves 99.95% accuracy on a meta-seq2seq version of the SCAN task. In this version, models are trained to learn how to generalize systematically across a number of variants of a compositional seq2seq problem. Here, we focus on the standard seq2seq version, which limits the model to one training set.

We also report the newer results of Li et al. (2019), which was work done concurrently with ours using a very similar approach. These results are very consistent with our own, and, taken together, lend support to the idea that separating mechanisms for learning syntactic information from mechanisms for learning the meanings of individual words can encourage systematicity in neural networks.

4.4.4 Additional experiments

To test our hypothesis that compositional generalization requires a separation between syntax (i.e., sequential information used for alignment), and semantics (i.e., the mapping from individual instruction words to individual actions), we conducted two more experiments:

- *Sequential semantics.* An additional biLSTM was used to process the semantics of the sentence: $m_j = [\overrightarrow{m_j}; \overleftarrow{m_j}]$, where $\overrightarrow{m_j}$ and $\overleftarrow{m_j}$ are the vectors produced for the input word x_j by a biLSTM on the forward and backward passes, respectively. These m_j replace those generated by the simple linear layer in the Syntactic Attention model (in equation (4.1)).

- *Syntax-action.* Syntactic information was allowed to directly influence the output at each time step in the decoder: $p(y_i|y_1, y_2, ..., y_{i-1}, \mathbf{x}) = f([d_i; c_i])$, where again f is parameterized with a linear function and a softmax output nonlinearity.

The results of the additional experiments (mean test accuracy (%) ± standard deviations) are shown in table 4.2. These results partially confirmed our hypothesis: performance on the add-jump test set was worse when the strict separation between syntax and semantics was violated by allowing sequential information to be processed in the semantic stream. In the *sequential semantics* experiment, the model performed comparably on the simple split (99.3 %) but performed worse on the compositional split even though we augmented its learning capacity by replacing a simple linear transformation with an RNN. This result suggests that this increase in capacity, which corresponded to a violation of the factorization assumption, allowed the model to memorize regularities in the dataset that prohibited systematic generalization during testing.

However, *syntax-action*, which included sequential information produced by a biLSTM (in the syntactic stream) in the final production of actions, maintained good compositional generalization performance. We hypothesize that this was because in this setup, it was easier for the model to learn to use the semantic information to directly translate actions, so it largely ignored the syntactic information. This experiment suggests that the separation between syntax and semantics does not have to be perfectly strict, as long as non-sequential semantic representations are available for direct translation.

Table 4.2: Results of additional experiments. Again star* indicates average of 25 runs with random initializations.

Model	**Add turn left**	**Add jump**
Sequential semantics	99.4 ± 1.1	42.3 ± 32.7
Syntax-action	98.2 ± 2.2	88.7 ± 14.2
Syntactic Attention	99.9 ± 0.16	78.4* ± 27.4

4.5 Related work

The principles of systematicity and compositionality have recently regained the attention of deep learning researchers (Bahdanau et al., 2019b; Lake et al., 2017; Lake & Baroni, 2018; Battaglia et al., 2018). In particular, these issues have been explored in the visual-question answering (VQA) setting (Andreas et al., 2016; Hudson & Manning, 2018; Yi et al., 2018). Many of the successful models in this setting learn hand-coded operations (Andreas et al., 2016), use highly specialized components (Hudson & Manning, 2018), or use additional supervision (Yi et al., 2018). In contrast, our model uses standard recurrent networks and simply imposes the additional constraint that mechanisms for syntax and semantics are separated.

Some of the recent research on compositionality in machine learning has had a special focus on the use of attention. For example, in the Compositional Attention Network, built for VQA, a strict separation is maintained between the representations used to encode images and the representations used to encode questions (Hudson & Manning, 2018). This separation is enforced by restricting them to interact only through attention distributions. Our model utilizes a similar restriction, reinforcing the idea that compositionality is enhanced when information from different modules are only allowed to interact through discrete probability distributions.

The results from the meta-seq2seq version of the SCAN task (Lake, 2019) suggest that meta-learning may be a viable approach to inducing compositionality in neural networks. Humans have ample opportunity through a long developmental trajectory to meta-learn the inductive biases that could facilitate compositional generalization, so this is a promising alternative to the work discussed here. However, a key difference in the particular implementation used in that study is that the additional training episodes explicitly demarcate the primitive verbs by permuting their meanings across episodes. In our work, the training is restricted to a single episode in which no such permutation occurs.

The work of Li et al. (2019) was done concurrently with ours; although their presentation is framed slightly differently, we believe very similar principles have motivated their model. There are few differences with our architecture, but their improved results on the SCAN task may be due to their use of additive noise during training. Future work will explore the exact differences with their model and analyze the important factors contributing to differences in results.

Finally, we note that the experiments presented here are limited to the SCAN dataset, which may not completely capture the kinds of compositional generalization that humans regularly manifest. This may be important, as recent work has shown that the extent to which models can generalize outside of their training distribution can depend heavily on the kind of environments in which they are trained (Hill et al., 2020). Recent work has experimented with other compositional generalization problems that may be more realistic

(Lake, 2019; Keysers et al., 2020). Future work will identify whether the principles developed in this paper can aid generalization performance in these other settings.

4.6 Discussion

The Syntactic Attention model was designed to incorporate principles from cognitive science and neuroscience as inductive biases into a neural network architecture: the mechanisms for learning rule-like or syntactic information are separated (or factorized (Behrens et al., 2018)) from mechanisms for learning semantic information. Our experiments confirm that this simple organizational principle encourages systematicity in recurrent neural networks in the seq2seq setting, as shown by the substantial improvement in the model's performance on the compositional generalization tasks in the SCAN dataset.

The model makes the assumption that the meanings of individual words should be independent of their alignment to actions in the target sequence (i.e., the attention weight applied to each word at each step in the action sequence). To this end, two separate encodings are produced for the words in the input: semantic representations in which each word is not influenced by other words in the sentence, and syntactic representations which are produced by an RNN that could capture temporal dependencies in the input sequence (e.g., modifying relationships, grammatical roles). The syntactic system alone has access to the sequential information in the inputs, but is constrained to influence actions through an attention mechanism only (see Figure 4.1). These constraints ensure that learning about the meanings of individual words happens independently of learning about the structured relationships *between* words. This encourages systematic generalization because, even if a word has only been encountered in a single context (e.g., "jump" in the add-jump split), as long as its syntactic role is known (e.g., that it is a verb that can be modified by adverbs such as "twice"), it can be used in many other constructions that follow the rules for that syntactic role. Additional experiments confirmed this intuition, showing that when sequential information is allowed to be processed by the semantic system (*sequential semantics*), systematic generalization performance is substantially reduced.

The paradigmatic example of systematicity is a symbolic system in which representational content (e.g., the value of a variable stored in memory) is maintained separately from the computations that are performed on that content. This separation ensures that the *manipulation* of the content stored in variables can be completely independent of the content itself, and will therefore generalize to arbitrary elements. Our model implements an analogous separation, but in a purely neural architecture that does not rely on hand-coded rules or additional supervision. In this way, it can be seen as transforming a difficult out-of-domain (o.o.d.) generalization problem into two separate i.i.d. generalization problems — one where the individual meanings

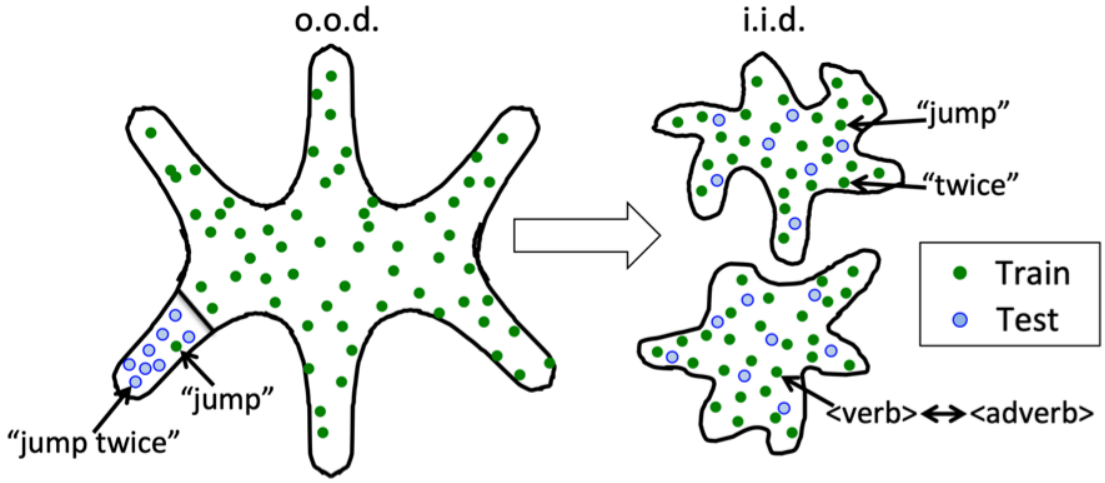

Figure 4.3: Illustration of the transformation of an out-of-domain (o.o.d.) generalization problem into two independent, identically distributed (i.i.d.) generalization problems. This transformation is accomplished by the Syntactic Attention model without hand-coding grammatical rules or supervising with additional information such as parts-of-speech tags.

of words are learned, and one where *how words are used* (e.g., how adverbs modify verbs) is learned (see Figure 4.3). This may be a useful approach to encouraging systematicity in neural networks, which are very good at i.i.d. generalization but generally fail when presented with o.o.d. problems.

Our work shows that a strict separation between syntax and semantics can be useful for encouraging systematicity and allowing for compositional generalization. It is unlikely that the human brain has such a strict separation, but our work builds on related ideas in neuroscience (Behrens et al., 2018) and suggests a useful framework for investigating whether a similar principle may be at work in the human brain. Future work will explore this principle in other settings, e.g. with transformer models (Vaswani et al., 2017), and investigate other ways in which such a separation can be softened while maintaining good compositional generalization performance.

4.7 Acknowledgments

We would like to thank reviewers for their thorough comments and useful suggestions and references. We would also like to thank all of the members of the Computational Cognitive Neuroscience Lab at UC Davis and the Analogy Group for helpful ongoing discussion on these topics. This work was supported by ONR N00014-19- 1-2684 / N00014-18-1-2116, ONR N00014-14-1-0670 / N00014-16-1-2128, and ONR N00014-18-C-2067.

Chapter 5

Systematicity Emerges in Transformers when Abstract Grammatical Roles Guide Attention

Ayush Chakravarthy* [1], Jacob Russin* [2], Randall C. O'Reilly[1,2] (* denotes equal contribution)

The original version of this article (Chakravarthy et al., 2022) was accepted for publication at the Student Research Workshop at the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2022). The opinions expressed here are the author's own and do not necessarily reflect the views of the conference, workshop, or publisher. The original version is available online at `https://aclanthology.org/2022.naacl-srw.1/`.

5.1 Abstract

Systematicity is thought to be a key inductive bias possessed by humans that is lacking in standard natural language processing systems such as those utilizing transformers. In this work, we investigate the extent to which the failure of transformers on systematic generalization tests can be attributed to a lack of linguistic abstraction in its attention mechanism. We develop a novel modification to the transformer by implementing two separate input streams: a role stream controls the attention distributions (i.e., queries and keys) at each layer, and a filler stream determines the values. Our results show that when abstract role labels are assigned to input sequences and provided to the role stream, systematic generalization is improved.

5.2 Introduction

Transformers have achieved state-of-the-art performance on many natural language processing (NLP) tasks (Brown et al., 2020; Devlin et al., 2019; Vaswani et al., 2017), but it has been suggested that they remain inferior to human language learners when it comes to sample efficiency (Linzen, 2020) and more difficult generalization problems (Baroni, 2020; Lake & Baroni, 2018; Lake et al., 2019a; Keysers et al., 2020). These architectures have proven to scale remarkably well (Brown et al., 2020), but may lack the strong inductive biases that contribute to these human abilities (Battaglia et al., 2018; Lake et al., 2017).

[1] Department of Computer Science, University of California, Davis
[2] Center for Neuroscience, University of California, Davis

Systematicity, or the capacity to leverage structural or grammatical knowledge to compose familiar concepts in novel ways (Fodor & Pylyshyn, 1988; Smolensky, 1990), has been highlighted as one potential inductive bias present in humans (Lake et al., 2019a; O'Reilly et al., 2021a) that deep learning architectures may lack (Lake & Baroni, 2018; Lake et al., 2017). It has been argued that in humans, the ability to understand sentences such as "John loves Mary" necessarily implies the ability to understand certain other sentences, e.g., those that are constructed from the same elements and grammatical relations such as "Mary loves John" (Fodor & Pylyshyn, 1988).

The SCAN dataset (Lake & Baroni, 2018) was introduced to evaluate the systematic generalization capabilities of deep neural networks. In SCAN, instructions generated from an artificial grammar must be translated into action sequences, and train-test splits require models to generalize to novel compositions of familiar words. Although deep learning models achieve good generalization performance when train and test data are split randomly, their performance suffers on these systematic generalization tests (Lake & Baroni, 2018), even though humans perform well on analogous generalization problems (Lake et al., 2019a).

The mechanisms underlying human systematicity remain unclear, but a number of candidates have been proposed, including tensor-product representations (Schlag et al., 2019; Smolensky, 1990) and specialized attention mechanisms (Goyal et al., 2019; Bengio, 2017; Russin et al., 2020a; Webb et al., 2021). Attention is central to the transformer architecture (Vaswani et al., 2017) and has been leveraged in mechanisms resembling systematic symbolic processing (Graves et al., 2014; Webb et al., 2021), thus making it a key potential target for encouraging systematicity (Russin et al., 2020a).

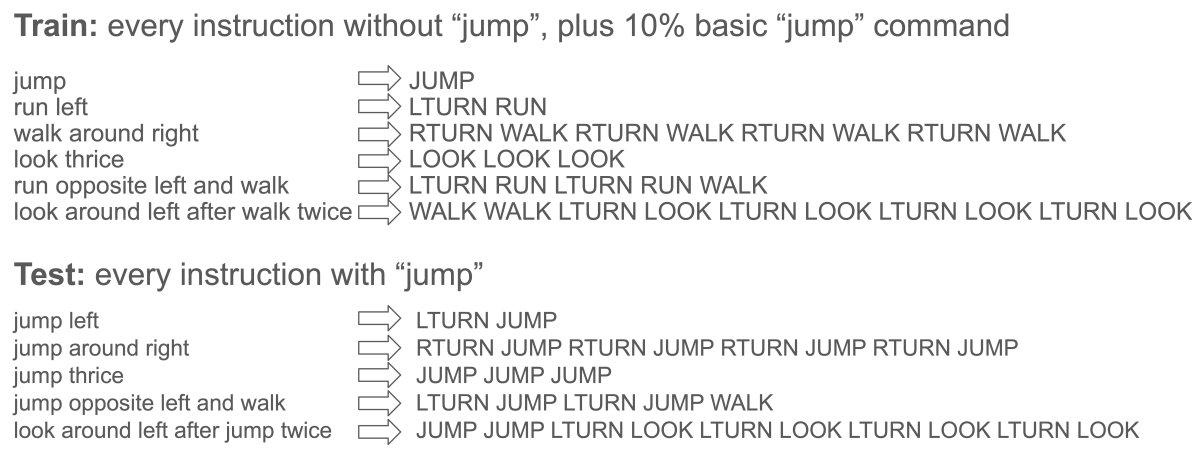

Figure 5.1: Examples from the add-jump split of SCAN. All except the simplest instructions with the word "jump" are held out of the training set, requiring models to generalize its usage to more complicated constructions.

In this work, we explore the connection between attention and systematicity using a novel transformer

architecture designed to leverage structural or abstract information in its attention mechanism. We hypothesized that systematicity would improve if attention distributions in the transformer were strictly determined from abstract inputs containing minimal token-specific information, as this may prevent memorization of spurious relationships in the training data. Previous work has experimented with incorporating additional linguistic inputs into NLP systems (e.g., Sachan et al., 2021), but here we propose a novel way of utilizing additional linguistic knowledge: a separate "role" input stream is introduced to the transformer, which determines the attention distributions at each layer but is kept separate from the typical ("filler") input stream used to directly generate outputs. Many kinds of information can be passed to the role input stream (including the original tokens themselves), thereby allowing us to explore the kinds of inputs that, when used to determine attention, result in improved systematicity. In our preliminary work, we explore the use of abstract grammatical roles to determine attention in the transformer on the SCAN dataset.

5.3 Related Work

5.3.1 SCAN

The SCAN dataset (see Figure 5.1) uses a simple finite phrase-structure grammar to generate instruction sequences that must be translated into sequences of actions (Lake & Baroni, 2018). In the *simple split*, train and test examples are sampled randomly from the set of all possible instructions. In the systematic generalization test called the *add-jump split*, all instruction sequences containing one of the primitive verbs ("jump") are systematically held out of the training set, except in its simplest form ("jump" $\rightarrow$ JUMP). The original work showed that recurrent neural networks such as long short-term memory (LSTM) succeed at the simple split but fail on the add-jump split (Lake & Baroni, 2018).

Subsequent work introduced a new framework for generating systematic generalization tests called distribution based compositionality assessment, and showed that transformers perform poorly on these tests in addition to the original add-jump split (Keysers et al., 2020). Although standard deep learning architectures consistently fail at this task, a number of non-standard approaches have demonstrated some success, including a meta-learning (Lake, 2019), recurrent networks that factorize alignment and translation (Russin et al., 2020a) or are designed for primitive substitution (Li et al., 2019), masked language model pretraining (Furrer et al., 2021); iterative back-translation (Guo et al., 2020), use of analytic expressions (Liu et al., 2020), and auxiliary sequence prediction (Jiang & Bansal, 2021). Our preliminary work presents a new approach that has many commonalities with these previous ideas.

5.3.2 Utilizing Linguistic Knowledge

Prior work has shown that a remarkable amount of linguistic structure emerges in the representations learned by large transformers self-supervised on natural language (Linzen & Baroni, 2021; Manning et al., 2020; Tenney et al., 2019), and that transformers can learn to approximate a compositional process for solving math problems (Russin et al., 2021a). These findings may cast doubt on the idea that injecting explicit linguistic structure will aid these models in producing the kinds of systematic behavior observed in human language learners. However, given their poor systematic generalization performance observed on tasks like SCAN (Lake & Baroni, 2018), and their reliance on certain syntactic heuristics that lead to predictable failures on challenging sentences (McCoy et al., 2019; Linzen & Baroni, 2021), it stands to reason that these models may benefit from access to explicit linguistic knowledge (Sachan et al., 2021).

Some work has attempted to incorporate linguistically-informed labels such as part-of-speech tags or syntactic parses into the inputs or training regiments of deep learning models (Sachan et al., 2021; Sennrich & Haddow, 2016; Strubell et al., 2018), showing some improvements on machine translation (Sennrich & Haddow, 2016) and semantic role labeling (Strubell et al., 2018). A number of methods have been used to inject linguistic knowledge into these models, including the use of graph neural networks (Marcheggiani & Titov, 2017; Sachan et al., 2021) and multi-task learning (Strubell et al., 2018). In this work, we develop a novel approach that attempts to establish an explicit link between linguistic structure and the attention mechanism of transformers to improve their systematic generalization capabilities.

5.4 Methods

5.4.1 Architecture

The transformer architecture (Vaswani et al., 2017) utilizes multi-head attention layers that take as input query (Q), key (K), and value (V) vectors:

$$\text{Attn}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{5.1}$$

where d_k is the dimension of the keys (K). Note that the probability distribution over the sequence length produced by the softmax is determined by the queries (Q) and keys (K) alone. We modified the existing transformer architecture by separating two streams of processing (see Figure 5.2): 1) the "filler" stream determines the values at each layer, which will be averaged according to the weights given by the attention distributions and contribute directly to the output of the model, and 2) the "role" stream determines at each layer the queries (Q) and keys (K) — and therefore the attention distributions — but otherwise does

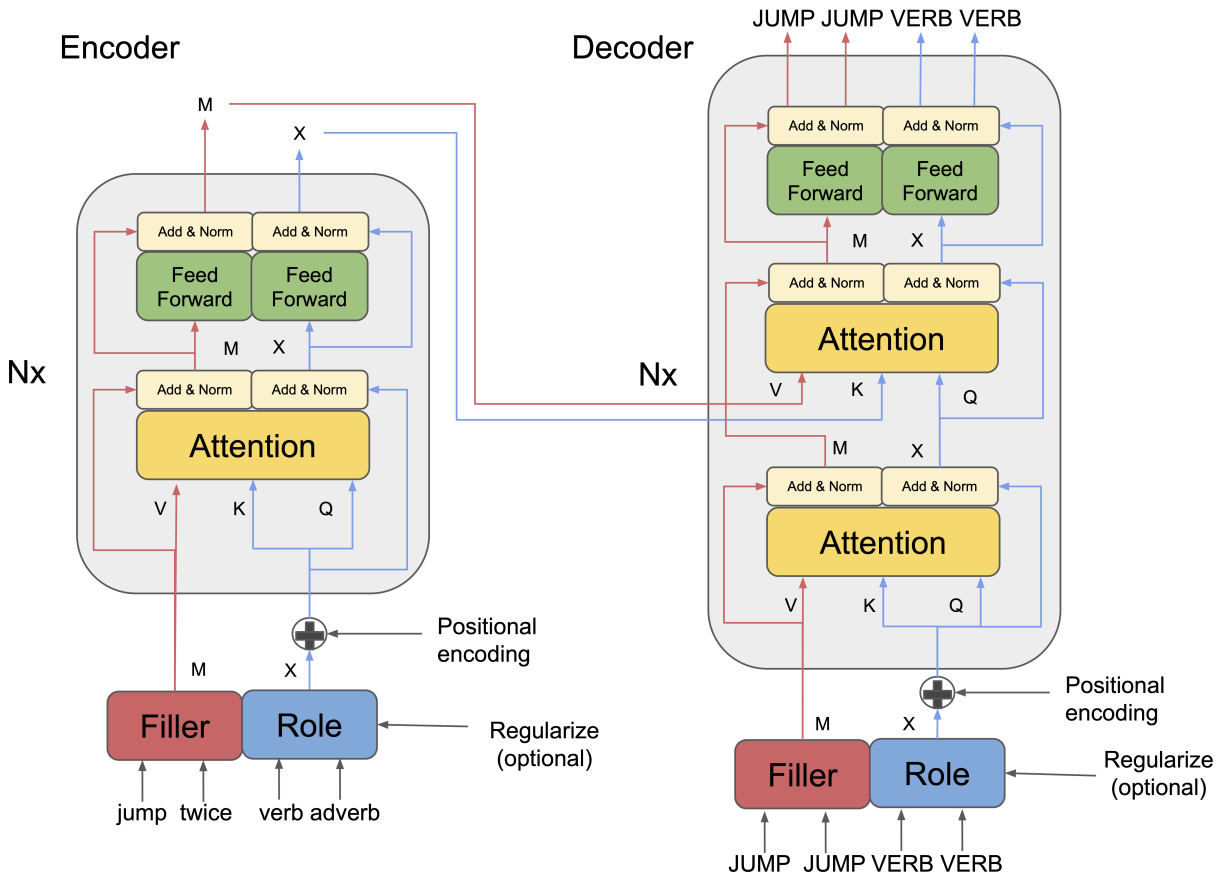

Figure 5.2: Modified transformer architecture. The architecture imposes two separate role and filler streams throughout the encoder (left) and decoder (middle). The filler stream determines the values (V) at each layer while the role stream determines the keys (K) and queries (Q), and therefore the attention distributions. This was accomplished by modifying the original attention mechanism (right).

not directly contribute to the output of the model. This was achieved by introducing a separate set of embeddings for each input stream (M for the fillers and X for the roles). The existing attention mechanism was modified so that the roles in layer $l+1$ are determined from a weighted combination of the keys in layer l:

$$\begin{aligned} M &= \text{Attn}(Q, K, V) \\ X &= \text{Attn}(Q, K, K) \end{aligned} \tag{5.2}$$

This ensures that no information from the filler stream can enter into the determination of the attention distributions at each layer, and that the roles can only affect the output of the model through their control over the attention, similar to Russin et al. (2020a). The attention at each layer can have multiple heads in the usual way (Vaswani et al., 2017), and the separation between the two streams is maintained throughout

both the encoder and the decoder (see Figure 5.2). Because the role stream determines the way information from the input tokens will be combined throughout the architecture (through its influence on the attention distributions), positional encodings are added to the role embeddings rather than the filler embeddings.

