Title
Leveraging deep neural networks to study human cognition

Author
Peterson, Joshua Caleb

Publication Date
2018

Peer reviewed|Thesis/dissertation

Leveraging deep neural networks to study human cognition

By

Joshua C. Peterson

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Psychology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Thomas L. Griffiths, Chair
Professor David Whitney
Professor Alexei A. Efros

Summer 2018

Abstract

Leveraging deep neural networks to study human cognition

by

Joshua C. Peterson

Doctor of Philosophy in Psychology

University of California, Berkeley

Professor Thomas L. Griffiths, Chair

The majority of computational theories of inductive processes in psychology derive from small-scale experiments with simple stimuli that are easy to represent. However, real-world stimuli are complex, hard to represent efficiently, and likely require very different cognitive strategies to cope with. Indeed, the difficulty of such tasks are part of what make humans so impressive, yet methodological resources for modeling their solutions are limited. This presents a fundamental challenge to the precision of psychology as a science, especially if traditional laboratory methods fail to generalize. Recently, a number of computationally tractable, data-driven methods such as deep neural networks have emerged in machine learning for deriving useful representations of complex perceptual stimuli, but they are explicitly optimized in service to engineering objectives rather than modeling human cognition. It has remained unclear to what extent engineering models, while often state-of-the-art in terms of human-level task performance, can be leveraged to model, predict, and understand humans.

In the following, I outline a methodology by which psychological research can confidently leverage representations learned by deep neural networks to model and predict complex human behavior, potentially extending the scope of the field. In Chapter 1, I discuss the challenges to ecological validity in the laboratory that may be partially circumvented by technological advances and trends in machine learning, and weigh the advantages and disadvantages of bootstrapping from largely uninterpretable models. In Chapter 2, I contrast methods from psychology and machine learning for representing complex stimuli like images. Chapter 3 provides a first case study of applying deep neural networks to predict whether objects in a large database of images will be remembered by humans. Chapter 4 provides the central argument for using representations from deep neural networks as proxies for human psychological representations in general. To do this, I establish and demonstrate methods for quantifying their correspondence, improving their correspondence with minimal cost, and applying the result to the modeling of downstream cognitive processes. Building on this, Chapter 5 develops a method for modeling human subjective probability over deep representations in order to capture multimodal mental visual concepts such as "landscape". Finally, in Chapter 6, I discuss the implications of the overall paradigm espoused in the current work, along with the most crucial challenges ahead and po-

tential ways forward. The overall endeavor is almost certainly a stepping stone to methods that may look very different in the near future, as the gains in leveraging machine learning methods are consolidated and made more interpretable/useful. The hope is that a synergy can be formed between the two fields, each bootstrapping and learning from the other.

In memory of Sayan Gul

Contents

Listing of figures

Listing of tables

Acknowledgments

Firstly, I would like to thank my advisors Stephen Palmer, who brought me to Berkeley, taught me how to be a scientist, and literally gave me shelter, David Wessel, who was one of the kindest mentors I've had in academia, and especially Tom Griffiths, whose generosity has always surprised me, and who somehow managed to take something I was already utterly fascinated by—human intelligence—and make it all the more enthralling by a much deeper characterization that I may have never known otherwise. I'd also like to thank the members of my thesis and qualifying exam committees: Tom Griffiths, Stephen Palmer, David Whitney, Alyosha Efros, and Bruno Olshausen.

My academic career has had its fair share of coin-flipping, and in the end, I've been incredibly lucky, with many friends, mentors, and collaborators tipping the balance. Thomas Langlois, who was my first close friend at Berkeley, and Rachit Dubey, who was once a mentee and now a mentor, have been with me through the thick of it, and we've come a long way together. The various people of the Palmer and Griffiths labs made Berkeley an exciting place to be. I'd like to thank all of my collaborators who made working on each project such a pleasure: Joshua Abbott, Ruairidh Battleday, David Bourgin, Dawn Chen, Rachit Dubey, Nori Jacoby, Thomas Langlois, Stephan Meylan, Aida Nematzadeh, Daniel Reichman, Paul Soulos, Jordan Suchow, and many more.

Lastly, I'd like to thank my wonderful family for indulging and supporting me, and my darling wife, Janell, who is the love of my life and has been by my side for over 12 years already.

Original Collaborators on Projects in this Thesis

Much of the work in this thesis is the product of collaborations with a number of talented people. The individuals listed as co-authors on the published works from which this thesis is derived are listed below.

- Chapter 3: Rachit Dubey, Aditya Khosla, Ming-Hsuan Yang, and Bernard Ghanem
- Chapter 4: Joshua T. Abbott and Thomas L. Griffiths
- Chapter 5: Jordan W. Suchow, Krisha Aghi, Alexander Y. Ku, and Thomas L. Griffiths

Funding Sources

The work in this thesis would not have been possible without serveral sources of funding, for which I am extremely grateful. They include the following:

- Grant number FA9550-13-1-0170 from the Air Force Office of Scientific Research
- Grant number 1718550 from the National Science Foundation
- Competitive research funding from King Abdullah University of Science and Technology

What I cannot create, I do not understand.

Richard Feynman

1 Introduction

Humans possess a remarkable ability to cope with complex inductive problems in the natural world. Faced with a massive stream of multi-sensory input, we are able to parse in large part the structure of our environment, and to locate, identify, and track an immense taxonomy of objects in that environment. To say nothing of the higher level, abstract reasoning we perform over the same represented world, no model or machine yet rivals the efficiency, robustness, and performance of humans on a number of key perceptual tasks (although we will review some recent practical—and not necessarily theoretical—breakthroughs in Chapter 2).

While it would be unreasonable to expect that every scientific model capture the complexity of the phenomenon of interest in its entirety, every scientific discipline must strive to obtain some level of generalization to the real world. This is much easier said than done, and if we've fallen short to some degree given time spent and methodologies available, it's not for a lack of interest or ambition. More often a technological advance comes along (e.g., higher resolution brain imaging techniques), that provides a new set of constraints for psychological theory. What I'd like to present in the current thesis is what I believe is such an advance—a set of recent, engineering motifs for relatively low-bias and data-driven learning (specifically *deep neural networks*), and what I think are reasonable schemas for their application to cognitive modeling. Like all new scientific tools, these methods are not meant to replace or suffice, but to allow for a fresh perspective on several primary challenges in psychological modeling.

1.1 Challenges to Classical Methods in Psychology

A staggering number of theories and insights have been born from small sets of simple (e.g., dot pattern) stimuli and handfuls of human participants in a laboratory. For example, the majority of seminal findings on human categorization behavior have be gleaned from relatively small, well-controlled experiments (e.g., Medin & Schaffer, 1978; Nosofsky, 1986; Posner & Keele, 1968; Reed, 1972; Rosch & Mervis, 1975). Several decades later, this trend is alive and well (see for example Vong, Hendrickson, Perfors, & Navarro, 2016). Such a productive paradigm should by all means continue forward, and is likely to yield further, invaluable insights into human behavior, but is not without important limitations.

We can get a sense of these limitations by considering an example rooted in the above findings. Exemplar models (Medin & Schaffer, 1978; Nosofsky, 1986) are a well-studied class of models that learn to categorize stimuli by comparing an input stimulus to be categorized to a set of other datapoints or *exemplars* in memory using a similarity function (e.g., $k(x_i, x_j)$; Shepard, 1987). This is in contrast to parametric models like the prototype model (Posner & Keele, 1968; Reed, 1972), which instead estimate the parameters θ of a categorization function $f(\theta, x_i)$, and often perform worse in predicting human categorization behavior. For most similarity functions k, we can often alternatively find a feature transformation ϕ of the input space that yields identical fit given even the simplest parametric models (Shawe-Taylor & Cristianini, 2004), or formally when

$$\phi(x_i) \cdot \phi(x_i) = k(x_i, x_j). \tag{1.1}$$

This implies that fit to human behavior in this case can potentially be determined almost entirely by the stimulus representation employed in the experiment (usually conceived and fixed *a priori* by the experimenter) rather than by an accurate model of the cognitive process. In fact, a similar identifiability problem should arise for nearly any representation-process pair, the building blocks of countless cognitive theories.

1.2 The Representation Problem

The question of how to best represent even simple stimuli is not always straightforward (Tversky, 1977). What type or form of representations should we use? How will we derive them in ways that parallel human bias and learning strategies? Which aspects of multiply-represented stimuli must we capture in order to study a phenomenon of interest? For more naturalistic stimuli (e.g.,

color photographs of objects and scenes), this becomes increasingly problematic. Although we may still ultimately be wrong, there are many fewer ways to represent stimuli like Gabor wavelets in ways that are obviously highly relevant to our experiments. When it comes to the complexity of the real world, human behavior seems all the more impressive, and our candidate models for such robust behavior are limited.

Austerweil and Griffiths (2013) provide an elegant account of rational feature learning, but full empirical validation is intractable given the complexity of real-world stimuli. The fact that this is true is not a valid criticism of the analysis, since the question of tractability is engaged at a different level of analysis (Marr, 1982), but the difficulty in engaging at this lower level is a formidable challenge to providing a complete scientific account of cognition. Only recently has a form of representation (and representation learning) emerged from researchers in computer science that allows for models that rival human performance on a number of complex perceptual tasks (see for example Krizhevsky, Sutskever, & Hinton, 2012; Long, Shelhamer, & Darrell, 2015), a topic to which we now turn.

1.3 The Ethos of Machine Learning & Real-World Performance

The landscape of machine learning research can look very different from psychology, even though they often share a common set of modeling tools (e.g., neural networks, statistical inference, etc). In fact, at least a handful of foundational methods in machine learning emerged from research in psychology (see Ackley, Hinton, & Sejnowski, 1985; Elman, 1990; Fukushima & Miyake, 1982; Rosenblatt, 1958; Rumelhart, Hinton, & Williams, 1986 for some notable examples). For our purposes, the differences will be of the most interest to consider.

Most (but not all) computational models of human cognition have been built alongside human datasets from well-controlled experiments, often in an artificial lab setting. By contrast, machine learning complements a pure theory element (in many respects insightful to formal theories of human learning) with a strong showing of highly practical applications aimed at performance/prediction in the face of considerable noise and complexity, and is ornamented by competitions, benchmarks, and formal challenges that reinforce those priorities.

This has proven extremely effective in recent years, blending classic methods with modern components and large datasets, and challenging human performance in previously difficult domains like object categorization (Krizhevsky et al., 2012), video game playing (Mnih et al., 2015), and a highly complex, millennia-old, abstract strategy board game (Silver et al., 2016). If we might hope for a proving ground outside of psychology for systems that could rival human intel-

ligence in a complex world—and may help explain something about general, animal, or human intelligence along the way—we could do much worse. In fact, little else comes to mind.

1.4 The Prediction-Interpretation Trade-off

Modern machine learning places a strong emphasis on highly expressive models and massive training datasets from which to learn. This is not a particularly dogmatic position—it is an obvious requirement of any learning system that must internalize a large, complex world, and a consequence of requiring that we make predictions for increasingly more data (human brains for example are drastically more massive, and accomplish much more).

The subset of these models that we will make use of in this thesis are flexible, parametric models with potentially tens or hundreds of millions of automatically learnable parameters or "knobs" that are obviously difficult to interpret, even if a high-performing solution is learned. Learning fully (or even partially) interpretable models this way is difficult to say the least, and a problem that is not likely to be solved quickly. This is a high price to pay for good predictions, and it seems likely that any compromise between classic model-building strategies and data-driven learning is subject to a tradeoff between prediction and explanation (see for example Plonsky, Erev, Hazan, & Tennenholtz, 2017).

To the engineer, the position along this continuum is dictated by the problem specification. To a scientist who values explanation above all else, the value in compromise is less clear. However, if scientific models are to accommodate greater complexity, and larger data spaces, they will not be exempt from this trade-off, and the emergent dynamics of even fully hand-built models will eventually become difficult to interpret, at least at the outset. It is also worth pointing out that machine learning may not have been such an extreme example just a few years ago; however, in the course of following what works, there has been a natural shift.

1.5 A Framework for Methodological Integration

A theme for the current work is to prevent ourselves from getting immediately bogged down by the obvious challenge of the Prediction-Interpretation Trade-off, partly due to its immense difficulty, but primarily to put forth what I think is a practical framework for making the best use of both methods in psychology, *right now*, and for bringing them into synergy as best one can at this stage. An interesting side-note is that our ability to exploit this potential synergy increases by the month, at least for some problems of interest, given the recent pace in machine learning

research. That is not to say that their limitations as say, surrogate (replacement) cognitive models necessarily decreases, but that their usefulness as complementary tools surely increases. I will make this point more concrete later on.

1.5.1 Black Boxes as Stepping Stones

To start, it is important to understand that a machine learning model does need to constitute a valid cognitive model in and of itself to be useful or informative, nor must it learn like humans, without their help, or with the same objectives (in many cases, it is probably enough that it solve a similar, or related problem, as we will see). That is not to say that these aren't also interesting questions/goals for particular models (e.g., to what extent could we show that a learned model captures what we had intended as an explanation of a cognitive process). In fact, these types of questions may bear fruit. However, it is more immediately important that such models *predictive* of human behavior. A model that is said to be "explanatory", but is not predictive, is no more valuable than a model that is neither explanatory nor predictive, since it simply does not describe the world. However, a model that is predictive, but not fully explanatory, can at least be argued to be relevant (even though we may ultimately reject it).

Moreover, a predictive model can be of use even if it is yet to be explained, or is not likely to be explained. In particular, if we like, we can position machine learning models as components of a larger scientific model. One way to do this is to model a prerequisite process with a one such "black box" model (and to suspend full understanding), assuming we can show that its behavior and performance are satisfactory, in service to a higher-level, or downstream cognitive process of interest. For example, understanding the process of memory search may not require the explicit integration of a representation learner (i.e., we can choose only to model memory search given an adequate representation, so long as our question does not primarily concern their interaction). Chapters 3, 4, and 5 can be interpreted either partially or fully in this way.

Another useful perspective is to think of machine-learned representations as imperfect, but ultimately useful initializations, surrogate representations that hover surprisingly near those of humans in a vast hypothesis space. Depending on just what the discrepancies are, one can answer certain questions that are invariant to some trivial differences. For example, a representation that entangles human-relevant features and their importance (salience) is more valuable than one that disentangles feature importance from a human-irrelevant representation (e.g., raw image pixels), because feature importance is more easily modeled through interpretable means. We will come across such examples in Chapters 4 and 5.

1.5.2 Iterative Bootstrapping

Ultimately, supposing we are successful in the sorts of strategies proposed above, one hopes that the insight into whatever natural processes that can now be studied can be in turn used to shape and constrain more human-relevant machine learning models. For example, we could ask which datasets to feed our black box learner in order to better support downstream modeling. Finding better surrogate representations means our cognitive models can obtain higher precision for complex, messy problems, which in turn might give us a better sense of what other prerequisites might further support our improved cognitive model. Put differently, the aim is to bootstrap from black box models, learning what we can from them, augmenting where possible, and iteratively exploiting the improvements they provide to constrain the next step.

1.6 An Overview of the Current Work

In the next chapter, I review the traditional role of representation in psychology, and modern tools for learning representations that we will exploit for the length of the thesis. The following chapters provide concrete case studies of the general framework espoused above. Vision is chosen as the primary superdomain of study, given what we will see is an existing toolkit of quite conveniently low-hanging fruit for which psychologists can make use, although one could reasonably expect some level of generalization to other perceptual domains (e.g., audition).

Chapter 3 presents a case study of the superiority in predicting human behavior with an off-the-shelf deep neural network. In particular, the objects that humans are likely to collectively remember in a large database of natural scene images can be predicted with a high degree of accuracy (even if say two objects from the same category and pose are remembered differently). In Chapter 4, I ask why these networks' representations should be expected to work (i.e., how are they like representations people have), identify minimal-cost corrections to apparent discrepancies, and assess the usefulness of these corrections in a subsequent discriminative modeling task. Finally, I show how another class of networks can be used, without any explicit augmentation, to capture human generative models (i.e., human category concepts as subjective probability distributions). In Chapter 6, I reflect on the practicality of sustaining such an overall paradigm, make clear some crucial and likely enduring limitations, and present an outline for future work.

Events and developments, such as... the Copernican Revolution... occurred only because some thinkers either decided not to be bound by certain "obvious" methodological rules, or because they unwittingly broke them.

Paul Feyerabend, *Against Method*

2 Background

The concept of **mental representation** stretches back to Aristotle's *De Anima* (350 BC), and has been the subject of considerable philosophical and scientific inquiry going forward (Cummins, 1989), but its most widespread application to psychology is rooted in the cognitive revolution. Before this shift, behaviorism—an early attempt to support a rigorous science of animal (and human) behavior—had taken an extreme methodological position. As a strong proponent, Skinner (1957, 1977) argued that when the description of mental models appears isomorphic to the contingencies in the environment (e.g., our apparent "cognitive associations" look suspiciously like associations in the world), there is little to be gained by doing anything more than enumerating those contingencies, which are the reason for that correspondence in the first place. The counterexamples to this generalization (see for example Chomsky, 1959) have rung loudly enough to eliminate such a constraint on psychology as a productive science.

In its place is indeed a collection of rich mental models with which we think of the mind as actively representing the world, sometimes apparently incorrectly, sometimes with a strong but justified bias, and often in service to other goals. In this chapter, we will review common ways of thinking about, inferring, or learning representations in both psychology and machine learning, which will lay the groundwork for their integration in the following chapters.

2.1 Representations in Psychology

A great deal of psychological research has been devoted to identifying the content of so-called *psychological representations*, their innate constraints, and the way in which they reflect the world. The principal challenge is that, unlike outward behavior, representations in our minds cannot be directly observed, although they can potentially be inferred through clever methods.

2.1.1 Inferring Representations

Shepard (1987) famously developed a method for inferring certain classes of representation with stunning detail, simply by making the assumption that human generalization behavior should be law-like (consistent). Previously, behavorist paradigms for studying generalization from one stimulus to another had created the illusion that functions describing the relationship between physical stimulus properties and generalization were highly inconsistent, and therefore nearly inexplicable. If instead the organism's mental representation of the stimulus was known, then perhaps human and animal behavior would seem less arbitrary. Indeed, using the generalization data themselves (pairwise confusions, similarity judgments, or recall order), and the quite weak assumption of monotonicity (i.e., differences in the mental representation of stimuli must always decrease generalization), complex psychological structure such as Newton's color wheel (Shepard, 1980) can be recovered.

This method, termed *Non-metric multidimensional scaling* (NMDS), uses gradient descent to infer representations in the form of manifolds embedded in geometric spaces. Points (stimuli) in this space preserve the ordination of the generalization data. More formally, the iterative algorithm minimizes an objective function referred to as *stress*, defined as

$$\sqrt{\frac{\sum_{i,j}\left(f_m(s_{ij}) - d_{ij}\right)^2}{\sum_{i,j}\left(d_{ij}\right)^2}}\,, \tag{2.1}$$

where f_m is a monotonically decreasing function of human-derived proximities (similarity or generalization data), and d_{ij} is the distance between current point representations of stimuli i and j in the inferred coordinate space. On each iteration, f_m is refit using monotonic regression (in psychology, often using a single-parameter exponential curve), and point representations are updated using gradient descent. The manifolds embedded in the resulting solutions can be mapped with complementary developments like the ISOMAP algorithm (Tenenbaum, De Silva, & Langford, 2000).

Where non-spatial representations are more appropriate (Tversky, 1977), alternative but analogous methods for fitting the stimulus generalization matrix have been developed (Shepard, 1980). *Additive clustering* (Shepard & Arabie, 1979) assumes that stimuli are represented by a discrete feature set, and that stimulus generalization is a weighted sum of shared features between two stimuli:

$$s_{ij} = \sum_k w_k f_{ik} f_{jk} \,. \tag{2.2}$$

where s_{ij} is the similarity between stimuli i and j. This model allows clusters to overlap (i.e., for features to be shared by any subset of stimuli), and can be viewed as a discrete analog of the eigenvalue decomposition of a covariance matrix (i.e., *principal component analysis*), where both binary features f_{ik} and weights w_k must be inferred. Additive clustering is general enough to encompass many potential discrete structures of interest, but this also implies a massive search problem due to the number of possible feature configurations (many heuristics are often used). For this reason, it is sometimes worth introducing additional constraints. For example, another popular form of discrete clustering, *hierarchical clustering*, is a special case of additive clustering where clusters are strictly nested (Shepard, 1980). The resulting familiar tree structure is often used to describe human knowledge about hierarchical taxonomies, but can also be to model phenomena such as generalization bias (Xu & Tenenbaum, 2007).

The problem of choosing what form of representation to infer from human behavior mirrors the problem that humans face as learners (i.e., how to best carve up the world). Several decades after most of the above methods were conceived and first applied, Kemp and Tenenbaum (2008) developed a method to infer both the optimal form (spatial, non-spatial, or other forms derivable through graph grammars), as well as the optimal structure for that form with a single procedure. To further avoid the problem of search over potentially biased, qualitative structural forms that are hand-specified by the algorithm designer, a sparsity prior is often the only necessary constraint to successfully infer relevant graphs (Lake, Lawrence, & Tenenbaum, 2018).

Inferred mental structures have been applied widely in psychological modeling, often yielding base representations on top of which to model phenomena such as categorization (Kruschke, 1992; Rips & Shoben, 1973), analogy (Ehresman & Wessel, 1978; Rumelhart & Abrahamson, 1973), and memory (Caramazza, Hersh, & Torgerson, 1976; Schwartz & Humphreys, 1973) to name just a few. The assumption (or the convenience) that a cognitive model can be factored into representation and process components is common in psychology, and although obviously reductive (see Goldstone, 1994a), it is understandably even more difficult to model representation-task dynamics given already shaky ground to stand on.

2.1.2 Limitations of the Psychologist's Current Toolkit

Despite great success, the methods above share a number of important limitations, since they are a function of the same form of human behavioral data and stimulus sets, specifically:

- Inferring representations directly from behavior this way is costly. Human generalization data is derived from the unique pairwise groupings $n(n-1)/2$ of all stimuli under consideration, which must be collected during an experiment for several human participants, whereas humans obviously learn from streams of stimuli by themselves.

- Fitting to stimulus generalization data implies that our algorithms must evaluate all stimulus pairings (as opposed to each stimulus by itself) during each iteration of the inference procedure, greatly effecting the ability of the algorithms to scale to large sets of stimuli. The computational complexity of a single iteration of NMDS for example is $O(n^2)$.

- Learning complex/high-dimensional representations for large naturalistic datasets would require lots of data, which is both experimentally and computationally costly as discussed above. This eliminates the possibility of inferring a staggeringly large class of human mental content.

- The representations learned from these methods are obviously sensitive to the stimulus set employed, especially since these sets are often small and considerably biased. Moreover, it is unclear whether it makes sense to infer a single representational picture of stimuli when they are more complex (i.e., we might think object taxonomies differently depending on the task at hand). If we must start with a single base representation, then we want something general and expressive enough to be selectively rescaled for particular contexts of interest.

- Directly inferred representations for a set of stimuli in relation to each other cannot by design be extended or generalize to unseen stimuli, since an explicit transformation of the raw stimuli into psychological structures is not learned. This also implies that we cannot make observations about the complexity of such a transformation, which would otherwise provide invaluable information about what inductive biases humans might be employing.

- Methods that infer representations directly miss out on the chance to learn them in ways that might mirror human learning. A model that successfully learns a good representation given the appropriate data is a validation of our grasp of the necessary constraints humans bring to the learning problem.

2.2 Representations in Machine Learning

2.2.1 The Similarity-Representation Duality

We saw above that Shepard exploited the relationship between similarity (inter-stimulus generalization) and representation in order to recover the latter from the former. This relationship, also demonstrated in Chapter 1 in the context of non-identifiability of categorization algorithms, is often explicitly leveraged in machine learning by use of what is referred to as the "kernel trick". For example, the Support Vector Machine (SVM) is a highly effective linear discriminator that can maximize the margin between classes in a dataset (Shawe-Taylor & Cristianini, 2004). It is easily regularizable, and the solution is a global optimum. Applying the kernel trick to SVMs allows one to learn nearly any nonlinear discriminator, corresponding to a linear discriminator in some space $\phi(x)$, without ever having to learn ϕ. From a theoretical perspective, there are little downsides to such a learning algorithm aside from having to identify sufficiently expressive kernels.

On the other side of the spectrum, *neural networks* are the most salient examples of a ϕ-learners (explicit representation learners). However, unlike kernel SVMs for example, the convergence properties of neural networks are less well-behaved (Choromanska, Henaff, Mathieu, Arous, & LeCun, 2015; Saad & Solla, 1995), which historically has been taken as reason enough not to use them. By practical reputation, neural networks can be both unstable and unreliable. Despite this seemingly damning property, neural networks have somewhat recently become the dominant method in machine learning. By luck, this has in one important way re-aligned many modern machine learning efforts with psychology. It's beyond a surface plausibility that humans engage in a fair share of rich representation learning (in lieu of kernels operating over raw input). In fact, the algorithm for training neural networks that is still in use today was conceived by psychologists (Rumelhart et al., 1986), constituting another psychological method that has been used to derive representations of simple stimuli (Kruschke, 1992). Some important reasons why such methods are now beginning to scale are discussed in the next section.

2.2.2 Deep Neural Networks

The re-emergence of artificial neural networks in machine learning comes in the form of *deep neural networks*. From their most typical description (see LeCun, Bengio, & Hinton, 2015), it would seem like there is nothing new under the sun, but there are in fact a number of very important distinctions that set them apart from their close cousins. By tradition, they are first

described as networks with many "layers" (a composition of many—jointly learned—parametric functions) that transform input(s) to output(s). We will proceed with this assumption for the time being, and then discuss some crucial differences.

2.2.2.1 *Multi-layer Perceptrons*

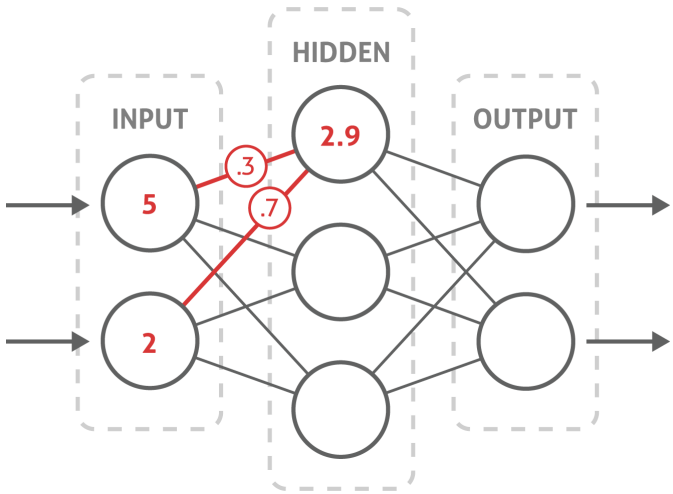

Figure 2.1: Multi-layer Perceptron. The output of each non-input layer is a function of a set of learned weights and the previous layer's output. An example for the first hidden node is shown in red: $(.3 \cdot 5) + (.7 \cdot 2) = 2.9$.

Multi-layer neural networks that can successfully be trained have been around for some time. Generalizations of single-layer *perceptrons* (Rosenblatt, 1958), multi-layer architectures are universal function approximators (Cybenko, 1989) given at least one ("hidden") layer between input and output layers, and as many hidden units as needed (and more practically, like all expressive learners, given enough data). More formally, a multi-layer perceptron (MLP) is a composition of functions, each of the form

$$o_j = \varphi(f_j(x, \theta_j)). \tag{2.3}$$

where the output o_i is determined by θ_j, a learned parameterization of a linear function f_j at layer j, and φ is any differentiable (usually fixed, and simple), nonlinear *activation* function (e.g., the sigmoid function). Since f is linear, each layer is an inner product between the input (or previous layer's output) and the set of parameters or *weights*, then passed through the activation function. Output at the final layer of the network with L layers is then a composition of the form

$$o_L = \varphi(f_L(\ldots \varphi(f_1(x, \theta_1),\ \theta_L)), \tag{2.4}$$

This implies several levels of representation, each layer representing the input in a different way, and as a function of the previous representation. Given such a simple composition, multi-layer perceptrons, including their modern, deeper instantiations, can be trained to minimize any of a number of typical loss function with a simple application of the chain rule, called *backpropagation* (i.e., of error) for MLPs (Rumelhart et al., 1986).

2.2.2.2 *Deep, Stable, & Appropriately Biased*

The most successful deep neural networks are indeed deep (and wide), but more importantly, they belong to a number of families of more specific architectures, such as *convolutional neural networks* (CNN; Krizhevsky et al., 2012), and *recurrent neural networks* (RNN; Graves, 2008). While any aspect of a network architecture (e.g., how many layers) can be considered an inductive bias, it has not surprisingly been more effective to impose much stronger constraints on modern networks, building in relatively weak (compared to hand-engineered expertise systems and cognitive models), but effective assumptions about the domain, and allowing for the flexibility of learning everything else.

These more specific architectures have risen in popularity alongside accompanying methods for eliminating more general issues with so-called "vanishing gradients" that plague deep networks (Glorot, Bordes, & Bengio, 2011; Hochreiter & Schmidhuber, 1997), as well as a number of innovative regularization and efficient ensembling techniques (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014), and methods for speeding up and stabilizing training (Ioffe & Szegedy, 2015). For that matter, increased physical computing power and access to large, fairly high-quality datasets have also played considerably large roles in their success. However, it can be argued that the right, architecturally implemented inductive biases are of the most value, the list of which continues to increase (see Battaglia et al., 2018 for a review of important recent additions). Given the focus on the visual domain for this thesis, we will review the corresponding architecture and bias in detail below.

2.2.2.3 *Convolutional Neural Networks*

Convolutional neural networks (see Figure 2.2) are specialized for visual tasks, and employ layers that make an amazingly obvious assumption about images, namely, that any sub-image pattern

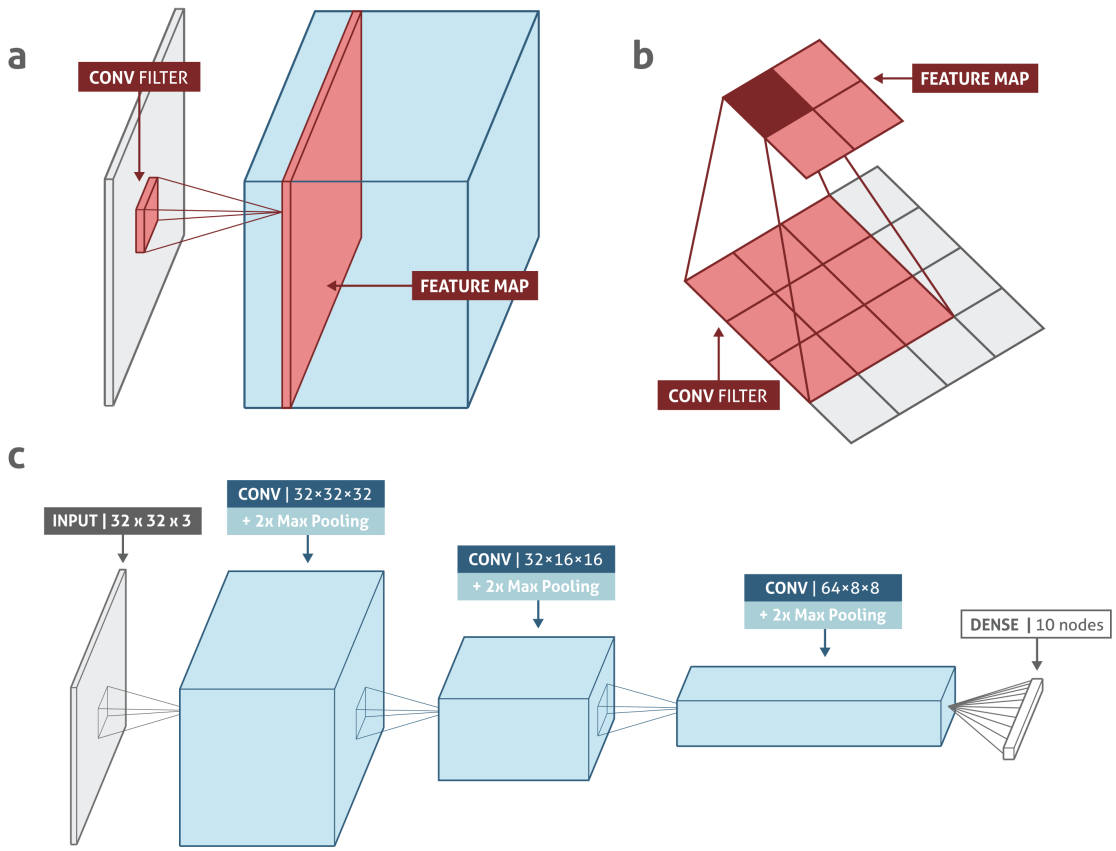

Figure 2.2: a. Convolutional layer. A single sliding convolutional filter produces a single output feature map within a larger volume corresponding to filter-map pairs. **b. Close-up of the 2D convolution operation.** A 3×3 filter (bottom light red) performs an inner product operation and produces a single scalar output (dark red square) for that location in the image (larger grey grid). Fitting the same filter into the remaining portions of the input will yield and 2×2 feature map. **c. A typical convolutional neural network.** A 3-channel color image is fed through three convolutional layers, each followed by pooling to shrink the spatial resolution by a factor of two. The channel dimension becomes a feature dimension, and is greatly expanded in the last conv layer. The final layer fully connects the last conv representation to a set of notes representing output classes, using a softmax function to compute classification probabilities.

can occur at any spatial position in the overall image (they are said to be *translation invariant*). The same goes for higher level patterns. That is, the co-occurrence of two eyes and a mouth can occur at any position in an image's reference frame (much like our eyes). These high-level patterns are simply patterns of low-level patterns (image patches), hence *deep* CNNs.

Now it is considerably less clear which specific patterns our models should look for, and this

is precisely the reason why one would want to learn them from a large set of examples. If we wanted, we could focus our efforts to find a sufficiently flexible model with enough training data to *learn* translation invariance (at least implicitly), but this is almost certainly a waste of data and model power (and it could take a very long time), although see Z. Lin, Memisevic, and Konda (2015) for an interesting experiment. The fact that we've found an assumption that gets us near human performance on certain tasks for the first time in history is good reason not to make the problem harder on ourselves.

CNNs excel at perceptual tasks such as object recognition (Krizhevsky et al., 2012) and image segmentation (Long et al., 2015), among many others. Early architectures were mainly aimed at classification (Figure 2.2c), often combining a fairly constrained recipe of convolutional layers, "pooling layers" (to reduce dimensionality at each layer), and traditional fully-connected layers, but subsequent work has successfully eschewed many of these non-focal components (Springenberg, Dosovitskiy, Brox, & Riedmiller, 2014), as we might expect given our argument about essential biases.

The essential component of CNNs, the convolutional layer (Figure 2.2a), is a learned set of patterns or *filters* (Figure 2.2b). Like fully-connected layers, the filters are sets of weights, but smaller than the input, and the inner product (for each filter in a layer) is computed at each position in a grid of some specified resolution across the original image. For example, if a subset of the image the same size as the filter has a similar appearance to the filter at the first layer, the inner product will be high, and the output feature-by-position activation map (a volume in this case) will reflect this. If we want instead to output an expanded spatial representation (in the case of outputting an image for example), we can use *transposed convolution* (Long et al., 2015), which can be thought of as "painting" an image using a filter as a "brush", often called *deconvolution*, or *upsampling* since the spatial resolution increases with additional layers. Like all neural networks, learning the parameters of each layer is done jointly, such that all parts of the network take each other into consideration when searching for a good solution.

2.2.3 Generative Neural Networks

Deep neural networks have also been used to learn generative models of the input domain. Specifically, the goal is to model $p(x)$, so that we can for example compute the probability of a particular datapoint x_i, or sample/synthesize new data. This also allows us to learn a representation of the data that is not dependant on any one particular task or set of tasks (a generic, and potentially optimal compression), a property that also appears in cognitive models (e.g.,

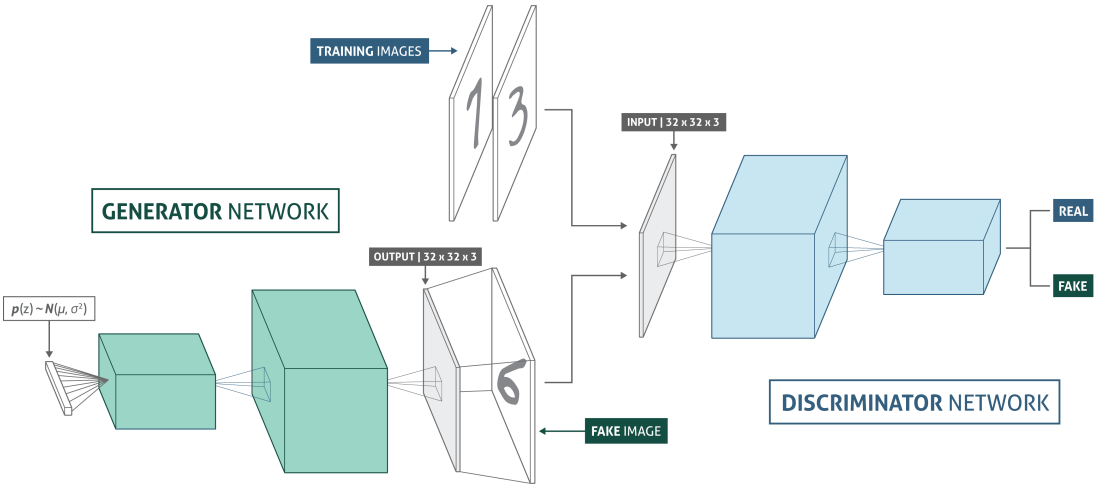

Figure 2.3: Typical generative adversarial network for images. The generator network G (green) maps some simple random variable (e.g., multivariate Gaussian) inputs to image-sized outputs. A discriminator network D (blue) learns to distinguish between generated (fake) images and images from a real dataset. The loss function for each network is symmetric: the discriminator maximizes the probability of real/fake guesses, which the generator simultaneously minimizes.

Austerweil & Griffiths, 2013; Kemp & Tenenbaum, 2008).

While a variety of comparble deep generative modeling approaches exist (e.g., Kingma & Welling, 2013; Oord, Kalchbrenner, & Kavukcuoglu, 2016), we focus on *generative adversarial networks* (GAN; Goodfellow et al., 2014), which when built with convolutional/deconvolutional layers (Radford, Metz, & Chintala, 2015), are arguably the most effective current method for subsecond generation of convincing image samples.

GANs approach the problem of learning $p(x)$ by teaching a *generator* network to mimic the training dataset. To do this, the data is represented by some prior $p_z(z)$, often a simple distribution (e.g., a multivariate Gaussian or uniform distribution). The learned generator $G(z; \theta_g)$ is a mapping from the latent encoding z to the data space x, where G is a differentiable function (e.g., a deconvolutional network) parameterized by θ_g. This defines a new distribution $p_g(x)$. Initially, p_g will be a poor approximation of $p(x)$, and feeding noise inputs $p_z(z)$ into G will yield random pixel configurations.

To train G, a second *discriminator* network $D(x; \theta_d)$, parameterized by θ_d, takes image-sized inputs, specifically both images from a dataset sampled from the true $p(x)$ and fake/synthesized images generated by G, and outputs a single probability that each image was indeed sampled from the true $p(x)$. D is trained to maximize the probability of assigning correct "real" labels to

samples from $p(x)$, and correct "fake" labels to samples from $p_g(x)$, and G is trained to minimize $\log(1 - D(G(z)))$, pitting the two networks directly against each other. This adversarial process is a two-player minimax game. During training, the game must be well-balanced (e.g., neither network should completely dominate the other), so that each provides a useful learning signal to the other. At the game's theoretical equilibrium, D is $\frac{1}{2}$ everywhere, and

$$p_g(x) = p(x) \tag{2.5}$$

when the image distribution has been captured. In practice, G is never perfect, but is powerful enough to compress realistic-looking images into a manageable encoding, sampled with plausible frequency (although mode collapse is a common issue in GANs). A schematic of the overall dual-network model and training process is illustrated in Figure 2.3.

2.3 Deep Neural Networks in Psychology

Despite bearing a lose analogy to human brains, and being rooted in classic psychology literature, machine learning is primarily focused on solving engineering problems, and not necessarily engineering human-like intelligence. However, machine learning models are almost always trained on datasets that were created, stratified, and annotated/labeled by humans. Data-driven language models are trained on human-written corpora, and speech production systems are trained on thousands of hours of human speech. It would not be particularly surprising if a model trained to explore the world, talk, reason, or understand like humans might reveal something about human cognition, since many of these tasks are difficult to model at all with any real success.

Lake, Zaremba, Fergus, and Gureckis (2015) were the first to propose an explicit, "synthetic psychology" of deep neural networks, mining them for insights about the cognition of concepts, demonstrating for example that output class probabilities from a deep CNN were predictive of human category typicality ratings. More recently, it was shown that human shape sensitivity for natural images could be explained well for the first time using a deep neural network model (Kubilius, Bracci, & Op de Beeck, 2016). A common theme across such works is to learn a successful (and often superior) model, and explain it after by further probing the behavioral and representational characteristics of the trained network. That is, it is more common to see deep neural networks used as cognitive model learners (in some ways not unlike how humans learn from data), as opposed to components of a larger cognitive modeling pipeline.

2.4 Summary

In this chapter, we reviewed classic psychological methods for inferring human mental representations, which while surprisingly innovative, do not seem well-suited to large, complex, naturalistic datasets. By sharp contrast, we can derive meaningful representations for complex inputs by learning them from weakly-constrained, but still very powerful machine learning tools, as long as we have enough data to feed them. However, these methods have largely progressed in parallel to efforts in cognitive science, with little integration between the two fields. Some initial work has demonstrated value in treating deep network models as cognitive models (perhaps that learn somewhat like we do), but as I have argued, this is not the only promising framework for integration. In the following chapters, I put all of the tools reviewed in this chapter to use in understanding human and machine alike.

Normal science, the activity in which most scientists inevitably spend almost all their time, is predicated on the assumption that the scientific community knows what the world is like.

Thomas S. Kuhn, *The Structure of Scientific Revolutions*

3

Predicting the memorability of objects in natural scenes

Recent large-scale web experiments on **image memorability** have shed light on what distinguishes the collective memorability of thousands of diverse, naturalistic images, a refreshing departure from the sorts of laboratory studies discussed in Chapter 1 that are built around small stimulus sets. However, they have often relied on image annotations and simple featural representations to make predictions and explore contributing factors. Partly for this reason, a more granular understanding, such as a knowledge of the contrasting memorability of specific objects in the image, especially within the same object class (e.g., similar objects with different surface appearance), remains illusive. This emerging, ecologically-focused study of a crucial cognitive process therefore presents a near-perfect case study for evaluating the utility of deep neural networks in predicting complex human behavior at scale.

Much of the content of this chapter was published in Dubey, Peterson, Khosla, Yang, and Ghanem (2015), the first-authorship for which was shared.

3.1 What Do We Remember When We Remember?

Early work on human memory capacity was unable to identify an upper bound—it appears to be effectively limitless—even when presented with streams of thousands of images (Standing, 1973). Although the proportion of images remembered decreases with additional stimuli, the total number remembered always increases. Brady, Konkle, Alvarez, and Oliva (2008a) later found that the information retained in such experiments is much more detailed than previously thought. Despite this, human memory is obviously imperfect, and much less work has been aimed at identifying exactly *what* we remember (or forget).

For example, consider the image on the left side of Figure 3.1. Even though the person on the right is comparable in size to the person on the left, he is remembered far less by human participants, indicated by their respective "memorability scores" (across-subject recall rates) of 0.18 and 0.64. Moreover, people tend to remember the person on the left and the fish in the center, even after three minutes and more than seventy additional visual stimuli have passed. Interestingly, despite vibrant colors and considerable size, the boat is far less memorable with a memorability score of 0.18. Why do we remember some of these objects better than others, and how do these objects influence the overall memorability of the scene?

Figure 3.1: Not all objects are equally remembered. The image on the left shows objects from a scene along with their respective memorability scores obtained from our experiment. Certain objects (fish and person on the left) are more memorable than other objects. The image on the right shows the ground truth map generated from the object segments and memorability scores, and the ultimate target for pixel-level prediction.

Some initial hints about the answers to these questions come from recent work on memorability at the image level (Bylinskii, Isola, Bainbridge, Torralba, & Oliva, 2015; Isola, Parikh, Torralba, & Oliva, 2011; Isola, Xiao, Parikh, Torralba, & Oliva, 2014; Isola, Xiao, Torralba, & Oliva, 2011; Khosla, Bainbridge, Torralba, & Oliva, 2013; Khosla, Xiao, Isola, Torralba, & Oliva, 2012).

The typical experimental design employed for these studies involves empirically estimating the memorability of a set of scenes (overall images). Whether an image was remembered due to a complex interaction between objects, or simply due to a single salient object for example is not immediately reflected in the recall rates for each image. Some of this knowledge can be potentially inferred from the memorability score of an image alone (Isola, Parikh, et al., 2011; Khosla, Xiao, Torralba, & Oliva, 2012). However, these methods will ultimately require ground truth object memorability data to be properly evaluated. Moreover, predicting such a detailed map of memorable content is likely to require innovative methods for representing complex image content.

3.2 Goals for the Chapter

To enable the direct study of content/object memorability by humans, we collect a large dataset of ground truth object-level memorability scores and conduct an extensive first empirical investigation of memorability at the object level. This allows for a simple yet powerful strategy that provides detailed answers to many interesting questions at hand. We then systematically explore the memorability of objects within individual images and shed light on the various factors that drive that memorability. In exploring the connection between object memorability, saliency, object categories, and image memorability, this chapter makes several distinct contributions:

- While previous work has tried to infer such knowledge computationally (Khosla, Xiao, Torralba, & Oliva, 2012), this work is the first to directly quantify and study what objects in an image humans actually remember, providing a ground truth with which to test predictive and explanatory models.
- We uncover the relationship between visual saliency and object memorability, and demonstrate those instances where visual saliency directly predicts object memorability and when/why it fails to do so. While recent work has explored the connection between image memorability and visual saliency (Bylinskii et al., 2015; Kim, Yoon, & Pavlovic, 2013; Mancas & Le Meur, 2013), our work is the first to explore the connection between object-level memorability and ground-truth human visual saliency.
- We make significant headway in disambiguating the link between image and object memorability, showing that in many cases, the memorability of an image is primarily driven by the memorability of its most memorable object, but not always.
- Finally, we show that deep neural networks appear to capture enough necessary information to beat out several competitive predictive models of object memorability.

3.3 Measuring Object Memorability

As a first step towards understanding the memorability of objects in images, we compile an image dataset containing a variety of objects from a diverse range of categories. We can then measure the probability that each object in each image will be remembered by a large group of participants after a single viewing, providing ground truth memorability maps for objects inside images (defined as image segments). This allows for a precise analysis and prediction of the memorable elements within an image.

Toward this, we utilized the PASCAL-S dataset (Li, Hou, Koch, Rehg, & Yuille, 2014), a fully segmented dataset built on the validation set of the PASCAL VOC 2010 (Everingham & Winn, 2010) segmentation challenge. To improve segmentation quality, we manually refined the segmentations from this dataset, removing all homogenous non-object or background segments (e.g. ground, grass, floor, and sky), along with imperceptible object fragments and excessively blurred regions. All remaining object segmentations were tested for memorability. Our final dataset comprises 850 images and $3,412$ object segmentations (i.e. an average of 4 objects per image), for which we gathered ground truth memorability through crowd sourcing on Amazon Mechanical Turk.

3.3.1 Object Memory Game

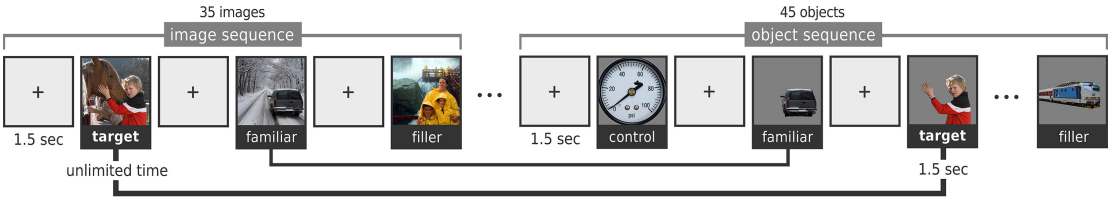

Figure 3.2: Object Memory Game. Participants viewed a series of images followed by a sequence of objects and were asked to indicate whether each object was seen in the earlier sequence of full images. Unfamiliar and singly-appearing filler images were used as spacing in the sequence, and reappearing control objects that were easy to identify ensured participants were paying attention.

To measure the memorability of individual objects in each image in the dataset, we created an alternate version of the Visual Memory Game through Amazon Mechanical Turk following the basic design in Isola, Xiao, et al. (2011), with the exception of a few key differences (refer to Figure 3.2). In our game, participants first viewed a sequence of 35 images, one at a time, with a 1.5 second interval between image presentations. Participants were then asked to remember

the contents and objects inside the images to the best of their ability. To ensure that participants would not heuristically inspect only the most salient and center-most objects, they were given unlimited time to freely view the images. Once they were done viewing an image, they could press any key to advance to the next image. After the initial image sequence, participants viewed a sequence of 45 objects, their task then being to indicate through a key press which of those objects was present in one of the previously shown images. Each object was displayed for 1.5 seconds, with a 1.5 second gap between each object in the sequence. Pairs of corresponding image and object sequences were broken up into 10 blocks. Each block consisted of 80 total stimuli (35 images and 45 objects), and lasted approximately 3 minutes. At the end of each block, the participant could take a short break. Overall, the experiment took approximately 30 minutes to complete.

Unknown to the participants, each sequence of images inside each block was pseudo-randomly generated to consist of 3 "target" images taken from the PASCAL-S dataset, whose objects were later presented to the participants for recall. The remaining images in the sequence consisted of 16 "filler" images and 16 "familiar" images. Filler images were randomly selected from the DUT-OMRON dataset (Yang, Zhang, Lu, Ruan, & Yang, 2013), while the familiar ones were randomly sampled from the MSRA dataset (Liu et al., 2011). In a similar fashion, the object sequence in each block was also generated pseudo-randomly to consist of 3 target objects (1 object taken randomly from each previously shown target image). The remaining objects in the sequence consisted of 10 control, 16 filler, and 16 familiar objects. Filler objects were sampled randomly from the 80 different object categories in the Microsoft COCO dataset (T.-Y. Lin et al., 2014), while the familiar objects were sampled from objects taken from the previously displayed familiar images in the image sequence. The familiars ensured that the participants were always engaged in the task and the fillers helped provide spacing between the target images and target objects. While the fillers and familiars (both images and objects) were taken from datasets of real-world scenes and objects, the control objects were artificial stimuli randomly sampled from the dataset proposed in Brady, Konkle, Alvarez, and Oliva (2008b). Control objects were meant to be easy to remember and served as a criteria to ensure quality (Brady et al., 2008b; Isola, Xiao, et al., 2011). Target images and their corresponding target objects were spaced $70 - 79$ stimuli apart, while familiar images and their objects were spaced $1 - 79$ stimuli apart.