Note that this setup allows us flexibility in terms of the kind of information that is passed to the role input stream. The original tokens themselves can be embedded separately and passed to the role stream, in which case the architecture becomes very similar to the original transformer, with the exception of the modification to the attention depicted in Figure 5.2. Here, we embed abstract roles for the tokens in the SCAN dataset to investigate the relationship between abstraction in the attention mechanism and systematic generalization behavior.

5.4.2 Role Auxiliary Loss

Each transformer layer returns two sets of vectors (X and M). The output of the filler stream (M) is a sequence of target predictions that are used to compute the usual cross entropy loss before backpropagation ("Filler loss"). The output of the role stream (X) can optionally be used in an auxiliary cross-entropy loss on the roles assigned to the target sequence ("Role loss"). We performed experiments with and without this auxiliary loss, and results are reported for both.

5.4.3 Thresholded Attention

Drawing inspiration from Rahaman et al. (2021), we also experimented with thresholding the encoder-decoder attention:

$$\text{threshold}(A_{ij}) = \begin{cases} A_{ij} & \text{if } A_{ij} > \tau \\ 0 & \text{otherwise} \end{cases} \tag{5.3}$$

Where τ is the attention threshold and $A = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})$. The thresholded attention matrix is then re-normalized and multiplied by the value matrix as in equation 5.1.

5.4.4 Implementation Details

The encoder and decoder had 2 layers with 8 attention heads and used a thresholding parameter (τ) of 0.08. The embedding dimension was 256, the hidden dimension was 512, and the dimension of the query, key and value vectors was 256. The model was optimized for 400 epochs using Adam (Kingma & Ba, 2015) with a learning rate of 2.5×10^{-4}. Experiments were performed using both absolute positional encodings (Vaswani et al., 2017) and relative positional embeddings (Dai et al., 2019); absolute positional encodings were found to lead to slightly better performance with reduced variance, so for simplicity we only report those results.

Model	Simple	Add jump
LSTM+Attn (Keysers et al., 2020)	99.9 ± 2.7	0.0 ± 0.0
Syntactic Attention (Russin et al., 2020a)	100.0 ± 0.0	78.4 ± 27.4
CGPS-RNN (Li et al., 2019)	99.9 ± 0.0	98.8 ± 1.4
T5-11B (Furrer et al., 2021)	X	98.3 ± 3.3
Semi-Sup (Guo et al., 2020)	X	100.0 ± 0.0
LANE (Liu et al., 2020)	100.0 ± 0.0	100.0 ± 0.0
Aux. seq. (Jiang & Bansal, 2021)	X	98.32 ± 0.3
Transformer	100.0 ± 0.0	0.19 ± 0.18
Filler loss, no thresh (ours)	99.9 ± 0.01	16.2 ± 25.1
Filler loss, thresh (ours)	99.9 ± 0.01	85.6 ± 1.15
Filler + Role loss, no thresh (ours)	99.9 ± 0.02	87.4 ± 5.6
Filler + Role loss, thresh (ours)	100.0 ± 0.0	92.7 ± 3.3

Table 5.1: Performance (average accuracy ± standard deviation) on the simple and add-jump splits of SCAN.

5.5 Experiments

To test our hypothesized link between attention, linguistic abstraction, and systematic generalization, we developed abstract roles to label each token in the SCAN vocabulary, and performed experiments testing our architecture with and without these abstract roles. We report results on the difficult add-jump split of the SCAN dataset, and compare against previous work. Our main purpose is to show that systematic generalization is improved in the transformer when linguistic abstractions are used as inputs to the role stream for determining attention, and that there is an asymmetry in the transformer such that these abstractions should be used to determine attention (i.e., keys and queries) and not to directly produce outputs (i.e., values).

5.5.1 SCAN Roles

The phrase-structure grammar used in SCAN is very simple, so the grammatical roles used as additional inputs were relatively straightforward to implement. In the case of the add-jump split, we hypothesized that the best abstract role scheme would be one that assigned all primitive verbs to a single role ("prim") in both the instructions (source) and the actions (target). Except where indicated (section 5.5.2.2), all results used this scheme.

5.5.2 Results

Our main results are shown in Table 5.1. We reproduce previous work and show that the baseline transformer (Vaswani et al., 2017) achieves perfect accuracy on the simple split of the SCAN dataset, but fails dramatically on the add-jump split testing its systematic generalization capabilities. Our architecture improves

performance on the add-jump split when the role labels are used as inputs to the role stream. Marginal improvement relative to baseline was observed without the use of attention thresholding and without back-propagating the auxiliary role loss ("Filler loss, no thresh"). Each of these two tweaks improved performance ("Filler loss, thresh", "Filler + Role loss, no thresh") and when both were used ("Filler + Role loss, thresh"), the architecture achieved 92.7% accuracy on the test set of the add-jump split.

5.5.2.1 Abstraction in Roles vs. Fillers

To further investigate the connection between attention and systematicity, we varied the inputs used in each of the filler and role streams of the architecture (see Table 5.2). When the filler tokens (i.e., the words from the original SCAN vocabulary) were used as inputs to both the role and filler streams, our architecture resembled the original transformer architecture, as these inputs were used to simultaneously determine the outputs (i.e., the values) and the attention (i.e., the keys and queries) at each layer. This was confirmed in the performance on the SCAN task, where using the fillers in both streams ("Fillers-Fillers") resulted in similar performance to the baseline transformer.

Model	**Simple**	**Add jump**
Transformer	100.0 ± 0.0	0.19 ± 0.18
Fillers-Fillers	100.0 ± 0.0	2.8 ± 1.6
Roles-Fillers	100.0 ± 0.0	0.22 ± 0.16
Fillers-Roles	100.0 ± 0.0	92.7 ± 3.3

Table 5.2: Performance on the add-jump split only improved when abstract annotations were used in the role stream ("Fillers-Roles").

As a sanity check, we also reversed the role and filler inputs, so that the role labels were inputs to the filler stream and the words from the original SCAN vocabulary were used as inputs to the role stream ("Roles-Fillers"). In this case, performance again matched the baseline transformer on the add-jump split, confirming our intuition that linguistic abstractions are best used to determine attention distributions, not values.

5.5.2.2 Varying the Level of Abstraction

We believe that the previous result highlights a strength of our setup, as it allows us the flexibility to diverge from the original transformer in a continuous way by varying the amount of abstraction used in the inputs to the role stream. For example, in a natural language task it would be possible to vary the kinds of abstract labels or annotations supplied as input to the role stream from highly abstract part-of-speech tags to more complex annotations from more sophisticated automated parses.

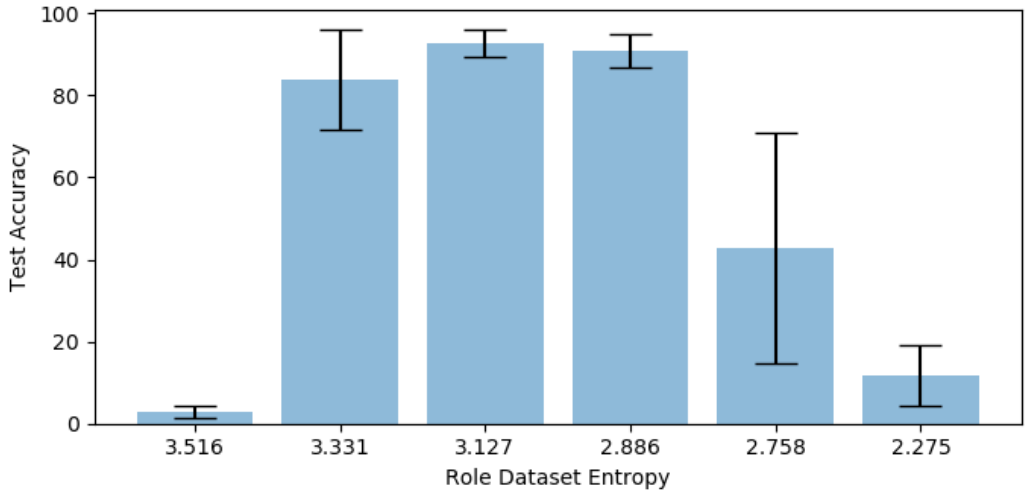

Figure 5.3: Add-jump performance varies with the level of abstraction in the inputs to the role stream (highest performance outlined in red).

To test this idea in the SCAN setting, we experimented with different schemes for assigning roles that varied in their level of abstraction, as measured by the empirical entropy of the resultant source role vocabulary (see Figure 5.3). After our initial role-assignment scheme, we made roles progressively more abstract by assigning additional instruction words to the same role (e.g., "left" and "right" to "dir", "twice" and "thrice" to "num", etc.). Results validated the assumption that the best scheme was one that used a single role for each of the primitive verbs, and assigned a different role to each of the other words (entropy = 3.127). This experiment shows that there is an ideal level of abstraction to use in the role stream: too much abstraction results in an inability to distinguish relevant distinctions, and too little results in the unsystematic memorization typical of the vanilla transformer.

5.6 Conclusion

Our preliminary work establishes a connection between linguistic abstraction, the attention mechanism used in transformers, and systematic generalization behavior as measured by performance on the SCAN dataset: when abstract roles are assigned to inputs and used to determine the attention at each layer, systematic generalization improves. We developed an architecture that may facilitate greater understanding of the original transformer (Vaswani et al., 2017) by allowing more precise investigation into the relative contributions of attention distributions and representation learning. Future work will test our setup on other compositional or systematic generalization tasks (Keysers et al., 2020; Kim & Linzen, 2020) and determine the kinds of linguistic abstraction that allows success on these tasks. In addition, future work will

experiment with using our novel architecture on natural language datasets using varying levels of linguistic abstraction.

The extent to which human-level language understanding requires stronger inductive biases than those currently implemented in deep learning systems remains an open question. Our work shows that utilizing linguistic abstraction in the attention mechanism of transformers may be a promising approach for improving the systematic generalization capabilities of deep neural networks.

5.7 Acknowledgements

We would like to thank the members of the Computational Cognitive Neuroscience lab at UC Davis, Paul Smolensky, Roland Fernandez, and other members of the Deep Learning Group at Microsoft Research, as well as reviewers for helpful comments and discussions. The work was supported by: ONR grants ONR N00014-20-1-2578, N00014-19-1-2684 / N00014-18-1-2116, N00014-18-C-2067.

Chapter 6

Complementary Structure-Learning Neural Networks for Relational Reasoning

Jacob Russin*[1], Maryam Zolfaghar*[1], Seongmin A. Park[2], Erie Boorman[2], Randall C. O'Reilly[1] (* denotes equal contribution)

The original version of this article (Russin et al., 2021b) was accepted for publication in the Proceedings for the 43rd Annual Meeting of the Cognitive Science Society (CogSci 2021). The opinions expressed here are the author's own and do not necessarily reflect the views of the conference, workshop, or publisher. The original version is available online at `https://arxiv.org/abs/2105.08944`.

6.1 Abstract

The neural mechanisms supporting flexible relational inferences, especially in novel situations, are a major focus of current research. In the complementary learning systems framework, pattern separation in the hippocampus allows rapid learning in novel environments, while slower learning in neocortex accumulates small weight changes to extract systematic structure from well-learned environments. In this work, we adapt this framework to a task from a recent fMRI experiment where novel transitive inferences must be made according to implicit relational structure. We show that computational models capturing the basic cognitive properties of these two systems can explain relational transitive inferences in both familiar and novel environments, and reproduce key phenomena observed in the fMRI experiment.

6.2 Introduction

Humans and non-human animals are capable of navigating efficiently in both novel and familiar environments. For example, in a well-learned environment like one's hometown, it is easy to navigate to new goal locations and plan novel routes. When traveling in a new city, it is also possible to navigate to a novel location by reasoning over recent experiences — even those accumulated on the same day. In both cases, efficiency requires processes or representations that allow generalization beyond previous experience. This kind of generalization has been a long-standing issue in cognitive science, and was integral to early arguments

[1] Center for Neuroscience, University of California, Davis
[2] Center for Mind and Brain, University of California, Davis

against behaviorism, where it was claimed that a simple stimulus-response mapping could not account for such behaviors (Tolman, 1948).

More recent work has investigated the computational and neural mechanisms underlying cognitive maps, or representations that capture the structure of the environment and thereby support generalization (Park et al., 2020b; Whittington et al., 2020; Behrens et al., 2018). This work has emphasized the importance of certain neocortical areas such as the entorhinal cortex (EC) for spatial reasoning and vector-based navigation (Moser et al., 2008). Furthermore, it has been argued that these structured spatial representations may be leveraged for other kinds of abstract relational reasoning in humans (Behrens et al., 2018). Relatedly, although neural networks have enjoyed massive success on difficult machine-learning tasks in recent years these models are known to fail on out-of-distribution or extrapolation problems (Lake et al., 2017) such as those requiring transitive inferences.

Here, we apply the well-supported complementary learning systems (CLS) framework (McClelland et al., 1995; O'Reilly et al., 2011) to explore two qualitatively different neural mechanisms underlying spatially-grounded relational reasoning abilities in novel and familiar environments. The CLS framework has emphasized the computational justification for learning mechanisms unfolding on two different timescales, as supported by separate brain areas. Slow learning in neocortex allows for the development of more abstract representations that integrate across many experiences and can be leveraged to make novel inferences. However, this kind of learning is not possible in naturalistic environments where sequences of events are not presented in an interleaved or random order, as when one explores only one part of an environment at a time. This is due to the well-known *catastrophic forgetting* phenomenon, where previous learning is erased by new experiences when learning occurs too quickly or training is not sufficiently interleaved (McClelland et al., 1995). The CLS framework proposes that fast learning can occur in the hippocampus due to its pattern-separated, sparse representations. These representations have little overlap across examples, and therefore allow fast learning of novel episodes, i.e., *episodic memory* (Yonelinas et al., 2019), to occur without catastrophic interference.

In the CLS framework, slow cortical learning is needed to build up structural or relational representations over time, which provide the foundation for systematic inferences. However, for more unfamiliar situations, rapid hippocampal learning is required. Previous work has found evidence suggesting a role for the hippocampus in rapid generalization (Eichenbaum, 2004; Zeithamova et al., 2012), and that a hippocampal model informed by the CLS framework can explain these findings when it is augmented with a recurrent similarity-based computation, proposed to be supported by "big-loop" recurrence between the hippocampus and the neocortex and within the hippocampus itself (Kumaran & McClelland, 2012).

Here, we build on this work and investigate the interplay between slow generalization in neocortex and

rapid generalization in the episodic memory system with computational models based on the principles of the CLS framework. Our model of the episodic memory system is similar to previous work (Kumaran & McClelland, 2012) in that it allows rapid generalization in unfamiliar environments, but relies on different computational mechanisms to do so (see Discussion). Our models of the cortical system and the episodic memory system were both tested on a novel non-spatial structure-learning paradigm from a recent fMRI experiment (Park et al., 2020b). Importantly, the task required transitive inferences based on learning over two different timescales: training experience over multiple days, and training examples given on the same day as the inference test. In the following, we briefly outline the key findings of the experiment and offer a conceptual framework that integrates them with the CLS perspective. We then describe the computational models that were built to capture the basic properties of the proposed conceptual framework, and show that these models are capable of performing transitive inferences in the same task and reproduce other key findings from Park et al. (2020b).

6.2.1 fMRI Experiment

Park et al. (2020b) studied the neural mechanisms underlying transitive inference performance on the structure-learning task illustrated in Figure 6.1. Participants learned to make judgments about the "popularity" or "competence" of 16 people through pair-wise comparisons along one of these two axes at a time. Unknown to the participants, these 16 faces were arranged in a 4x4 2D grid, and were implicitly separated into two groups. In the first two days of training participants only learned about within-group pairs that different by a rank of 1 (see Figure 6.1A). On the third day of the experiment, participants learned about between-group pairs containing certain faces that acted as hubs between the two groups (see Figure 6.1B). This training provided sufficient evidence to allow participants to integrate their previously separated cognitive maps, but was conducted on the same day as fMRI scanning. In the scanner, participants performed a transitive inference test in which unseen pairs of faces from different groups were compared (see Figure 6.1C). For each of these test pairs, one of two corresponding hubs could be used to make the transitive inference. The results we focused on in our work can be summarized as follows:

1. Participants exhibited good transitive inference performance, achieving 93.6% mean accuracy on the unseen pairs tested in day 3.
2. Map-like representations were found in several brain areas, including ventromedial prefrontal cortex (vmPFC) and entorhinal cortex (EC). Patterns of activity in these areas demonstrated sensitivity to the ground-truth Euclidean distances between faces in the implicit grid. However, these effects were significantly reduced when the analysis was restricted to between-group pairs that were not encountered

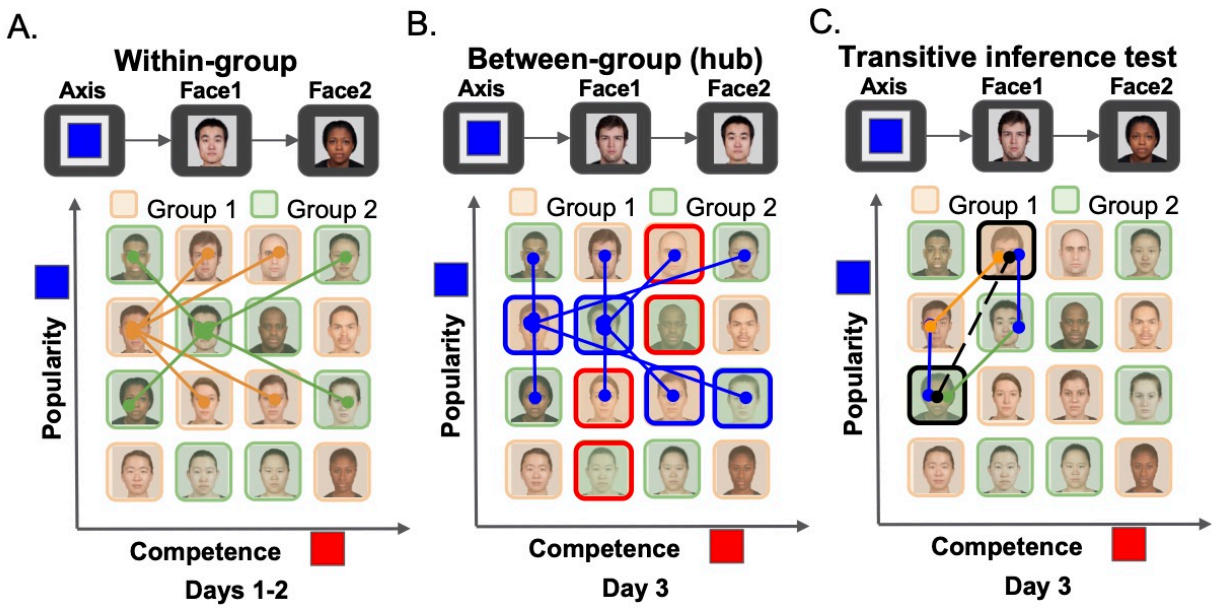

Figure 6.1: Experimental paradigm used in Park et al. (2020b). Participants learned the relative ranks of pairs of faces on an implicit grid with two axes: competence and popularity. These faces were split into two groups (shown in green and orange). **A)** Over the first two days, participants were trained on within-group pairs that differed by a rank of 1 along each designated axis. Example pairs on the popularity axis are shown with orange lines (within group 1) and green lines (within group 2). **B)** On the third day, participants learned between-group pairs containing exactly one hub linking the two groups. There were a total of 8 hubs, and each was associated with a certain axis (shown by red and blue outlines). Each hub was paired with the 4 faces from the other group that differed by a rank of 1 along the designated axis. Examples of such pairs are shown for two hubs with blue lines (indicating the popularity axis). **C)** The third day included the fMRI experiment and transitive inference test. Participants were tested on pairs of faces from different groups that could be connected through one of the two hubs on the appropriate axis. Green, orange, and blue lines indicate the training pairs (2 within-group and 2 between-group, which are shown in A and B) that could be used to make the transitive inference for the pair indicated by the black dotted line.

during training.

3. A repetition-suppression analysis in hippocampus suggested that one of the two relevant hubs was retrieved from episodic memory at the time of inference.

Taken together, these findings suggest that cortical learning systems in vmPFC and EC were able to integrate across the pairs of faces encountered during training to form map-like representations that would be useful for making transitive inferences within groups. However, the effects in these areas were reduced when the analysis was restricted to novel between-group pairs, and participants seemed to retrieve the relevant hubs from episodic memory in hippocampus during the transitive inference test. Thus, although the within-group pairs were well-learned over the first two days of training, these groups may not have been fully integrated into a single coherent cognitive map at the time of testing. This may have forced

participants to rely instead on hippocampal retrieval of recently-learned between-group training episodes (which always included a hub) to generalize during the transitive inference test. Thus, there appear to be two separable cognitive mechanisms that allow for relational transitive inferences to be made in this task: 1) if given enough training time, cortical areas such as vmPFC and EC can learn representations that reflect the implicit relational structure of the grid, and 2) an episodic retrieval mechanism can ensure good transitive inference performance with pairs that were seen only on the same day as the test. Below we outline a general framework that integrates these findings, and the apparent redundancy in these two systems, with the CLS perspective.

6.3 Complementary Structure-Learning Systems

The CLS framework explains how the brain can support integrative representation learning without suffering from catastrophic forgetting (McClelland et al., 1995; O'Reilly et al., 2011). However, the CLS framework also emphasizes other important reasons for fast learning in an episodic memory system. In particular, slow cortical learning may be insufficient to allow for efficient adaptation in relatively unfamiliar environments (Kumaran & McClelland, 2012). The findings from Park et al. (2020b) suggest that humans are capable of making novel transitive inferences using experiences acquired on the same day. Furthermore, they show that these inferences are mediated by hippocampal retrieval of the intermediate states (i.e., hubs) that would allow such inferences to occur. Taken together, these findings suggest that the dual-process view emphasized in CLS may explain the apparent redundancy in structure-learning mechanisms studied in neuroscience and psychology (see Table 6.1).

Table 6.1: Complementary structure-learning systems.