All images and objects appeared only once, and each participant was tested on only one object from each target image to prevent objects from priming memory of other objects in the scene. Objects were centered within the image they originated from and non-object pixels were set to grey. Participants were required to complete the entire task, which included 10 blocks (~ 30

minutes) and could not participate in the experiment a second time. The maximum time that participants could take to finish the experiment was 1 hour. After collecting the data, we assigned a memorability score to each target object in our dataset, defined as the percentage of correct detections by participants (refer to Figure 3.1 for an example). Strict criteria was undertaken to screen participants' performance and to ensure that our final dataset consisted of quality participants. We discarded all participants whose accuracy on the control objects was below 70%. The accuracy of these participants on filler objects and familiar objects was greater than chance ($> 75\%$) demonstrating that our data consists of participants who were paying attention to the task. The mean time taken by the participants to view an image was 2.2 seconds with a standard deviation of 1.6 seconds. In total, we had $1,823$ workers from Mechanical Turk each having at least 95% approval rating in Amazon's system. On average, each object was scored by 16 participants and the average memorability score was 0.33 with a standard deviation of 0.28.

3.3.2 Human Inter-rater Reliability Analysis

To assess human inter-rater reliability in remembering objects, we repeatedly divided our entire participant pool into two equal halves and quantified the degree to which memorability scores for the two sets of participants were in agreement using Spearman's rank correlation (ρ), a non-parametric measure for testing the monotonic relationship between two variables. We computed the average correlation over 25 of these random split iterations, yielding an average correlation of $\rho = 0.76$. This high reliability in object memorability indicates that, like full images, object memorability is a shared property across participants. That is, people tend to remember (and forget) the same objects in images, and exhibit similar performance in doing so. Thus memorability of objects in images can potentially be predicted with high accuracy, and we can compare prediction performance to human consistency. In the next section, we study the various factors that drive object memorability in images.

3.4 Understanding Object Memorability

In this section, we aim to better understand how object memorability is influenced by a number of visual factors. Specifically, we study the relationship between object memorability and simple color features, visual saliency, and object semantics.

3.4.1 Can simple features explain memorability?

While simple low-level image features are traditionally poor predictors of image memorability (Isola, Xiao, et al., 2011), and with good reason (Konkle, Brady, Alvarez, & Oliva, 2010), the question arises whether such features play any role in determining *object* memorability in images. To address this question and following a similar strategy as in Isola et al. (2014), we compute the mean and variance of each HSV color channel for each object in our dataset, and compute the Spearman rank correlation with the corresponding object memorability score (refer to Figure 3.3).

We find that the mean ($\rho = 0.10$) and variance ($\rho = 0.25$) of the value channel correlates weakly with object memorability, suggesting that brighter and higher contrast objects may be more memorable. On the other hand, essentially no relationship exists between memorability and either the hue or saturation channels. This deviates slightly from the findings in Isola, Xiao, et al. (2011), which show mean hue to be weakly predictive of image memorability. This difference could be due to the fact that the dataset in Isola, Xiao, et al. (2011) contains blue and green outdoor landscapes that are less memorable than the warmly colored human faces and indoor scenes. In contrast, outdoor scene-related segments such as sky and ground were not included as objects in our dataset. From these results, we see that, like image memorability, simple pixel statistics do not play a significant role in determining object memorability in images.

3.4.2 What is the role of saliency in memorability?

Intuitively, we expect that objects within an image that are most salient are likely to be remembered, since they tend to draw a viewer's attention, i.e. a majority of his/her eye fixations will lie within those object regions. On the other hand, it is conceivable that some visually appealing regions will not be memorable, especially since aesthetic images are known to be less memorable (Isola et al., 2014; Isola, Xiao, et al., 2011). When can visual saliency predict object memorability and what are the possible differences between the two? Studying the relationship between saliency and memorability is paramount for understanding object memorability in greater depth.

To address this question, we utilize the eye fixation data made available for the PASCAL-S dataset (Li et al., 2014). First, we compute the number of unique fixation points within the image segment of each object and the correlation between this metric and the object's memorability score (refer to Figure 3.4, *left*). We find this correlation to be positive and considerably high ($\rho = 0.71$), suggesting that fixation count and visual saliency may drive object memorability considerably. However, the large concentration of points on the bottom left part of the scatter plot

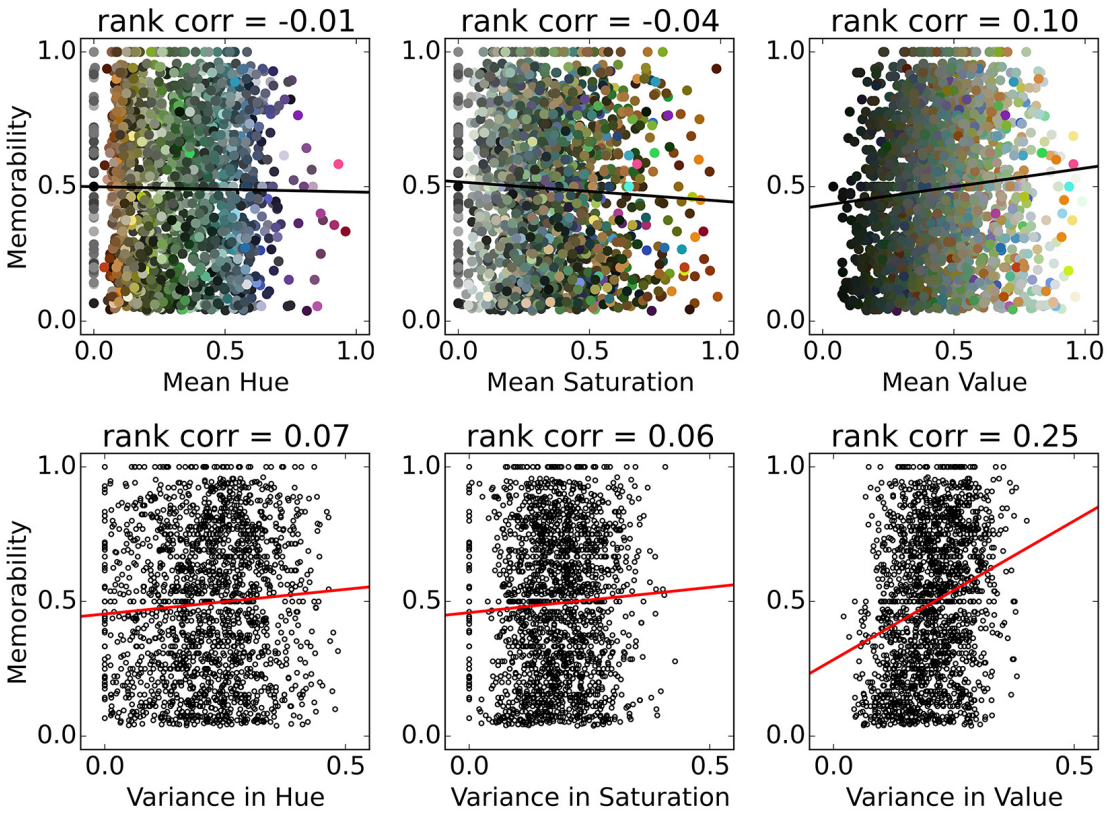

Figure 3.3: Simple color features do not explain object memorability. Correlations of object memorability scores with hue and saturation are near zero. Only value shows a weak correlation.

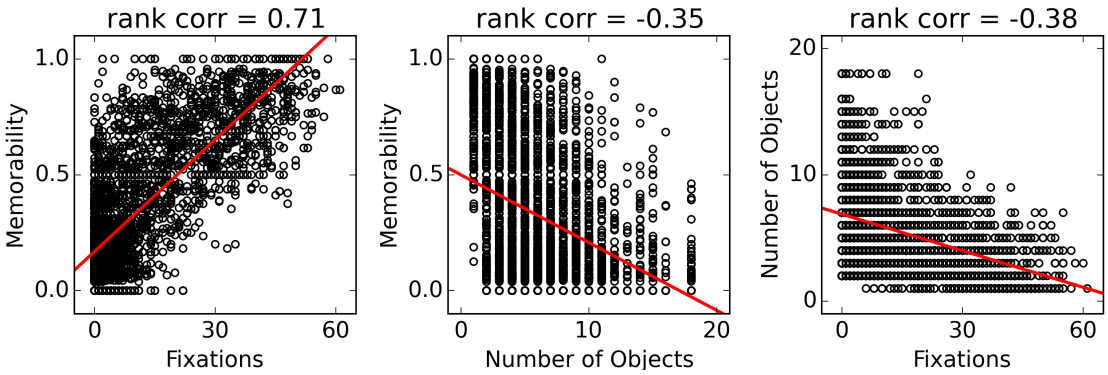

Figure 3.4: Correlations between memorability, fixation count, and number of objects. Left: Memorability and fixation counts correlate positively. Middle: Memorability and number of objects are negatively correlated. Right: Fixations and object counts are weakly negatively correlated.

in Figure 3.4 (left panel) suggests that part of the reason for this high correlation is that objects that have not been viewed (i.e. no fixation points associated with them) at all have essentially no memorability, and therefore will always imply correlation. If we remove these simple cases, we can examine whether or not the full range of memorability scores is predicted by fixation count.

To investigate this, we plot the change in correlation between object memorability and fixations as the minimum number of fixations inside objects increases. For each minimum fixation count, we compute the memorability-fixation correlation again, but *only* using objects that contain at least this number of fixations (refer to right panel of Figure 3.5). The decreasing trend in correlation indicates that as the number of fixations inside an object increases, the predictive ability diminishes significantly, indicating that the full range of memorability scores are not well predicted. In addition, Figure 3.5 (*left*) plots this correlation as a function of total number of objects in an image. Interestingly, as the number of objects in an image increases, the correlation between saliency, i.e. number of fixations, and memorability decreases sharply. The two remaining scatter plots in Figure 3.4 (*middle*) and (*right*) provide additional clues about the relationship between memorability and fixation count. Note that object count is negatively correlated with both memorability and fixation count. This makes sense, since people have more to look at in an image when more objects are present. In this case, they tend to look less at any single object, especially if some of these objects compete for saliency, and therefore may have a more difficult time remembering those objects.

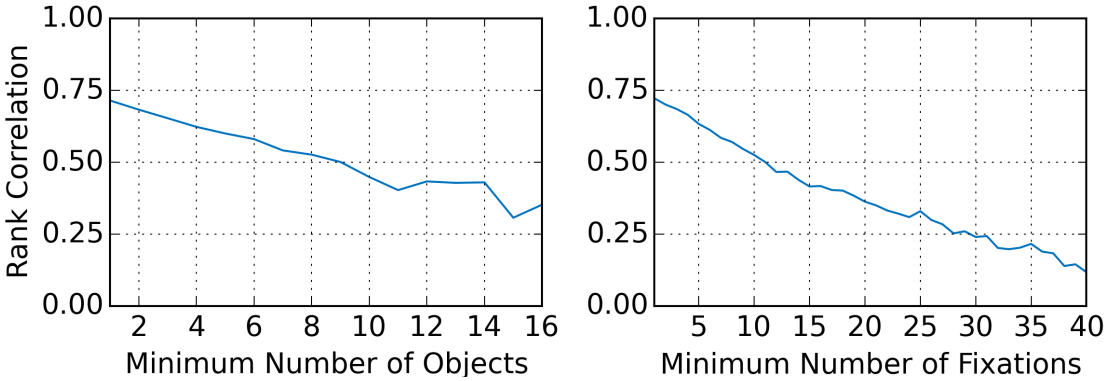

Figure 3.5: Correlation between object memorability and object fixation count as a function of minimum number of objects (left) and minimum number of fixations (right).

In summary, saliency is a good predictor of object memorability in simple contexts with few objects or when an object has few interesting points, but it is a much weaker predictor of memorability in complex scenes containing multiple objects that have many points of interest.

Figure 3.6: Memorability prediction by saliency in complex scenes. Top row: the memorability of the dog is low even though humans fixate on it. Bottom row: Humans look at the person more than the horse, but the horse is more memorable than the person.

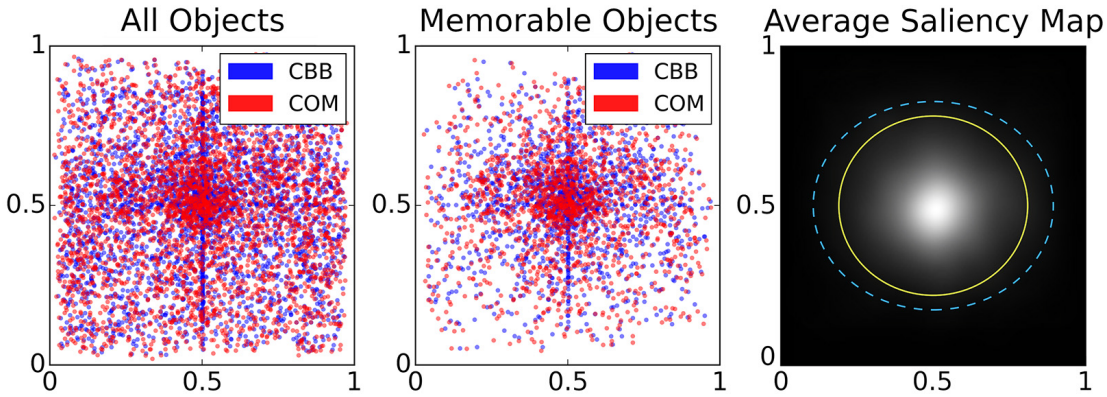

Figure 3.7: Memorable objects and fixation locations. Left: Normalized locations for all objects in the dataset. Both center of object bounding boxes (CBB, blue) and object center of mass (COM, red) are shown. Middle: Locations for memorable objects only. Right: Average ground truth saliency map across the entire dataset. The solid yellow line marks the region containing 95% of all normalized fixation locations. The dashed blue line marks the region with above-median memorable objects. Center bias is more strongly expressed in the fixation locations.

3.4.3 Center Bias

Figure 3.7 illustrates another example where saliency and memorability diverge. Previous studies related to visual saliency have shown that saliency is heavily influenced by center bias (Judd, Ehinger, Durand, & Torralba, 2009; Zhang, Tong, Marks, Shan, & Cottrell, 2008), primarily due to photographer bias (also evident in the left-most panel of Figure 3.7) and viewing strategy (Tseng, Carmi, Cameron, Munoz, & Itti, 2009). Since our data collection experiment tries to control for viewing strategy, memorability in our dataset exhibits comparatively less center bias than saliency. This is most apparent when considering the difference between the solid ellipse in the right-most panel of Figure 3.7), which shows where 95% of fixations are located, and the dashed ellipse, which shows where 95% of the above-median memorable objects are located.

3.4.4 How do object-level statistics affect memorability?

In the previous section, we explored the relationship between visual saliency and object memorability. Now, we explore how object-level information such as category labels and co-occurrence influences the probability of remembrance.

3.4.4.1 Are some object categories more memorable?

For this analysis, three in-house annotators manually labeled the object segmentations in our dataset. The annotators were provided the original image (for reference) and the object segmentation and asked to assign a single category to the segment out of 7 possible categories: animal, building, device, furniture, nature, person, and vehicle. We chose these categories so that a wide range of object classes could be covered. For example, the category "device" includes objects like utensils, bottles, and televisions, while "nature" includes objects like trees, mountains, and flowers etc. Figure 3.8 shows the distribution of the memorability scores of all 7 object categories in our dataset.

Results in Figure 3.8 give a sense of how memorability changes across different object categories. Animal, person, and vehicle are all highly memorable classes, each associated with an average memorability score greater than or close to 0.5. Interestingly, all other categories have an average memorability lower than 0.25, indicating that humans do not remember objects from these categories very well. In particular, furniture is the least memorable category with an average score of only 0.14. This is possibly due to the fact that most objects in the furniture, nature, and building categories either appear mostly in the background or are occluded, which likely de-

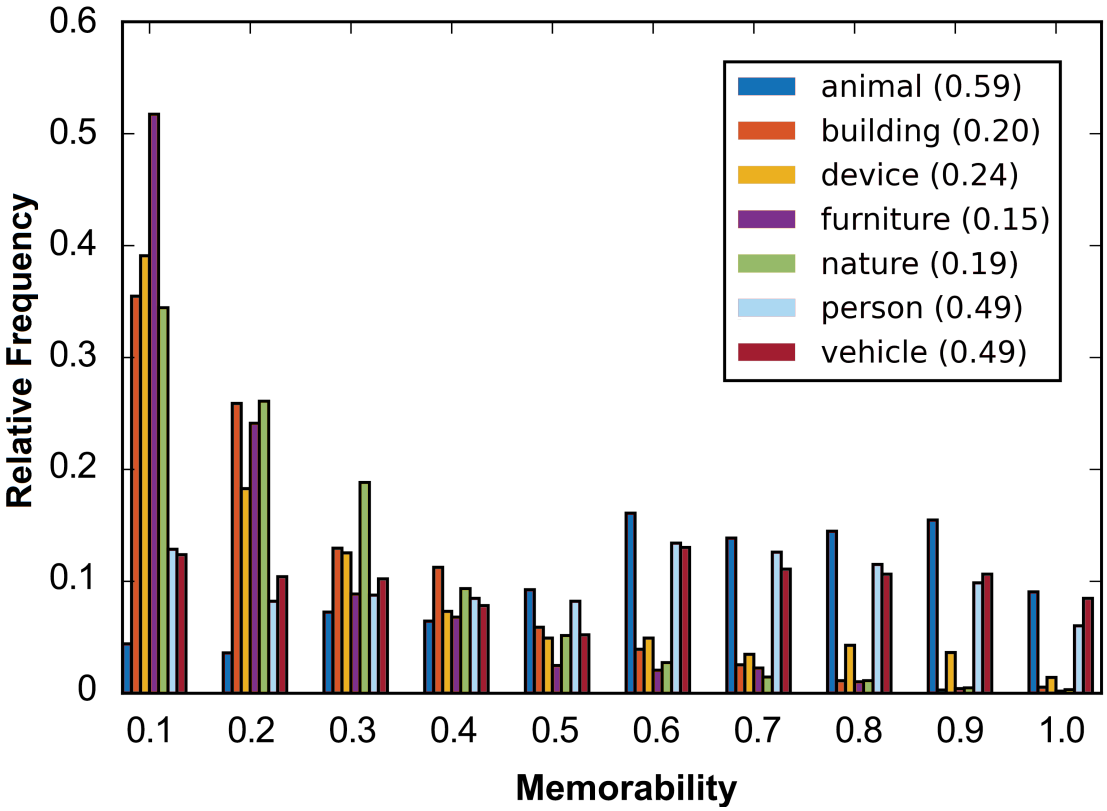

Figure 3.8: Some object categories are more memorable than others. Categories like furniture, nature, building, and device tend to have a large majority of objects with very low memorability scores. Objects belonging to animal, person, and vehicle categories are remembered more often.

creases their memorability significantly. In contrast, objects from the animal, person, and vehicle categories appear mostly in the foreground, leading to a higher memorability score on average. Interestingly, the most memorable objects from the building, furniture, and nature categories tend to have an average memorability in the range of $0.4 - 0.8$, whereas the score of the most memorable objects from person, animal and vehicle is higher than 0.9. While the differences in the memorability of different object categories could be driven by factors like occlusion, size, background/foreground, or photographic bias, the distribution in Figure 3.8 suggests that humans remember some object classes such as person, animal, and vehicle irrespective of external nuisance factors and these categories are intrinsically more memorable than others.

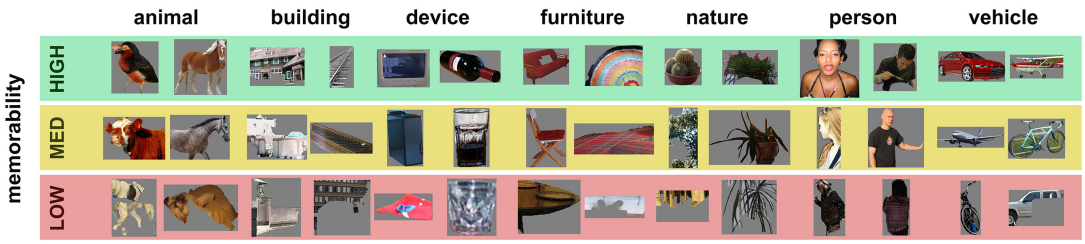

Figure 3.9: Memorability of object categories. Most memorable, medium memorable and least memorable objects from each of the 7 categories.

3.4.4.2 *Exploring category-specific memorability*

As demonstrated above, some object categories (i.e. animal, person, and vehicle) tend to be more memorable than others. However, not all objects in the same category are equally memorable. The examples in Figure 3.9 show the most memorable, medium memorable, and least memorable objects for each category. Across categories, medium to high memorability objects tend to have little to no occlusion, and low memorability objects tend to be both occluded and darker. What other category-related factors could influence the memorability of objects? Among the possible factors, we explore how category-specific object memorability is influenced by **(i)** the number of objects in an image and **(ii)** the presence of other object categories.

Number of objects: Figure 3.10 shows the change in average memorability for the different categories when the minimum number of objects within an image is increased. Results indicate that the number of objects present in an image is an important factor in determining memorability. For example, as the number of objects in an image increases, the memorability of animals and vehicles decreases sharply, most likely a result of competition for attention. Although the memorability of vehicles starts to show a slight increase for objects greater than 8, this arises only due to insufficient data (number of images is less than 30). Interestingly, the memorability of the person category does not change significantly when an increasing number of objects are present in the image. This suggests that people are not only one of the most memorable object categories, but that their memorability is the least sensitive to the presence of object clutter in an image.

Inter-category memorability: How much is the memorability of a particular object category affected when it co-occurs with another object category (or another instance of the same category)? To quantify the effect of one category on another, we consider each pairwise combination

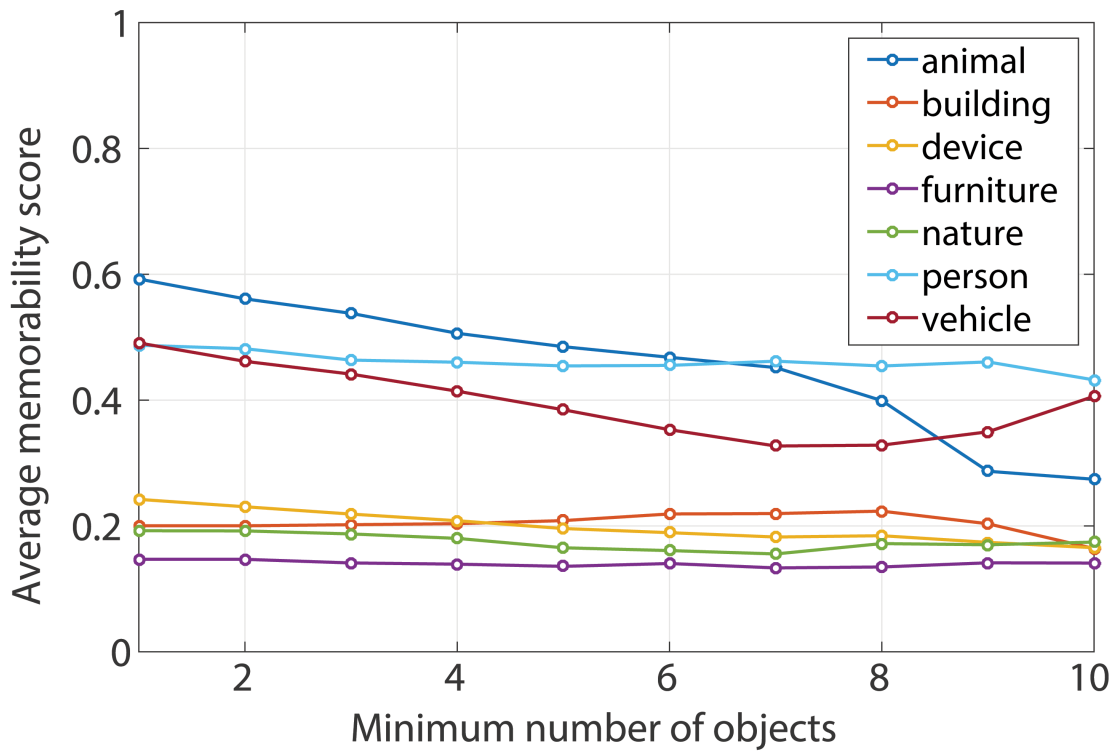

Figure 3.10: Object number affects category-specific memorability. For each category, a curve shows the change in average memorability as the number of additional objects in the images increase. The memorability of objects belonging to categories like animals and vehicles decreases significantly with an increase in object number.

of categories and gather all images that contain at least one object from both categories. By taking one category as the *reference* and the other as the *distractor*, we compute the average memorability score $m_{R|D}$ of the reference in the images common to the reference and distractor. To isolate the effect of the distractor, we compute the memorability difference

$$\Delta m = m_{R|D} - m_R, \tag{3.1}$$

where m_R is the memorability score of the reference in all images where it exists. Figure 3.11 shows Δm for all possible reference and distractor pairs. It is clear that Δm for low-memorability categories (i.e. nature, furniture, device, and building) is not significantly affected by the presence of other categories.