System	Properties
Cortical learning	• Learns slowly through small, incremental weight changes • Inference is fast and less effortful with map-like representations
Episodic memory	• Learning can be fast due to sparse, pattern-separated representations • Inference is slower, requiring cognitive control for deliberate, goal-directed retrieval

In the case of spatial navigation, slow cortical learning can integrate across many experiences to form map-like representations. This system is capable of directly utilizing its integrative representations without further processing, and can thus make inferences rapidly. However, this system would not be able to make

inferences in a newly learned environment if it did not have time to integrate across particular episodes (Kumaran & McClelland, 2012). This may have been the case in the transitive inference test conducted on the same day as the between-group training in the fMRI study (Park et al., 2020b). Fast episodic learning, on the other hand, can immediately store memories of individual experiences, allowing inferences to be made in unfamiliar environments based on few such experiences. However, the episodic nature of its representations do not allow the sort of direct inferences that are available to the cortical system. Instead, transitive inferences require a slower, more deliberate process of goal-directed retrieval and further processing of the stored memories (Zeithamova et al., 2012). An organism equipped with both systems would be capable of making novel inferences in both familiar and unfamiliar environments. In the following, we provide evidence from models that capture, on a *computational* level, the basic properties of the proposed complementary structure-learning systems, and show that these systems reproduce key findings from Park et al. (2020b).

6.4 Modeling Framework

We simulated[3] each of our models on the training and testing procedure used in the task, including its within-group and between-group structure and transitive-inference test. In particular, each trial consisted of a presentation of two faces and the axis along which the judgment should be made (i.e., "competence" or "popularity"). The models were required to make a binary judgment about whether the first face ranked higher or lower than the second face along the specified axis.

6.4.1 Cortical Map-Building

The cortical representation-learning system should accumulate small updates over many trials to build map-like representations that can be directly utilized to make transitive inferences. We modeled this process with a simple feedforward neural network with two convolutional layers (see right side of Figure 6.2). Face images were taken from the same database used in the fMRI experiment (Strohminger et al., 2016), and were downsampled to 64x64 and grayscaled for faster simulation. The within-group and between-group hub samples were all trained simultaneously (i.e., the pairs that were trained on different days of the fMRI experiment were trained simultaneously in the model). This is because the purpose of our model of the cortical system was to show that, if given enough training time, it could perform transitive inferences based on its learned representations, and allow fast inference in familiar environments. Each face was processed with the same convolutional neural network, and the axis variable, encoded as a one-hot vector, was embedded with a linear layer. These three embeddings were then concatenated and passed through a multi-layer

[3]All data and code used for experiments and analyses are available at https://github.com/MaryZolfaghar/CSLS

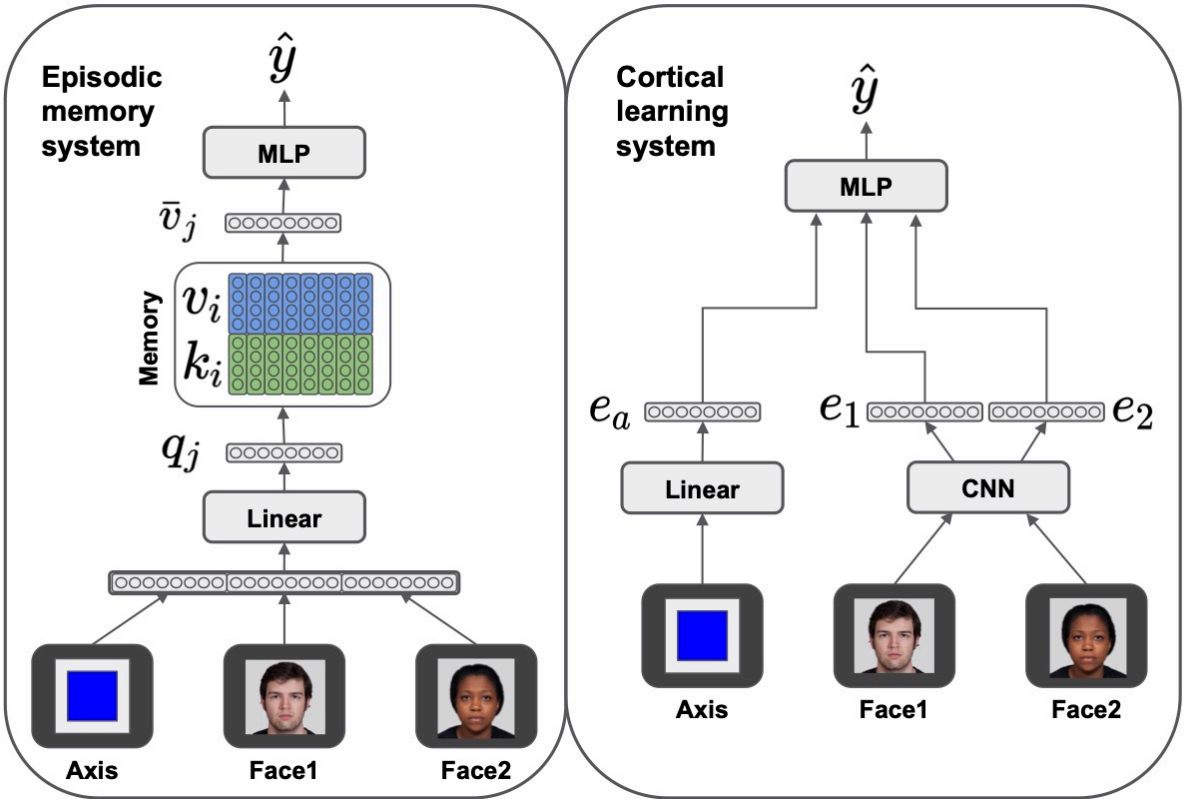

Figure 6.2: Model architecture. (Left) The episodic memory system stores representations of individual training trials in a key-value memory. New inferences are made by querying the memory to retrieve the relevant trials, which are then processed by an MLP to generate an answer. (Right) The cortical learning system was modeled as a simple feedforward network with convolutional layers to process the images. This system relies on its learned representations to perform transitive inferences.

perceptron (MLP) with rectified linear unit (ReLU) activation functions. This network captures the basic properties of slow cortical learning in that it accumulates small updates to its synaptic weights over many trials, and makes inferences directly based on its learned representations of each face.

6.4.2 Goal-Directed Episodic Memory Retrieval

The episodic memory system should learn quickly by storing individual training episodes, and make inferences by retrieving the previous trials that are relevant to the current one (McClelland et al., 1995; Kumaran & McClelland, 2012). For this purpose, we used a neural memory system (see left side of Figure 6.2) with a soft retrieval mechanism (Botvinick et al., 2019a). This memory system immediately stores each trial (x_i) seen during training as a key, value pair: $k_i = W_k x_i$, $v_i = W_v x_i$, where k_i is the key, v_i is the value, and x_i is the trial, which is a concatenation of a one-hot encoding of each face, the axis variable (a), and the correct answer (y) of the ith trial. One-hot encodings were used for faces under the assumption that what is stored

in the episodic memory system should be a highly processed, sparse encoding (McClelland et al., 1995). To make an inference, the model generates a query according to the current pair of faces: $q_j = W_q x_j^- + b_q$, where x_j^- indicates the jth test trial with the same components but excludes the correct answer (y). This query is then used to retrieve the memories most relevant to the current trial:

$$\bar{v}_j = \text{softmax}\left(q_j K^T\right) V \tag{6.1}$$

where K and V are matrices containing all of the stored memories. Finally, the retrieved memories $\bar{v}_j$ are passed through an MLP to produce the final answer: $\hat{y_j} = \text{MLP}(\bar{v}_j)$. This network captures the basic properties of a fast-learning episodic memory system in that each training episode can be stored in memory immediately upon presentation, and must later be retrieved in a goal-directed way to make a transitive inference.

An interesting problem in modeling episodic memory concerns the learning mechanisms involved in goal-directed memory retrieval. We assume that the human participants recruited for the Park et al. (2020b) study had extensive prior experience with goal-directed memory retrieval and everyday transitive inferences. We therefore adopted a meta-learning strategy (Santoro et al., 2016) to model this prior experience, and pretrained the episodic memory system to learn to solve new transitive inference problems sampled from a distribution of such tasks. This pretraining consisted of slow, incremental changes to the weights responsible for mapping into and out of the episodic memory itself, and should thus be thought of as occurring in memory-related cortical areas rather than in the hippocampus proper (McClelland et al., 1995). The system was pretrained on a distribution that was generated by permuting the positions of each face in the 4x4 grid. For each new task, the memory system stored training samples in its memory and used them to make transitive inferences in the testing phase, where it accumulated errors that were then used to update its learnable parameters. The model was then tested on how well it could generalize with a new configuration of faces it had never seen before.

This kind of meta-learning strategy was adopted from previous work (Lake, 2019), and shares with it the limitation that the pretraining tasks are unrealistically similar to the final test — future work will examine the extent to which the model can generalize when trained on substantially different goal-directed retrieval and transitive inference tasks. Additionally, although the resulting goal-directed retrieval mechanism in this model does not capture the hypothesized properties of being deliberative and requiring cognitive control (thus making inferences slower), a more biologically grounded approach involving frontal cortical executive function systems, planned for future work, would do so. Our purpose in the current study was to show that this system was capable of making transitive inferences in a structured environment.

6.4.3 Implementation Details

Models were built using PyTorch. Models were trained with a cross-entropy loss function and Adam optimizer (Kingma & Ba, 2015) with a batch size of 32 and a learning rate of 0.001.[1] The cortical system was trained for 100 epochs with a batch size of 32. The axis embedding (e_a) had 32 dimensions. Convolutional layers had no padding, a kernel size of 3, a stride of 2, and 4 and 8 channels in the first and second layers, respectively. Each convolutional layer was followed by a max-pooling layer with a kernel size of 2. The CNN contained a linear layer to produce flat 72-dimensional vectors e_1 and e_2, which were passed to the final MLP, which had 128 hidden units. The episodic memory system was pre-trained on 10,000 permutations. Queries, keys, and values were all 32-dimensional, and the final MLP had 64 hidden units.

6.5 Results

Both systems proved to be capable of performing transitive inferences in the task environment from Park et al. (2020b): each system achieved 100% accuracy on the held-out test set in which unseen between-group pairs were tested. This validates the idea that the two qualitatively different kinds of learning system outlined above are capable of reproducing human transitive inference performance on the task. To investigate how these qualitative differences might have affected each model's inference strategy, we performed analogues of key analyses done in the experiment (Park et al., 2020b) to interpret the behavior of each system, and to evaluate them against empirical results obtained in the fMRI experiment.

6.5.1 Cortical Representations Reflect Task Structure

To understand how the cortical system had learned to represent each of the faces, we conducted analyses on the embeddings of each face obtained from the CNN. Visualization of these embeddings with principal components analysis (PCA; see Figure 6.3) showed that the cortical system had learned to represent the faces in terms of their structured relationships, i.e., it had learned map-like representations. These top two principal components explained 95.1 % of the variance in the embeddings, indicating that the model had learned to represent the faces on a near two-dimensional grid.

In addition to the PCA, we conducted an analysis similar to those done in the fMRI experiment (Park et al., 2020b), where patterns of activity in vmPFC and EC were found to be sensitive to Euclidean distances in the ground-truth grid. We measured the Pearson correlation between ground-truth Euclidean distances in the grid and the observed distances between each pair of embeddings. A strong correlation was observed

[1] Note that the "learning rate" for the episodic memory refers to the weight updates in the pre-training phase. During training, it immediately stored experiences upon presentation.

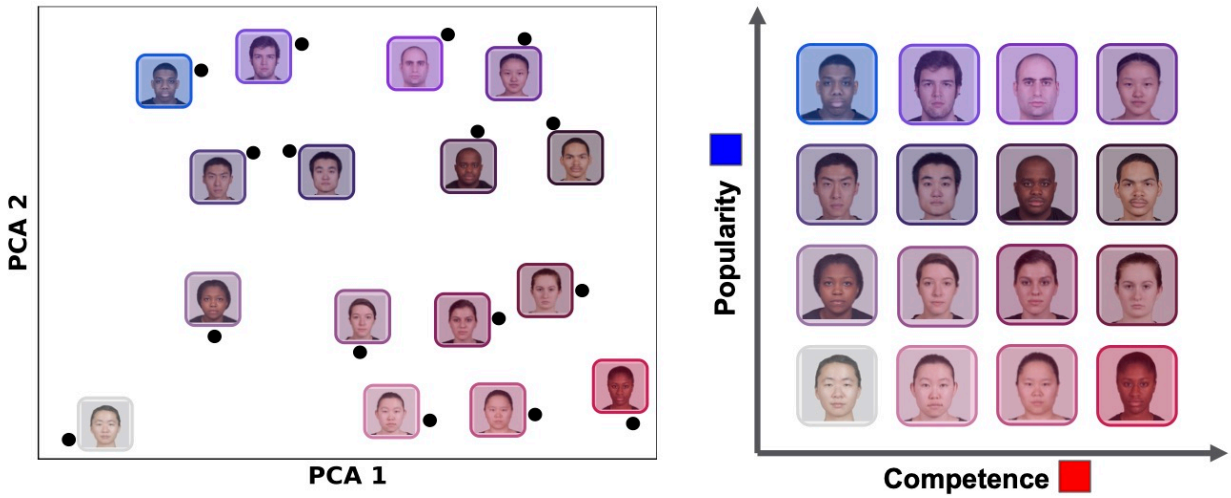

Figure 6.3: Visualization of embeddings learned by the cortical system. Embeddings for each face was projected into two dimensions using PCA, and then rotated by a fixed angle for illustration purposes. The relative positions of the representations indicate that the model has learned to represent the faces in terms of their implicit relational structure.

($r(118) = .910$, $p < 0.001$), indicating the same sensitivity to structured relationships in the grid.

6.5.2 Episodic Memory System Retrieves Hubs

In the original fMRI experiment, a repetition-suppression analysis suggested that participants were retrieving the relevant hubs from hippocampus during the transitive inference test (see Figure 6.1C). Although the episodic memory model did not have analogous neural adaptation dynamics that would allow us to model repetition suppression, we conducted an analysis on the retrieved memories to see how the hubs were being used to make transitive inferences. The soft episodic retrieval mechanism shown in equation (6.1) uses a softmax to produce a probability distribution over all of the items in memory. For each test trial, we directly analyzed the weights applied to the memories for the relevant hub trials and compared these weights to the irrelevant memories (see Figure 6.4). Memories were counted as relevant if they included one of the two possible between-group hubs for the given pair of faces, and connected this hub to one of the two faces from the current trial (see Figure 6.1C). This revealed that the weights applied to the relevant hub memories were usually the largest (i.e., the hub trials were retrieved more than the irrelevant trials). Furthermore, an additional analysis found that in every test trial, one of the two possible "paths" connecting the first face to the second face (e.g., in Figure 6.1C, the path through the blue line and green line or the path through the blue line and orange line) was in the top 5% of retrieved memories.

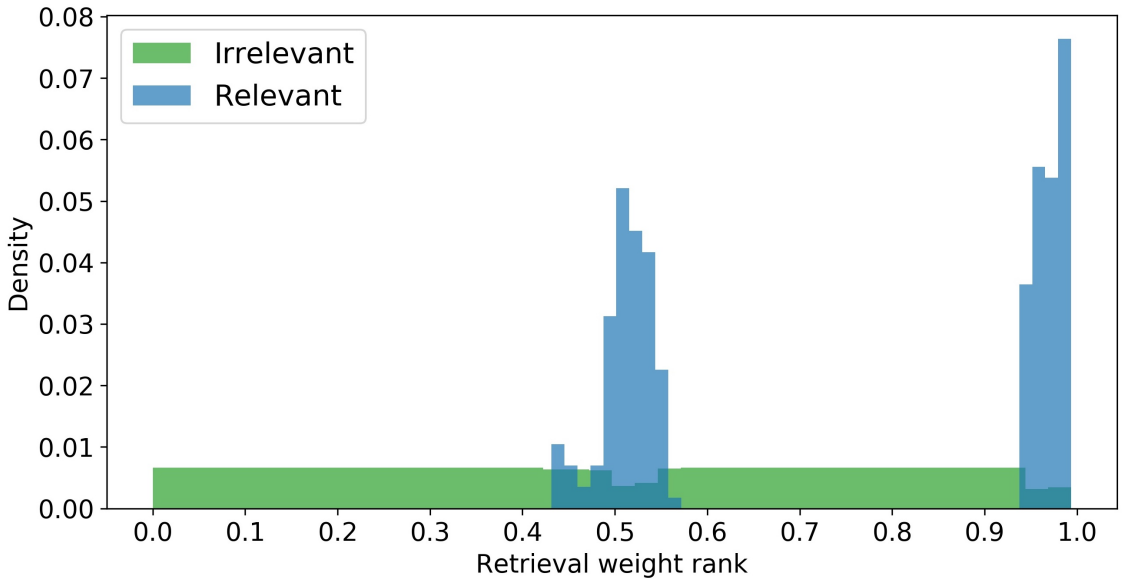

Figure 6.4: Histograms of relevant and irrelevant trials retrieved from the episodic memory during testing. Relevant memories, which always contained a hub, made up the majority of those with the highest weights. This reproduces the fMRI finding that the hubs were retrieved during the inference test. Note that it was not necessary to retrieve every relevant memory to get the correct answer, which may be why the relevant memories were not always retrieved with the highest weights. Counts were normalized to probability densities.

6.6 Discussion

The CLS perspective emphasizes the need for two qualitatively different learning systems in the brain: fast learning can occur in the hippocampus due to its pattern-separated representations, while learning in the neocortex must be slow due to its overlapping representations (McClelland et al., 1995). Here, we investigate this conceptual framework in the domain of structure-learning and relational transitive inference (Kumaran & McClelland, 2012), and propose an analogous distinction. The episodic memory system can learn quickly and generalize in relatively unfamiliar environments, but requires a more deliberate goal-directed retrieval process. The cortical system learns slowly but can make fast inferences in familiar environments from its learned representations. As in the traditional CLS framework, an organism equipped with both systems would retain the benefits of each, allowing generalization in both novel and familiar environments. Our computational models provide evidence that each of the two proposed systems are able to perform well on a difficult relational transitive-inference test under different circumstances: the cortical system can make these inferences once extensive experience with an environment has been accumulated, while the episodic system can do so quickly, as long as it has had sufficient prior exposure to similar tasks. Our models also reproduce the basic findings from a human fMRI experiment (Park et al., 2020b): the cortical system learns map-like representations that encode the implicit relational structure of the grid, while the episodic memory system

learns to query its memory for the appropriate hubs connecting the two groups.

Kumaran & McClelland (2012) investigate rapid generalization in a hippocampal model based on the principles of the CLS framework. The model allows retrieval-based inferences to be made — despite the nature of its pattern-separated representations — by incorporating a recurrent similarity computation that can perform associative linking (Eichenbaum, 2004; O'Reilly & Rudy, 2001). This computation is hypothesized to be supported by "big-loop" recurrence (Koster et al., 2018). Our model of the episodic memory system is not inconsistent with hippocampal retrieval-based inferences based on dynamic similarity computation, and in fact the fMRI experiment showed evidence of the presence of such similarity structure in the hippocampus (Park et al., 2020b). In addition, the strategy used by our model to solve transitive inference problems appeared consistent with the associative linking exhibited by the model of Kumaran & McClelland (2012), as shown by the retrieval of hubs linking the two groups (see Figure 6.4). However, in our model this strategy emerged over the course of (meta-)learning the structure of transitive inference problems, suggesting a more general mechanism that could be applied to goal-directed retrieval tasks that are not solvable with an associative linking strategy. This learning mechanism has been shown to be useful in the context of one-shot learning (Santoro et al., 2016), and compositional generalization (Lake, 2019). More work is needed to investigate whether hippocampal involvement in rapid generalization occurs when such a strategy is not possible, and whether our model would benefit from the recurrent computation intrinsic to the model of Kumaran & McClelland (2012).

Our modeling framework shares important properties with the Tolman-Eichenbaum Machine (TEM) (Whittington et al., 2020), which also incorporates meta-learning and models structure-learning in EC. A critical difference between these two models is that in TEM, structure-learning depends on backpropagating error signals through the hippocampus, whereas the CLS framework holds that slow cortical learning can operate independent of the hippocampus to facilitate inferences, consistent with the remarkably intact abilities of early developmental amnesics (Vargha-Khadem et al., 1997).

Our proposed framework integrates ongoing empirical findings about cognitive maps with the CLS perspective, but it also shares some similarities to other prominent dual-process views in cognitive science. For example, prominent theories emphasize a distinction between habitual and controlled processing (O'Reilly et al., 2020), fast and slow thinking (Kahneman, 2011) and model-free and model-based RL (Botvinick et al., 2019a). Our conceptual framework proposes a similar distinction between the deliberative, goal-directed retrieval that must occur in the episodic memory system to make transitive inferences, and the more automatic or vector-based generalization that can occur in the cortical system in familiar environments.

There are some important limitations of our current computational models that must be addressed in future work. First, although the two proposed cognitive systems are hypothesized to be realized in the hip-

pocampus and cortical areas such as EC, we have not focused on the interactions that should occur between the two systems. For example, the representations stored in episodic memory should be directly informed by the slowly changing representations learned in cortex, reflecting cortical inputs to the hippocampus. The fMRI study found that map-like representations were also present in the hippocampus (Park et al., 2020b), perhaps due to interactions with nearby cortical areas (Kumaran & McClelland, 2012). A more integrated model would show how map-like representations in cortex can influence hippocampal processing, and how reliance on the episodic memory early in learning shifts to reliance on the cortical system later in learning. This shift may occur due to the cognitive demands imposed on an episodic retrieval mechanism required to reason over individual past experiences. The current episodic memory system does not capture the cognitive control hypothesized to be required for inferences to be made; future work will address this with a more integrated model that deploys an episodic retrieval mechanism with costly sequential processing. Finally, the neural networks used in our models biologically implausible in a number of ways, e.g., the use of a slot-based episodic memory and the standard backpropagation algorithm. Future work will focus on more biologically plausible learning algorithms and more detailed biology of the neocortex and hippocampus.

6.7 Acknowledgments

We would like to thank the members of the Computational Cognitive Neuroscience lab and the Learning and Decision Making lab, as well as reviewers for helpful comments and discussions. The work was supported by: ONR grants ONR N00014-20-1-2578, N00014-19-1-2684 / N00014-18-1-2116, N00014-18-C-2067. J.R. was supported by the NIMH under Award Number T32MH112507. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Chapter 7

The Geometry of Map-Like Representations under Dynamic Cognitive Control

Maryam Zolfaghar*[1], Jacob Russin*[1], Seongmin A. Park*[2], Erie Boorman[2], Randall C. O'Reilly[1] (* denotes equal contribution)

The original version of this article (Zolfaghar et al., 2022) was accepted for publication in the Proceedings for the 44th Annual Meeting of the Cognitive Science Society (CogSci 2022). The opinions expressed here are the author's own and do not necessarily reflect the views of the conference, workshop, or publisher. The original version is available online at `https://escholarship.org/uc/item/28j425kf`.

7.1 Abstract

Recent work has shown that the brain organizes abstract, non-spatial relationships between entities into map-like representations. However, an animal's objectives often depend on only a subset of the features of the environment. Under these circumstances, cognitive control – the capacity to flexibly select the features most relevant in the current context – becomes paramount. Here, we explore the relationship between cognitive control and the geometry of map-like representations by combining fMRI with neural network modeling. We find that brain areas including hippocampus and entorhinal cortex spontaneously organize pairwise relationships into 2D map-like representations, and that this 2D structure was controlled by compressing task-irrelevant dimensions in areas of prefrontal and parietal cortex. Our neural network model reproduced these findings and additionally predicted warping in the geometry along a context-invariant axis. This prediction was confirmed with fMRI, which showed that the degree of warping was correlated with individual differences in cognitive control.

7.2 Introduction

Substantial evidence suggests the brain organizes incoming relational information into cognitive maps (O'Keefe & Nadel, 1978; Moser et al., 2008) – even when relations are abstract or non-spatial (Behrens et al., 2018; Bernardi et al., 2020; Garvert et al., 2017; Knudsen & Wallis, 2021; Park et al., 2020b; Stachenfeld et al.,

[1] Center for Neuroscience, University of California, Davis
[2] Center for Mind and Brain, University of California, Davis

2017). For example, when humans are trained on stimulus features that vary systematically (e.g., the lengths of a bird's neck or the competence and popularity of people in a social hierarchy), brain areas such as the medial temporal lobe and medial parietal cortex efficiently encode these features with map-like representations (Constantinescu et al., 2016; Park et al., 2021, 2020b). Just as geographic maps depict true distances between locations in the world, the geometry of these "map-like" representations reflects the latent structure of their underlying feature spaces, such that distances in the representational space are consistent with distances in the feature space. This map-like quality is thought to facilitate relational reasoning behavior, allowing animals to generalize to novel or unseen stimuli (Behrens et al., 2018; O'Reilly et al., 2021a; Summerfield et al., 2020; Whittington et al., 2020).

Most models of cognitive map formation consider cases where animals must navigate or reason within a single context or goal. However, animals are often faced with scenarios where multiple possible objectives can determine the subset of stimulus features that are important at any given time. These scenarios require cognitive control, or the capacity to flexibly select or attend the features of the environment that are most relevant to the current context or goal. Classic theories of cognitive control hypothesize that top-down attention mechanisms in the prefrontal cortex can dynamically modulate the processing in posterior brain regions in order to meet the demands of a current goal (Miller & Cohen, 2001; Herd et al., 2006; Rougier et al., 2005). Computational models of these processes have been used to successfully explain cognitive and neural phenomena and have emphasized their functional benefits for managing the interference caused by conflict or incongruence (Miller & Cohen, 2001; Musslick et al., 2017; Shenhav et al., 2013). These studies have focused on classic cognitive control tasks such as the Stroop task (Stroop, 1935) that use discrete exogenous stimulus features such as color or orthographic features (Cohen et al., 1990; Herd et al., 2006; Rougier et al., 2005). However, less is known about how such processes might be used to control endogenous map-like representations retrieved from memory, such as those observed in the medial temporal lobe (MTL) or parietal cortex during abstract relational reasoning tasks.

Here, we combined fMRI with a neural network model to investigate the relationship between cognitive control and endogenous map-like representations. We tested human participants and the neural network on the same task, which was designed to facilitate the learning of map-like representations while simultaneously requiring the use of cognitive control as a function of the current task context. Using parallel analyses of the neural network and human fMRI data, we observed three key phenomena related to cognitive maps, cognitive control, and their relationship:

1. Learning in both the human brain and neural network models sculpted unitary map-like representations that integrated information across multiple contexts to capture the latent relational structure of the

task space.

2. To effectively resolve the interference caused by incongruence in the task, these map-like representations were dynamically modulated by cognitive control processes such that irrelevant dimensions of the representational space were compressed according to the current context.

3. This interference, and the resultant demand on cognitive control, was related to congruence effects in the map-like representations learned over the course of the task, as measured by the degree of warping in their geometry: pattern dissimilarity between congruent stimulus pairs was greater than that of incongruent stimulus pairs.

7.3 Methods

7.3.1 Experimental Task

Participants learned about the relative ranks of 16 hypothetical people on two social hierarchy dimensions: "competence" and "popularity." Unknown to the participants, these 16 people were arranged in a 4x4 grid along these two axes (see Figure 7.1). On each trial, a cue indicating the axis was presented, followed by two images showing the faces of two of the people in the hierarchy. The stimuli consisted of 16 grayscale photographic images of faces (Strohminger et al., 2016) and two colored cues (red and blue squares). Participants were instructed to select which of the two people ranked higher on the indicated axis. During training, participants only saw pairs of faces that differed by one rank along the appropriate axis (and were instructed that this was the case). Participants were then tested in the scanner on pairs with rank differences greater than or equal to one, requiring them to make novel transitive inferences.

It was hypothesized that over the course of training, participants would learn the latent 4x4 grid structure of the hierarchy. However, participants could only learn this structure through pairwise comparisons, allowing us to investigate how the brain learns to encode stimuli into map-like representations.

In addition to facilitating an examination of cognitive maps, the task probed the interaction between these map-like representations and cognitive control processes. Cognitive control is required to selectively attend to goal-relevant features in the presence of interference (Miller & Cohen, 2001). In this task, the relevant axis of the social hierarchy is cued, but the irrelevant axis may cause interference when the pair of faces is incongruent, i.e., when each of the two faces being compared ranks higher on one of the two axes, such that the correct answer depends on which axis was cued (see Figure 7.1). This is analogous to classic cognitive control tasks such as Stroop (Stroop, 1935), where control demands are increased when the stimulus features (e.g., ink color and color word) conflict.

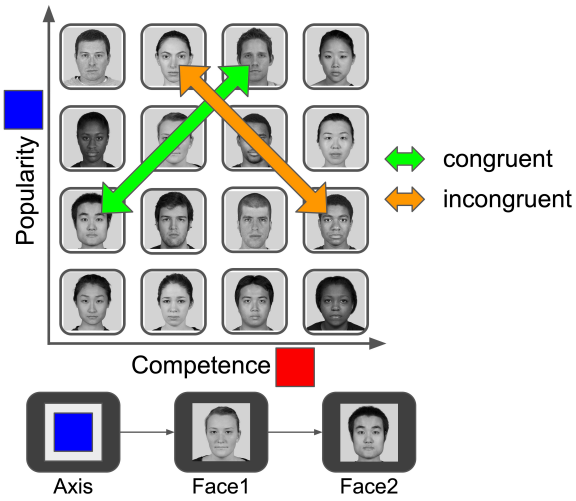

Figure 7.1: Experimental task design. Participants made decisions about which of two people ranked higher on one of two social hierarchy dimensions. The dimension ("Axis") was cued at the start of each trial. Unknown to the participants, the 16 hypothetical people in the hierarchy were arranged in a 4x4 grid along these two dimensions. Some pairs were congruent (example shown with green arrow), in the sense that the correct face ranked higher on both dimensions, and some pairs were incongruent (example shown with orange arrow), in the sense that each of the two faces ranked higher on one of the two dimensions, thereby requiring the "Axis" cue to disambiguate the higher-ranking face.

7.3.2 Participants

A total of 33 participants (16 female, age range: 19–23, normal or corrected to normal vision) were recruited for this study via an online recruitment system. Six participants were excluded due to strong head movements larger than the voxel size of 3mm. In total, 27 participants entered the analysis (mean age: 19.37 $\pm$ 0.26, standard error mean (SEM)). The study was approved by the local ethics committee, all relevant ethical regulations were followed, and participants gave written consent before the experiment.

7.4 Neural Network Model

The model was trained and tested on the same task used in the fMRI experiment, including its 4x4 grid structure and transitive inference test. On each trial, the model was presented with two 64x64 grayscale images of faces (x_1 and x_2), along with the context indicating the axis along which they should be compared (x_a - represented as a 1-hot vector). The model was trained to select which of the two faces ranked higher on the appropriate axis through supervised feedback on the correct answers during training. As in the fMRI experiment, the model did not have access to these rankings or to the latent structure of the 4x4 grid and had to learn these through trial-and-error on pairwise comparisons.

Our goal was to explore the effects of dynamic cognitive control on the geometry of map-like representa-

tions, so we developed a recurrent neural network model in order to capture the dynamics of representations unfolding over the course of each trial.

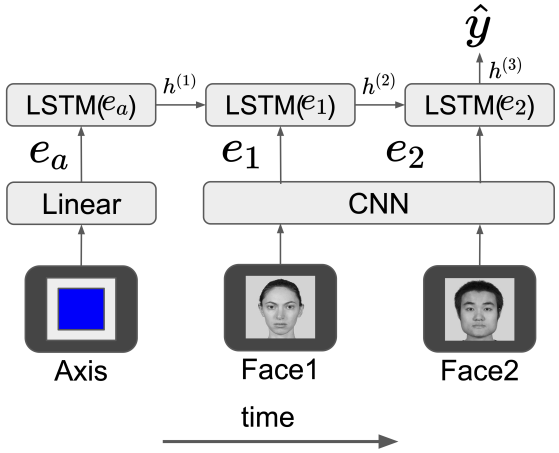

Figure 7.2: Neural network model architecture. The model used a convolutional neural network (CNN) to process images and a long short-term memory (LSTM) to process the sequence of inputs over time.

7.4.1 Model Architecture

The neural network model (see Figure 7.2) was composed of standard building blocks including a convolutional neural network (CNN) and a long short-term memory (LSTM). The CNN processed the images of faces, and a linear embedding layer processed the axis. The same CNN was used to process both faces on each trial. The result of these initial operations was a single vector encoding each of the axis (e_a), Face 1 (e_1), and Face 2 (e_2), each with the same dimension:

$$e_a = W_a x_a \quad e_1 = CNN(x_1) \quad e_2 = CNN(x_2) \tag{7.1}$$

where x_a is a 2-dimensional (one for each context) one-hot vector indicating the current context, and x_1 and x_2 are the Face 1 and Face 2 images. The LSTM processed the embeddings (e_a, e_1, e_2) in sequence:

$$h^{(t)} = \text{LSTM}(e^{(t)}, h^{(t-1)}) \tag{7.2}$$

$$\hat{y} = W_o h^{(3)} \tag{7.3}$$

where $h^{(t)}$ is the hidden state of the LSTM at time-step t, $e^{(t)}$ is the input embedding at time step t (e_a, e_1, or e_2), and W_o is a linear output layer that produces a prediction $\hat{y}$ about the answer from the final (third) hidden state of the LSTM ($h^{(3)}$).

7.4.2 Implementation Details

All modeling experiments were implemented using PyTorch. The model was trained using standard optimization techniques, including the backpropagation algorithm. The model was optimized using Adam (Kingma & Ba, 2015) with a learning rate of 0.001 and a batch size of 32 for 1000 gradient steps. For each simulation, 20 runs were performed with different random initializations.

The CNN included two convolutional layers with kernel sizes (3, 3), strides (2, 2) and number of channels (4, 8). Max pooling followed each convolutional layer with kernel sizes (2, 2) and strides (2, 2). The CNN also included a final linear layer to map the output of the last pooling operation to a single vector with 32 dimensions. The axis embedding e_a was also 32-dimensional, and the LSTM had a hidden layer size of 128.

7.5 Results

Participants performed well on the unseen pairs of faces tested in the scanner (93.6% mean accuracy), indicating good transitive inference performance. To test our main hypotheses, we conducted representational similarity analyses (RSA) Kriegeskorte et al. (2008) to characterize the representations in the brain, and performed analogous tests on the representations gathered from the model throughout training (see Figure 7.3). In the following, we describe the results of the analyses pertaining to each of our main hypotheses, with particular emphasis on the warped representational geometry, as this was the most novel aspect of our investigation.

7.5.1 Map-Like Representations

Although neither the model nor the human participants were explicitly instructed on the underlying structure of the 4x4 grid, representations in hippocampus (HC), entorhinal cortex (EC), orbitofrontal cortex (OFC), and in hidden layers of the model captured this basic structure (see Figure 7.3A). In both the model and these brain regions, similarity between representations was correlated with 2D Euclidean distances between faces' position in the 4x4 social hierarchy. In the fMRI data, this was shown with both whole-brain searchlight-based and anatomically-defined ROI-based RSA. Importantly, ROI analyses revealed that idealized 2D representations explained pattern similarity significantly better than the alternative hypothesis of two separate one-dimensional maps in HC, EC, and OFC. In the model, a similar analysis revealed a significant correlation between representational distances and pairwise Euclidean distances between faces in the 4x4 grid ($p < 0.05$). This relationship emerged early and was maintained throughout training. These results are consistent with the hypothesis that the human brain spontaneously organizes incoming relational information into map-like representations (Behrens et al., 2018; Park et al., 2020b), and show that these

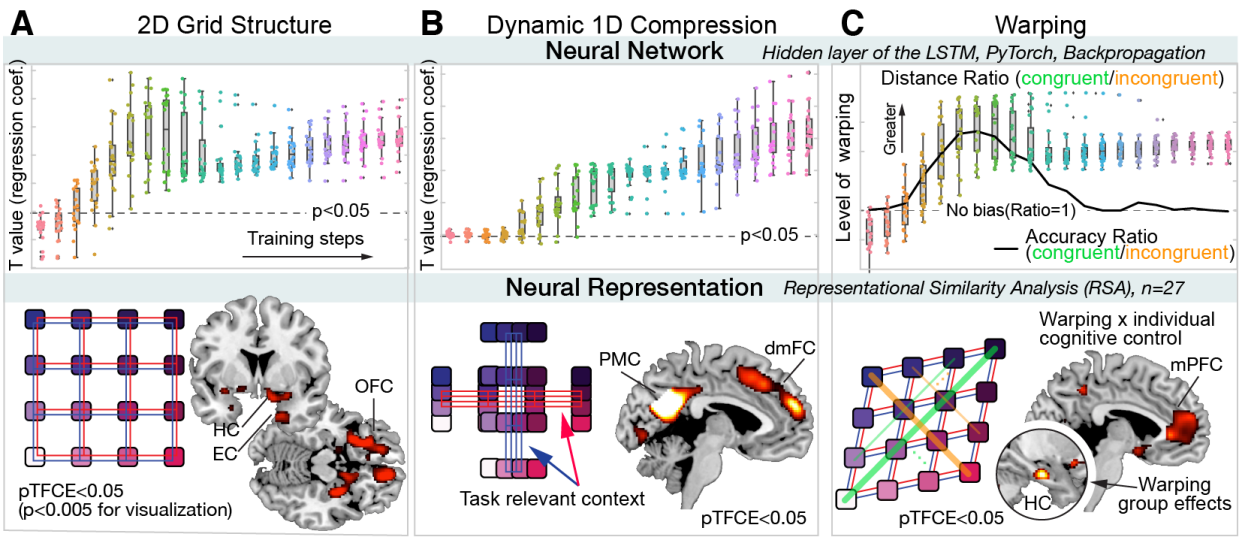

Figure 7.3: Results of analyses testing three main hypotheses. Top panels show results of analyses on the neural network model, which are given by statistics calculated on its representations over the course of training. Box plots show the variance in these statistics over 20 runs. Bottom panels show results of fMRI analyses, along with diagrams depicting idealized representations of the grid used as regressors in the RSA. **A)** Evidence of map-like representations was found in both the model and the brain. In the model, this is shown by a significant relationship ($p< 0.05$) between distances between representations and their corresponding distances in the underlying grid. In the fMRI analysis, map-like representations were found in hippocampus (HC), entorhinal cortex (EC) and orbitofrontal cortex (OFC). pTFCE: threshold-free cluster enhancement. **B)** Evidence of dynamic selection of the task-relevant dimension was found in both the model and the brain. Representations were expanded along the task-relevant axis (or equivalently, compressed along the task-irrelevant axis). This effect emerged early in the model's training and was significant in posterior and medial parietal cortex (PMC) and dorsomedial frontal cortex (dmFC). **C)** Evidence of warped representational geometry was observed in both the model and the brain. Distances between representations of congruent pairs of faces (along the green diagonal) were larger than those of incongruent pairs (along the orange diagonal), causing a consistent warping of the space along the context-invariant axis (i.e., along the congruent diagonal). In the model, this is visualized with the ratio of average Euclidean distances between congruent pairs of faces to average Euclidean distances between incongruent pairs of faces, which is consistently larger than 1 throughout training. This was associated with the corresponding accuracy ratio (black line in the same plot). Warping was observed on a group-level in HC, and warping in the medial prefrontal cortex (mPFC) and posterior cingulate cortex (PCC, not labeled in figure) was correlated with individual differences in cognitive control (as measured by differences in reaction times between congruent and incongruent trials.)

emerge in a neural network model without any specialized components that were explicitly designed to do so.

7.5.2 Dynamic Selection of Task-Relevant Dimension

This basic 2D representational geometry was dynamically modified by the current context: in both the neural network model and in brain regions including dorsomedial frontal cortex (dmFC) and posterior and medial parietal cortex (PMC) distances along the irrelevant axis were compressed (see Figure 7.3B). In the fMRI data, this was again shown using a searchlight-based multiple regression RSA that included the representational dissimilarity matrix (RDM) capturing 2D Euclidean distances between face pairs and a RDM capturing task-relevant 1D distances that assumed the currently irrelevant dimension of the grid was compressed (see Figures 7.3A and 7.3B, respectively). In the model, a regression revealed a significant relationship between the pairwise 1D rank-differences between faces in the task-relevant dimension and the distances between corresponding representations (over and above their 2D structure, which was also included as an independent variable in the regression). These findings indicate that the task-irrelevant features of the map-like representations stored in memory were compressed relative to the task-relevant features. This relationship emerged early in training and was maintained throughout the entire training period. These findings are consistent with previous experiments in tasks with exogenous sensory features in these regions (Mante et al., 2013; Takagi et al., 2021; Flesch et al., 2022a), and suggests how cognitive control can operate on 2D map-like representations from memory by accentuating the task-relevant dimensions of a learned representational space.

7.5.3 Warped Representational Geometry

In the fMRI experiment, reaction time (RT) was faster on congruent than incongruent trials regardless of the distance between locations of faces ($p< 0.01$), providing a measure of individual differences in cognitive control. We also tested for an effect of congruence on the map-like representations in both the model and the human participants by comparing distances between representations of congruent and incongruent pairs of faces sampled from different trials across blocks. Analyses of both the hidden layers and fMRI data consistently showed warping along the congruent compared to the incongruent axis – i.e., a stronger relationship between pattern similarity and Euclidean distance in congruent pairs of faces than incongruent pairs of faces (see Figure 7.3C). In the model, a regression revealed that distances between representations of congruent pairs of faces were significantly larger than those of incongruent pairs ($p < 0.05$, see top panel of Figure 7.3C). Group-level fMRI analyses revealed the same warped geometry in representations in brain areas including HC (see bottom panel of Figure 7.3C). Additionally, the levels of warping observed in the amygdala, medial prefrontal cortex (mPFC), and posterior cingulate cortex (PCC) were correlated with individual differences in our behavioral measure of cognitive control (i.e., the difference in RT between congruent and

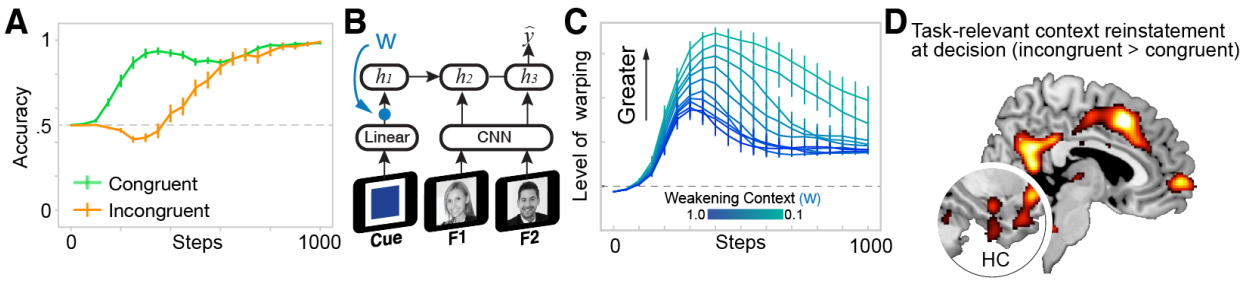

Figure 7.4: Results of additional analyses addressing the causal relationship between warping and cognitive control. **A)** The model improved its accuracy on congruent trials before incongruent trials. This difference coincided with the emergence of warping in its representations (see accuracy ratio vs. warping in previous figure), suggesting that warping is associated with a decreased reliance on contextual information and cognitive control. Error bars show standard error of the mean (SEM). The dotted horizontal line indicates chance performance (50% accuracy). **B)** To directly investigate the relationship between the strength of contextual information and the degree of warping, we performed simulated ablations of the axis embedding in the model by multiplying (e_a) by values (w) ranging from 0 to 1. **C)** Results of this ablation experiment showed increases in warping (again measured by the ratio of congruent distances to incongruent distances) when context information was inhibited. The dotted horizontal line indicates a ratio of 1 (i.e., no warping). Error bars again indicate SEM. **D)** A related fMRI analysis showed stronger reactivation of the current context on incongruent trials than congruent trials in hippocampus (HC), medial prefrontal cortex (mPFC) and posterior cingulate cortex (PCC) at decision time (pTFCE$<$ 0.05; p$<$ 0.005 for visualization).

incongruent trials). This result suggests that individuals with greater representational warping in these brain regions also experienced a greater demand on cognitive control on incongruent trials (as measured in their behavior).

Finally, we found that the warping in the neural network model was linked to an initial tendency to ineffectively utilize context information that would mitigate the interference on incongruent trials. On congruent trials, context information does not strictly need to be maintained because the correct answer does not depend on the current context (see Figure 7.1). However, on incongruent trials context information is required to mitigate the interference caused by the fact that each face ranks higher on one of the two axes. We observed that early in training, the model performed well on congruent trials but poorly on incongruent trials (see Figure 7.4A), and that this difference emerged simultaneously with the warping in its representational geometry (see accuracy ratio in Figure 7.3C). We reasoned that the warping may be related to the model's capacity to manage the interference on incongruent trials by utilizing context (axis) information. We therefore simulated an "ablation" of the contextual input by down-scaling the embedding vector (e_a) by multiplying it by a factor less than 1 (see Figure 7.4B). Results of this simulated ablation showed that warping increased when context was inhibited (see Figure 7.4C), confirming the link between the strength of context information in the model and the degree of warping in its representations. Inspired

by this observation in the model, an fMRI analysis found stronger reactivation of the current context on incongruent trials than congruent trials in HC, mPFC and PCC at decision time (see Figure 7.4D), suggesting greater reinstatement of the behavioral context during decisions in the presence of incongruence.

7.6 Discussion

It has been suggested that a key to the power of human intelligence is the capacity to integrate sensory information into cognitive maps or representations that capture the structure of the environment (Behrens et al., 2018; O'Reilly et al., 2021a). The map-like quality of these representations is thought to empower deliberative or model-based reasoning capabilities and to facilitate generalization to unseen stimuli (Behrens et al., 2018; Vikbladh et al., 2019). However, humans are not only capable of systematically encoding relevant structural information into map-like representations, but can flexibly deploy them to meet the demands of multiple contexts or goals (Miller & Cohen, 2001; Musslick et al., 2017). This flexibility is thought to emerge in part from cognitive control mechanisms in the prefrontal cortex (Herd et al., 2006; Miller & Cohen, 2001; Rougier et al., 2005), which may endow humans with the specialized circuitry necessary for systematic generalization (Russin et al., 2020b).