Also, the memorability of the animal category maintains its high score in the presence of other categories, except vehicles, people, and itself, where it decreases substantially. The memorability

Figure 3.11: Inter-category object memorability relationship. Effect of distractor categories on the memorability of reference categories.

of people tends to be unaffected by the presence of most other categories including itself. However, it decreases in the presence of vehicles and buildings. This could be due to the fact that people in images containing vehicles or buildings are usually zoomed out and smaller in size (refer to Figure 3.12). The memorability of the vehicle category is strongly affected by the presence of other object categories. In particular, it drops significantly in the presence of other vehicles, people, and animals.

In summary, when an animal, vehicle, or person co-occur in the same image, the memorability of all three categories usually decreases. However, this pattern of change in memorability is category-specific in general. For example, when a vehicle and animal are present in the same image, the animal is generally more memorable, even though both their memorability scores drop significantly. When a vehicle or an animal co-occurs with a person, the person is generally more memorable (also shown in Figure 3.12).

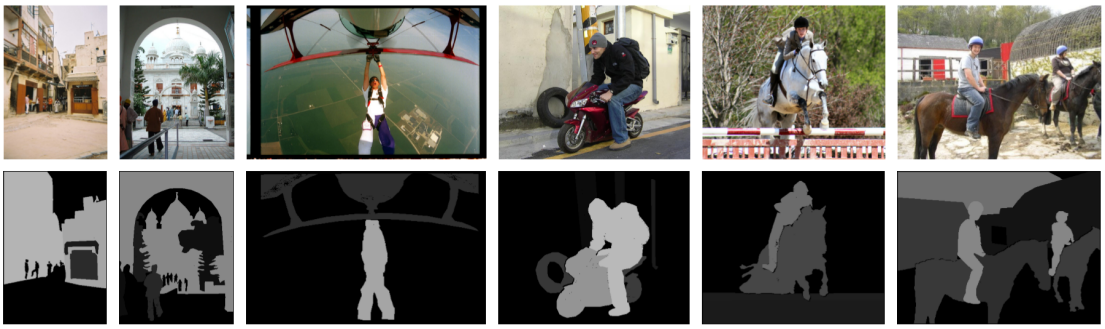

Figure 3.12: Memorability of people in the presence of other categories. Top row: Images where a person co-occurs with other categories. Bottom row: Ground truth object memorability maps. In the presence of buildings, the memorability of a person can drop. In the presence of a vehicle or animal, the person is usually more memorable.

3.4.5 How are object & image memorability related?

Until now, we have studied what objects people remember and the factors that influence their memorability, but to what extent does the memorability of individual objects affect the overall memorability of an image? Moreover, if an image is highly memorable, what can we say about the memorability of the objects inside those images (and vice versa)? To shed light on these questions, we conducted a second large-scale experiment on Amazon Mechanical Turk for all images in our dataset to gather their respective *image* memorability scores. For this experiment, we followed the same strategy as the memory game experiment proposed in Khosla, Raju, Torralba, and Oliva (2015). A series of images from our dataset and the Microsoft COCO dataset (T.-Y. Lin et al., 2014) (i.e. 'filler' images) were flashed for 1 second each, and participants were instructed to press a key whenever they detected a repeat presentation of an image. A total of 350 workers participated in this experiment with each image being viewed 80 times on average. The rank correlation after averaging over 25 random splits of the participants' memorability scores was 0.70, indicating high inter-rater reliability.

Using results from the previous experiments, we computed the correlation between the scores of the single most memorable *object* in each image and the memorability score of each *image*. This correlation is moderately high with $\rho = 0.40$, suggesting that the most memorable object in an image plays a crucial role in determining the overall memorability of an image. To investigate this finding in relation to some extreme cases, we repeated the same analysis as above but on a subset of the data containing the 100 most memorable images and the 100 least memorable images. The correlation between maximum object memorability and image memorability for this subset

of images increased significantly to $\rho = 0.62$. This means that maximum object memorability serves as a strong indicator of whether an image is *highly* memorable or *not* memorable at all. In other words, images that are highly memorable contain at least one highly memorable object and images with low memorability usually do not contain a single highly memorable object (refer to Figure 3.13).

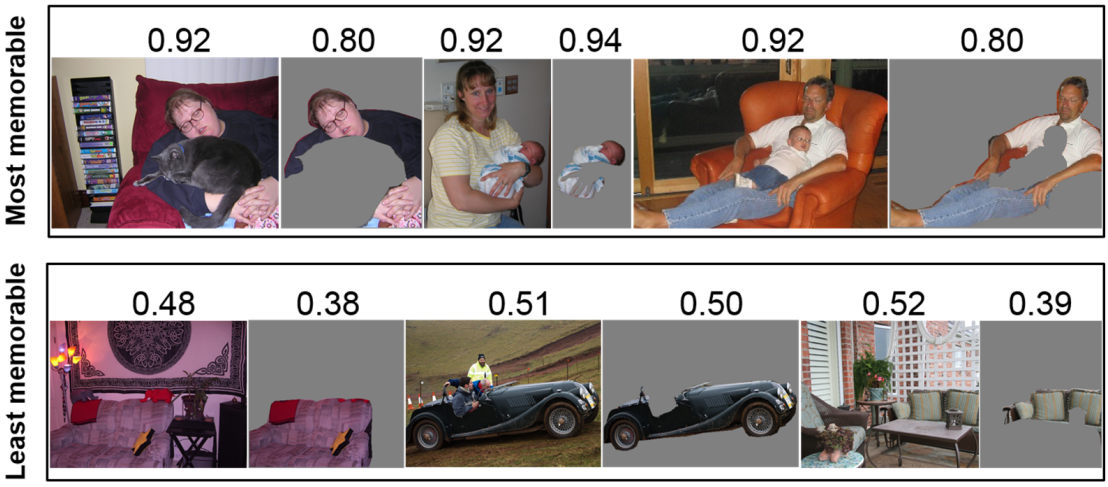

Figure 3.13: Max object memorability predicts image memorability. Top row: Most memorable images taken from our dataset along with their highest memorable object and their respective memorability scores. Bottom row: least memorable images in the dataset along with their most memorable object and their respective memorability scores. Maximum object memorability correlates strongly with image memorability in both cases.

To study the relationship between maximum object memorability and image memorability conditioned on the object category, we computed the correlation between maximum object and image memorability for each individual object category separately. The resulting correlations for each category are: animal ($\rho = 0.38$), building ($\rho = 0.22$), device ($\rho = 0.47$), furniture ($\rho = 0.53$), nature ($\rho = 0.64$), person ($\rho = 0.54$), and vehicle ($\rho = 0.30$), indicating that certain categories are more susceptible to the effect than others. For example, images containing animals, buildings, or vehicles as the most memorable objects tend to have varying degree of image memorability (indicated by their lower ρ values). On the other hand, device, furniture, nature, and person are strongly correlated with image memorability, indicating that if an image's most memorable object belongs to one of these categories, the object memorability score is strongly predictive of the image memorability score. We can imagine scenarios in which this information is potentially useful. For example, in vision systems that are tasked to predict scene memorability, a *single* object and its category can serve as a strong prior in predicting this score.

3.5 Predicting Object Memorability

This work makes available the very first dataset containing ground truth memorability of constituent objects from a highly diverse image set. In this section, we show that our dataset can be used to benchmark computational models of object memorability.

3.5.1 Memorability Prediction with Deep Neural Networks

To predict object memorability using representations learned using a deep neural network, we first generated object segments by using multiscale combinatorial grouping (MCG), a generic object proposal method proposed in Arbelaez, Pont-Tuset, Barron, Marques, and Malik (2014). We then extract feature representations for each object proposal using the CaffeNet implementation of AlexNet (Jia et al., 2014; Krizhevsky et al., 2012), a deep convolutional network classifier that was pretrained to classify 1000 object classes of the ImageNet dataset (Russakovsky et al., 2015). While this network was not explicitly trained to predict saliency or memorability, object categories play an important role in determining object memorability (Section 3.4.4).

Next, a support vector regressor (SVR) was trained using 6-fold cross-validation to map our extracted deep object features to memorability scores. This model was used to predict memorability scores for the top $K = 20$ object segments obtained using the MCG algorithm, as well as the original object segmentations. After predicting these memorability scores, memorability maps were generated by averaging the scores of these top K segments at the pixel level (DL-MCG).

3.5.2 Baseline Models

Feature-based Models: Since image features like Scale-invariant Feature Transform (SIFT; Lowe, 2004) and Histogram of Oriented Gradients (HOG; Dalal & Triggs, 2005) have previously been shown to achieve good performance in predicting image memorability (Isola et al., 2014; Isola, Xiao, et al., 2011), a single baseline model was built using both of these features for comparison. Training and testing of this model was performed similarly to our deep network model.

Saliency Models: Given the uncovered correlation between what is remembered and what participants fixate on, we employed eight state-of-the-art saliency methods (top performing methods according to benchmarks in Borji, Sihite, and Itti (2012, 2013)): GB (Harel, Koch, & Perona, 2006), AIM (Bruce & Tsotsos, 2006), DV (Hou & Zhang, 2009), IT (Itti, Koch, & Niebur,

1998), GC (Cheng, Zhang, Mitra, Huang, & Hu, 2011), PC (Margolin, Tal, & Zelnik-Manor, 2013), SF (Perazzi, Krahenbuhl, Pritch, & Hornung, 2012), and FT (Achanta, Hemami, Estrada, & Susstrunk, 2009). Each algorithm produces a pixel-level map of predicted saliency.

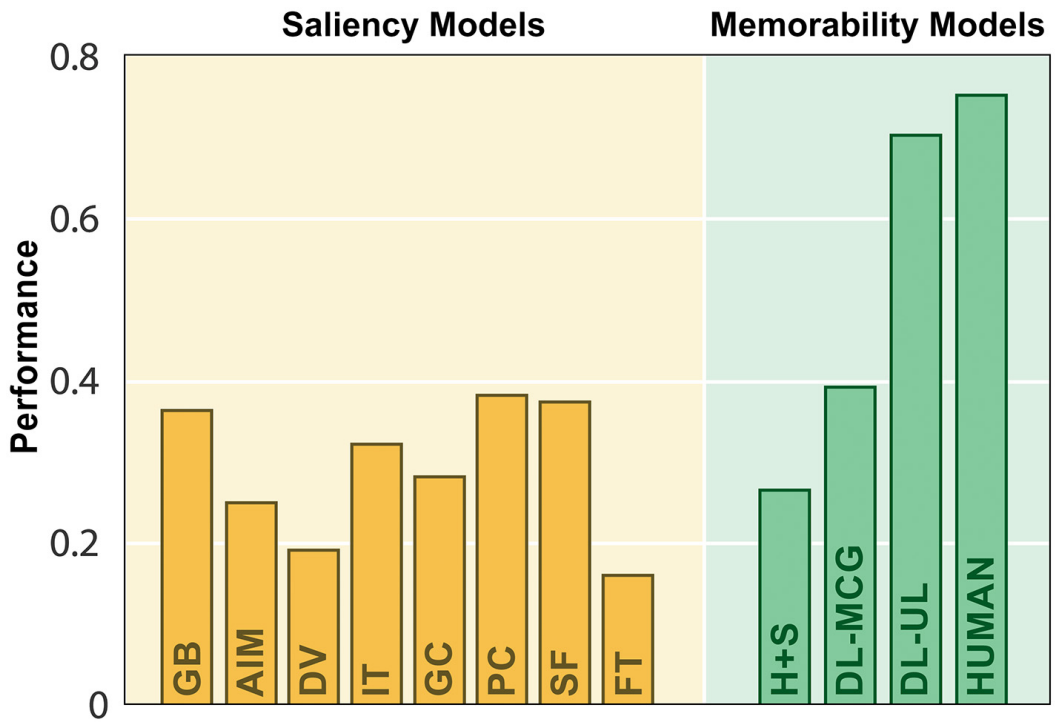

Figure 3.14: Rank correlation of predicted object memorability. Accuracy of the saliency and feature-based models on proposed benchmark.

3.5.3 Results

To evaluate the accuracy of the predicted object memorability maps, we computed the rank correlation between the mean predicted memorability score inside each of the object segments and their ground truth memorability scores. These results are reported in Figure 3.14. Clearly, the deep network model, DL-MCG, performs considerably well ($\rho = 0.39$). In contrast, the baseline trained using HOG and SIFT (H+S) exhibits much weaker performance ($\rho = 0.27$).

Figure 3.14 shows that the H+S baseline is also outperformed by most saliency prediction methods. Thus, even though models using SIFT and HOG have previously demonstrated high predictive power for image memorability, they may not be as well suited for the task of predicting object memorability. The deep network model (DL-MCG) performs better than all other

saliency methods with only PC ($\rho = 0.38$), SF ($\rho = 0.37$), and GB ($\rho = 0.36$) showing comparable performance. A common factor between these saliency methods is that they explicitly add center bias to their implementation. Although object memorability exhibits less center bias when compared to eye fixations, it still tends to be biased somewhat towards the center due to photographer bias (see Section 3.4.2), which could be a reason for the high performance of these saliency methods.

While DL-MCG performed favorably in predicting object memorability, its accuracy is highly dependent on the quality of the segmentations used. To illustrate this fact, we redo the prediction task but with the ground truth segments replacing the MCG segments. The resulting baseline is referred to as DL-UL, which can be considered the gold standard or the upper bound on automated object memorability prediction. Its correlation score is very high and close to human inter-rater relability ($\rho = 0.70$), which suggests that the deep network model does have high predictive ability, but that it is sensitive to the image segmentations it is applied to. This is not that surprising given that MCG does not always propose object-like segmentation bounds, and sometimes groups multiple objects and background elements together by mistake. It should be noted that this problem might be solved with better segmentation methods that also make use of deep neural networks (Long et al., 2015).

3.6 Conclusion

In this chapter, we focused on the problem of understanding the memorability of objects in natural images. To this end, we obtained ground truth data that helps to study and analyze this problem in depth. We show that the category of an object is a good index of its memorability, and that visual saliency can predict object memorability to some degree. Moreover, we studied the relationship between image and object memorability and compiled a benchmark dataset for object memorability prediction.

In the end, deep networks were most successful in making predictions about exactly which objects will be remembered in an image. Human image annotations can tell us when a person is present in an image for example, and this information is surprisingly predictive of memorability, but it does not not include the sorts of rich representational detail that deep networks appear to have captured. However, in the end, what can we claim beyond predictive superiority? When deep features fail to predict human behavior for certain images, can we know why? If we can, can we explicitly optimize them to avoid this? In the next chapter, we will provide an initial set of answers to these questions.

Science is an essentially anarchic enterprise: theoretical anarchism is more humanitarian and more likely to encourage progress than its law-and-order alternatives... The only principle that does not inhibit progress is: anything goes.

Paul Feyerabend, *Against Method*

4

Predicting human similarity judgments for natural images

Deep neural networks continue to excel in new tasks that have been historically very difficult for computers to solve, including problems in vision (LeCun et al., 2015), natural language processing (Collobert et al., 2011) and reinforcement learning (Mnih et al., 2015). In the last chapter, we found similar success in applying deep networks to predict human behavior directly, as opposed to accidentally (e.g., needing humans to label the structure in the world given no other feasible alternatives). Indeed, our model did not learn what should be remembered in the world given certain model constraints, but what people *do* remember in the world.

We further accomplished this without explicitly re-training the network for our own prediction problem—we used the feature representation that was learned to support object recognition. This opens up a number of interesting questions about the generality and usefulness of the learned representation. For example, can we think of these representations as a generally good compression of the images, not unlike what humans might employ to solve several types of problems? Further, to the extent that our model was imperfect, what representational differences were to blame? If instead we found a perfectly predictive model for this particular task, does that

Much of the content of this chapter was published in Peterson, Abbott, and Griffiths (2018).

mean the network represents those stimuli much like people do? In this chapter, we explore how well the representations discovered by deep convolutional network classifiers align with human psychological representations of natural images, show how they can be adjusted to increase this alignment, and demonstrate that the resulting representations can be used to predict complex human behaviors such as learning novel categories.

4.1 Comparing Representations

4.1.1 Deep Neural Networks and the Brain

Following the success of DNNs in computer vision, recent work has begun to compare the properties of these networks to psychological and neural data. Much of the initial work in comparing deep neural network representations to those of humans comes from neuroscience. For example, early work found that neural network representations beat out 36 other popular models from neuroscience and computer vision in predicting IT cortex representations (Khaligh-Razavi & Kriegeskorte, 2014), and later work found a similar primacy of these representations in predicting voxel-wise activity across the visual hierarchy (Agrawal, Stansbury, Malik, & Gallant, 2014). However, neural representations are not necessarily the gold standard for capturing all of the complex structure of human mental representations. Human similarity judgments for a set of objects encode representational detail that cannot be estimated by inferotemporal cortex representations, which are more similar to monkey inferotemporal cortex than to human psychological representations (Mur et al., 2013). For this reason, estimating human behavior directly may also be fruitful, and possibly more informative.

4.1.2 Deep Neural Networks and Human Behavior

Several recent studies have seen some initial success in applying representations from deep neural networks to psychological tasks, including predicting human typicality ratings (Lake, Zaremba, et al., 2015), as well as the work presented on memorability in Chapter 3, for natural object images. More recently, it was shown that human shape sensitivity for natural images could be explained well for the first time using deep neural networks (Kubilius et al., 2016), which now constitute a near essential baseline for emerging models of human shape perception (Erdogan & Jacobs, 2017). A follow-up to our own previous work (Peterson, Abbott, & Griffiths, 2016) showed that important categorical information is missing from deep representations (Jozwik, Kriegeskorte, Storrs, & Mur, 2017).

4.2 Overview of the Chapter

As we saw in Chapter 2, human psychological representations cannot be observed directly, and so comparing them to representations formed by deep neural networks is challenging. Our approach is to solve this problem by exploiting the close relationship between *representation* and *similarity* discussed in Chapters 1 and 2 (i.e., every similarity function over a set of pairs of data points corresponds to an implicit representation of those points). This provides an empirical basis for the first detailed evaluation of DNNs as an approximation of human psychological representations. To do this, we can subject both DNN and human similarities to an ensemble of classic psychological methods for probing the spatial and taxonomic information they encode. This identifies aspects of human psychological representations that are captured by DNNs, but also significant ways in which they seem to differ. We can then consider whether a better model of human representations can be efficiently bootstrapped by transforming the deep representations. The resulting method opens the door to ecological validation of decades of psychological theory using large datasets of highly complex, natural stimuli, which is demonstrated by predicting the difficulty with which people learn natural image categories.

4.3 Experiment 1: Evaluating the Correspondence Between Representations

Human psychological representations are not directly observable, and cannot yet be inferred from neural activity (Mur et al., 2013). However, psychologists have developed methods for inferring representations from behavior alone. Human similarity judgments capture stimulus generalization behavior (Shepard, 1987) and have been shown to encode the complex spatial, hierarchical (Shepard, 1980), and overlapping (Shepard & Arabie, 1979) structure of psychological representations, around which numerous psychological models of categorization and inference are built (Goldstone, 1994b; Kruschke, 1992; Nosofsky, 1987). If we can capture similarity judgments, we will have obtained a considerably high resolution picture of human psychological representations. Experiment 1 evaluated the performance of deep neural networks in predicting human similarity judgments for six large sets of natural images drawn from a variety of visual domains: animals, automobiles, fruits, furniture, vegetables, and a set intended to cross-cut visual categories (referred to below as "various").

4.3.1 Methods

4.3.1.1 Stimuli

Stimuli were hand-collected for each of the six domains, digital photos that were meant to exhibit wide variety in object pose, camera viewpoint, formality, and subordinate class. Each domain contained 120 total images, each cropped to a square aspect ratio and resized to 300×300 pixel dimensions. An example subset of these images for each dataset is provided in Figure 4.1, and the full sets are provided in Appendix A.

4.3.1.2 Procedure

For all six stimulus categories, pairwise image similarity ratings (within each category) were collected from human participants on Amazon Mechanical Turk. Participants were paid $\$0.02$ to rate the similarity of four pairs of images within one of the six categories on a scale from 0 ("not similar at all") to 10 ("very similar"). They could repeat the task as many times as they wanted, but were not allowed to repeat ratings of the same unique image pair. The experiment obtained exactly 10 unique ratings for each pair of images ($7{,}140$ total) in each category, yielding $71,400$ ratings per category ($428{,}400$ total ratings), from over $1{,}200$ unique participants. The result is six 120×120 similarity matrices after averaging over individual judgments, for which each entry represents human psychological similarity between a pair of objects. The raw similarity matrices are included in Appendix A.

4.3.1.3 Deep neural network representations

To obtain image representations from our deep neural networks, each input image is fed through each network. The nodes that comprise the network obtain different activation values for each image after each layer performs a transformation. We can take these activation values as a vector of "features" representing the image (for the entire network, or for a particular layer). These feature vectors can be collected into a feature matrix $\mathbf{F}$, which specifies a multidimensional feature representation (columns) for each image (rows). A similarity matrix $\hat{\mathbf{S}}$, in which the entry $\hat{s_{ij}}$ gives the similarity between images i and j in the network's representation space, can then be computed by the matrix product

$$\hat{\mathbf{S}} = \mathbf{F}\mathbf{F}^T, \tag{4.1}$$

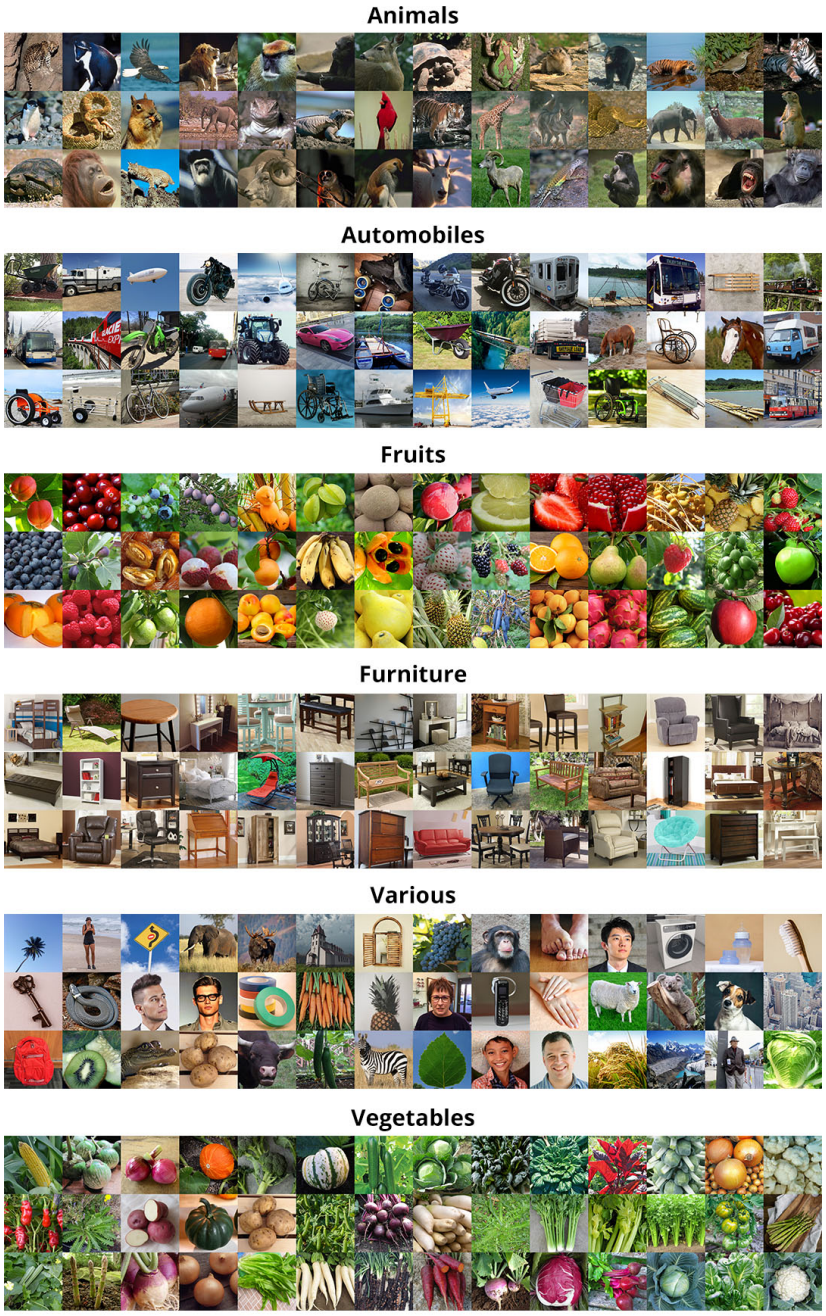

Figure 4.1: Example image stimuli from our six domains.

modeling $\hat{s}_{ij}$ as the inner product of the vectors representing images i and j. Given human similarity judgments $\mathbf{S}$ and an artificial feature representation $\mathbf{F}$, we can evaluate the correspondence between the two by computing the correlation between the entries in $\mathbf{S}$ and $\hat{\mathbf{S}}$.