In this work we explore the relationship between cognitive control and the geometry of map-like representations by integrating fMRI with neural network models (Flesch et al., 2018, 2022a). Consistent with previous work (Russin et al., 2021b), we find that when a neural network was trained on our task, it developed map-like representations that captured the latent 2D structure of the task, qualitatively reproducing phenomena observed with fMRI in HC, EC and OFC. These representations emerged despite the fact that only one of the two dimensions of the grid was cued at a time and both the model and the human participants learned from pairwise comparisons alone.

As with all neural network simulations, our model could not be supplied with the wealth of experience that human participants bring to laboratory tasks such as ours, and therefore may not capture the breadth of the processes that are likely involved in the formation of cognitive maps in humans. We expect that prior experience with 2-dimensional spaces allowed participants to leverage existing knowledge while they learned the task. However, we emphasize that the latent 4x4 configuration of the faces was completely arbitrary, and that despite its lack of prior experience, the model captured the map-like qualities observed in the brain's representations.

The emergence of map-like representations allowed us to explore their interaction with cognitive control. In particular, we investigated whether their geometry was dynamically modulated according to the current context. Parallel analyses revealed dynamic compression of the irrelevant axis (or equivalently, expansion

of the relevant axis) in the representations of both the neural network and brain regions including PMC and dmFC, consistent with previous experimental findings using a different task (Flesch et al., 2022a). This phenomenon emerged in the dynamics of the model through learning: the model was not explicitly designed with a capacity to scale its representations or implement a specific mechanism for cognitive control. However, these emergent dynamics are consistent with previous neural network models of cognitive control, which implement a top-down attention mechanism to modulate specific stimulus features in posterior representations (Herd et al., 2006; Rougier et al., 2005). Flesch et al. (2022a) found that a similar compression emerged in neural networks in the "rich" regime, where they were initialized with small weights. We did not test our models with different initialization schemes, but we expect that the defaults we used would put them in the "rich" regime. Future work will test the extent to which our results depend on the magnitudes of initial weights.

The model also revealed another way in which cognitive control can affect the geometry of map-like representations: the 2D structure of the representations in the model was warped along the context-invariant axis (i.e. the congruent diagonal, see Figure 7.3C). Again, this phenomenon occurred without any specific modification to the model, suggesting it is an emergent property of representations learned by neural networks trained on the task. This prediction of the model was confirmed in the fMRI experiment: group-level analyses revealed significant warping in HC, and warping in mPFC and PCC was found to be correlated with individual differences in cognitive control.

One of the strengths of pairing a computational model with any experimental approach is the ability to simulate experiments that would be difficult or impossible with real subjects. Our simulations offered insight into the relationship between warping and the dynamics of cognitive control. Early in training, the model performed better on congruent than incongruent trials; because congruent trials did not require contextual information to be used, we reasoned that warping in the model's 2D map-like representations was related to its capacity to utilize the current context. When we simulated an ablation of contextual information, warping increased and was maintained for a longer time period throughout training (see Figure 7.4). This suggests that warping may be a natural way for the brain to compensate for a relatively weak capacity for cognitive control to orthogonalize representations according to the current context. Without strong contextual information, learning may opportunistically seize on relations between congruent pairs, which do not require context to disambiguate correct responses.

If no contextual information was available whatsoever, the best an agent could do would be to collapse the 2D grid into an integrated 1D ranking, resulting in a perfectly "warped" map projecting each face onto the congruent (bottom-left to top-right) diagonal. Thus, one explanation for our findings is that warping in HC, as well as mPFC and PCC compensated for imperfect cognitive control in the human participants by shifting

their representations to approximate this idealized 1D ranking. This is consistent with the finding that the degree of warping found in mPFC and PCC was correlated with individual differences in cognitive control, as measured by the difference in RT between congruent and incongruent trials. This explanation is also consistent with a further fMRI analysis that found that the current task-relevant context was more strongly reinstated in HC, mPFC and PCC on incongruent trials compared to congruent trials. An alternative way of interpreting our results is that participants who developed more warping in their representational geometry could not utilize context information as effectively during learning, which in turn led to greater difficulty in overcoming interference on incongruent trials. Our results, although they are suggestive, cannot definitively establish the direction of causality between representational geometry and individual differences in cognitive control. We leave it to future work to more thoroughly investigate the causal structure of the link between these phenomena that we establish here.

Taken together, our results reveal an intricate relationship between cognitive control during cognitive map formation, the resulting representational geometry, and its role on subsequent control during decisions. We found evidence of complementary representational geometries for efficiently encoding abstract relational information and flexibly selecting behaviorally relevant attributes from those representations in both neural networks and human brains. The findings further cast cognitive control in a new light, whereby an individual's representational geometry is both sculpted by and used for cognitive control when retrieving representations with endogenous feature dimensions from memory. Furthermore, our work demonstrates the virtues of integrating a neural network modeling approach with neuroimaging, and may help to address current limitations of modern neural networks used for artificial intelligence (Russin et al., 2020b).

7.7 Acknowledgments

We would like to thank the members of the Computational Cognitive Neuroscience lab and the Learning and Decision Making lab, as well as reviewers for helpful comments and discussions. The work was supported by: ONR grants ONR N00014-20-1-2578, N00014-19-1-2684 / N00014-18-1-2116, N00014-18-C-2067, as well as NSF CAREER Award 1846578, and NIH R56 MH119116.

Chapter 8

A Neural Network Model of Continual Learning with Cognitive Control

Jacob Russin[1], Maryam Zolfaghar[1], Seongmin A. Park[2], Erie Boorman[2], Randall C. O'Reilly[1]

The original version of this article (Russin et al., 2022) was accepted for publication in the Proceedings for the 44th Annual Meeting of the Cognitive Science Society (CogSci 2022). The opinions expressed here are the author's own and do not necessarily reflect the views of the conference, workshop, or publisher. The original version is available online at `https://escholarship.org/uc/item/3gn3w58z`.

8.1 Abstract

Neural networks struggle in continual learning settings from catastrophic forgetting: when trials are blocked, new learning can overwrite the learning from previous blocks. Humans learn effectively in these settings, in some cases even showing an advantage of blocking, suggesting the brain contains mechanisms to overcome this problem. Here, we build on previous work and show that neural networks equipped with a mechanism for cognitive control do not exhibit catastrophic forgetting when trials are blocked. We further show an advantage of blocking over interleaving when there is a bias for active maintenance in the control signal, implying a tradeoff between maintenance and the strength of control. Analyses of map-like representations learned by the networks provided additional insights into these mechanisms. Our work highlights the potential of cognitive control to aid continual learning in neural networks, and offers an explanation for the advantage of blocking that has been observed in humans.

8.2 Introduction

Neural networks have shown impressive performance in many domains in machine learning (ML), where they are typically trained on batches of data that are independent and identically distributed (Hadsell et al., 2020). However, agents learning about the world in real time experience streams of data that are not independent (e.g., a human may spend a few hours exploring one part of an unfamiliar city). The neural networks that have driven recent success in artificial intelligence perform poorly in these continual-learning

[1] Center for Neuroscience, University of California, Davis
[2] Center for Mind and Brain, University of California, Davis

settings because of the well known phenomenon of catastrophic forgetting/interference (McClelland et al., 1995; McCloskey & Cohen, 1989). When samples or trials are blocked, learning in new blocks overwrites the learning that occurred in previous blocks. Humans and other animals do not exhibit such extreme forgetting (McClelland et al., 1995), and in some cases even demonstrate an *advantage* when trials are blocked (Carvalho & Goldstone, 2014; Flesch et al., 2018; Noh et al., 2016; Wulf & Shea, 2002), suggesting there are mechanisms in the brain that mitigate catastrophic forgetting and can even reverse it, making learning easier when experiences are correlated over time.

A number of strategies for overcoming catastrophic forgetting in neural networks have been proposed in both computational neuroscience (Flesch et al., 2018, 2022b; McClelland et al., 1995) and ML (Botvinick et al., 2019b; Hadsell et al., 2020; Mnih et al., 2013; Velez & Clune, 2017). Complementary learning systems (CLS) theory emphasizes that catastrophic forgetting arises when learning occurs too quickly in overlapping representations (McClelland et al., 1995; O'Reilly et al., 2011), and that the episodic memory system in the hippocampus plays an important role in learning representations that are sparse or pattern-separated, allowing rapid learning to take place. However, constraining patterns of activity to be sparse is not the only way to ensure they will not overlap and interfere with each other. Theories of cognitive control in the prefrontal cortex (PFC) emphasize that a crucial function of control is to selectively modulate activity in other brain areas in order to coordinate a response that aligns with the current context or goal (Herd et al., 2014; Miller & Cohen, 2001; Rougier et al., 2005). Cognitive control may therefore play an important role in regulating learning so that patterns of activity do not overlap across different contexts or goals (Rougier et al., 2005; Tsuda et al., 2020).

Here, we build on this work and test neural networks in conditions where trials are either blocked or interleaved, showing how even in the absence of a hippocampal episodic memory system, cognitive control can help to mitigate catastrophic forgetting in the blocked condition. We further hypothesized that in some cases learning across blocked trials is *superior* to interleaving because of an internal bias of the PFC to maintain its activity over time, creating a cost to rapidly switching between contexts or goals (Blackwell et al., 2014; Herd et al., 2014; O'Reilly & Frank, 2006). This idea fits well with a general framework where the cost of switching must be traded off against the strength of control: stronger control results in less catastrophic forgetting, but more difficulty switching (Herd et al., 2006; Shenhav et al., 2013). We perform our simulations on a task designed to induce learning of map-like representations (Park et al., 2021, 2020b; Russin et al., 2021b) so that we could additionally investigate how cognitive control affected the model's representations.

8.2.1 Task

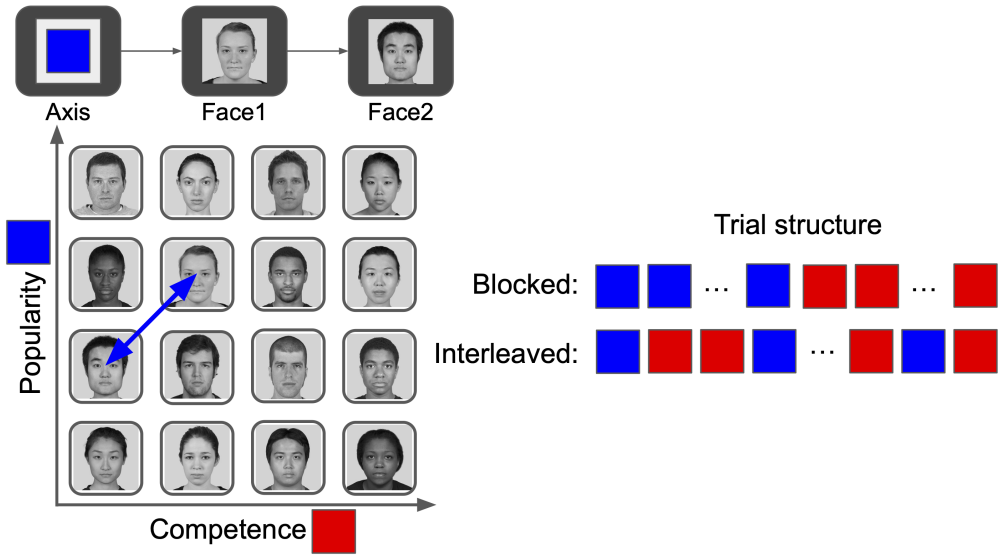

Figure 8.1: Task structure. The model learned the relative ranks of people along two social hierarchy dimensions: popularity and competence. The model learned through trial and error to select which of two faces ranked higher along one of the two dimensions (indicated by a cue). Trials were either interleaved, where cues were randomly shuffled, or blocked, where one dimension was learned at a time.

We trained neural network models on an existing task taken from an fMRI experiment (Park et al., 2020b; Russin et al., 2021b). Participants in the experiment learned about the relative ranks of 16 people in a hypothetical social hierarchy along two separate social dimensions: "popularity" and "competence" (see Figure 8.1). On each trial, the participants predicted which of two people ranked higher on one of the two dimensions, indicated by a cue. Unknown to the participants, these faces were organized into a 4x4 grid along the two dimensions; the participants were not instructed on the structure of the grid, and had to infer this structure from trial-and-error learning over pairwise comparisons.

During training, participants only saw pairs of faces that differed by one rank on the given dimension. Then in the scanner they performed a transitive inference test where comparisons were made between faces more than one rank apart. Intriguingly, the researchers found in pilot experiments that participants learned better when trials were blocked (i.e., one dimension learned at a time). This is consistent with previous results showing that learning is improved when trials are blocked (Flesch et al., 2018). This task allowed us to explore the learning dynamics in our models, but because it was designed to investigate cognitive maps in the brain, we were also able to make concrete predictions about the representations that would be learned under different conditions.

We tested neural networks on the same task structure, including its 4x4 grid and transitive inference test.

However, we introduced two training conditions to compare the learning behavior of models when trials were blocked vs. interleaved (see Figure 8.1). In the interleaved condition, popularity and competence trials were shuffled randomly, but in the blocked condition the models were trained on one of the two dimensions at a time. This allowed us to investigate the potential for cognitive control and gating mechanisms to alleviate the effects of catastrophic forgetting, as has been observed in humans learning certain tasks (Carvalho & Goldstone, 2014; Flesch et al., 2018; Noh et al., 2016; Wulf & Shea, 2002).

8.3 Neural Network Model

We designed a neural network that leveraged the principles of cognitive control in the PFC, including active maintenance and selective modulation according to the current context or goal. To test our hypotheses, we implemented models 1) with and without PFC gating, 2) with different levels of active maintenance, and 3) with different levels of control strength.

8.3.0.1 Base Model

To start, we built a simple base neural network with a multi-layer perceptron (MLP) for learning the relationships between the faces in the task (see Figure 8.2). The base model takes three one-hot vectors representing the context cue ("Axis", 2 dimensions) and each of the two faces ("Face1" and "Face2", 16 dimensions each) as inputs, and returns a prediction for which face ranked higher on the appropriate dimension. Each of these three inputs were embedded with linear layers , concatenated, and fed into an MLP with one hidden layer:

$$e_a = W_a x_a + b_a \quad e_1 = W_f x_1 + b_f \quad e_2 = W_f x_2 + b_f \tag{8.1}$$

$$h = \text{ReLU}(W_h[e_a e_1 e_2] + b_h) \tag{8.2}$$

$$\hat{y} = W_y h + b_y \tag{8.3}$$

where x_a, x_1, x_2 and e_a, e_1, e_2 are the one-hot vectors and embeddings representing the axis cue, face 1, and face 2, respectively, h is the hidden representation of the MLP, and $\hat{y}$ is the output. Brackets denote concatenation, and ReLU() is the rectified linear unit activation function.

8.3.0.2 Prefrontal Cortex for Cognitive Control

In further simulations the base MLP was augmented with a PFC layer that received the context as input and controlled the units in the hidden layer of the MLP with a gating mechanism:

$$g = c \odot h \tag{8.4}$$

where c is a control signal vector generated from the axis cue, and $\odot$ signifies element-wise multiplication. The output layer of the MLP then acted on the gated hidden layer, rather than the hidden layer itself (replacing equation 8.3 above):

$$\hat{y} = W_y g + b_y \tag{8.5}$$

Note the PFC was responsible for modulating activity according to the current context, as the MLP no longer received e_a as input. This mechanism is largely consistent with classic neural network models of cognitive control (Cohen et al., 1990; Miller & Cohen, 2001; Rougier et al., 2005), which emphasize the role of the PFC in modulating and regulating the flow of activity in posterior areas through top-down attentional control according to the current goal.

The control signal was determined from the axis cue according to a simple scheme: half of the units in the hidden layer were gated in response to one of the cues, and the other half of the units were gated in response to the other cue.

$$c = \begin{cases} [11...100...0] \cdot \gamma & \text{if axis} = 0 \\ [00...011...1] \cdot \gamma & \text{if axis} = 1 \end{cases} \tag{8.6}$$

where γ determines the strength of the control signal's influence on the hidden units. Note that there was no learning in the PFC: in this work we were interested in the effects of cognitive control and gating on learning in the MLP when trials were blocked or interleaved. Future work will explore methods for introducing learning into the PFC (Flesch et al., 2022b; Tsuda et al., 2020; Wang et al., 2018).

8.3.0.3 Active Maintenance

We also implemented a parameter λ that controlled a default bias in the PFC layer to maintain its activity over time:

$$s^{(t)} = \sigma(c^{(t)} + \sigma(s^{(t-1)} - 1 + \lambda)) \tag{8.7}$$

where t indicates time, λ determines the degree to which the previous control signal is added to the current one on each time step, σ is a rectified linear function that returns 0 for inputs less than 0 and 1 for inputs

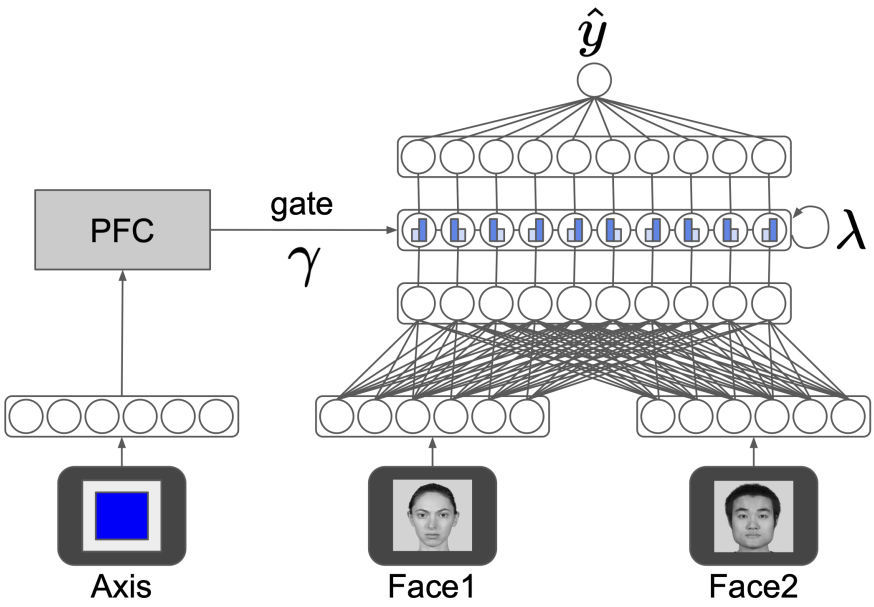

Figure 8.2: Model architecture. The model was trained to predict which of two faces ranked higher on the cued dimension ("Axis"). Inputs were embedded and passed through an MLP. The units in the hidden layer were modulated by a PFC module, which could gate them via element-wise multiplication by numbers from 0 to 1 (shown for illustration purposes as unit-wise probabilities of gating vs. not gating). Additional parameters γ and λ determined the strength of the control signal and the active maintenance, respectively.

greater than 1, and now the new variable s integrates the control signal over time and acts on the hidden state of the MLP (replacing Equation 8.4):

$$g^{(t)} = s^{(t)} \odot h^{(t)} \tag{8.8}$$

The maintenance parameter (λ) allowed us to control the degree to which the control signal was biased to maintain its activity over time, which introduces a cost when the context (i.e., the axis cue) was switched from trial to trial due to interference from the previous control signal ($s^{(t-1)}$). The bias to actively maintain patterns of activity in PFC is well established (O'Reilly & Frank, 2006), and is fundamental to the important role the PFC plays in working memory, executive functioning, and planning. We hypothesized that these dynamics would be relevant to our setting because when trials are interleaved the switch cost may have negative effects on learning. We used a particularly simple implementation to capture this basic dynamic, but future work will investigate whether its effects on learning play out in more realistic implementations (O'Reilly & Frank, 2006).

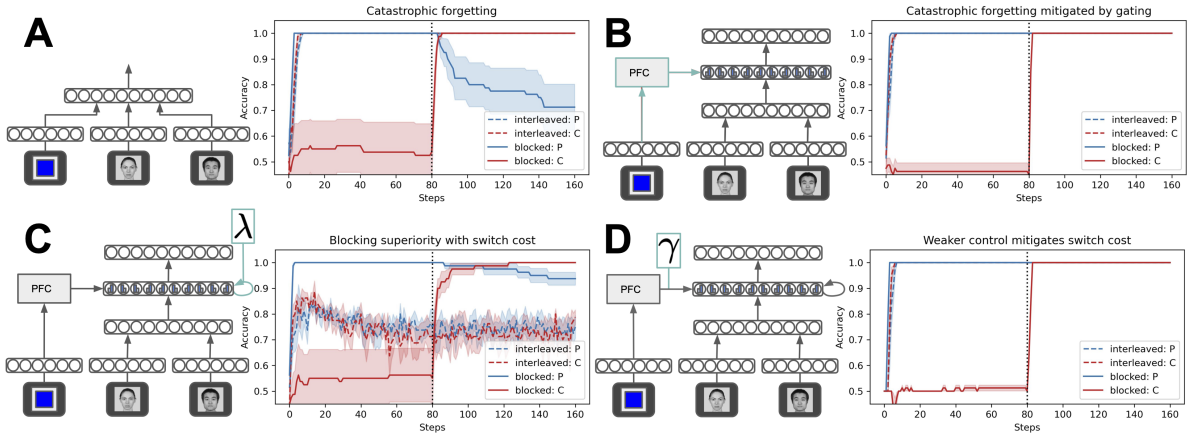

Figure 8.3: Accuracy results. Each plot shows accuracy (y-axis) over the course of training steps (x-axis) for a configuration of the model, depicted by a diagram next to the associated plot. In each experiment, trials were split on the test set by the relevant axis cue (popularity = P, shown in blue, and competence = C, shown in red), and accuracy was measured separately for each in order to show the effects of blocking. Each simulation included 5 runs where trials were interleaved (dashed lines) and 5 runs where trials were blocked (solid lines). Solid areas show SEM across runs. **A)** Catastrophic forgetting occurred in the base MLP model when it was trained on the blocked condition. **B)** Catastrophic forgetting was alleviated by the addition of a control signal from the PFC module (highlighted in cyan). **C)** The model's performance on the interleaved condition suffered when a default active maintenance was introduced in the control signal (shown by self-connection highlighted in cyan), inducing a cost to switching between contexts. **D)** This switch cost was eliminated when the control strength was reduced (shown by γ highlighted in cyan), demonstrating the tradeoff between control strength and switch cost.

8.3.0.4 Implementation Details

Models were built in PyTorch, and were supervised on correct responses with a cross entropy loss function. Models were optimized using backpropagation and Adam (Kingma & Ba, 2015) with a learning rate of 0.001. Embedding vectors had 32 dimensions, and there were 128 units in the hidden layer. For each simulation, 5 runs with different random initializations were performed.

8.4 Results

All versions of the model were trained on both blocked and interleaved conditions. In particular, we explored our hypotheses by testing the model with different configurations of the parameters described above. Accuracy on the test set was evaluated for each social dimension separately in order to assess forgetting in the blocked condition.

8.4.1 Catastrophic Forgetting when Trials are Blocked

First, we reproduced catastrophic forgetting in the model by training the base MLP (without a PFC) on both the blocked and interleaved conditions of the task (see Figure 8.3A). When trials were interleaved, the base MLP model had no problem learning the task, and quickly achieved 100% accuracy on the test set. However, when trials were blocked, we observed catastrophic forgetting: after initially performing well on the first block, over the course of the second block performance progressively declined, indicating increasing forgetting of the relationships along the first dimension that were learned in the preceding block. This result can be understood in the context of CLS theory (McClelland et al., 1995), which suggests that catastrophic forgetting occurs whenever overlapping patterns interfere with each other.

8.4.2 Cognitive Control Mitigates Forgetting

To establish that gating in the PFC can mitigate interference and reduce catastrophic forgetting, we trained the model equipped with a PFC on the same set of conditions (see Figure 8.3B). For the purposes of this experiment, we removed the internal dynamics of the PFC, setting the λ parameter to 0 (no maintenance) and the γ parameter to 1.0. When this model was trained on the task, its performance on interleaved trials was unaffected, and quickly rose to 100% accuracy. However, when it was trained on blocked trials, the catastrophic forgetting observed in the previous experiment was alleviated, and the model was capable of retaining what it had learned in the first block through the subsequent block.

This finding is consistent with the basic principles of CLS (McClelland et al., 1995): when the overlap between patterns of activity in the hidden layer is reduced, interference and forgetting are alleviated. However, CLS theory holds that the hippocampus reduces overlap in its representations with mechanisms that promote sparsity, whereas here we show that a PFC equipped with a dynamic gating mechanism can accomplish a similar goal. This is consistent with the results of previous computational models (Rougier et al., 2005; Tsuda et al., 2020) showing that adaptive gating can offer an alternative mechanism for reducing the overlap between patterns of activity, thereby reducing interference and forgetting.

8.4.3 Blocking Advantage with a Switch Cost

The results above and the results of previous models (Rougier et al., 2005; Tsuda et al., 2020) show that catastrophic forgetting can be reduced when learning occurs in non-overlapping patterns of activity across a layer, thereby explaining the reduced effects of interference observed in humans and other animals as compared with standard neural network models. However, in certain cases human performance has been shown to be *superior* when trials are blocked compared with when they are interleaved (Carvalho & Goldstone,

2014; Flesch et al., 2018; Noh et al., 2016). We hypothesized that this reversal of the catastrophic forgetting phenomenon may be due to the internal dynamics of cognitive control processes (Flesch et al., 2022b), and in particular due to the bias in neurons in the PFC to actively maintain their activity over time (O'Reilly & Frank, 2006). To explore this hypothesis, we implemented a control model with simple recurrent dynamics (see Equation 8.7), keeping the γ parameter at 1.0 but setting the λ parameter to 0.9 (i.e., 90% of the previous control signal is maintained at each time step). The resultant dynamics can be thought of as exhibiting a switch cost (Blackwell et al., 2014; Hyafil et al., 2009), wherein rapidly switching the context or goal (in this case the relevant social dimension) introduces interference due to the ongoing maintenance of the previous context. Note that the cognitive cost of task switching is usually measured in increased reaction times or errors, but here we study it in the context of its effects on learning.

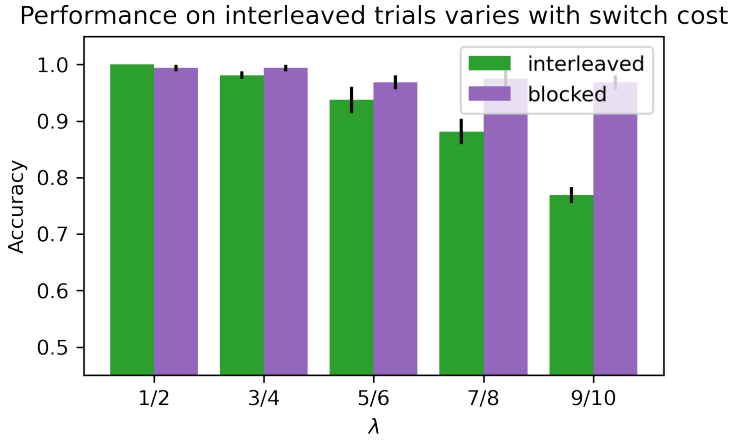

Figure 8.4: Effect of maintenance parameter (λ) on performance. In the blocked condition (purple), accuracy on the test set does not depend much on maintenance. However, as maintenance increases, the cost to switching worsens and performance in the interleaved condition declines.

When these dynamics were introduced, the model was relatively unaffected when trials were blocked, but exhibited a consistent reduction in performance when trials were interleaved (see Figure 8.3C). When trials were interleaved, many switches between contexts occurred throughout training, thereby introducing interference in the control signal, causing processing to be ineffectively modulated according to the current context. We also performed simulations where we systematically varied the λ parameter (see Figure 8.4), showing consistent reductions in performance on the interleaved condition with increased active maintenance.

8.4.4 Tradeoff between Control Strength and Switch Cost

Previous work has suggested a natural tradeoff between the strength of cognitive control and the cost incurred when a context or task-set is switched (Herd et al., 2014): stronger control would be more effective in coordi-

nating activity in other brain regions according to the current goal, but may make rapid switching between task sets or goals more difficult. To demonstrate this tradeoff, we tested a model with the maintenance (λ) kept at 0.9, but reduced the value of γ (control strength) to 0.1. In this case, the model still performed well when trials were blocked, but the reductions in performance when trials were interleaved disappeared (see Figure 8.3D). This shows that weakening the control signal can reduce the switch cost, aiding performance when there are many switches. Our results are consistent with a tradeoff between the strength of control and the switch cost: without control, catastrophic forgetting is detrimental to performance when trials are blocked, but when control is too strong, interference hurts performance when trials are interleaved.

8.4.5 Analysis of Learned Representations

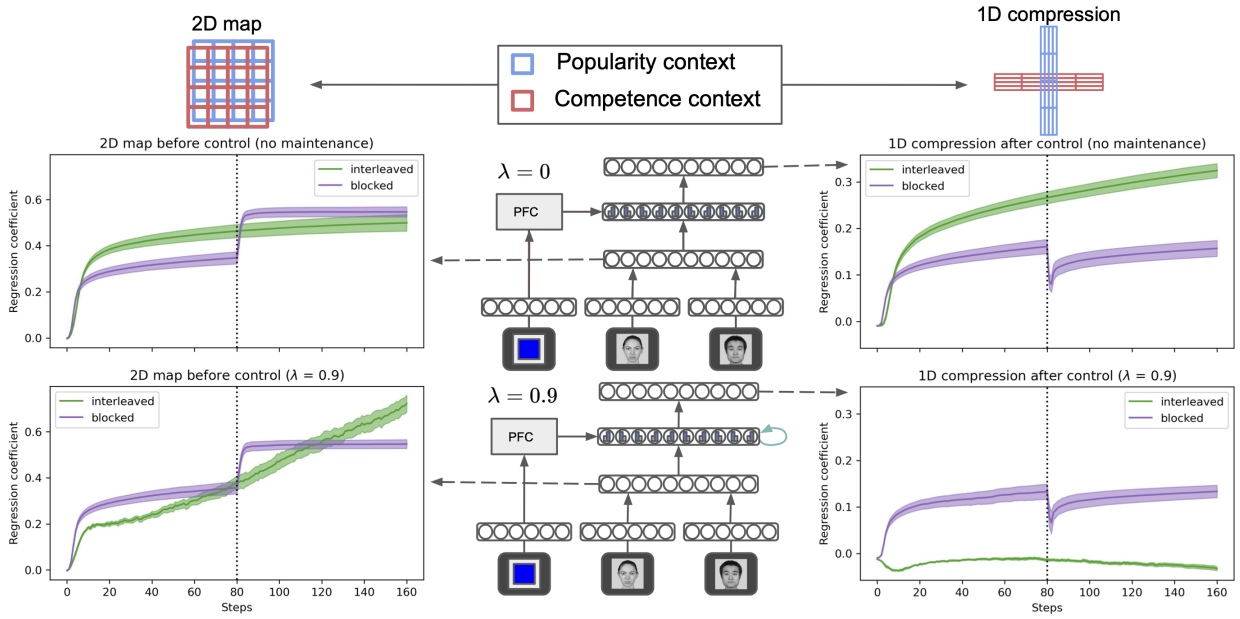

Figure 8.5: Results of analyzing the learned representations of the model. Representations were analyzed in terms of how well they captured the 2D structure of the 4x4 grid (left) and how much the irrelevant dimension of the grid was compressed on each trial (right). Idealized grids depicting these two predictions are shown on the top, where the red grid indicates idealized spacing between representations extracted during trials on which competence was cued, and the blue grid indicates the same for popularity trials. Plots show the beta coefficients over training from performing the relevant regressions. These were conducted on hidden representations either before (left) or after (right) control was applied, on two configurations of the model - one with $\lambda = 0$ (no maintenance) and one with $\lambda = 0.9$. Regression results revealed strong 2D map-like structure in the hidden layer before control was applied, and strong 1D compression of the irrelevant dimension after control was applied. However, when the active maintenance was too strong ($\lambda = 0.9$), the compression effect disappeared in the interleaved condition, indicating a failure to modulate representations according to the current context. Vertical lines indicate the switch in the blocked condition.

The grid structure of our task allowed us to make concrete predictions about the representations that would be learned in the hidden layers of the network (Park et al., 2020b; Russin et al., 2021b). In particular, we tested whether the model formed 2D map-like representations that captured the basic structure of the grid (Constantinescu et al., 2016; Park et al., 2020b, 2021), and whether these 2D map-like representations were modulated by the current context. Previous work has shown that on a similar task, 2D structure was modulated by the current context, compressing the irrelevant dimension (Flesch et al., 2022a).

Figure 8.5 shows the results of performing a regression on the representations from the hidden layer with hypothetical distance matrices (depicted as idealized map-like representations) as the predictors. We compared the results of this regression throughout training when the maintenance parameter (λ) was set to 0 and 0.9, and when the hidden representations were extracted before and after the control signal was applied (i.e., g and h in equation 8.8). γ was fixed at 1.0.

The model reliably learned the 2D structure of the grid in its hidden representations regardless of the maintenance, as can be seen in the results from the hidden representations before the control signal was applied. This 2D structure was modulated by the current control signal, which had the effect of compressing the currently irrelevant dimension (or equivalently, expanding the relevant dimension). This suggests that the effect of the control signal was to allow the model to generate its response based on the relevant dimension, and to appropriately facilitate learning in the neurons coding for that dimension. However, when trials were interleaved and maintenance (λ) was set to 0.9, the model did not show this compression pattern after control was applied, indicating a failure to modulate its representations according to the current context. This confirmed the idea that the poor performance on interleaved trials when the switch cost was high (see Figure 8.3C) was caused by interference in the control signal.

8.5 Discussion

The neural networks driving current ML research do not perform well in continual-learning settings where incoming data is blocked or otherwise correlated over time (Hadsell et al., 2020). Humans do not exhibit the catastrophic forgetting that plagues these neural networks in these settings (McClelland et al., 1995), and in some cases even show a learning advantage when trials are blocked (Carvalho & Goldstone, 2014; Flesch et al., 2018). In this work, we built on previous computational frameworks (Flesch et al., 2018; Rougier et al., 2005; Tsuda et al., 2020), and investigated the potential for cognitive control mechanisms in the PFC to induce non-overlapping patterns of activity in order to mitigate interference. Consistent with previous studies (Tsuda et al., 2020), our simulations suggest that these mechanisms can aid learning when trials are blocked over time.

In addition to pattern-separation mechanisms in the hippocampus proposed in CLS (McClelland et al., 1995), and the gating mechanism in PFC proposed here and elsewhere (Rougier et al., 2005; Tsuda et al., 2020) a number of alternative mechanisms for alleviating catastrophic forgetting in neural networks have been explored (Flesch et al., 2018; Kirkpatrick et al., 2017; Velez & Clune, 2017). In particular, Flesch et al. (2018) show that forgetting was reduced on a similar task when their network was augmented with a good inductive prior. However, they did not show an advantage to blocking over interleaving, although they observed this effect in their human experiments. While our approach is not incompatible with the idea that good inductive priors can mitigate catastrophic forgetting, we also show that a bias to maintain activity in the PFC leads to an advantage of blocking over interleaving, providing an explanation for some of the results observed by Flesch et al. (2018) and others.

In work developed concurrently with ours, Flesch et al. (2022b) show an advantage of blocking in a model based on very similar principles. In their framework, a neural network equipped with a context-gating mechanism was modified to have "sluggish" units that maintain information from previous trials, inducing a switch cost that degrades performance when trials are interleaved. Although there were some slight differences in implementation and in interpretation, we believe the broad convergence between this work and ours highlights the potential of these principles for explaining the advantage of blocking observed in humans.

The advantage of blocking can seem to contradict the well-established principles of CLS (McClelland et al., 1995). However, we show here that a "cortex-like" neural system equipped with mechanisms for cognitive control and active maintenance can enter a different regime than those typically considered in the CLS framework, wherein a reliance on control exposes the system to interference in the control signal caused by rapid context switches. We speculate that in the brain, pattern-separation mechanisms in the hippocampus are usually sufficient to ensure effective learning regardless of whether experiences are correlated over time, but animals such as humans that rely heavily on cognitive control may in some cases *require* learning experiences to be correlated over time due to the bias for active maintenance in the cognitive controller. In our simulations, we introduced this bias to show how it could lead to a learning advantage of blocking, but of course there was no real need for active maintenance in the task (as shown by the good performance of the base MLP when trials were interleaved). We expect that there are good computational reasons that the PFC would have a bias to maintain its activity over time (e.g., related to its role in working memory and planning), and that these may be unrelated to the demands of this particular task. We leave it to future work to show that a system augmented with cognitive control and active maintenance is superior in an absolute sense to one without these mechanisms.

Our simulations were also inspired by the idea that active maintenance engenders a cost to switching

between contexts, which must be traded off against the strength with which control can be applied (Herd et al., 2014). The presence of this tradeoff means the cognitive system as a whole must optimize the strength of its control signal according to constraints imposed by learning as well as the current need for control (Shenhav et al., 2013). This optimization may have taken place over the course of evolution (Herd et al., 2014), but it may also occur in real time according to the task at hand (O'Reilly et al., 2020).

Representational analyses showed that cognitive control can act on 2D map-like representations to modulate them according to the current context by compressing irrelevant dimensions and allowing learning to take place in non-overlapping patterns. However, a strong bias to maintain activity over time leads to interference in the control signal, reducing this effect and leading to poor performance when trials are interleaved. Flesch et al. (2022a) also showed compression along currently irrelevant dimensions in representations of a neural network trained on a similar task. In particular, this occurred in a "rich" regime when their neural network was initialized with small weights. The default random initializations we used in our model were likely small enough to put them in the "rich" regime, but future work will assess the extent to which our results depend on initialization.

Intelligent systems should be capable of continually learning in settings where data is not independently sampled over time. Our simulations demonstrate computational principles that may underlie human continual learning, and help to explain behavioral phenomena observed in human experiments.

8.6 Acknowledgments

We would like to thank the members of the Computational Cognitive Neuroscience and Learning and Decision Making labs at UC Davis, as well as reviewers for helpful comments and discussions. The work was supported by: ONR grants ONR N00014-20-1-2578, N00014-19-1-2684 / N00014-18-1-2116, N00014-18-C-2067, as well as NSF CAREER Award 1846578, and NIH R56 MH119116.

References

Andreas, J., Rohrbach, M., Darrell, T., & Klein, D. *2016*, Neural Module Networks, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Las Vegas, NV, USA: IEEE), 39–48

Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. *2017*, *Deep Reinforcement Learning: A Brief Survey*, IEEE Signal Processing Magazine, 34, 26, doi: 10.1109/MSP.2017.2743240

Bahdanau, D., Cho, K., & Bengio, Y. *2014*, *Neural Machine Translation by Jointly Learning to Align and Translate*, arXiv:1409.0473 [cs, stat]. http://ascl.net/1409.0473

Bahdanau, D., de Vries, H., O'Donnell, T. J., et al. *2019a*, *CLOSURE: Assessing Systematic Generalization of CLEVR Models*, arXiv:1912.05783 [cs]. http://ascl.net/1912.05783

Bahdanau, D., Murty, S., Noukhovitch, M., et al. *2019b*, Systematic Generalization: What Is Required and Can It Be Learned?, in 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019 (OpenReview.net)

Baroni, M. *2020*, *Linguistic Generalization and Compositionality in Modern Artificial Neural Networks*, Philosophical Transactions of the Royal Society B: Biological Sciences, 375, 20190307, doi: 10.1098/rstb.2019.0307

—. *2021*, *On the Proper Role of Linguistically-Oriented Deep Net Analysis in Linguistic Theorizing*, arXiv:2106.08694 [cs]. http://ascl.net/2106.08694

Barrett, D. G. T., Hill, F., Santoro, A., Morcos, A. S., & Lillicrap, T. *2018*, *Measuring Abstract Reasoning in Neural Networks*, arXiv:1807.04225 [cs, stat]. http://ascl.net/1807.04225

Barto, A. G., & Mahadevan, S. *2003*, *Recent Advances in Hierarchical Reinforcement Learning*, Discrete Event Dynamic Systems, 13, 341, doi: 10.1023/A:1025696116075

Bastings, J., Baroni, M., Weston, J., Cho, K., & Kiela, D. *2018*, Jump to Better Conclusions: SCAN Both Left and Right, in Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (Brussels, Belgium: Association for Computational Linguistics), 47–55

Battaglia, P. W., Hamrick, J. B., Bapst, V., et al. *2018*, *Relational Inductive Biases, Deep Learning, and Graph Networks*, arXiv:1806.01261 [cs, stat]. http://ascl.net/1806.01261

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., et al. *2018*, *What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior*, Neuron, 100, 490, doi: 10.1016/j.neuron.2018.10.002

Bengio, Y. *2017*, *The Consciousness Prior*, arXiv:1709.08568 [cs, stat]. http://ascl.net/1709.08568

—. *2019*, From System 1 Deep Learning to System 2 Deep Learning

Bengio, Y., Courville, A., & Vincent, P. *2014*, Representation Learning: A Review and New Perspectives, arXiv, doi: 10.48550/arXiv.1206.5538

Bengio, Y., Deleu, T., Rahaman, N., et al. *2019*, A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms, in ICLR'2020

Berg, E. A. *1948*, *A Simple Objective Technique for Measuring Flexibility in Thinking*, The Journal of General Psychology, 39, 15, doi: 10.1080/00221309.1948.9918159

Bernardi, S., Benna, M. K., Rigotti, M., et al. *2020*, *The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex*, Cell, 183, 954, doi: 10.1016/j.cell.2020.09.031

Bhagavatula, C., Bras, R. L., Malaviya, C., et al. *2019*, *Abductive Commonsense Reasoning*, arXiv:1908.05739 [cs]. http://ascl.net/1908.05739

Blackwell, K. A., Chatham, C. H., Wiseheart, M., & Munakata, Y. *2014*, *A Developmental Window into Trade-Offs in Executive Function: The Case of Task Switching versus Response Inhibition in 6-Year-Olds*, Neuropsychologia, 62, 356, doi: 10.1016/j.neuropsychologia.2014.04.016

Bottou, L. *2011*, *From Machine Learning to Machine Reasoning*, arXiv:1102.1808 [cs]. http://ascl.net/1102.1808

Botvinick, M., Niv, Y., & Barto, A. C. *2009*, *Hierarchically Organized Behavior and Its Neural Foundations: A Reinforcement Learning Perspective*, Cognition, 113, 262, doi: 10.1016/j.cognition.2008.08.011

Botvinick, M., Ritter, S., Wang, J. X., et al. *2019a*, *Reinforcement Learning, Fast and Slow*, Trends in Cognitive Sciences, 23, 408, doi: 10.1016/j.tics.2019.02.006

—. *2019b*, *Reinforcement Learning, Fast and Slow*, Trends in Cognitive Sciences, 23, 408, doi: 10.1016/j.tics.2019.02.006

Botvinick, M., & Weinstein, A. *2014*, *Model-Based Hierarchical Reinforcement Learning and Human Action Control*, Phil. Trans. R. Soc. B, 369, 20130480, doi: 10.1098/rstb.2013.0480

Botvinick, M. M. *2008*, *Hierarchical Models of Behavior and Prefrontal Function*, Trends in cognitive sciences, 12, 201, doi: 10.1016/j.tics.2008.02.009

Botvinick, M. M., & Cohen, J. D. *2014*, *The Computational and Neural Basis of Cognitive Control: Charted Territory and New Frontiers*, Cognitive Science, 38, 1249, doi: 10.1111/cogs.12126

Brown, T. B., Mann, B., Ryder, N., et al. *2020*, Language Models Are Few-Shot Learners

Buchsbaum, B. R., Greer, S., Chang, W.-L., & Berman, K. F. *2005*, *Meta-Analysis of Neuroimaging Studies of the Wisconsin Card-Sorting Task and Component Processes*, Human Brain Mapping, 25, 35, doi: 10.1002/hbm.20128

Buckner, C., & Garson, J. *2019*, *Connectionism*, The Stanford Encyclopedia of Philosophy

Cadieu, C. F., Hong, H., Yamins, D. L. K., et al. *2014*, *Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition*, PLoS Computational Biology, 10, e1003963, doi: 10.1371/journal.pcbi.1003963

Carvalho, P. F., & Goldstone, R. L. *2014*, *Putting Category Learning in Order: Category Structure and Temporal Arrangement Affect the Benefit of Interleaved over Blocked Study*, Memory & Cognition, 42, 481, doi: 10.3758/s13421-013-0371-0

Caucheteux, C., & King, J.-R. *2022*, *Brains and Algorithms Partially Converge in Natural Language Processing*, Communications Biology, 5, 1, doi: 10.1038/s42003-022-03036-1

Chakravarthy, A., Russin, J., & O'Reilly, R. C. *2022*, Systematicity in Transformers by Leveraging Linguistic Abstraction, in Proceedings of the 58th Annual Meeting of the As- Sociation for Computational Linguistics: Student Research Workshop,

Chalmers, D. J. *1990*, Why Fodor and Pylyshyn Were Wrong: The Simplest Refutation, in Proceedings of the Twelfth Annual Conference of the Cognitive Science Society, 8

Chomsky, N., ed. 1957, Syntactic Structures (The Hague: Mouton & Co.)