For each image in all six categories, deep feature representations were extracted using four highly popular convolutional neural network image classifiers that were pretrained in Caffe (Jia et al., 2014) on ILSVRC12, a large dataset of 1.2 million images taken from 1000 objects categories in the ImageNet database (Deng et al., 2009). This dataset serves as a central benchmark in the computer vision community. Our own image datasets were not explicitly sampled from categories in ILSVRC12 and likely diverge to some degree. For example, of the 1000 ILSVRC12 classes, 120 are different dog breeds, whereas our animal set contains no dogs. The networks, in order of depth, are AlexNet (Krizhevsky et al., 2012, 7 layers), VGG (Simonyan & Zisserman, 2014, 19 layers), GoogLeNet (Szegedy et al., 2014, 22 layers), and ResNet (He, Zhang, Ren, & Sun, 2016, 152 layers), three of which are ILSVRC12 competition winners. VGG, GoogLeNet, and ResNet all achieve at least half the error rate of AlexNet.

Images are fed forward through each network as non-flattened tensors, and activations are recorded at each layer of the network. For most of our analyses besides the AlexNet layer analysis, activations at the final hidden layer only of each network are extracted. For AlexNet and VGG, this is a 4096-dimensional fully-connected layer, while the last layers in GoogleNet and ResNet are 1024- and 2048-dimensional pooling layers respectively. As an example, feature extraction for the animals training image set provides a 120×4096 matrix. All feature sets were then z-score normalized.

4.3.1.4 Unsupervised Baseline Representations

Another model of interest was a recent state-of-the-art unsupervised network (Donahue, Krähenbühl, & Darrell, 2016; Dumoulin et al., 2016), a generative model trained to capture the distribution of the entire ILSVRC12 dataset. This network (BiGAN) is a bidirectional variant of a Generative Adversarial Network (Goodfellow et al., 2014) that can both generate images from a uniform latent variable and perform inference to project real images into this latent space. This 200-dimensional latent encoding was used as the image representation for this network. As an additional baseline, two forms of shallow (non-deep) feature sets were also included, both being previously popular methods from computer vision: the Scale-invariant feature transform (SIFT; Lowe, 2004), using the bag-of-words technique trained on a large image database, and Histogram of Oriented Gradients (HOG; Dalal & Triggs, 2005), with a best-performing bin size of 2×2.

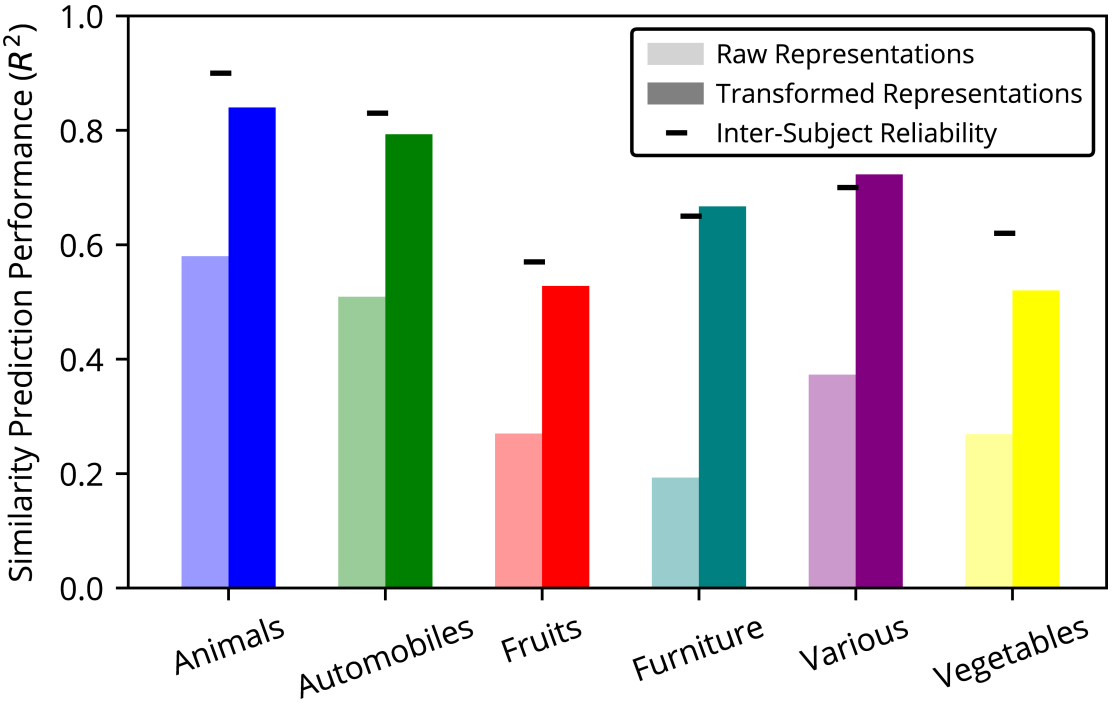

Figure 4.2: Model performance (proportion of variance accounted for, R^2) in predicting human similarity judgments for each image set using the best raw (light colors) and best transformed (dark colors) DNN representations.

Table 4.1: Variance explained in human similarity judgments for raw and transformed representations for the best performing network (VGG).

Dataset	Raw R^2	Transformed R^2	CV Control R^2	Human Inter-reliability
Animals	0.58	0.84	0.74	0.90
Automobiles	0.51	0.79	0.58	0.83
Fruits	0.27	0.53	0.36	0.57
Furniture	0.19	0.67	0.35	0.65
Various	0.37	0.72	0.54	0.70
Vegetables	0.27	0.52	0.35	0.62

4.3.2 Results and Discussion

The variance explained in human similarity judgments by the best performing DNN architecture (this was VGG in all cases) is plotted in Figure 4.2 (lighter colors) and given in Table 4.1 ("raw"),

and indicates that the raw deep representations provide a reasonable first approximation to human similarity judgments, although the level of precision depends on the domain. Animals were the best approximated of the six image sets, reaching up to nearly 60% variance explained. Alternative metrics such as Euclidean distance yielded essentially identical results (not shown).

4.3.2.1 Visualizating Representations

To better understand how DNNs succeed and fail to reproduce the structure of psychological representations, we can apply two classic psychological tools: non-metric multidimensional scaling, which converts similarities into a spatial representation, and hierarchical clustering, which produces a tree structure (dendrogram) (Shepard, 1980). For our NMDS analysis, the scikit-learn Python library was used to obtain only two-dimensional solutions, with a maximum iteration limit of 10,000 in fitting the models through gradient descent, and a convergence tolerance of 1e-100. Embeddings were first initialized with standard metric MDS, then taking the best fitting solution of four independent initializations. For HCA, the scipy Python library was used along with a centroid linkage function in all models.

The results for the best-performing DNN on the animals stimuli are shown in Figure 4.3, and point out the most crucial differences in these two representations. Human representations exhibit highly distinguished clusters in the spatial projections and intuitive taxonomic structure in the dendrograms, neither of which are present in the DNN representations. This gives us an idea of what relevant information is missing from the deep representations in order to fully approximate human representations.

4.3.2.2 Predictive Variability Across Network Architectures

Beyond identifying the DNN that best captures human similarity judgments, it is useful to understand how competing networks compare in their predictive ability. Figure 4.4 shows the results of comparing the representations from all four classification networks, as well as a recent high-performing unsupervised deep architecture (BiGAN; Donahue et al., 2016; Dumoulin et al., 2016) and two older, non-deep standards from computer vision: HOG (Lowe, 2004) and SIFT (Dalal & Triggs, 2005) features. Most classification networks perform similarly, yet VGG is slightly better on average. Surprisingly, representations from the BiGAN, while useful for machine object classification (Donahue et al., 2016), don't seem to correspond as well to human representations, and are even less effective than shallow methods like HOG+SIFT.

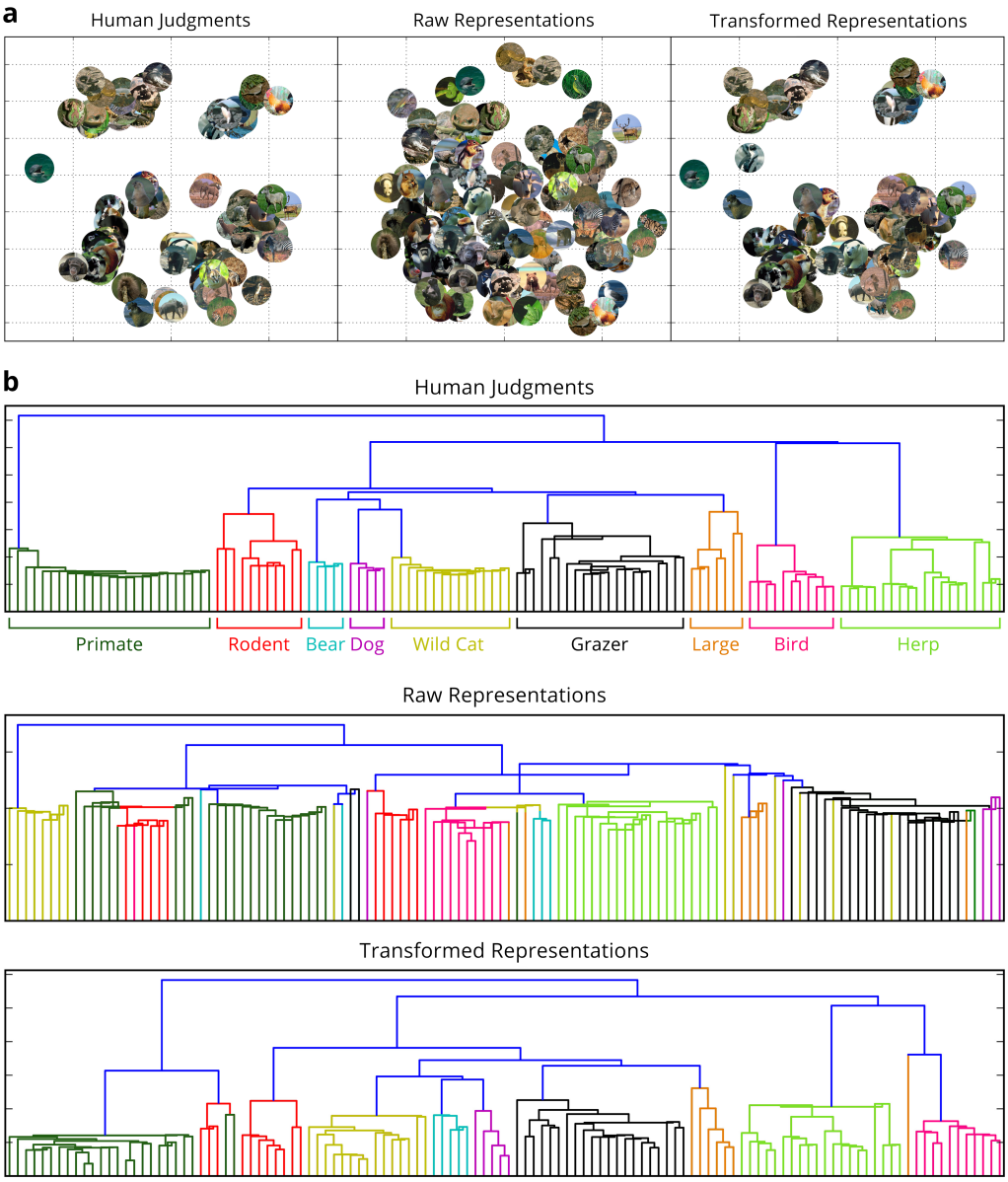

Figure 4.3: Representations of Animals. (a) Non-metric multidimensional scaling solutions for human similarity judgments (left), raw DNN representations (middle), and transformed DNN representations (right). (b), Dendrograms of hierarchical clusterings (centroid method) for human similarity judgments (top), raw DNN representations (middle), and the transformed DNN representations (bottom).

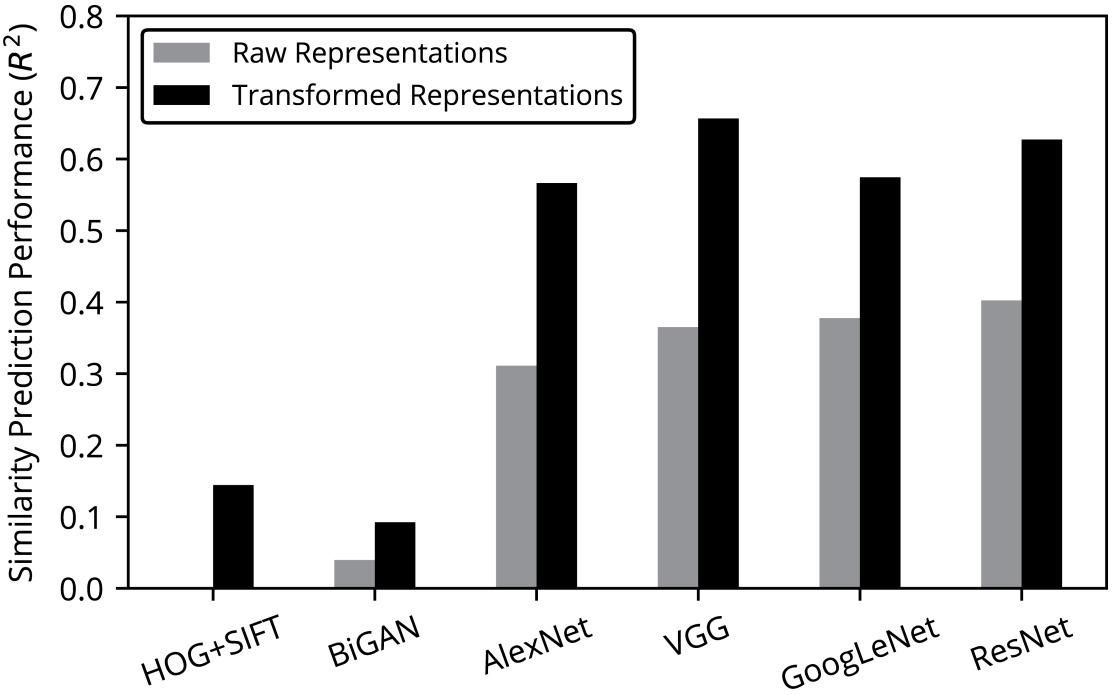

Figure 4.4: Similarity prediction performance using the best weighted representations from four popular deep classifiers, an unsupervised network (BiGAN), and a non-deep baseline (HOG+SIFT). Results are averaged across all six image sets.

4.3.2.3 Representational Abstraction Analysis

Using AlexNet, which has a manageable yet still large number of layers, performance at each layer of the network was examined, including final class probabilities from the softmax layer and discrete "one-hot" labels for the predicted most probable class. Since early layers represent lower-level features, and later layers represent increasingly abstract structure, we can ask which level of abstraction best fits our human judgments. As Figure 4.5 shows, performance climbs as the depth of the network increases, but falls off near the end when the final classification outputs near. For all datasets, the best layer was the final hidden layer, yielding a 4096-dimensional vector, as opposed to the classification layer which by design must shrink to merely 1000 dimensions. This indicates that relatively high-level, yet non-semantic information is most relevant to the human judgments obtained.

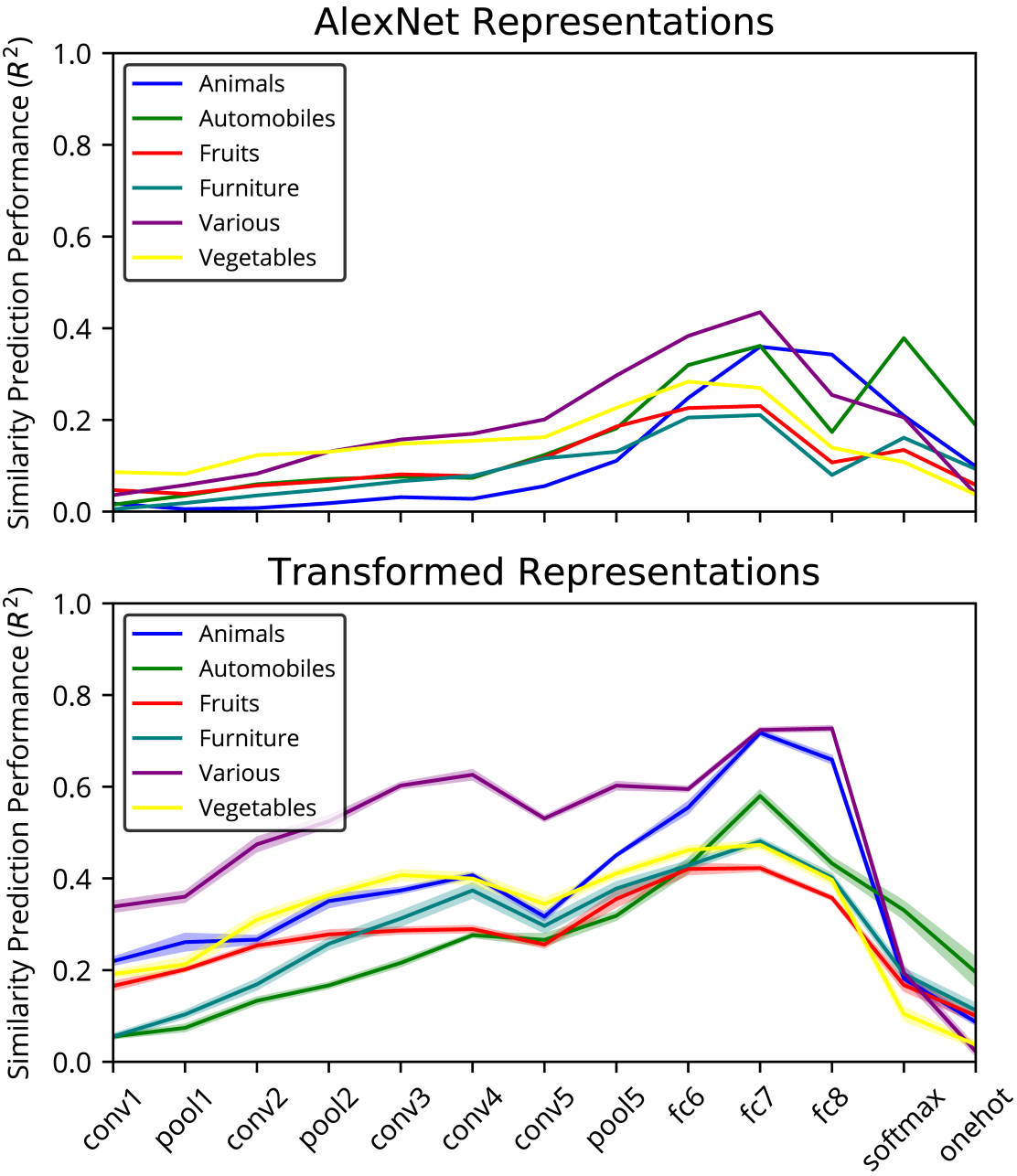

Figure 4.5: Similarity prediction performance using transformed representations at each layer of AlexNet for each dataset ("softmax" is predicted class probabilities, and "one-hot" is predicted class labels).

4.4 Transforming Deep Representations

Experiment 1 showed that the raw representations discovered by deep neural networks perform reasonably well as predictors of human similarity judgments. This correspondence suggests that deep neural networks could potentially provide an indispensable tool to psychologists aiming to test theories with naturalistic stimuli. Even a crude approximation of a complex representation may vastly outperform classic low-level features often used to characterize natural stimuli (e.g., Gabor wavelet responses). More importantly, having a representation that approximates human similarity judgments provides a starting point for identifying representations that are even more closely aligned with people's intuitions. This section explores how DNN representations can be transformed to increase the alignment with psychological representations.

4.4.1 Transforming representations

Formally, given the ground truth human similarity kernel $s(x_i, x_j)$ for stimuli x_i and x_j, and some starting set of deep features $\mathbf{F}$, our goal is to find an additional transformation ϕ, such that

$$s(x_i, x_j) = \phi(F_i) \cdot \phi(F_j), \tag{4.2}$$

where F_i is row i of feature matrix $\mathbf{F}$. The space of possible ϕ transformations to search is massive, and one would hope that most of the "work" has already been done by the deep network, such that ϕ is simple and easy to find. This would also be an indication that our deep network is already a good approximation. This assumption is built in to the model formulation below.

The model of similarity judgments given in Equation 4.1 can be augmented with a set of weights on the features used to compute similarity, with

$$\mathbf{S} = \mathbf{F}\mathbf{W}\mathbf{F}^T, \tag{4.3}$$

where $\mathbf{W}$ is a diagonal matrix of dimension weights. This formulation is similar to that employed by additive clustering models (Shepard & Arabie, 1979), wherein $\mathbf{F}$ represents a binary feature identity matrix, and is similar to Tversky's classic model of similarity (Navarro & Lee, 2004; Tversky, 1977). Concretely, it provides a way to specify the relationship between a feature representation and stimulus similarities. When used with continuous features, this approach is akin to factor analysis.

Given an existing feature-by-object matrix $\mathbf{F}$, we can show that the diagonal of $\mathbf{W}$, the vector

of weights $\mathbf{w}$, can be expressed as the solution to a linear regression problem where the predictors for each similarity s_{ij} are the (elementwise) product of the values of each feature for objects i and j (i.e. each row $\mathbf{X_i}$ of the regression design matrix $\mathbf{X}$ can be written as $\mathbf{F_i} \circ \mathbf{F_j}$, where $\circ$ is the Hadamard product). The predicted similarity $\hat{s_{ij}}$ between objects i and j is therefore

$$\hat{s}(x_i, x_j) = \hat{s_{ij}} = \sum_k w_k f_{ik} f_{jk}, \tag{4.4}$$

where f_{ik} is the kth feature of image i and w_k is its weight. The squared error in reconstructing the human similarity judgments can be minimized by convex optimization. Gershman and Tenenbaum (2015) proposed a similar method using a full $\mathbf{W}$ matrix, which is a more expressive model, but requires fitting more parameters. The current model employs a diagonal $\mathbf{W}$ matrix to minimize the amount of data and regularization needed to find a good solution, and assumes that the needed transformation is as simple as possible.

The resulting alignment method is akin to metric learning methods in machine learning (Kulis, 2013). Estimating both the features and the weights that contribute to human similarity judgments, even for simple stimuli, is a historically challenging problem (Shepard & Arabie, 1979). Our main contribution is to propose that $\mathbf{F}$ be substituted by features from a deep neural network, and only $\mathbf{w}$ be learned. This both coheres with our comparison framework and greatly simplifies the problem of estimating human representations.

If $\mathbf{w}$ is also constrained to be nonnegative, then the square root of these weights can be interpreted as a multiplicative rescaling of the feature space:

$$\hat{s_{ij}} = \sum_k \sqrt{w_k} f_{ik} \cdot \sqrt{w_k} f_{jk}, \tag{4.5}$$

where each $\sqrt{w_k}$ is a scaling factor for feature k. This makes it possible to directly construct transformed spatial representations of stimuli, and also implies that ϕ is linear. Since a direct feature transformation is not necessary for the present evaluations, no such constraint is included in the results that follow. However, it should be noted that this variation allows for applications where it is essential that transformed features be exposed (i.e., when similarities will not suffice).

4.4.2 Learning the transformations

Freely identifying the $\mathbf{w}$ that best predicts human similarity judgments runs the risk of overfitting, since our DNNs generate thousands of features. To address this, all of our models use L2

regularization on $\mathbf{w}$, penalizing models for which the inner product $\mathbf{w}^T\mathbf{w}$ is large. Minimizing the squared error in the reconstruction of s_{ij} with L2 regularization on $\mathbf{w}$ results in a convex optimization problem that is equivalent to ridge regression (Friedman, Hastie, & Tibshirani, 2001), and our loss function becomes

$$\mathcal{L} = (\sum_k w_k f_{ik} f_{jk} - s_{ij})^2 + \lambda \sum_k (w_k)^2, \tag{4.6}$$

Given the size of the problem, $\mathbf{w}$ can be found by gradient descent on an objective function combining the squared error and $\mathbf{w}^T\mathbf{w}$, with the latter weighted by a regularization parameter λ. To accomplish this, we used the ridge regression implementation in the scikit-learn Python library with a stochastic average gradient solver in order to reduce high memory consumption during fitting. We use 6-fold cross-validation to find the best value for this regularization parameter, optimizing generalization performance on held-out data. We chose 6 folds as a rule of thumb, although the results did not appear to be largely dependent on the number of folds used. We report variance explained only for models predicting non-redundant similarity values (only the lower triangle of the similarity matrix, excluding the diagonal).

4.4.3 Improvements through feature adaptation

We applied the method for adapting the DNN representations outlined above to the human similarity judgments and network representations used in Experiment 1. The best λ values for each dataset were comparable, in the range of $2000 - 9000$. After learning the best cross-validated weights $\mathbf{w}$ that map these features to human similarity judgments, the new representation that emerges explained nearly twice the variance for all datasets after cross-validating predictions (Figs. 4.2 and 4.4, darker colors). We also provide the raw scores for the best performing model (VGG) in Table 4.1, along with the results of a control cross-validation ("CV Control") scheme in which no single image occurred in both the training fold sets and test folds (as opposed to exclusivity with respect only to *pairs* of images). The MDS and dendrogram plots for the transformed representations in Figure 4.3 show a strong resemblance to the original human judgments. Notably, taxonomic structure and spatial clustering is almost entirely reconstructed, effectively bridging the gap between human and deep representations with only linear corrections.

4.4.4 Additional baseline models

As an additional check for overfitting, we constructed baseline models for each set of deep representations for each image dataset in which either (1) the rows, (2) the columns (separately for each row), or (3) both row and columns of the regression design matrix $\mathbf{X}$ were randomly permuted. The order of the target similarities from $\mathbf{S}$ remained unchanged. When all three models were subject to the same cross-validation procedure as the unshuffled models, variance explained (R^2) never reached or exceeded 0.01. This confirms that our regularization procedure was successful in controlling overfitting.

4.4.5 Inter-domain transfer

The transformations learned are highly contingent on the domain, and do not generalize well to others (e.g., a transformation trained on fruits is not effective when tested on animals). Table 4.2 shows the performance of the best DNN representations for each domain when applied to each other domain. The correlations are relatively poor, and worse than those produced by the best untransformed representations.

This pattern of poor inter-domain transfer is to be expected, since the number of DNN features is large and each domain only covers a small subset of the space of images and thus only provides information about the value of a small subset of features. However, it is possible to use the same adaptation method to produce a more robust transformation of the DNN representations for the purposes of predicting human similarity judgments. To do so, we learned a transformation using all six domains at once. This can also be thought of as a test of the robustness of our method when provided with an incomplete similarity matrix, specifically one containing only within-domain comparisons, yet still using all domains to constrain the ultimate model solution. This also allows for larger sets of images to be leveraged simultaneously for better learning.

We found this method to be highly effective, doubling the variance explained in human similarity judgments by the DNN representations from 30% to 60% after the transformation. A leave-one-out procedure in which every combination of five domains predicted the sixth provided similar improvements, as shown in Table 4.3. This is a strong control given that no images (and no similar images) are shared between the training and test sets in this formulation.

Table 4.2: Inter-domain generalization of best performing DNN transformations

Training Set	Test Set	R^2
Animals	Fruits	0.11
Animals	Furniture	0.02
Animals	Vegetables	0.11
Animals	Automobiles	0.17
Animals	Various	0.12
Fruits	Animals	0.14
Fruits	Furniture	0.12
Fruits	Vegetables	0.14
Fruits	Automobiles	0.25
Fruits	Various	0.13
Furniture	Animals	0.20
Furniture	Fruits	0.07
Furniture	Vegetables	0.11
Furniture	Automobiles	0.10
Furniture	Various	0.06
Vegetables	Animals	0.30
Vegetables	Fruits	0.10
Vegetables	Furniture	0.11
Vegetables	Automobiles	0.21
Vegetables	Various	0.08
Automobiles	Animals	0.36
Automobiles	Fruits	0.11
Automobiles	Furniture	0.07
Automobiles	Vegetables	0.13
Automobiles	Various	0.12
Various	Animals	0.41
Various	Fruits	0.05
Various	Furniture	0.06
Various	Vegetables	0.11
Various	Automobiles	0.21

Table 4.3: Generalization performance leaving out a single domain and training on the remaining five.

Leave-out	R^2
Animals	0.53
Automobiles	0.57
Fruits	0.63
Furniture	0.62
Various	0.59
Vegetables	0.63

4.5 Experiment 2: Predicting the Difficulty of Learning Categories of Natural Images

A simple linear transformation was able to adapt DNN representations to predict human similarity judgments at a level that is close to the inter-rater reliability. The transformed representation also corrected for the qualitative differences between the raw DNN representation and psychological representations. These results indicate that the rich features formed by DNNs can be used to capture psychological representations of natural images, potentially making it possible to run a much wider range of psychological experiments with natural images as stimuli.

The value of these representations for broadening the scope of psychological research can only be assessed by establishing that they generalize to new stimuli, and are predictive of other aspects of human behavior. To further explore the generalizability and applicability of this approach, we applied the learned transformation to the DNN representations (from VGG) of six new 120-image sets drawn from the same domains and assessed the ease with which people could learn categories constructed from the raw and transformed similarities.

Since our transformation is applied to the representational similarity measure (weighted inner product) as opposed to the feature space itself, we constructed categories via k-means clustering based on the rows of either the raw or transformed similarities (representing images as vectors of similarities to other images), ensuring that each category consisted of a coherent group of images as assessed by the appropriate similarity measure. Consequently, we should expect the ease of learning these category constructions to reflect the extent to which people's sense of similarity has been captured by the underlying models. In addition, traditional image features such as HOG and SIFT should make category learning more difficult than using DNN features, given the mismatch between representations observed in our previous analyses.

4.5.1 Methods

4.5.1.1 *Stimuli*

Using the best performing network and layer for each image dataset, we applied the learned transformation to a second set of 120 new images in each category (see Appendix A). This produced six predicted similarity matrices for each set. Using the rows of these matrices as image representations, we calculated k-means clusterings where the number of clusters (k) was either 2, 3, or 4. We repeated this process using the untransformed representations, for which similarities were simply inner products. This resulted in the following between-subjects conditions for our experiment: space (transformed, raw) $\times$ k (2,3,4) $\times$ domain (e.g., animals). We also replicated these experiments using baseline HOG+SIFT representations, yielding a total of 72 between-subjects conditions. An example of the clusterings used in the animal experiments where $k = 3$ are shown in Figure 4.6.

4.5.1.2 *Procedure*

A total of $2,880$ participants (40 per condition) were recruited on Amazon Mechanical Turk, paid \$1.00, and were not allowed to participate in multiple conditions. Participants in each condition were shown a single random sequence of the images from the dataset corresponding to their assigned condition and were instructed to press a key to indicate the correct category (where the correct category was the pre-defined cluster). Subjects could take as much time as they wanted to make their decisions. If a participant guessed incorrectly, an "incorrect" message was shown for 1.5 seconds. If they guessed correctly, this message read "correct". Initially, participants performed poorly as they had little information to associate keys with clusters, but showed consistent progress after a few examples from each cluster.

4.5.2 Results and Discussion

Figure 4.7 shows the difference in the ease with which people learned 2-, 3-, and 4-category partitions derived from the raw and transformed similarities. Using DNN features, categorization performance is higher for categories derived from the transformed spaces, and a three-way ANOVA ($k \times$ image set $\times$ transformation, see Table 4.4) confirmed that this effect was statistically significant ($F_{1,1404} = 66.28, p < .0001$). Participants also performed worse in the HOG+SIFT baseline condition, confirmed by a large main effect of feature set in a model including both feature sets ($F_{1,2845} = 3833.35, p < .0001$, see Table 4.5). Notably, the effect of the transformation

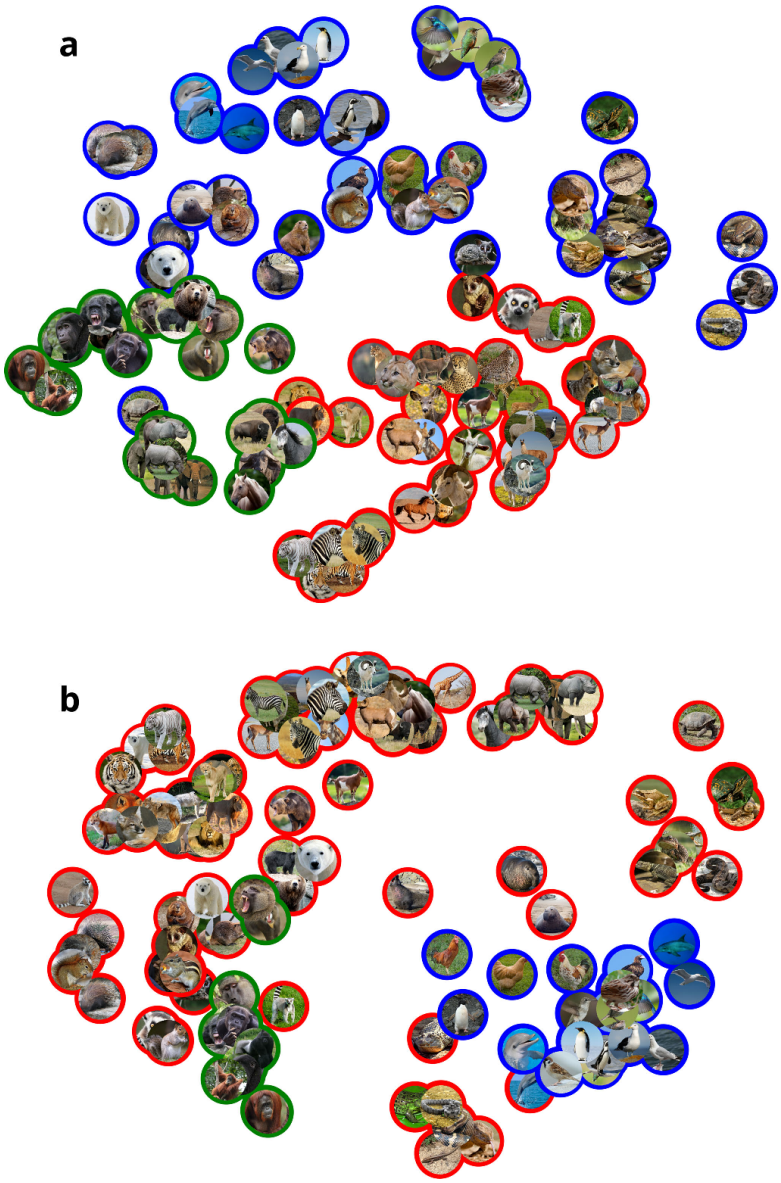

Figure 4.6: Examples of animal clusterings used in our categorization experiments where $k = 3$ for (a) the raw deep representations, and (b) the transformed deep representations. The transformation was learned on a different set of animal images, and appears to improve clustering in some aspects of the space. For example, the transformation makes primates more unique (i.e., not grouped with quadrupeds), and doesn't group small land and marine animals.

was reversed for the baseline features, confirmed by a significant interaction between feature set and transformation ($F_{5,2845} = 65.22, p < .0001$, see Table 4.6), indicating that HOG+SIFT feature tuning may not generalize, in sharp contrast with the DNN features. To assess learning effects, we grouped trials into five learning blocks. Average learning curves for the experiments using DNN features are shown in Figure 4.8. An ANOVA with learning block as a factor in Table 4.7 confirms a large main effect of block ($F_{4,5616} = 752.91, p < .0001$), and an interaction between block and transformation ($F_{4,5616} = 5.96, p < .0001$), likely due to the more rapid increase in performance in the first block for the transformed representation condition.

Table 4.4: ANOVA results for Experiment 2 using only DNN features.

	df	*F*	*p*
k	2	614.95	< 0.0001
image set	5	137.52	< 0.0001
transformation	1	66.28	< 0.0001
k × image set	10	7.14	< 0.0001
k × transformation	2	3.42	< 0.01
image set × transformation	5	29.20	< 0.0001
k × image set × transformation	10	3.17	< 0.001

Table 4.5: ANOVA results for Experiment 2 using feature set as a factor.

	df	*F*	*p*
k	2	2021.39	< 0.0001
image set	5	169.89	< 0.0001
transformation	1	139.96	< 0.0001
feature set	1	3833.35	< 0.0001
k × image set	10	14.96	< 0.0001
k × transformation	2	35.86	< 0.0001
k × feature set	2	13.38	< 0.0001
set × transformation	5	65.22	< 0.0001
image set × feature set	5	64.19	< 0.0001
transformation × feature set	1	645.71	< 0.0001

Table 4.6: ANOVA results for Experiment 2 using only baseline HOG+SIFT features.

	df	*F*	*p*
k	2	3005.96	< 0.0001
image set	5	108.98	< 0.0001
transformation	1	1767.70	< 0.0001
k × image set	10	25.67	< 0.0001
k × transformation	2	101.38	< 0.0001
image set × transformation	5	123.82	< 0.0001
k × image set × transformation	10	27.85	< 0.0001

Table 4.7: ANOVA results for Experiment 2 using only DNN features and learning block as a factor.

	df	*F*	*p*
k	2	605.49	< 0.0001
image set	5	137.10	< 0.0001
transformation	1	66.86	< 0.0001
block	4	752.91	< 0.0001
k × image set	10	7.23	< 0.0001
k × transformation	2	3.68	< 0.001
k × block	8	39.32	< 0.0001
image set × transformation	5	29.17	< 0.0001
image set × block	20	9.51	< 0.0001
transformation × block	4	5.96	< 0.0001

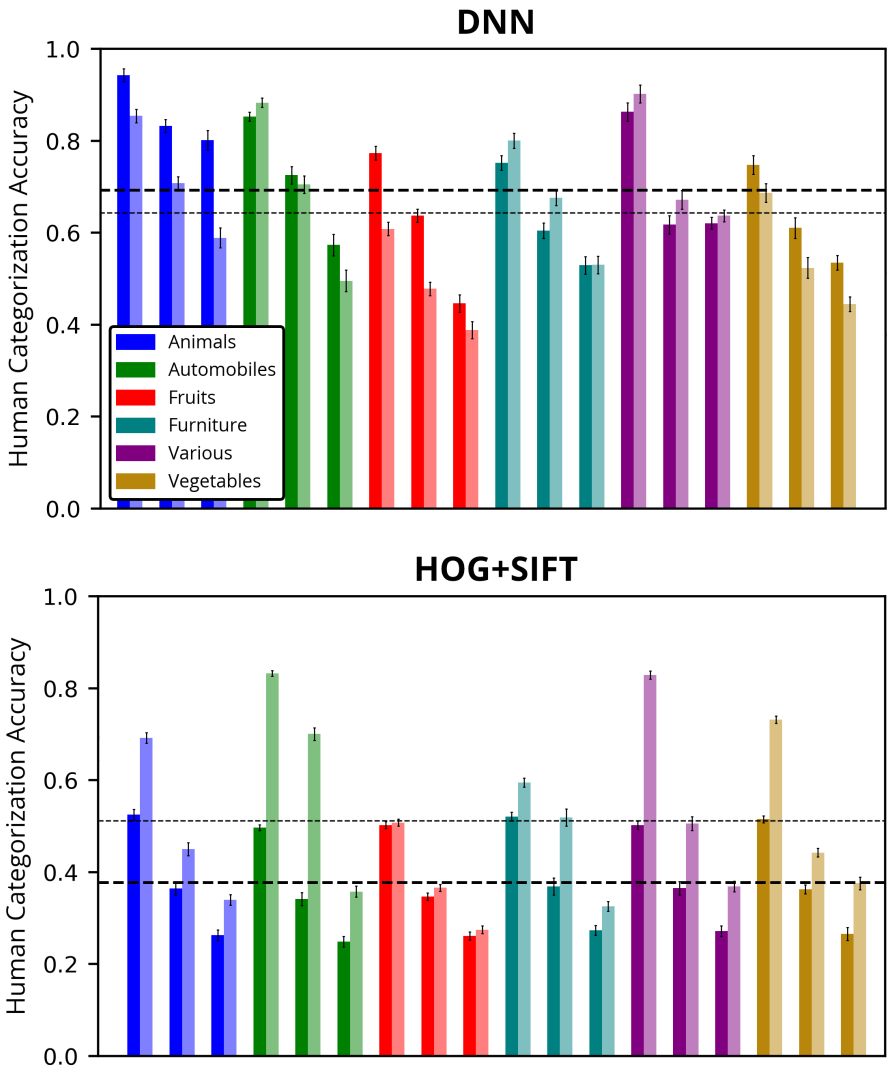

Figure 4.7: Average human categorization performance on each image set using raw and transformed DNN representations (top) and baseline HOG+SIFT features (bottom). Darker colors represent transformed versions of the raw representations (lighter colors). The three sets of bars for each image set represent 2-, 3-, and 4-category versions of the experiment. Thick dashed lines represent average accuracy for the raw representations, and thick dashed lines represent average accuracy for the transformed representations.

4.6 General Discussion

The framework presented here, inspired by classic psychological methods, is the first comprehensive comparison between modern deep neural networks and human psychological represen-

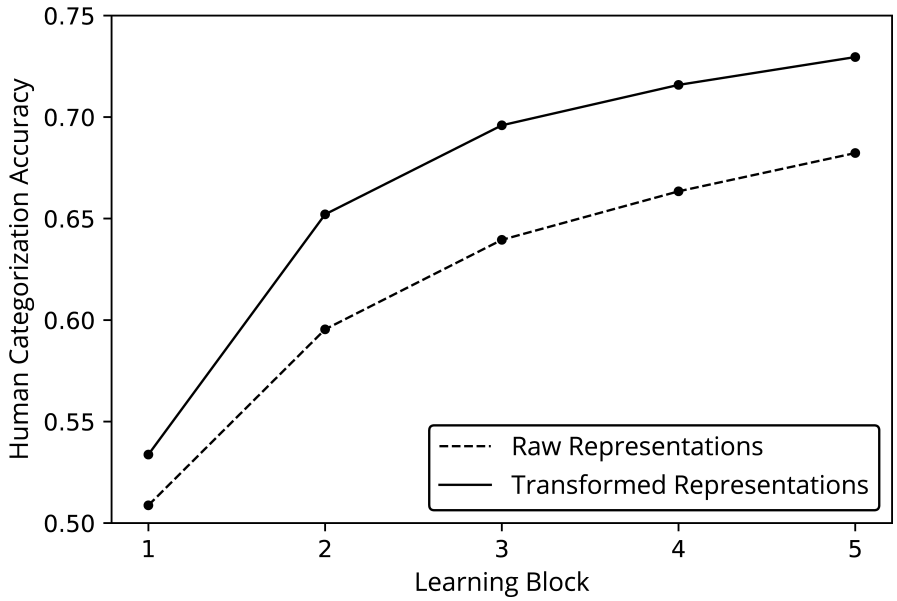

Figure 4.8: Average human categorization performance for each of five learning blocks.

tations. These artificial neural networks appear to make surprisingly good approximations to human similarities. Importantly, they also diverge in systematic ways (e.g., lacking taxonomic representational information) (Mur et al., 2013). However, the representations formed by these networks can easily be transformed to produce extremely good predictions of human similarity judgments for natural images. The resulting models transfer to new stimuli, and can be used to predict complex behaviors such as the ease of category learning. Since these representations and artificial networks are easy and cheap to manipulate, they present a valuable resource for rapidly probing and mimicking human-like representations and a potential path towards studying human cognition using more naturalistic stimuli.

Were these deep representations different enough from humans (i.e., requiring nonlinear transformations and therefore additional complex feature learning), adapting them to people would require either vastly more human judgments or significantly revised network architectures, the former being quite costly and the latter presenting a massive search problem. The method we propose to transform representations is extremely effective despite being constrained to a simple reweighting of the features. The linear transformation learned can be interpreted as an analogue of dimensional attention (Nosofsky, 1987), highlighting the possibility that the gap between these two sets of representations may be even smaller than we think. In fact, given that

our stimulus sets are mostly restricted to single domains (e.g., fruits), whereas the DNN classifiers make all output discriminations with respect to 1000 highly diverse object classes, one would expect that certain features should become more salient, while still others should be suppressed when making judgments in context (an important real-life situation not often incorporated in machine learning models). Finally, the ability of these adapted representations to predict human categorization behavior with novel stimuli demonstrates their applicability to studying downstream cognitive processes that rely on these representations, and may have applications in the optimal design of learning software.

4.7 Conclusion

The proliferation of machine learning methods for representing complex stimuli is likely to continue. We can think of the present approach as a way to leverage these advances and combine them with decades of research on psychological methods to shed light on questions about human cognition. This allows us to learn something about the potential weaknesses in artificial systems, and inspires new ideas for engineering those systems to more closely match human abilities (e.g., incorportating taxonomic information). Most significantly, it provides a way for psychologists to begin to answer questions about the exercise of intelligence in a complex world, abstracting over the representational challenges that can make it difficult to identify higher-level principles of cognition (Shepard, 1987) in natural settings.

Unanticipated novelty, the new discovery, can emerge only to the extent that his anticipations about nature and his instruments prove wrong.

Thomas S. Kuhn, *The Structure of Scientific Revolutions*

5

Estimating Categories in Deep Feature Spaces

Probabilistic models of cognition provide important, high-level abstractions for thinking about the problems that humans solve involving subjective probability, or inherent uncertainty (Chater, Tenenbaum, & Yuille, 2006). Inductive problems, as introduced in Chapter 1, are of this kind. Shepard for example reposed the problem of similarity as one of probabilistic generalization (i.e., what is the probability we will generalize a property of stimulus x to stimulus y?) to derive his method for inferring psychological representations. Like all modeling problems we have considered so far, inferring human subjective probability is difficult for naturalistic stimuli like images, since representing distributions over pixels directly is intractable.

In this chapter, we continue with our familiar example of categorization, and present a method for integrating deep generative neural networks with human-in-the-loop experimental designs for inferring subjective probability. In our case, the goal is to capture to human visual concepts or *categories*. Along the way, we will try to quantify our progress, but more interestingly, the immediate experimental result leaves us with a model that can dream up new images with a likeness closely mirroring human imagination, a feat not previously possible with classic methods alone, and an important qualitative test for truly human-like models.

Much of this chapter was published in Peterson, Suchow, Aghi, Alexander, and Griffiths (2018).

5.1 The Categorization Problem

Human visual category knowledge is inherently fuzzy (Reed, 1972). That is, we cannot observe every possible instance of a cat (such that any subsequent cat stimulus to identify is not new) without infinite time. Instead, given a new animal we might encounter, with both dog-like and cat-like features, we can only assign a probability that the animal is indeed a cat, based on the cats we have seen and conceptualized in the past.

More formally, given an encounter with stimulus x, we can ask about the probability that it belongs to the "cat" category, i.e., $p(\text{cat}|x)$. To solve this problem, we can make use of Bayes' theorem:

$$p(c|x) = \frac{p(x|c)p(c)}{p(x)} \tag{5.1}$$

The category label prior $p(c)$ is not particularly interesting or problematic to model, since it just describes for example how likely we are to encounter a cat as opposed to a dog etc. Similarly, the marginal $p(x)$ is just a normalization constant that can be computed if we know the value of the numerator for all categories (which we will indeed need to know). Of most interest is the likelihood $p(x|c)$. To see why this is the case, note that in order to make a categorization decision (i.e., when $p(\text{cat}|x)$ is higher than $p(\text{dog}|x)$), we can simply evaluate the ratio

$$\frac{p(\text{cat}|x)}{p(\text{dog}|x)}, \tag{5.2}$$

where a value of the ratio larger than 1 indicates that the stimulus is more likely a cat. A value of exactly 1 indicates that we can do no better than guess uniformly randomly. When $p(c)$ is uninformative, and we expand the above ratio using Bayes' rule, note that $p(x)$ is eliminated, leaving

$$\frac{p(x|\text{cat})}{p(x|\text{dog})} \tag{5.3}$$

Indeed, $p(x|\text{cat})$ is a crucial quantity because it describes the probability of all cat stimuli in the world, which undoubtedly overlaps with perhaps smaller, more exotic dog breeds. Not surprisingly, models of category learning can be formalized as methods for estimating the class-conditional density $p(x|c)$ given some training examples, or samples from the distribution (Ashby & Alfonso-Reese, 1995).

5.2 Representing Categories

Modeling how humans efficiently estimate categories from complex inputs or testing the density estimation hypothesis is is a formidable challenge of its own (a deep neural network classifier is an interesting model of this phenomenon). The issue is that such models cannot be evaluated, because understanding precisely what category knowledge humans have learned has never been fully accomplished for natural images. That is, we have no way of knowing if $p_D(x|\text{cat})$ as estimated by a deep CNN classifier is equivalent or not to $p_H(x|\text{cat})$ estimated by a human, even though the former is readily available. For this reason, we need a method for estimating $p_H(x|\text{cat})$ in a tractable way. While generic methods exist for such an experiment, they are limited to stimuli that can be easily represented, parameterized, and generated.

In what follows, a method is proposed that uses a human in the loop to estimate arbitrary distributions over complex feature spaces, adapting an existing experimental paradigm to exploit advances in deep architectures to capture the precise structure of human category representations, and iteratively sharpen them. Such knowledge is crucial to forming an ecological theory of intelligent categorization behavior and to providing a ground-truth benchmark to guide future work in cognitive modeling and machine learning.

5.3 Estimating the Structure of Human Categories

Methods for estimating human category templates have existed for some time. In psychophysics, the most popular and well-understood method is known as *classification images* (CI; Ahumada, 1996).