Cohen, J. D., Dunbar, K., & McClelland, J. L. *1990*, *On the Control of Automatic Processes: A Parallel Distributed Processing Model of the Stroop Effect*, Psychological Review, 97, 332

Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. J. *2016*, *Organizing Conceptual Knowledge in Humans with a Gridlike Code*, Science, 352, 1464, doi: 10.1126/science.aaf0941

Corkery, M., Matusevych, Y., & Goldwater, S. *2019*, Are We There yet? Encoder-decoder Neural Networks as Cognitive Models of English Past Tense Inflection, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Florence, Italy: Association for Computational Linguistics), 3868–3877

Corneil, D., Gerstner, W., & Brea, J. *2018*, *Efficient Model-Based Deep Reinforcement Learning with Variational State Tabulation*, arXiv:1802.04325 [cs, stat]. http://ascl.net/1802.04325

Crescentini, C., Seyed-Allaei, S., Pisapia, N. D., et al. *2011*, *Mechanisms of Rule Acquisition and Rule Following in Inductive Reasoning*, Journal of Neuroscience, 31, 7763, doi: 10.1523/JNEUROSCI.4579-10.2011

Dai, Z., Yang, Z., Yang, Y., et al. *2019*, Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context, arXiv, doi: 10.48550/arXiv.1901.02860

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. *2011*, *Model-Based Influences on Humans' Choices and Striatal Prediction Errors*, Neuron, 69, 1204, doi: 10.1016/j.neuron.2011.02.027

Daw, N. D., Niv, Y., & Dayan, P. *2005*, *Uncertainty-Based Competition between Prefrontal and Dorsolateral Striatal Systems for Behavioral Control*, Nature Neuroscience, 8, 1704, doi: 10.1038/nn1560

Dessì, R., & Baroni, M. *2019*, CNNs Found to Jump around More Skillfully than RNNs: Compositional Generalization in Seq2seq Convolutional Networks, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Florence, Italy: Association for Computational Linguistics), 3919–3923

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. *2019*, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in Proc. of the 2019 Conf. of the NA Chapt. of the Assoc. for Comp. Ling., ed. J. Burstein, C. Doran, & T. Solorio (Minneapolis, MN, USA: Association for Computational Linguistics), 4171–4186

Domenech, P., & Koechlin, E. *2015*, *Executive Control and Decision-Making in the Prefrontal Cortex*, Current Opinion in Behavioral Sciences, 1, 101, doi: 10.1016/j.cobeha.2014.10.007

Donoso, M., Collins, A. G. E., & Koechlin, E. *2014*, *Human Cognition. Foundations of Human Reasoning in the Prefrontal Cortex*, Science (New York, N.Y.), 344, 1481, doi: 10.1126/science.1252254

Duncan, J. *1986*, *Disorganisation of Behaviour after Frontal Lobe Damage*, Cognitive Neuropsychology, 3, 271, doi: 10.1080/02643298608253360

Eichenbaum, H. *2004*, *Hippocampus: Cognitive Processes and Neural Representations That Underlie Declarative Memory*, Neuron, 44, 109, doi: 10.1016/j.neuron.2004.08.028

Elman, J. L., Bates, E. A., Johnson, M. H., et al. 1997, Rethinking Innateness: A Connectionist Perspective on Development, new edition edn. (Cambridge (Mass.): A Bradford Book / The MIT Press)

Falk, D., Hominin Paleoneurology. 2012, in Progress in Brain Research, Vol. 195 (Elsevier), 255–272

Feinberg, V., Wan, A., Stoica, I., et al. *2018*, *Model-Based Value Estimation for Efficient Model-Free Reinforcement Learning*, arXiv:1803.00101 [cs, stat]. http://ascl.net/1803.00101

Feng, S. F., Schwemmer, M., Gershman, S. J., & Cohen, J. D. *2014*, *Multitasking versus Multiplexing: Toward a Normative Account of Limitations in the Simultaneous Execution of Control-Demanding Behaviors*, Cognitive, Affective & Behavioral Neuroscience, 14, 129, doi: 10.3758/s13415-013-0236-9

Finn, C., Abbeel, P., & Levine, S. *2017*, *Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks*, arXiv:1703.03400 [cs]. http://ascl.net/1703.03400

Finn, C., & Levine, S. *2017*, Deep Visual Foresight for Planning Robot Motion, in 2017 IEEE International Conference on Robotics and Automation (ICRA), 2786–2793

Flesch, T., Balaguer, J., Dekker, R., Nili, H., & Summerfield, C. *2018*, *Comparing Continual Task Learning in Minds and Machines*, Proceedings of the National Academy of Sciences, 115, E10313, doi: 10.1073/pnas.1800755115

Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., & Summerfield, C. *2022a*, *Orthogonal Representations for Robust Context-Dependent Task Performance in Brains and Neural Networks*, Neuron, S0896, doi: 10.1016/j.neuron.2022.01.005

Flesch, T., Nagy, D. G., Saxe, A., & Summerfield, C. *2022b*, *Modelling Continual Learning in Humans with Hebbian Context Gating and Exponentially Decaying Task Signals*, arXiv:2203.11560 [cs, q-bio]. http://ascl.net/2203.11560

Fodor, J. *1997*, *Connectionism and the Problem of Systematicity (Continued): Why Smolensky's Solution Still Doesn't Work*, Cognition, 62, 109, doi: 10.1016/s0010-0277(96)00780-9

Fodor, J., & McLaughlin, B. P. *1990*, *Connectionism and the Problem of Systematicity: Why Smolensky's Solution Doesn't Work*, Cognition, 35, 183, doi: 10.1016/0010-0277(90)90014-B

Fodor, J. A., & Pylyshyn, Z. W. *1988*, *Connectionism and Cognitive Architecture: A Critical Analysis*, Cognition, 28, 3, doi: 10.1016/0010-0277(88)90031-5

Frank, M. J., & Badre, D. *2012*, *Mechanisms of Hierarchical Reinforcement Learning in Corticostriatal Circuits 1: Computational Analysis*, Cerebral Cortex (New York, N.Y.: 1991), 22, 509, doi: 10.1093/cercor/bhr114

Furrer, D., van Zee, M., Scales, N., & Schärli, N. *2021*, *Compositional Generalization in Semantic Parsing: Pre-training vs. Specialized Architectures*, arXiv:2007.08970 [cs]. http://ascl.net/2007.08970

Fuster, J. M., Prefrontal Cortex. 2009, in Encyclopedia of Neuroscience, ed. L. R. Squire (Oxford: Academic Press), 905–908

Fuster, J. M., & Alexander, G. E. *1971*, *Neuron Activity Related to Short-Term Memory*, Science (New York, N.Y.), 173, 652, doi: 10.1126/science.173.3997.652

Garvert, M. M., Dolan, R. J., & Behrens, T. E. *2017*, *A Map of Abstract Relational Knowledge in the Human Hippocampal–Entorhinal Cortex*, eLife, 6, e17086, doi: 10.7554/eLife.17086

Goel, V. *2007*, *Anatomy of Deductive Reasoning*, Trends in Cognitive Sciences, 11, 435, doi: 10.1016/j.tics.2007.09.003

Goldman-Rakic, P. S. *1995*, *Cellular Basis of Working Memory.*, Neuron, 14, 477

Goldstein, A., Zada, Z., Buchnik, E., et al. *2022*, *Shared Computational Principles for Language Processing in Humans and Deep Language Models*, Nature Neuroscience, 25, 369, doi: 10.1038/s41593-022-01026-4

Goodfellow, I., Bengio, Y., & Courville, A. 2016, Deep Learning, Adaptive Computation and Machine Learning (Cambridge, Massachusetts: The MIT Press)

Goyal, A., Lamb, A., Hoffmann, J., et al. *2019*, *Recurrent Independent Mechanisms*, arXiv:1909.10893 [cs, stat]. http://ascl.net/1909.10893

Graves, A., Wayne, G., & Danihelka, I. *2014*, *Neural Turing Machines*, arXiv:1410.5401 [cs]. http://ascl.net/1410.5401

Graves, A., Wayne, G., Reynolds, M., et al. *2016*, *Hybrid Computing Using a Neural Network with Dynamic External Memory*, Nature, 538, 471, doi: 10.1038/nature20101

Griffiths, T. L., Callaway, F., Chang, M. B., et al. *2019*, *Doing More with Less: Meta-reasoning and Meta-Learning in Humans and Machines*, Current Opinion in Behavioral Sciences, 29, 24, doi: 10.1016/j.cobeha.2019.01.005

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. *2018*, Colorless Green Recurrent Networks Dream Hierarchically, in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (New Orleans, Louisiana: Association for Computational Linguistics), 1195–1205

Guo, Y., Zhu, H., Lin, Z., et al. *2020*, *Revisiting Iterative Back-Translation from the Perspective of Compositional Generalization*, arXiv:2012.04276 [cs]. http://ascl.net/2012.04276

Ha, D., Dai, A., & Le, Q. V. *2016*, *HyperNetworks*, arXiv:1609.09106 [cs]. http://ascl.net/1609.09106

Hadsell, R., Rao, D., Rusu, A. A., & Pascanu, R. *2020*, *Embracing Change: Continual Learning in Deep Neural Networks*, Trends in Cognitive Sciences, 24, 1028, doi: 10.1016/j.tics.2020.09.004

Hagendorff, T., & Wezel, K. *2019*, *15 Challenges for AI: Or What AI (Currently) Can't Do*, AI & SOCIETY, doi: 10.1007/s00146-019-00886-y

Hampshire, A., Thompson, R., Duncan, J., & Owen, A. M. *2011*, *Lateral Prefrontal Cortex Subregions Make Dissociable Contributions during Fluid Reasoning*, Cerebral Cortex (New York, N.Y.: 1991), 21, 1, doi: 10.1093/cercor/bhq085

Hamrick, J. B. *2019*, *Analogues of Mental Simulation and Imagination in Deep Learning*, Current Opinion in Behavioral Sciences, 29, 8, doi: 10.1016/j.cobeha.2018.12.011

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. *2017*, *Neuroscience-Inspired Artificial Intelligence*, Neuron, 95, 245, doi: 10.1016/j.neuron.2017.06.011

Hayworth, K. J., & Marblestone, A. H. *2018*, *How Thalamic Relays Might Orchestrate Supervised Deep Training and Symbolic Computation in the Brain*, bioRxiv, 304980, doi: 10.1101/304980

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. *2007*, *Towards an Executive without a Homunculus: Computational Models of the Prefrontal Cortex/Basal Ganglia System.*, Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 362, 1601

Herd, S. A., Banich, M. T., & O'Reilly, R. C. *2006*, *Neural Mechanisms of Cognitive Control: An Integrative Model of Stroop Task Performance and fMRI Data.*, Journal of Cognitive Neuroscience, 18, 22

Herd, S. A., O'Reilly, R. C., Hazy, T. E., et al. *2014*, *A Neural Network Model of Individual Differences in Task Switching Abilities*, Neuropsychologia, 62, 375, doi: 10.1016/j.neuropsychologia.2014.04.014

Hill, F., Lampinen, A., Schneider, R., et al. *2020*, *Environmental Drivers of Systematicity and Generalization in a Situated Agent*, arXiv:1910.00571 [cs]. http://ascl.net/1910.00571

Hinzen, W., Machery, E., & Werning, M., eds. 2012, The Oxford Handbook of Compositionality (Oxford University Press)

Hofstadter, D. R. *1982*, *Metamagical Themas*, Scientific American, 247, 20

Hudson, D. A., & Manning, C. D. *2018*, Compositional Attention Networks for Machine Reasoning, in 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings (OpenReview.net)

Hudson, D. A., & Manning, C. D. *2019*, *GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering*, arXiv:1902.09506 [cs]. http://ascl.net/1902.09506

Hunt, L. T., & Hayden, B. Y. *2017*, *A Distributed, Hierarchical and Recurrent Framework for Reward-Based Choice*, Nature Reviews Neuroscience, 18, 172, doi: 10.1038/nrn.2017.7

Hunt, L. T., Malalasekera, W. M. N., de Berker, A. O., et al. *2018*, *Triple Dissociation of Attention and Decision Computations across Prefrontal Cortex*, Nature Neuroscience, 21, 1471, doi: 10.1038/s41593-018-0239-5

Hupkes, D., Dankers, V., Mul, M., & Bruni, E. *2020*, *Compositionality Decomposed: How Do Neural Networks Generalise?*, arXiv:1908.08351 [cs, stat]. http://ascl.net/1908.08351

Hyafil, A., Summerfield, C., & Koechlin, E. *2009*, *Two Mechanisms for Task Switching in the Prefrontal Cortex*, Journal of Neuroscience, 29, 5135, doi: 10.1523/JNEUROSCI.2828-08.2009

Jiang, Y., & Bansal, M. *2021*, Inducing Transformer's Compositional Generalization Ability via Auxiliary Sequence Prediction Tasks, in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (Online and Punta Cana, Dominican Republic: Association for Computational Linguistics), 6253–6265

Johnson, J., Hariharan, B., van der Maaten, L., et al. *2017*, CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning, in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1988–1997

Kahneman, D. 2011, Thinking, Fast and Slow (New York: Farrar, Straus and Giroux)

Kansky, K., Silver, T., Mély, D. A., et al. *2017*, *Schema Networks: Zero-Shot Transfer with a Generative Causal Model of Intuitive Physics*, arXiv:1706.04317 [cs]. http://ascl.net/1706.04317

Keysers, D., Schärli, N., Scales, N., et al. *2020*, Measuring Compositional Generalization: A Comprehensive Method on Realistic Data

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. *2018*, *Deep Neural Networks in Computational Neuroscience*, bioRxiv, 133504, doi: 10.1101/133504

Kim, N., & Linzen, T. *2020*, COGS: A Compositional Generalization Challenge Based on Semantic Interpretation, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Online: Association for Computational Linguistics), 9087–9105

Kingma, D. P., & Ba, J. *2015*, Adam: A Method for Stochastic Optimization, in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, ed. Y. Bengio & Y. LeCun

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., et al. *2017*, *Overcoming Catastrophic Forgetting in Neural Networks*, Proceedings of the National Academy of Sciences, 114, 3521, doi: 10.1073/pnas.1611835114

Kirov, C., & Cotterell, R. *2018*, *Recurrent Neural Networks in Linguistic Theory: Revisiting Pinker and Prince (1988) and the Past Tense Debate*, Transactions of the Association for Computational Linguistics, 6, 651, doi: 10.1162/tacl_a_00247

Knudsen, E. B., & Wallis, J. D. *2021*, *Hippocampal Neurons Construct a Map of an Abstract Value Space*, Cell, 184, 4640, doi: 10.1016/j.cell.2021.07.010

Koster, R., Chadwick, M. J., Chen, Y., et al. *2018*, *Big-Loop Recurrence within the Hippocampal System Supports Integration of Information across Episodes*, Neuron, 99, 1342, doi: 10.1016/j.neuron.2018.08.009

Krawczyk, D. C., Michelle McClelland, M., & Donovan, C. M. *2011*, *A Hierarchy for Relational Reasoning in the Prefrontal Cortex*, Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 47, 588, doi: 10.1016/j.cortex.2010.04.008

Kriegeskorte, N., Mur, M., & Bandettini, P. *2008*, *Representational Similarity Analysis - Connecting the Branches of Systems Neuroscience*, Frontiers in Systems Neuroscience, 2

Kriete, T., Noelle, D. C., Cohen, J. D., & O'Reilly, R. C. *2013*, *Indirection and Symbol-like Processing in the Prefrontal Cortex and Basal Ganglia*, Proceedings of the National Academy of Sciences, 110, 16390, doi: 10.1073/pnas.1303547110

Kubota, K., & Niki, H. *1971*, *Prefrontal Cortical Unit Activity and Delayed Alternation Performance in Monkeys*, Journal of Neurophysiology, 34, 337, doi: 10.1152/jn.1971.34.3.337

Kumaran, D., & McClelland, J. L. *2012*, *Generalization through the Recurrent Interaction of Episodic Memories: A Model of the Hippocampal System*, Psychological Review, 119, 573, doi: 10.1037/a0028681

Lake, B. M. *2019*, *Compositional Generalization through Meta Sequence-to-Sequence Learning*, arXiv:1906.05381 [cs]. http://ascl.net/1906.05381

Lake, B. M., & Baroni, M. *2018*, Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks, in Proceedings of Machine Learning Research, Vol. 80, Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, ed. J. G. Dy & A. Krause (PMLR), 2879–2888

Lake, B. M., Linzen, T., & Baroni, M. *2019a*, Human Few-Shot Learning of Compositional Instructions, in Proceedings of the 41th Annual Meeting of the Cognitive Science Society, CogSci 2019: Creativity + Cognition + Computation, Montreal, Canada, July 24-27, 2019, ed. A. K. Goel, C. M. Seifert, & C. Freksa (cognitivesciencesociety.org), 611–617

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. *2015*, *Human-Level Concept Learning through Probabilistic Program Induction*, Science, 350, 1332, doi: 10.1126/science.aab3050

—. *2019b*, *The Omniglot Challenge: A 3-Year Progress Report*, arXiv:1902.03477 [cs]. http://ascl.net/1902.03477

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. *2017*, *Building Machines That Learn and Think like People*, The Behavioral and Brain Sciences, 40, e253, doi: 10.1017/S0140525X16001837

Lansdell, B. J., & Kording, K. P. *2019*, *Towards Learning-to-Learn*, Current Opinion in Behavioral Sciences, 29, 45, doi: 10.1016/j.cobeha.2019.04.005

Lara, A. H., & Wallis, J. D. *2015*, *The Role of Prefrontal Cortex in Working Memory: A Mini Review*, Frontiers in Systems Neuroscience, 9

LeCun, Y., Bengio, Y., & Hinton, G. *2015*, *Deep Learning*, Nature, 521, 436, doi: 10.1038/nature14539

Lee, M. J., & DiCarlo, J. J. *2023*, An Empirical Assay of View-Invariant Object Learning in Humans and Comparison with Baseline Image-Computable Models, Preprint, Animal Behavior and Cognition, doi: 10.1101/2022.12.31.522402

Lepori, M. A., Serre, T., & Pavlick, E. *2023*, Break It Down: Evidence for Structural Compositionality in Neural Networks, arXiv, doi: 10.48550/arXiv.2301.10884

Levine, B., Stuss, D. T., Milberg, W. P., et al. *1998*, *The Effects of Focal and Diffuse Brain Damage on Strategy Application: Evidence from Focal Lesions, Traumatic Brain Injury and Normal Aging*, Journal of the International Neuropsychological Society: JINS, 4, 247

Lewis, M., Nayak, N. V., Yu, P., et al. *2023*, Does CLIP Bind Concepts? Probing Compositionality in Large Image Models, arXiv, doi: 10.48550/arXiv.2212.10537

Li, Y., Zhao, L., Wang, J., & Hestness, J. *2019*, Compositional Generalization for Primitive Substitutions, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (Hong Kong, China: Association for Computational Linguistics), 4293–4302

Linzen, T. *2020*, How Can We Accelerate Progress Towards Human-like Linguistic Generalization?, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Online: Association for Computational Linguistics), 5210–5217

Linzen, T., & Baroni, M. *2021*, *Syntactic Structure from Deep Learning*, Annual Review of Linguistics, 7, null, doi: 10.1146/annurev-linguistics-032020-051035

Linzen, T., Dupoux, E., & Goldberg, Y. *2016*, *Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies*, arXiv:1611.01368 [cs]. http://ascl.net/1611.01368

Lipton, Z. C. *2017*, *The Mythos of Model Interpretability*, arXiv:1606.03490 [cs, stat]. http://ascl.net/1606.03490

Liu, Q., An, S., Lou, J.-G., et al. *2020*, *Compositional Generalization by Learning Analytical Expressions*, arXiv:2006.10627 [cs]. http://ascl.net/2006.10627

Loula, J., Baroni, M., & Lake, B. *2018*, Rearranging the Familiar: Testing Compositional Generalization in Recurrent Networks, in Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (Brussels, Belgium: Association for Computational Linguistics), 108–114

Luckett, P., Lee, J. J., Park, K. Y., et al. *2020*, *Mapping of the Language Network With Deep Learning*, Frontiers in Neurology, 11

MacDonald, C., & MacDonald, G., eds. 1991, Connectionism: Debates on Psychological Explanation, Volume 2, 1st edn. (Oxford, UK ; Cambridge, Mass., USA: Wiley-Blackwell)

Mahowald, K., Ivanova, A. A., Blank, I. A., et al. *2023*, Dissociating Language and Thought in Large Language Models: A Cognitive Perspective, arXiv, doi: 10.48550/arXiv.2301.06627

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. *2020*, *Emergent Linguistic Structure in Artificial Neural Networks Trained by Self-Supervision*, Proceedings of the National Academy of Sciences, doi: 10.1073/pnas.1907367117

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. *2013*, *Context-Dependent Computation by Recurrent Dynamics in Prefrontal Cortex*, Nature, 503, 78, doi: 10.1038/nature12742

Marblestone, A. H., Wayne, G., & Kording, K. P. *2016*, *Toward an Integration of Deep Learning and Neuroscience*, Frontiers in Computational Neuroscience, 10, doi: 10.3389/fncom.2016.00094

Marcheggiani, D., & Titov, I. *2017*, Encoding Sentences with Graph Convolutional Networks for Semantic Role Labeling, in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (Copenhagen, Denmark: Association for Computational Linguistics), 1506–1515

Marcus, G. *2018*, Deep Learning: A Critical Appraisal

—. *2020*, *The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence*, arXiv:2002.06177 [cs]. http://ascl.net/2002.06177

Marcus, G., & Davis, E. 2020, Rebooting AI: Building Artificial Intelligence We Can Trust (New York: Vintage)

Marcus, G. F. *1998*, *Rethinking Eliminative Connectionism*, Cognitive Psychology, 37, 243, doi: 10.1006/cogp.1998.0694

Martin, A. E., & Baggio, G. *2020*, *Modelling Meaning Composition from Formalism to Mechanism*, Philosophical Transactions of the Royal Society B: Biological Sciences, 375, 20190298, doi: 10.1098/rstb.2019.0298

McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., & Schütze, H. *2020*, *Placing Language in an Integrated Understanding System: Next Steps toward Human-Level Performance in Neural Language Models*, Proceedings of the National Academy of Sciences, 117, 25966, doi: 10.1073/pnas.1910416117

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. *1995*, *Why There Are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory.*, Psychological Review, 102, 419

McClelland, J. L., Rumelhart, D. E., & the PDP Research Group, eds. 1986, Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2: Psychological and Biological Models (MIT Press)

McCloskey, M., & Cohen, N. J., Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. 1989, in The Psychology of Learning and Motivation, Vol. 24, ed. G. H. Bower (San Diego, CA: Academic Press), 109–164

McCoy, R. T., Frank, R., & Linzen, T. *2020*, *Does Syntax Need to Grow on Trees? Sources of Hierarchical Inductive Bias in Sequence-to-Sequence Networks*, arXiv:2001.03632 [cs]. http://ascl.net/2001.03632

McCoy, R. T., Pavlick, E., & Linzen, T. *2019*, *Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference*, arXiv:1902.01007 [cs]. http://ascl.net/1902.01007

McLaughlin, B. P. *1993*, *The Connectionism/Classicism Battle to Win Souls*, Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 71, 163

—, Can an ICS Architecture Meet the Systematicity and Productivity Challenges? 2014, in The Architecture of Cognition: Rethinking Fodor and Pylyshyn's Systematicity Challenge, ed. P. Calvo & J. Symons (The MIT Press), 0

Medler, D. A. *1998*, *A Brief History of Connectionism*, Neural Computing Surveys, 1, 18

Menon, V. *2016*, *Memory and Cognitive Control Circuits in Mathematical Cognition and Learning*, Progress in brain research, 227, 159, doi: 10.1016/bs.pbr.2016.04.026

Mian, M. K., Sheth, S. A., Patel, S. R., et al. *2014*, *Encoding of Rules by Neurons in the Human Dorsolateral Prefrontal Cortex*, Cerebral Cortex (New York, NY), 24, 807, doi: 10.1093/cercor/bhs361

Miller, E. K., & Cohen, J. D. *2001*, *An Integrative Theory of Prefrontal Cortex Function.*, Annual Review of Neuroscience, 24, 167

Miller, E. K., & Desimone, R. *1994*, *Parallel Neuronal Mechanisms for Short-Term Memory.*, Science (New York, N.Y.), 263, 520

Milner, B. *1963*, *Effects of Different Brain Lesions on Card Sorting: The Role of the Frontal Lobes*, Archives of Neurology, 9, 90, doi: 10.1001/archneur.1963.00460070100010