5.3.1 Classification Images

In the classification images experimental procedure, a human participant is presented with an image from one of a set of categories (e.g., A and B), each with white noise overlaid, and asked to identify the true category. On most trials, the participant will obviously select the correct category. However, if the added white noise significantly perturbs features of the image that are important to making the distinction, they may fail. Exploiting this, we can estimate the decision boundary from a number of these trials using the simple formula:

$$(n_{AA} + n_{BA}) - (n_{AB} + n_{BB}), \tag{5.4}$$

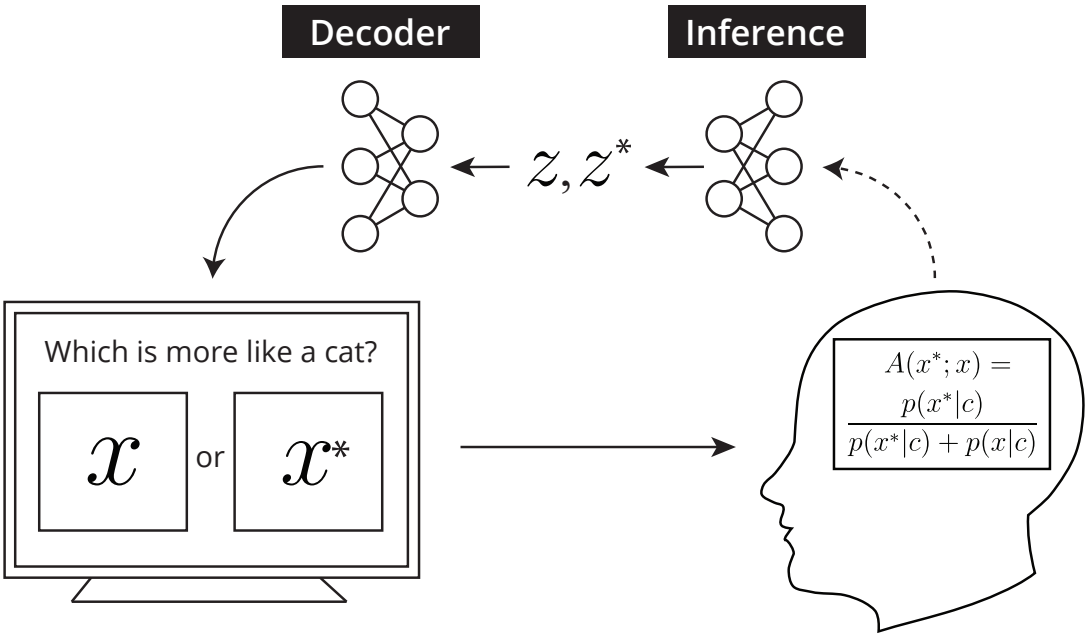

Figure 5.1: Deep MCMCP. A current state z and proposal z^* (top middle) are fed to a pretrained deep image generator/decoder network (top left). The corresponding decoded images x and x^* for the two states are presented to human raters on a computer screen (leftmost arrow and bottom left). Human raters then view the images in an experiment (bottom middle arrow) and act as part of an MCMC sampling loop, choosing between the two states/images in accordance with the Barker acceptance function (bottom right). The chosen image can then be sent to the inference network (rightmost arrow) and decoded in order to select the state for the next trial, however this step is unnecessary when we know exactly which states corresponds to which images.

where n_{XY} is the average of the noise across trials where the correct class is X and the observer chooses Y. Because the boundary is a difference of two class means, it can be regarded as a nearest-mean classifier. Interpreted probabilistically, the method assumes that the likelihood $p(c|x)$ for each category is Gaussian distributed with equal, spherical variance. Even if these assumptions are reasonable for a particular domain, the estimate may be biased depending by the experimenter's choice of base stimuli with which to overlay noise.

Vondrick, Pirsiavash, Oliva, and Torralba (2015) used a variation on classification images using deep image representations that could be inverted back to images using an external algorithm. In order to avoid dataset bias introduced by perturbing real class exemplars, white noise in the feature space was used to generate stimuli. In this special case, category templates become

$$n_A - n_B \,. \tag{5.5}$$

On each trial of the experiment, participants were asked to select which of two images (inverted from feature noise) most resembled a particular category. Because the feature vectors for all trials were random, thousands of stimuli could be rendered in advance of the experiment using relatively slow methods that require access to large datasets. This early inversion method was applied to mean feature vectors for thousands of positive choices in the experiments and yielded qualitatively decipherable category template images, as well as better objective classification decision boundaries that were guided human bias. However, this variant of CI requires an even more massive number of trials to be successful.

5.3.2 Estimating Arbitrary Category Structures by Sampling from People

To explicitly make use of rich probabilistic information as opposed to class boundaries, we need to turn to procedures for *sampling*. The following section reviews one such procedure, and how it has been integrated into experiments with human participants.

5.3.2.1 Markov Chain Monte Carlo

Only a handful of simple types of probability distributions have direct methods for sampling (e.g., Gaussians, Gaussian mixtures, etc). To sample from arbitrary distributions, one must often turn to one of a family of Monte Carlo methods, in our case, a popular method called Markov chain Monte Carlo (MCMC). This method does not require that the distribution in question be normalized (i.e., it can be multiplied by an unknown constant), as is often the case in its

application.

A *Markov chain* is a sequence of random variables that take on the Markov assumption, namely that the value of each variable in the chain depends only on the previous value from the previous variable (i.e., ignoring the values from the rest of the chain at all times):

$$p(x_t|x_{t-1}, x_{t-2}, ..., x_{t-n}) = p(x_t|x_{t-1}) \tag{5.6}$$

By repeatedly sampling from the above, we generate a sequence of *states* of the chain. The transition probabilities of moving between particular states are what differentiate a chain, and fully define what is referred to as the *stationary* or *target* distribution of the chain.

5.3.2.2 *The Metropolis Method*

A popular procedure for the construction Markov chains is the Metropolis method (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953), in which the transition probabilities are jointly defined by a *proposal distribution* $q(x^*|x)$, which proposes subsequent states x^* of the chain conditioned on the current state of the chain x, and an *acceptance function* $A(x^*; x)$, which gives the probability of accepting a proposal. In the simplest case, the proposal distribution $q(x^*|x)$ can be any symmetric distribution, meaning that

$$q(x^*|x) = q(x|x^*) \tag{5.7}$$

for all possible states x and x^*, and the acceptance function is given by

$$A(x^*; x) = \min\left(\frac{p(x^*)}{p(x)}, 1\right), \tag{5.8}$$

requiring that all proposals with a higher probability than the current state are automatically accepted, and all others with probability

$$\frac{p(x^*)}{p(x)}. \tag{5.9}$$

Since we only need know this ratio, any multiplicative constant, such as a normalization constant, will cancel out (including it would make no difference, and so it is therefore not required). For non-symmetric q distributions, this method is further generalized to the *Metropolis–Hastings* method (Hastings, 1970).

5.3.2.3 *Sampling with (from) People*

Markov Chain Monte Carlo with People (MCMCP; Sanborn & Griffiths, 2007), an alternative to classification images, is an experimental procedure in which humans act as a valid acceptance function in the Metropolis-Hastings algorithm, exploiting the fact that Luce's choice axiom (Luce, 1963), a well-known model of human choice behavior, is equivalent to a valid acceptance function, the Barker acceptance function

$$\frac{p(x^*)}{p(x^*) + p(x)}. \tag{5.10}$$

Sanborn, Griffiths, and Shiffrin (2010) also provide a proof that a rational Bayesian learner should behave in this way, and provide extensive empirical validation.

On the first trial of an MCMCP experiment, a stimulus x is drawn arbitrarily from the parameter space and compared to a new proposed stimulus x^* that is nearby in that parameter space. The participant makes a forced choice as to which is the better exemplar of some category (e.g., dog), acting as the acceptance function $A(x^*; x)$. If the initial stimulus is chosen, the Markov chain remains in that state. If the proposed stimulus is chosen, the chain moves to the proposed state. The process then repeats until the chain converges to the target category distribution $p(x|c)$. In practice, convergence is assessed heuristically, or limited by the number of human trials that can be practically obtained.

MCMCP has been successfully employed to capture mental categories from a number of domains, such as parameterized stick figure animals (Sanborn & Griffiths, 2007), and emotional faces (Martin, Griffiths, & Sanborn, 2012), and though these spaces are higher-dimensional than those in previous laboratory experiments, they are still relatively small and artificial compared to real images. Unlike classification images, this method makes no assumptions about the structure of the category distributions and thus can estimate means, variances, and higher order moments. Therefore, we take it as a starting point for the current method.

5.4 MCMCP in Deep Feature Spaces

Typical MCMCP experiments, as well both variants of the classification images procedure discussed earlier, are effective so long as noise can be added to dimensions in the stimulus parameter space to create meaningful changes in content. In the case of natural images, noise in the space of all pixel intensities is very unlikely to modify the stimulus in meaningful ways. Sanborn et

al. (2010) avoid this problem by representing stimuli that have a well-known structure (and can be easily rendered in pixels for participants to view), and Martin et al. (2012) use an image domain (faces) where image content is easy to align (reducing the dimensionality considerably), but neither of these strategies is immediately applicable to more complex stimuli.

As a first approach to solving this problem, we propose to perturb images in the latent space of a deep generative adversarial network (Goodfellow et al., 2014). This latent encoding z, is much lower-dimensional than pixel space, and captures only essential variation in the image content. Further, the generator network G can systematically translate these image parameterizations in real time into images viewable by the participant.

The mapping from features to images learned by a GAN is deterministic, and therefore MCMCP in low-dimensional feature space approximates the same process in high-dimensional image space. The resulting human judgments (accepted samples), with respect to images parameterized by z, can be used to approximate arbitrary category distributions. Given the true class-conditional likelihoods for a set of categories, the Bayes' optimal category boundaries are implicitly defined where the likelihood ratio is 1, and unlike CI, can be both linear or nonlinear. Moreover, samples from each density over z along with the generator network G define an approximation of the human generative model for each concept (from which we can sample new examples).

Since trials in an MCMCP experiment are not independent, we employ real-time, web-accessible generative adversarial networks to render high quality inversions from their latent features during online experiments. A schematic of the overall procedure is illustrated in Figure 5.1, and the algorithm is given below.

Algorithm 1: MCMCP using a a deep generator network G

```
Initialize starting state z ← N(0, I);
while trial < num_trials do
    z* ← z* + N(0, I) * σ;
    x ← G(z);
    x* ← G(z*);
    if x* a better example? then
        z ← z*;
    end
    Store z in list of trials
end
```

There are several theoretical and practical advantages to our method over previous efforts. First, MCMCP can capture arbitrary distributions, so it is not as sensitive to the structure of the underlying low-dimensional feature space and should provide better category boundaries than classification images when required. This is important when using various deep features spaces that were learned with different training objectives and architectures. MCMC inherently spends less time in low probability regions and should in theory waste fewer trials. Having generated the images online and as a function of the participantss decisions, there is no dataset or sampling bias, and auto-correlation can be addressed by removing temporally adjacent samples from the chain. Finally, using a deep generator provides drastically clearer samples than shallow reconstruction methods, and can be trained end-to-end with an inference network that allows us to categorize new images using the learned distribution.

Importantly, the proposed method of capturing categories in most ways avoids the identifiability problem discussed in Chapter 1. That is, while the distribution of features in the latent space of the GAN may be a distortion or warping of psychological space, MCMC ensures that the probability mass will be distributed across that space in such a way so as to yield identical values of $p(x|c)$ given a corresponding (x_i, z_i) pair. This could for example allow us to eventually test for Bayes' optimal categorization behavior in humans, simply by having a complex image representation that is relevant to humans, even though the correspondence may not be linear (although note we could apply the mapping strategy from Chapter 4 if it is linear).

5.5 Experiments

For our experiments, we explored two image generator networks trained on various datasets. Since even relatively low-dimensional deep image embeddings are large compared to controlled laboratory stimulus parameter spaces, we use a hybrid proposal distribution in which a Gaussian with a low variance is used with probability P and a Gaussian with a high variance is used with probability $1 - P$. This allows participants to both refine and escape nearby modes, but is simple enough to avoid excessive experimental piloting that more advanced proposal methods often require.

Participants in all experiments completed exactly 64 trials (image comparisons), collectively taking about 5 minutes, containing segments of several chains for multiple categories. The order of the categories and chains within those categories were always interleaved. Each participant's set of chains for each category were initialized with the previous participant's final states, resulting in large, multi-participant chains. All experiments were conducted on Amazon Mechanical Turk.

If a single image did not load for a single trial, the data for the subject undergoing that trial was completely discarded, and a new subject was recruited to continue on from the original chain state.

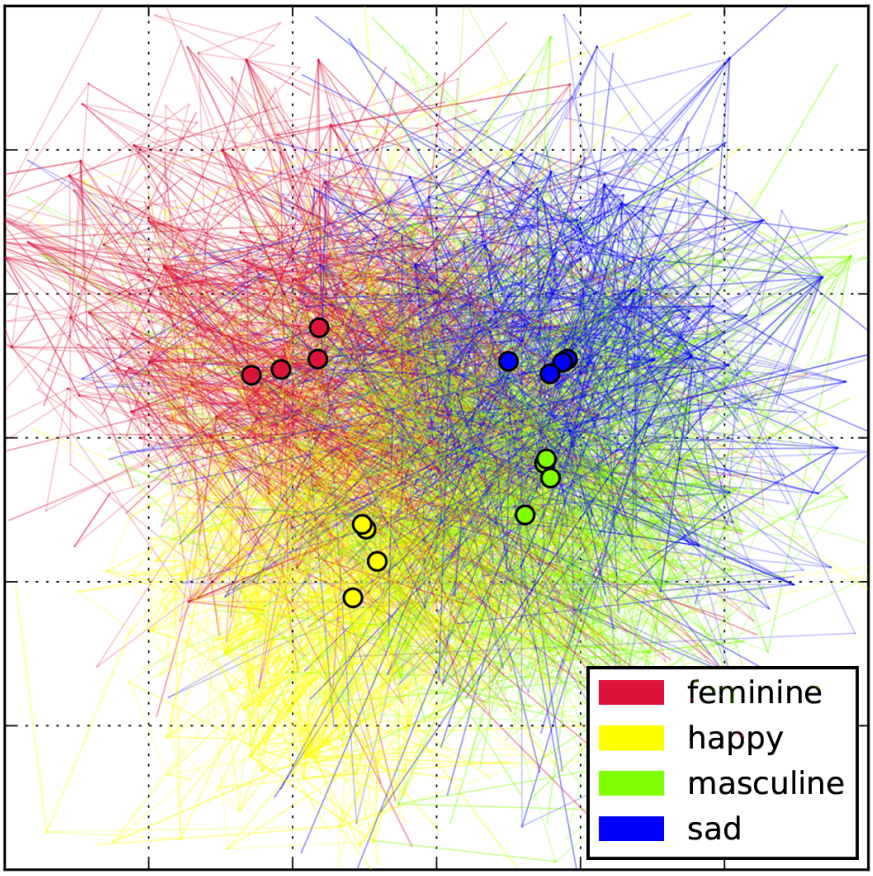

Figure 5.2: Visualizing MCMCP chains for faces. Fisher Linear Discriminant projections of all four MCMCP chains for each of the four face categories are shown. The four sets of chains overlap to some degree, but are also well-separated overall. Means of individual chains are closer to other means from the same class than to those of other classes.

5.5.1 Experiment 1: Initial test with face categories

5.5.1.1 Methods

We first test our method using DCGAN (Radford et al., 2015) trained on a large dataset of asian faces. We chose this dataset because it requires a deep architecture to produce reasonable samples (unlike MNIST, for example), yet it is constrained enough to test-drive our method using a relatively simple latent space. Four chains for each of four categories (male, female, happy, and sad) were used. Proposals were generated from an isometric Gaussian with a standard deviation of 0.25 50% of the time, and 2 otherwise. In addition, an analogous experiment was run using the classification images method. The final dataset contained 50 participants and over $3{,}200$ trials (samples) in total for all chains. The baseline classification images (CI) dataset contained the same number of trials and participants.

5.5.1.2 Results

An example of trial-level choices for the "person " category from a single subject is given in Figure 5.6. Full MCMCP chains are visualized using Fisher Linear Discriminant Analysis in Figure 5.2, along with the resulting averages for each chain and each category. Chain means within a category show interesting variation, yet converge to similar regions in the latent space as expected. Figure 5.3 also shows visualizations of the mean faces for both methods in the final two columns. MCMCP means appear to have converged quickly, whereas CI means only moderately resemble their corresponding category (e.g., the MCMCP mean for "happy" is fully smiling, while the CI mean barely reveals teeth). All four CI means appear closer to a mean face, which is what one would expect from averages of noise. We validated this improvement with a human experiment in which 30 participants made forced choices between CI and MCMCP means. The results are reported in Figure 5.4. MCMCP means are consistently highly preferred as representations of each category as compared to CI. This remained true even when an additional 50 participants (total of 100) completed the CI task, obtaining twice as many trials as MCMCP.

5.5.2 Experiment 2: Larger networks & larger spaces

The results of Experiment 1 show that reasonable category templates can be obtained using our method, yet the complexity of the stimulus space used does not rival that of large object classification networks. In Experiment 2, we tackled a more challenging (and interesting) form of the problem. To do this, we employed a bidirectional generative adversarial network (BiGAN;

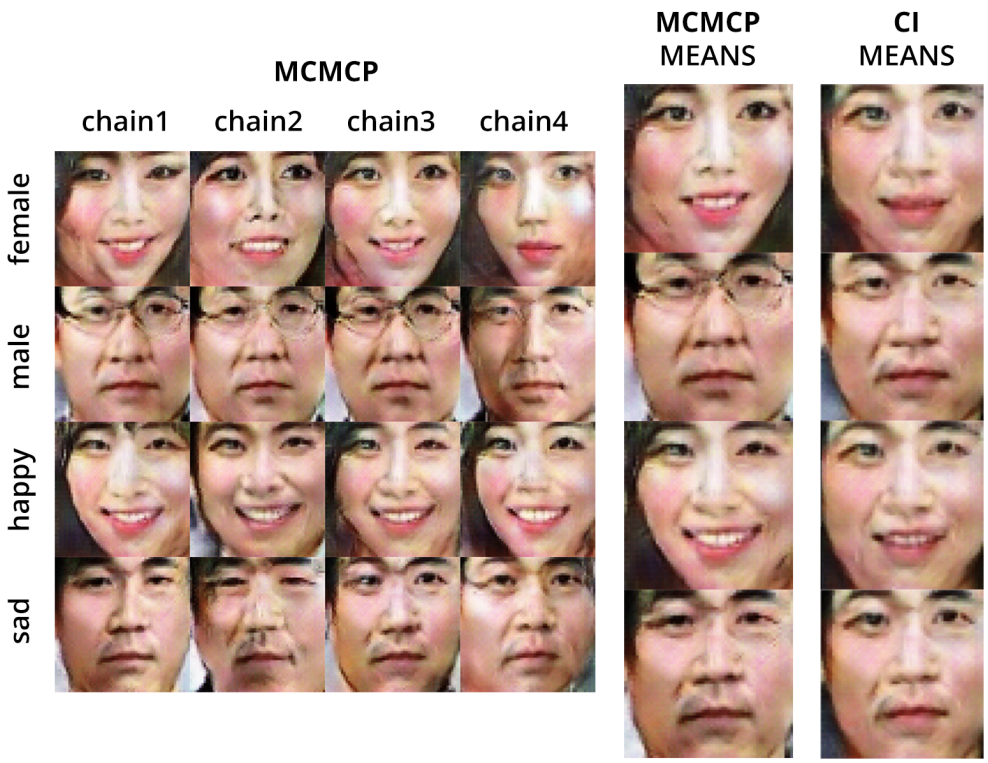

Figure 5.3: Visualizing captured representations. Individual MCMCP chain means (4×4 grid) and overall category means (second to last) are visualized as images using the generator network from our GAN (overall CI means are also shown for comparison in the final column). MCMCP means are much more differentiated than CI means, and better resemble the category in question

Donahue et al., 2016) trained on the 1.2 million-image ILSVRC dataset (64×64 center-cropped). BiGAN includes an inference network, which regularizes the rest of the model and produces unconditional samples competitive with the state-of-the-art. This also allows for the later possibility of comparing human distributions with other networks as well as assessing machine classification performance with new images based on the granular human biases captured. To give a sense of the expressive capability of this network, random samples from the network we used are shown in Figure 5.5.

5.5.2.1 *Methods*

Our generator network was trained given uniform rather than Gaussian noise, which allows us to avoid proposing highly improbable stimuli to participants. Additionally, we avoid proposing states outside of this hypercube by forcing z to wrap around (proposals that travel outside of z are injected back in from the opposite direction by the amount originally exceeded). In particular, we run our MCMC chains through an unbounded state space by redefining each bounded dimension z_k as

$$z'_k = \begin{cases} -sgn(z_k) \times [1 - (z_k - \lfloor z_k \rfloor)], & \text{if } |z| > 1 \\ z_k, & \text{otherwise.} \end{cases} \tag{5.11}$$

Proposals were generated from an isometric Gaussian with a standard deviation of 0.1 60% of the time, and 0.7 otherwise.

We use this network to obtain large chains for two groups of five categories. Group 1 included *bottle*, *car*, *fire hydrant*, *person*, and *television*, following Vondrick et al. (2015). Group 2 included

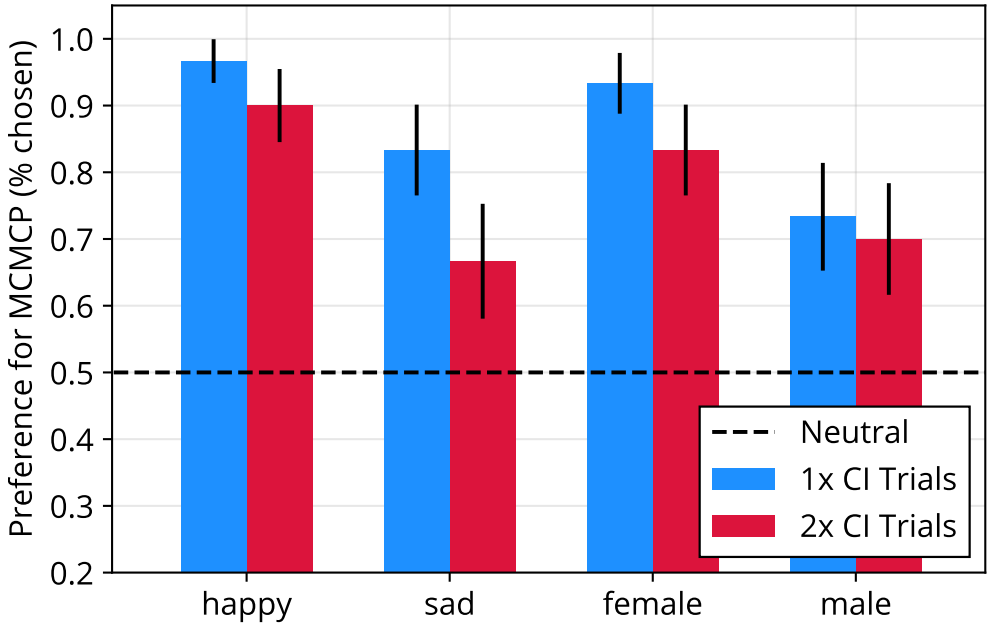

Figure 5.4: Human two-alternative forced-choice tasks reveal a strong preference for MCMCP means as representations of a category, when twice as many trials are used for CI.

Figure 5.5: Random samples from BiGAN trained on 1000 ImageNet classes.

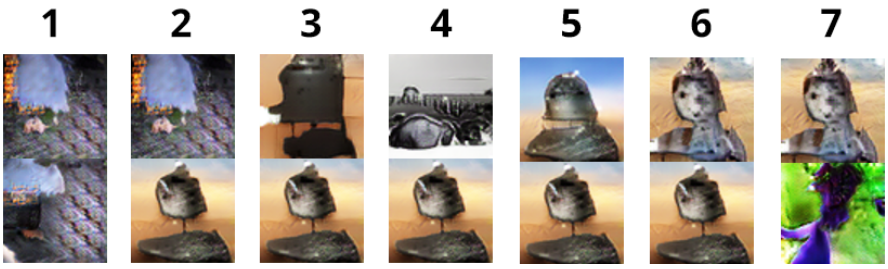

Figure 5.6: Examples from seven comparisons in the first few hundred trials of a "person" chain. Finding a reasonable first result took subjects over 200 trials, which may help to indicate burn-in. In each set of images (proposal and current, randomized order), it can be plainly observed which image is chosen and reused for the next trial. Trials 3-5 make no changes, while trial 6 refines the human bust shape with facial features.

bird, *body of water*, *fish*, *flower*, and *landscape*. Each chain was approximately $1{,}040$ states long, and four of these chains were used for each category (approximately $4{,}160$). In total, across both groups of categories, we obtained exactly $41{,}600$ samples from 650 participants.

To demonstrate the efficiency and flexibility of our method compared to alternatives, we obtained an equivalent number of trials for all categories using the variant of classification images introduced in Vondrick et al. (2015), with the exception that we used our BiGAN generator instead of the offline inversion previously used. This also serves as an important baseline against which to quantitatively evaluate our method because it estimates the simplest possible template.