Mnih, V., Kavukcuoglu, K., Silver, D., et al. *2013*, *Playing Atari with Deep Reinforcement Learning*, arXiv:1312.5602 [cs]. http://ascl.net/1312.5602

Momennejad, I., Russek, E. M., Cheong, J. H., et al. *2017*, *The Successor Representation in Human Reinforcement Learning*, Nature Human Behaviour, 1, 680, doi: 10.1038/s41562-017-0180-8

Moser, E. I., Kropff, E., & Moser, M.-B. *2008*, *Place Cells, Grid Cells, and the Brain's Spatial Representation System*, Annual Review of Neuroscience, 31, 69, doi: 10.1146/annurev.neuro.31.061307.090723

Müller, M. G., Papadimitriou, C. H., Maass, W., & Legenstein, R. *2016*, *A Model for Structured Information Representation in Neural Networks*, arXiv:1611.03698 [q-bio]. http://ascl.net/1611.03698

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. *2019*, *Interpretable Machine Learning: Definitions, Methods, and Applications*, Proceedings of the National Academy of Sciences, 116, 22071, doi: 10.1073/pnas.1900654116

Musslick, S., Saxe, A. M., Dey, B., Henselman, G., & Cohen, J. D. *2017*, Multitasking Capability Versus Learning Efficiency in Neural Network Architectures, in Proceedings for the 39th Annual Meeting of the Cognitive Science Society, London, UK, 6

Nagabandi, A., Kahn, G., Fearing, R. S., & Levine, S. *2018*, Neural Network Dynamics for Model-Based Deep Reinforcement Learning with Model-Free Fine-Tuning, in 2018 IEEE International Conference on Robotics and Automation (ICRA), 7559–7566

Nayebi, A., Sagastuy-Brena, J., Bear, D. M., et al. *2021a*, Goal-Driven Recurrent Neural Network Models of the Ventral Visual Stream, Preprint, Neuroscience, doi: 10.1101/2021.02.17.431717

Nayebi, A., Attinger, A., Campbell, M. G., et al. *2021b*, Explaining Heterogeneity in Medial Entorhinal Cortex with Task-Driven Neural Networks, bioRxiv, doi: 10.1101/2021.10.30.466617

Newell, A. 1990, Unified Theories of Cognition (Cambridge, MA: Harvard University Press)

Newell, A., & Simon, H. *1956*, *The Logic Theory Machine–A Complex Information Processing System*, IRE Transactions on Information Theory, 2, 61, doi: 10.1109/TIT.1956.1056797

Noh, S. M., Yan, V. X., Bjork, R. A., & Maddox, W. T. *2016*, *Optimal Sequencing during Category Learning: Testing a Dual-Learning Systems Perspective*, Cognition, 155, 23, doi: 10.1016/j.cognition.2016.06.007

Oberauer, K., & Kliegl, R. *2006*, *A Formal Model of Capacity Limits in Working Memory*, Journal of Memory and Language, 55, 601, doi: 10.1016/j.jml.2006.08.009

O'Keefe, J., & Nadel, L. 1978, The Hippocampus as a Cognitive Map (Oxford, England: Oxford University Press)

O'Reilly, R. C. *1996*, *Biologically Plausible Error-Driven Learning Using Local Activation Differences: The Generalized Recirculation Algorithm*, Neural Computation, 8, 895, doi: 10.1162/neco.1996.8.5.895

—. *2010*, *The What and How of Prefrontal Cortical Organization*, Trends in Neurosciences, 33, 355, doi: 10.1016/j.tins.2010.05.002

O'Reilly, R. C., Bhattacharyya, R., Howard, M. D., & Ketz, N. *2011*, *Complementary Learning Systems.*, Cognitive Science, Epub ahead of print

O'Reilly, R. C., & Frank, M. J. *2006*, *Making Working Memory Work: A Computational Model of Learning in the Prefrontal Cortex and Basal Ganglia.*, Neural Computation, 18, 283

O'Reilly, R. C., Munakata, Y., Frank, M. J., Hazy, T. E., & Contributors. 2012, Computational Cognitive Neuroscience (Wiki Book, 1st Edition, URL: http://ccnbook.colorado.edu)

O'Reilly, R. C., Nair, A., Russin, J. L., & Herd, S. A. *2020*, *How Sequential Interactive Processing Within Frontostriatal Loops Supports a Continuum of Habitual to Controlled Processing*, Frontiers in Psychology, 11, doi: 10.3389/fpsyg.2020.00380

O'Reilly, R. C., Petrov, A. A., Cohen, J. D., et al., How Limited Systematicity Emerges: A Computational Cognitive Neuroscience Approach. 2014, in The Architecture of Cognition: Rethinking Fodor and Pylyshyn[1]s Systematicity Challenge, ed. I. P. Calvo & J. Symons (Cambridge, MA: MIT Press), 191–225

O'Reilly, R. C., Ranganath, C., & Russin, J. L. *2021a*, *The Structure of Systematicity in the Brain*, arXiv:2108.03387 [q-bio]. http://ascl.net/2108.03387

O'Reilly, R. C., & Rudy, J. W. *2001*, *Conjunctive Representations in Learning and Memory: Principles of Cortical and Hippocampal Function.*, Psychological Review, 108, 311

O'Reilly, R. C., Russin, J. L., Zolfaghar, M., & Rohrlich, J. *2021b*, *Deep Predictive Learning in Neocortex and Pulvinar*, Journal of Cognitive Neuroscience, 33, 1158, doi: 10.1162/jocn_a_01708

O'Reilly, R. C., Wyatte, D. R., & Rohrlich, J. *2017*, *Deep Predictive Learning: A Comprehensive Model of Three Visual Streams*, arXiv:1709.04654 [q-bio]. http://ascl.net/1709.04654

Park, S. A., Miller, D. S., & Boorman, E. D. *2020a*, *Novel Inferences in a Multidimensional Social Network Use a Grid-like Code*, bioRxiv, 2020.05.29.124651, doi: 10.1101/2020.05.29.124651

—. *2021*, *Inferences on a Multidimensional Social Hierarchy Use a Grid-like Code*, Nature Neuroscience, 24, 1292, doi: 10.1038/s41593-021-00916-3

Park, S. A., Miller, D. S., Nili, H., Ranganath, C., & Boorman, E. D. *2020b*, *Map Making: Constructing, Combining, and Inferring on Abstract Cognitive Maps*, Neuron, 107, 1226, doi: 10.1016/j.neuron.2020.06.030

Pavlick, E. *2022*, *Semantic Structure in Deep Learning*, Annual Review of Linguistics, 8, 447, doi: 10.1146/annurev-linguistics-031120-122924

Perez, E., Strub, F., de Vries, H., Dumoulin, V., & Courville, A. C. *2018*, FiLM: Visual Reasoning with a General Conditioning Layer, in Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, ed. S. A. McIlraith & K. Q. Weinberger (AAAI Press), 3942–3951

Perret, E. *1974*, *The Left Frontal Lobe of Man and the Suppression of Habitual Responses in Verbal Categorical Behaviour*, Neuropsychologia, 12, 323, doi: 10.1016/0028-3932(74)90047-5

Petri, G., Musslick, S., Dey, B., et al. *2021*, *Universal Limits to Parallel Processing Capability of Network Architectures*, Nature Physics, 17, doi: 10.1038/s41567-021-01170-x

Piantadosi, S. *2023*, Modern Language Models Refute Chomsky's Approach to Language, LingBuzz

Pinker, S., & Prince, A. *1988*, *On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition*, Cognition, 28, 73, doi: 10.1016/0010-0277(88)90032-7

Place, U. T. *1992*, *Eliminative Connectionism: Its Implications for a Return to an Empiricist/Behaviorist Linguistics*, Behavior and Philosophy, 20, 21

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. *2022*, *The Best Game in Town: The Re-Emergence of the Language of Thought Hypothesis Across the Cognitive Sciences*, The Behavioral and Brain Sciences, 1, doi: 10.1017/S0140525X22002849

Rahaman, N., Gondal, M. W., Joshi, S., et al. *2021*, Dynamic Inference with Neural Interpreters, in Advances in Neural Information Processing Systems

Ranganath, C., & Ritchey, M. *2012*, *Two Cortical Systems for Memory-Guided Behaviour*, Nature Reviews Neuroscience, 13, 713, doi: 10.1038/nrn3338

Richards, B. A., Lillicrap, T. P., Beaudoin, P., et al. *2019*, *A Deep Learning Framework for Neuroscience*, Nature Neuroscience, 22, 1761, doi: 10.1038/s41593-019-0520-2

Rilling, J. K. *2006*, *Human and Nonhuman Primate Brains: Are They Allometrically Scaled Versions of the Same Design?*, Evolutionary Anthropology: Issues, News, and Reviews, 15, 65, doi: 10.1002/evan.20095

Rougier, N. P., Noelle, D., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. *2005*, *Prefrontal Cortex and the Flexibility of Cognitive Control: Rules Without Symbols.*, Proceedings of the National Academy of Sciences, 102, 7338

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. *1986a*, *Learning Representations by Back-Propagating Errors*, Nature, 323, 533

Rumelhart, D. E., & McClelland, J. L., On Learning the Past Tenses of English Verbs. 1986, in Parallel Distributed Processing. Volume 2: Psychological and Biological Models, ed. J. L. McClelland, D. E. Rumelhart, & P. R. Group (Cambridge, MA: MIT Press), 216–271

Rumelhart, D. E., McClelland, J. L., & PDP Research Group, CORPORATE., eds. 1986b, Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2: Psychological and Biological Models (Cambridge, MA, USA: MIT Press)

Rushworth, M. F. S., & Behrens, T. E. J. *2008*, *Choice, Uncertainty and Value in Prefrontal and Cingulate Cortex*, Nature Neuroscience, 11, 389, doi: 10.1038/nn2066

Russin, J., Fernandez, R., Palangi, H., et al. *2021a*, Compositional Processing Emerges in Neural Networks Solving Math Problems, in Proceedings for the 43rd Annual Meeting of the Cognitive Science Society

Russin, J., Jo, J., O'Reilly, R. C., & Bengio, Y. *2019*, *Compositional Generalization in a Deep Seq2seq Model by Separating Syntax and Semantics*, arXiv:1904.09708 [cs, stat]. http://ascl.net/1904.09708

Russin, J., Jo, J., O'Reilly, R. C., & Bengio, Y. *2020a*, Systematicity in a Recurrent Neural Network by Factorizing Syntax and Semantics, in Proceedings for the 42nd Annual Meeting of the Cognitive Science Society, 7

Russin, J., O'Reilly, R. C., & Bengio, Y. *2020b*, Deep Learning Needs a Prefrontal Cortex, in Bridging AI and Cognitive Science (BAICS) Workshop, ICLR 2020, 11

Russin, J., Zolfaghar, M., Park, S. A., Boorman, E., & O'Reilly, R. C. *2021b*, Complementary Structure-Learning Neural Networks for Relational Reasoning, in Proceedings for the 43rd Annual Meeting of the Cognitive Science Society

Russin, J., Zolfaghar, M., Park, S. A., Boorman, E., & O'Reilly, R. C. *2022*, A Neural Network Model of Continual Learning with Cognitive Control, in Proceedings for the 44th Annual Meeting of the Cognitive Science Society

Sachan, D. S., Zhang, Y., Qi, P., & Hamilton, W. *2021*, *Do Syntax Trees Help Pre-trained Transformers Extract Information?*, arXiv:2008.09084 [cs]. http://ascl.net/2008.09084

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., & Lillicrap, T. *2016*, Meta-Learning with Memory-Augmented Neural Networks

Sarafyazd, M., & Jazayeri, M. *2019*, *Hierarchical Reasoning by Neural Circuits in the Frontal Cortex*, Science, 364, doi: 10.1126/science.aav8911

Saxe, A., Nelli, S., & Summerfield, C. *2021*, *If Deep Learning Is the Answer, What Is the Question?*, Nature Reviews Neuroscience, 22, 55, doi: 10.1038/s41583-020-00395-8

Saxton, D., Grefenstette, E., Hill, F., & Kohli, P. *2019*, Analysing Mathematical Reasoning Abilities of Neural Models, in 7th Intern. Conf. on Lear. Repr. (new orleans, LA, USA: OpenReview.net)

Schlag, I., Smolensky, P., Fernandez, R., et al. *2019*, *Enhancing the Transformer with Explicit Relational Encoding for Math Problem Solving*, arXiv:1910.06611 [cs, stat]. http://ascl.net/1910.06611

Schoenemann, P. T., Sheehan, M. J., & Glotzer, L. D. *2005*, *Prefrontal White Matter Volume Is Disproportionately Larger in Humans than in Other Primates*, Nature Neuroscience, 8, 242, doi: 10.1038/nn1394

Schölkopf, B., & Smola, A. J. 2002, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, Adaptive Computation and Machine Learning (Cambridge, Mass: MIT Press)

Seidenberg, M. S., & Plaut, D. C. *2014*, *Quasiregularity and Its Discontents: The Legacy of the Past Tense Debate*, Cognitive Science, 38, 1190, doi: 10.1111/cogs.12147

Semendeferi, K., Armstrong, E., Schleicher, A., Zilles, K., & Van Hoesen, G. W. *2001*, *Prefrontal Cortex in Humans and Apes: A Comparative Study of Area 10*, American Journal of Physical Anthropology, 114, 224, doi: 10.1002/1096-8644(200103)114:3<224::AID-AJPA1022>3.0.CO;2-I

Sennrich, R., & Haddow, B. *2016*, Linguistic Input Features Improve Neural Machine Translation, in Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers (Berlin, Germany: Association for Computational Linguistics), 83–91

Serre, T. *2019*, *Deep Learning: The Good, the Bad, and the Ugly*, Annual Review of Vision Science, 5, 399, doi: 10.1146/annurev-vision-091718-014951

Shallice, T. *1982*, *Specific Impairments of Planning*, Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 298, 199, doi: 10.1098/rstb.1982.0082

Shallice, T., & Burgess, P. W. *1991*, *Deficits in Strategy Application Following Frontal Lobe Damage in Man*, Brain: A Journal of Neurology, 114 (Pt 2), 727, doi: 10.1093/brain/114.2.727

Shenhav, A., Botvinick, M. M., & Cohen, J. D. *2013*, *The Expected Value of Control: An Integrative Theory of Anterior Cingulate Cortex Function*, Neuron, 79, 217, doi: 10.1016/j.neuron.2013.07.007

Silver, D., Huang, A., Maddison, C. J., et al. *2016*, *Mastering the Game of Go with Deep Neural Networks and Tree Search*, Nature, 529, 484, doi: 10.1038/nature16961

Smittenaar, P., FitzGerald, T. H., Romei, V., Wright, N. D., & Dolan, R. J. *2013*, *Disruption of Dorsolateral Prefrontal Cortex Decreases Model-Based in Favor of Model-Free Control in Humans*, Neuron, 80, 914, doi: 10.1016/j.neuron.2013.08.009

Smolensky, P. *1987*, *The Constituent Structure of Connectionist Mental States: A Reply to Fodor and Pylyshyn*, Southern Journal of Philosophy, 137

—. *1988*, *Connectionism, Constituency, and the Language of Thought*, Computer Science Technical Reports, 31

—. *1990*, *Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Systems*, Artificial Intelligence, 46, 159, doi: 10.1016/0004-3702(90)90007-M

Sommer, M. A., & Wurtz, R. H. *2000*, *Composition and Topographic Organization of Signals Sent from the Frontal Eye Field to the Superior Colliculus.*, Journal of Neurophysiology, 83, 1979

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. *2017*, *The Hippocampus as a Predictive Map*, Nature Neuroscience, 20, 1643, doi: 10.1038/nn.4650

Strohminger, N., Gray, K., Chituc, V., et al. *2016*, *The MR2: A Multi-Racial, Mega-Resolution Database of Facial Stimuli*, Behavior Research Methods, 48, 1197, doi: 10.3758/s13428-015-0641-9

Stroop, J. R. *1935*, *Studies of Interference in Serial Verbal Reactions.*, Journal of Experimental Psychology, 18, 643

Strubell, E., Verga, P., Andor, D., Weiss, D., & McCallum, A. *2018*, Linguistically-Informed Self-Attention for Semantic Role Labeling, in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (Brussels, Belgium: Association for Computational Linguistics), 5027–5038

Summerfield, C., Luyckx, F., & Sheahan, H. *2020*, *Structure Learning and the Posterior Parietal Cortex*, Progress in Neurobiology, 184, 101717, doi: 10.1016/j.pneurobio.2019.101717

Sutton, R. S., Precup, D., & Singh, S. *1999*, *Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning*, Artificial Intelligence, 112, 181, doi: 10.1016/S0004-3702(99)00052-1

Szabó, Z. G., Compositionality. 2020, in The Stanford Encyclopedia of Philosophy, fall 2020 edn., ed. E. N. Zalta (Metaphysics Research Lab, Stanford University), NA

Takagi, Y., Hunt, L. T., Woolrich, M. W., Behrens, T. E., & Klein-Flügge, M. C. *2021*, *Adapting Non-Invasive Human Recordings along Multiple Task-Axes Shows Unfolding of Spontaneous and over-Trained Choice*, eLife, 10, e60988, doi: 10.7554/eLife.60988

Tenney, I., Das, D., & Pavlick, E. *2019*, BERT Rediscovers the Classical NLP Pipeline, in Proc. of the 57th Conf. of the Asso. for Comp. Ling., ed. A. Korhonen, D. R. Traum, & L. Màrquez (Florence, Italy: Association for Computational Linguistics), 4593–4601

Thompson-Schill, S. L., Dissecting the Language Organ : A New Look at the Role of Broca ' s Area in Language Processing. 2005, in Twenty-First Century Psycholinguistics, 1st edn., Vol. 1 (Routledge), 1–18

Tolman, E. *1948*, *Cognitive Maps in Rats and Men.*, Psychological Review, 55, 189

Tsuda, B., Tye, K. M., Siegelmann, H. T., & Sejnowski, T. J. *2020*, *A Modeling Framework for Adaptive Lifelong Learning with Transfer and Savings through Gating in the Prefrontal Cortex*, Proceedings of the National Academy of Sciences of the United States of America, 117, 29872, doi: 10.1073/pnas.2009591117

van der Maaten, L., & Hinton, G. *2008*, *Visualizing Data Using T-SNE*, Journal of Machine Learning Research, 9, 2579

Vargha-Khadem, F., Gadian, D. G., Watkins, K. E., et al. *1997*, *Differential Effects of Early Hippocampal Pathology on Episodic and Semantic Memory*, Science, 277, 376, doi: 10.1126/science.277.5324.376

Vaswani, A., Shazeer, N., Parmar, N., et al. *2017*, Attention Is All You Need, in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, ed. I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett, 5998–6008

Velez, R., & Clune, J. *2017*, *Diffusion-Based Neuromodulation Can Eliminate Catastrophic Forgetting in Simple Neural Networks*, PLOS ONE, 12, e0187736, doi: 10.1371/journal.pone.0187736

Vendrell, P., Junqué, C., Pujol, J., et al. *1995*, *The Role of Prefrontal Regions in the Stroop Task*, Neuropsychologia, 33, 341, doi: 10.1016/0028-3932(94)00116-7

Vikbladh, O. M., Meager, M. R., King, J., et al. *2019*, *Hippocampal Contributions to Model-Based Planning and Spatial Memory*, Neuron, 102, 683, doi: 10.1016/j.neuron.2019.02.014

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., & Wierstra, D. *2017*, *Matching Networks for One Shot Learning*, arXiv:1606.04080 [cs, stat]. http://ascl.net/1606.04080

Wallis, J. D., Anderson, K. C., & Miller, E. K. *2001*, *Single Neurons in Prefrontal Cortex Encode Abstract Rules.*, Nature, 411, 953

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., et al. *2018*, *Prefrontal Cortex as a Meta-Reinforcement Learning System*, Nature Neuroscience, 21, 860, doi: 10.1038/s41593-018-0147-8

Webb, T. W., Sinha, I., & Cohen, J. D. *2021*, *Emergent Symbols through Binding in External Memory*, arXiv:2012.14601 [cs]. http://ascl.net/2012.14601

Webson, A., Loo, A. M., Yu, Q., & Pavlick, E. *2023*, Are Language Models Worse than Humans at Following Prompts? It's Complicated, arXiv, doi: 10.48550/arXiv.2301.07085

Weiss, K., Khoshgoftaar, T. M., & Wang, D. *2016*, *A Survey of Transfer Learning*, Journal of Big Data, 3, 9, doi: 10.1186/s40537-016-0043-6

Weston, J., Bordes, A., Chopra, S., et al. *2015*, *Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks*, arXiv:1502.05698 [cs, stat]. http://ascl.net/1502.05698

Whittington, J. C. R., & Bogacz, R. *2019*, *Theories of Error Back-Propagation in the Brain*, Trends in Cognitive Sciences, 23, 235, doi: 10.1016/j.tics.2018.12.005

Whittington, J. C. R., Muller, T. H., Mark, S., et al. *2020*, *The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation*, Cell, 183, 1249, doi: 10.1016/j.cell.2020.10.024

Wiemerslage, A., Dudy, S., & Kann, K. *2022*, *A Comprehensive Comparison of Neural Networks as Cognitive Models of Inflection*, undefined

Wilcox, E. G., Futrell, R., & Levy, R. *2022*, *Using Computational Models to Test Syntactic Learnability*, Linguistic Inquiry, 1, doi: 10.1162/ling_a_00491

Wulf, G., & Shea, C. H. *2002*, *Principles Derived from the Study of Simple Skills Do Not Generalize to Complex Skill Learning*, Psychonomic Bulletin & Review, 9, 185, doi: 10.3758/BF03196276

Xu, K., Ba, J., Kiros, R., et al. *2016*, *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*, arXiv:1502.03044 [cs]. http://ascl.net/1502.03044

Xu, K., Li, J., Zhang, M., et al. *2020*, WHAT CAN NEURAL NETWORKS REASON ABOUT?

Yamins, D. L. K., Hong, H., Cadieu, C. F., et al. *2014*, *Performance-Optimized Hierarchical Models Predict Neural Responses in Higher Visual Cortex*, Proceedings of the National Academy of Sciences, 111, 8619, doi: 10.1073/pnas.1403112111

Yi, K., Wu, J., Gan, C., et al., Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. 2018, in Advances in Neural Information Processing Systems 31, ed. S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Curran Associates, Inc.), 1031–1042

Yonelinas, A., Ranganath, C., Ekstrom, A., & Wiltgen, B. *2019*, *A Contextual Binding Theory of Episodic Memory: Systems Consolidation Reconsidered*, Nature reviews. Neuroscience, 20, 364, doi: 10.1038/s41583-019-0150-4

Zador, A., Escola, S., Richards, B., et al. *2023*, *Catalyzing Next-Generation Artificial Intelligence through NeuroAI*, Nature Communications, 14, 1597, doi: 10.1038/s41467-023-37180-x

Zeithamova, D., Schlichting, M. L., & Preston, A. R. *2012*, *The Hippocampus and Inferential Reasoning: Building Memories to Navigate Future Decisions*, Frontiers in Human Neuroscience, 6, doi: 10.3389/fnhum.2012.00070

Zolfaghar, M., Russin, J., Park, S. A., Boorman, E. D., & O'Reilly, R. C. *2022*, The Geometry of Map-Like Representations under Dynamic Cognitive Control, in Proceedings for the 44th Annual Meeting of the Cognitive Science Society