5.5.2.2 *Results*

The acceptance rate was approximately 50% for both category groups. The samples for all ten categories are shown in Figure 5.7B and D using Fisher Linear Discriminant Analysis. Similar to the face chains, the four chains for each category converge to similar regions in space, largely away from other categories. In contrast, classification images shows little separation with so few trials (5.7C and D). Previous work suggests that at least an order of magnitude higher number of comparisons may be needed for satisfactory estimation of category means. Our method estimates well-separated category means in a manageable number of trials, allowing for the method to scale greatly. This makes sense given that unbiased CI must find a signal in arbitrary noise images, potentially wasting many trials. Beyond yielding a decision rule, our method additionally produces a density estimate of the entire category distribution. In classification images, only mean template images can be viewed, while we are able to visualize several modes in the category distribution. Figure 5.8 visualizes these modes using the means of each component in a

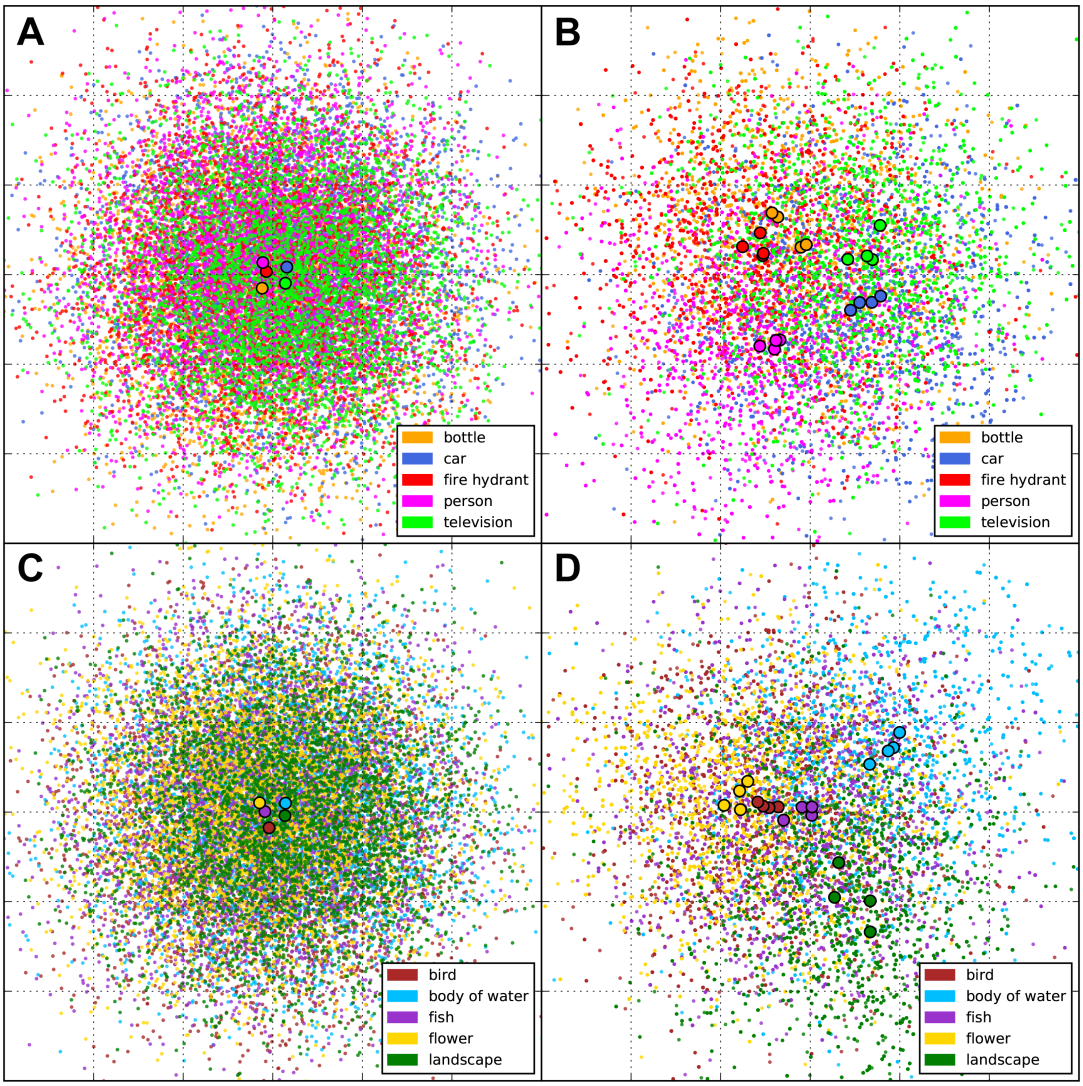

Figure 5.7: Categories are better separated by MCMCP representations. Fisher Linear Discriminant projections of **A**. CI comparisons for each category of group 1, **B.** samples for MCMCP chains for category group 1, **C.** CI comparisons for each category of group 2, and **D.** samples for MCMCP chains for category group 2. For A and C, large dots represent category means. For B and D, large dots represent chain means.

mixture of Gaussians density estimate. This produces realistic-looking multi-modal mental category templates, which to our knowledge has never been accomplished with respect to natural image categories.

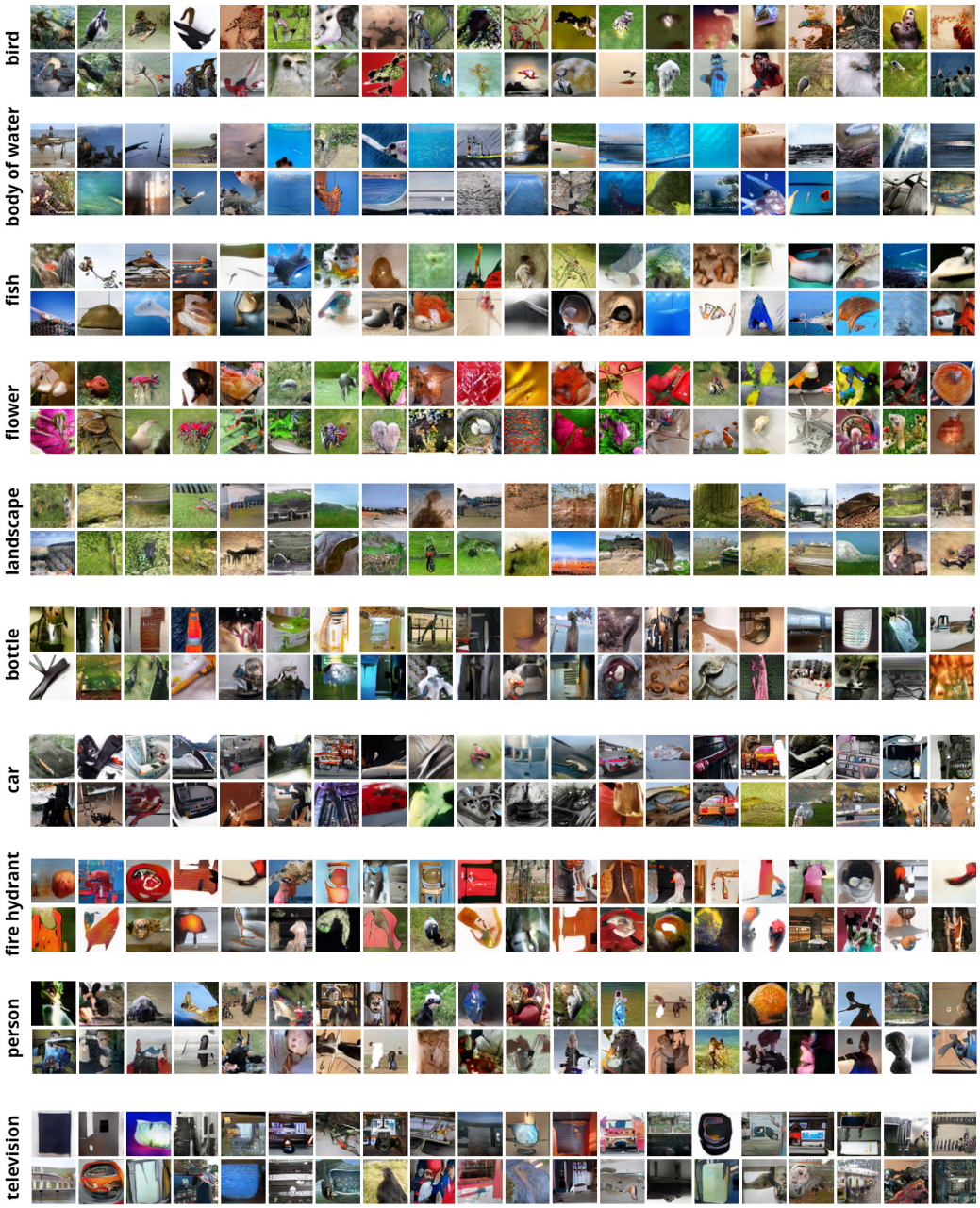

Figure 5.8: 40 most interpretable mixture component means (modes) taken from the 50 largest mixture weights for category.

Table 5.1: Classification performance compared to chance for both category sets (chance is 0.20).

	bird	**body of water**	**fish**	**flower**	**landscape**	**all**
MCMCP Mean	.33	.28	.01	.57	.67	.37
MCMCP Density	.23	.31	.18	.44	.73	**.38**
CI Mean	.23	.30	.2	.24	.52	.30
	bottle	**fire hydrant**	**car**	**person**	**television**	**all**
MCMCP Mean	.15	.11	.32	.77	.73	**.42**
MCMCP Density	.25	.26	.56	.19	.50	.35
CI Mean	.28	.15	.62	.12	.13	.26

5.5.3 Efficacy in classifying real images

Improvements of MCMCP over classification images may be both perceptible and detectable, but their practical differences are also worth considering — do they differ significantly on real-world tasks? Moreover, if the representations we learn through MCMCP are good approximations of people, we would expect them to perform reasonably well in categorizing real images. For this reason, we provide an additional quantitative assessment of the samples we obtained and compare them to classification images (CI) using an external classification task.

To do this, we scraped approximately 500 images from Flickr for each of the ten categories, which was used for a classification task. To classify the images using our human-derived samples, we used (1) the nearest-mean decision rule, and (2) a decision rule based on the highest log-probability given by our ten density estimates. For classification images, only a nearest-mean decision rule can be tested. In all cases, decision rules based on our MCMCP-obtained samples overall outperform a nearest-mean decision rule using classification images (see Table 5.1). In category group 1, the MCMCP density performed best and was more even across classes. In category group 2, nearest-mean using our MCMCP samples did much better than a density estimate or CI-based nearest-mean.

5.6 Discussion

Our results demonstrate the potential of our method, which leverages both psychological methods and deep surrogate representations to make the problem of capturing human category representations tractable. The flexibility of our method in fitting arbitrary generative models allows us

to visualize multi-modal category templates for the first time, and improve on human-based classification performance benchmarks. It is difficult to guarantee that our chains explored enough of the relevant space to actually capture the concepts in their entirety, but the diversity in the modes visualized and the improvement in class separation achieved are positive indications that we are on the right track. Further, the framework we present can be straightforwardly improved as generative image models advance, and a number of known methods for improving the speed, reach, and accuracy of MCMC algorithms can be applied to MCMCP to make better use of costly human trials.

There are several obvious limitations of our method. First, the structure of the underlying feature spaces used may either lack the expressiveness (some features may be missing) or the constraints (too many irrelevant features) needed to map all characteristics of human mental categories in a practical number of trials. Even well-behaved spaces are very large and will require many trials to adequately cover. Addressing this will require continuing exploration of a variety of generative image models. We see our work as part of an iterative refinement process that can yield more granular human observations and inform new deep network objectives and architectures, both of which may yet converge on a proper, yet tractable model of real-world human categorization.

6
Conclusion

The present work has attempted to forge powerful machine learning methods into tools that psychologists can apply to studying complex phenomena that are often out of reach. This is in contrast to thinking of these models as useful if and only if they are valid cognitive (or for that matter, abstract biological) models *in toto*. Instead, machine learning tools solve problems that we can fix as constants in a larger system of complex cognitive components, the interactions and entanglement of which in the real world are exceedingly hard to grasp with scientific precision. However, the practical scope of such a program has not yet been demonstrated through broad proofs-of-concept. The present goal of this thesis has been to fill this gap, and to jump-start an ecological revolution of sorts, imperfect as it may be, as a complementary new paradigm for the rigorous study of intelligent human behavior.

6.1 Summary of the Current Work

In Chapter 1, I reanimated the classic problem of external validity in the context of psychology, and saw that our explanations of human behavior depend on what assumptions we are forced to make in the face of imperfect methodological tools. A precise science of human behavior is particularly threatened by the fact that observations of the brain do not guarantee that we will learn

anything (at least immediately) about mental objects and human computational abstractions—the mind is not directly observable.

Interestingly, partially rooted in concepts from scientific psychology, machine learning practitioners have inadvertently, in the course of focusing on surprisingly different goals, returned the favor by stress testing some of our own models in the real world (or something more like it). It just so happens that what these resulting models now learn and do is harder to hold in mind—to explain in a fully, scientifically (and traditionally) satisfying way. In retrospect, the idea that an anything-goes approach to practically solving the problem of robustly detecting an object in the environment like humans might yield insights about humans doing the exact same thing is not that surprising. In any case, it is natural to ask how psychologists can best leverage this outcome, which I have tried my best to motivate in a meaningful way.

In Chapter 2, we discussed the innovative approaches from classic cognitive psychology to inferring mental content from behavior alone, and a few satisfyingly general assumptions. Despite this, we do not have the human participant and physical computational power to regularly sustain these methods for complex domains like vision, which might otherwise bring these efforts to greater fruition. At least for certain perceptual problems (that are certainly of interest), machine learning methods that focus on the abstract problem being solved as opposed to laboratory phenomenon have successfully scaled. We then reviewed the primary tools from modern computer vision, and the well-chosen translation invariance bias that is largely to thank (along with the internet-led abundance of training data and increased computational resources). Some psychologists have already begun to experiment with deep neural networks, and with success, lending further motivation to the current efforts.

In Chapter 3, we set out to answer a question about human behavior (i.e., what are people likely to remember in busy, natural scenes), with three important departures from the standard laboratory paradigm. The first was to literally eschew the laboratory in order to obtain a large sample of human participants. The second was to study the phenomenon of interest with the largest image (stimulus) dataset we could find that met our requirements. Lastly, and most importantly, after testing a number of hypotheses about aspects of stimulus content that we both know how to measure and think might contribute to memorability, we turned to the task of maximizing predictability of human memory for objects in natural scenes. We learned for example that simple image category labels and visual saliency (where we are drawn to look in an image) are broadly explanatory across much (but not all) of the image dataset, but were also able to demonstrate in parallel and without additional cost that knowledge internalized in deep neural networks about object discrimination was sufficient to yield superior prediction of our human

data. This was true compared to both the explanatory factors we revealed as well as several competitive computational baselines. We also came away with a benchmark dataset inspired by the culture and productivity of machine learning competitions, but focused on rich human behavior instead of more weakly aligned general goals.

For memory, at least to some degree, deep representations are apparently quite relevant, but how can we start to get a sense of when we should expect success or failure? In Chapter 4, we turned to understanding the generality and quantifying the utility of such representations, and ask a crucial question: can deep representations compete with, or better yet approximate the sorts of rich content that we often obtain with human similarity judgments in psychology? Further, how easily and rapidly can they be adapted to a particular task, new set of stimuli, or alternative cognitive context? We found that raw deep representations are already good approximations of human psychological content as typically measured, although they are also understandably more perceptual in nature, and lacking higher level taxonomic distinctions. Encouragingly, we found that human representations can be better approximated by solving a simple convex problem, essentially scaling importance of the deep network's features, and indicating that most of the necessary information is already contained in the network. Lastly, we showed that our simple correction generalized to a different context involving semi-novel category learning. The resulting image dataset and corresponding human judgments were similar in size to those employed in Chapter 3, and can also serve as a potential benchmark for explaining mental content directly, especially before being applied to a particular cognitive model.

Finally, in Chapter 5, we set ourselves to the task of re-purposing deep neural networks to aid in studying an essential component of modern cognitive science—subject probability distributions that help us describe how humans reason under considerable uncertainty. In particular, we revisit the topic of categorization, and ask with what resolution we might capture human knowledge about complex visual concepts. By making use of a recent, and successful deep generative network to both parameterize and synthesize high-dimensional images, as well as modern innovations in psychological methods, we appear to have made some strides, and can both intuitively visualize and quantitatively evaluate the results. Arguably, there is a good deal of room for improvement, but the limitations of this approach are some of the most likely to be improved by already-nearing developments in machine learning, namely better image renderers.

6.2 Limitations of Applying Deep Neural Networks

We took categorization as a phenomenon of interest in Chapters 4 and 5 because engaging with classic theories of category learning with naturalistic stimuli has traditionally been so difficult. However, this class of models can be thought of as being highly constrained to the simplest form of categorization behavior (and in some sense, missing the point entirely). Murphy and Medin (1985) argued that the coherence of a concept should not be limited in this way, since humans clearly possess a more structured understanding of the world that transcends its feature correlations, and goes a long way in describing more intelligent human behavior. That is, many of the still popular models discussed in Chapter 1 are feature-based accounts — they only require that stimuli be represented by a fixed set of continuous or discrete features. The utility in using deep networks in the present work was indeed such feature representations. Can we also hope to capture more complex behavior as well?

Lake, Ullman, Tenenbaum, and Gershman (2017) intentionally echo Murphy and Medin (1985) in arguing that modern machine learning breakthroughs are limited to function approximation, pattern recognition, and feature learning, and do not support a more structured understanding of the world that includes intuitive theories and casual reasoning. One response to this is to note that exploiting correlational structure in the world helps guide and interacts with higher level processes, for example by providing heuristics for fast processing, suggesting that the two are entangled.

However, it is also the case that, as machine learning practitioners begin to feel for the edges of current applications, and consequently current limitations, tools for more inherently structured domains are emerging (Battaglia et al., 2018; Graves, Wayne, & Danihelka, 2014). To the extent that these developing models continue to solve certain new aspects of human-relevant problems, the representations they learn may be useful to psychology (e.g., program primitives and abstractions learned by neural program induction models).

6.3 Directions for Future Work

6.3.1 The gift that keeps on giving

If a given type of deep network is found to be of use for a particular cognitive modeling context, it is likely to become more useful with time. The reason is that, due to the culture of standardized datasets and benchmarking, the machine learning community tends to produce regular im-

provements (by the month or year) to model architectures and training regimes, such as better classification accuracy, image compression, or image rendering.

While improvements to object classifiers may be less drastic (they are already quite similar to human abilities) and therefore less useful, image generation for example has a long way to go. For example, in Chapter 4, the depth and overall rank accuracy of the classifier had little correlation with the weak variation in fit to human representations, but human category approximations in Chapter 4 were clearly limited considerably by the quality of the image generator network (only orange blobs could be found in the space to represent fire hydrants). To our luck, since the MCMCP experiments in the current work were conducted, at least a handful of superior networks that produce larger and higher quality renderings for the same object dataset, and other interesting datasets, have been developed (see Karras, Aila, Laine, & Lehtinen, 2017, for just one stunning example).

6.3.2 From exploitation to intervention

Because psychologists are perhaps the most aware of how deep networks deviate from human behavior, we may be particularly well-positioned to suggest improvements or interesting variants. Many of these suggestions are likely to be simple augmentations — a source of low-hanging fruit for perhaps the next few years.

For example, most object classifiers are trained on the same ILSVRC competition set of ImageNet (Deng et al., 2009), and the data source of most of the tools used in this thesis, which has a curious yet unchanging stratification. For example, it is heavily biased toward dog breeds (about 20% of the total number of classes), but also contains other animals, different types of automobiles, and household objects etc. Altering this stratification, and resampling from the much larger parent ImageNet dataset is a simple change that might yield very different representations that are useful for different modeling tasks. Another idea is to alter the level (or multiplicity) of abstraction of the training labels (e.g., learning to classify "dogs" instead of "Dalmatians"). Some initial results from my own recent follow-up on this topic has yielded similarly useful representations as those in Chapter 4, and is a much better fit to human generalization out-of-the-box (Peterson, Soulos, Nematzadeh, & Griffiths, 2018). Moreover, in Chapter 4, I proposed a method for adapting deep representations to a particular domain or context using human generalization data, but could we derive such a transformation by principle, without supervision? I propose a potential starting point in Peterson and Griffiths (2017).

Other strategies for engaging our cousins in machine learning are more difficult, but have po-

tential for bringing larger rewards. One interesting phenomenon is the reformation of questions and criticisms into challenges. Lake, Salakhutdinov, and Tenenbaum (2015) proposed a stronger focus on human-like abilities to learn quickly, and proposed a challenge dataset to measure it, now a standard benchmark that has spawned a number of extremely competitive new model architectures (see Finn, Abbeel, & Levine, 2017; Koch, Zemel, & Salakhutdinov, 2015; Vinyals, Blundell, Lillicrap, & Wierstra, 2016, although there are many more).

Because such practical challenges are taken seriously, a number of fascinating new tools (and perhaps even more likely candidates for human cognitive models) have been put forward. If we can turn reform our observations and criticisms into tangible challenges, we can effectively outsource some of the highly complex engineering that such a diverse field can provide, and provide interesting subject matter for machine learning research at the same time.

6.3.3 Moving Beyond Vision

The bulk of the work I have presented here is limited to vision, and more generally perception, but as we have already pointed out, there have been interesting developments in other domains as well, such as reinforcement learning (Mnih et al., 2015), program induction (Graves et al., 2014), language (Luong, Kayser, & Manning, 2015), and relational reasoning (Santoro et al., 2017) to name just a few examples. Some of my own joint work is in the initial stages of extending the present framework to non-perceptual domains such as human analogy-making (Chen, Peterson, & Griffiths, 2017) for example, but there is a great deal more to be done.

6.3.4 More Immediate Questions

Testing Theories of Categorization. In Chapter 1, we introduced an identifiability problem of categorization models to motivate the difficulty in modeling processes that make use of complex, unobserved representations. However, while we made progress both in capturing human-like representational spaces and subjective probability information over these spaces that represent category information, we are still left with a question of which categorization strategy, namely abstraction versus a memory-based search-and-compare method, is a better characterization (if either) of human behavior with naturalistic stimuli. If we can use deep networks to capture enough information about mental content, we can potentially make a good inference about which models are a better fit.

Scope and Generality of Approximated "Psychological Representations". The selection of image stimuli under consideration in Chapter 4 were obviously small compared to the say roughly 1.2 million images in ILSVRC (Deng et al., 2009), but even more pressing is the generality of the learned representations to different modeling tasks. To what extent can we predict other cognitive processes that operate on top of such representations, for example by improving predictions for object memorability scores in the large dataset obtained in Chapter 3? Are there important contexts where these stimulus characterizations will utterly fail?

Unfactoring Representation and Process. A more fundamental limit to the above inquiry is the fact that we have proceeded with the simplifying assumption that a relatively fixed representation can be learned to support downstream usage and processing. Deep classifiers used in this work are themselves and interesting case that breaks this assumption, since the representations are learned through the pressure to categorize complex stimuli, and not through some general learning process than occurs in advance (e.g. like many in psychology; Austerweil & Griffiths, 2013; Kemp & Tenenbaum, 2008). In fact, it has not yet been demonstrated that deep unsupervised feature learning can catch up to deep supervised methods (Donahue et al., 2016). Can the current framework be extended to modeling dynamic and complex changes in representational content as additional learning and task pressures are experienced by a learner?

6.4 Concluding Remarks

The level of explanation and precision that we wish to obtain as psychologists is ultimately our decision, but in any case, it is important to consider that the reason that human intelligence is so fascinating to us, and worth understanding and applying in machines, is that it is surprisingly complex, efficient, and seemingly ever-expanding (as we continue to use our faculties to grasp ever more about ourselves and the world in which we find ourselves). Understanding such a grand natural system is going to take all of the tricks that we have at our disposal, and if something comes along that looks more like us than our own explicit models of ourselves, we ought to take a closer look, and possibly even borrow from the parts that work best. What I have presented here is merely a fragment of what might be possible in mining automated learning systems that abstract knowledge from large data sources, but it is perhaps also an important and necessary demonstration that will expand the scope of the field and the sorts of questions we can answer as a rigorous science of behavior.

References

Achanta, R., Hemami, S., Estrada, F., & Susstrunk, S. (2009). Frequency-tuned salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1597–1604).

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines. *Cognitive Science*, *9*(1), 147–169.

Agrawal, P., Stansbury, D., Malik, J., & Gallant, J. L. (2014). Pixels to voxels: Modeling visual representation in the human brain. *arXiv preprint arXiv:1407.5104*.

Ahumada, A. J., Jr. (1996). Perceptual classification images from vernier acuity masked by noise. *Perception*, *25*, 2–2.

Arbelaez, P., Pont-Tuset, J., Barron, J., Marques, F., & Malik, J. (2014). Multiscale combinatorial grouping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 328–335).

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*(2), 216–233.

Austerweil, J. L., & Griffiths, T. L. (2013). A nonparametric bayesian framework for constructing flexible feature representations. *Psychological Review*, *120*(4), 817.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., … Pascanu, R. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*.

Borji, A., Sihite, D. N., & Itti, L. (2012). Salient object detection: A benchmark. In *Proceedings of the IEEE European Conference on Computer Vision* (pp. 414–429).

Borji, A., Sihite, D. N., & Itti, L. (2013). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, *22*(1), 55–69.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008a). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325–14329.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008b). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325–14329.

Bruce, N., & Tsotsos, J. (2006). Saliency based on information maximization. In *Advances in Neural Information Processing Systems* (pp. 155–162).

Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., & Oliva, A. (2015). Intrinsic and extrinsic effects on image memorability. *Vision Research*, *116*, 165–178.

Caramazza, A., Hersh, H., & Torgerson, W. S. (1976). Subjective structures and operations in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *15*(1), 103–117.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287–291.

Chen, D., Peterson, J. C., & Griffiths, T. L. (2017). Evaluating vector-space models of analogy. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Cheng, M.-M., Zhang, G.-X., Mitra, N. J., Huang, X., & Hu, S.-M. (2011). Global contrast based salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 409–416).

Chomsky, N. (1959). A review of bf skinner's verbal behavior. *Language*, *35*(1), 26–58.

Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., & LeCun, Y. (2015). The loss surfaces of multilayer networks. In *Artificial intelligence and statistics* (pp. 192–204).

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*, 2493–2537.

Cummins, R. C. (1989). *Meaning and mental representation*. Bradford Books / MIT Pres.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *2*(4), 303–314.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 1, pp. 886–893).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255).

Donahue, J., Krähenbühl, P., & Darrell, T. (2016). Adversarial feature learning. *arXiv preprint arXiv:1605.09782*.

Dubey, R., Peterson, J. C., Khosla, A., Yang, M.-H., & Ghanem, B. (2015). What makes an object memorable? In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1089–1097).

Dumoulin, V., Belghazi, I., Poole, B., Mastropietro, O., Lamb, A., Arjovsky, M., & Courville, A. (2016). Adversarially learned inference. *arXiv preprint arXiv:1606.00704*.

Ehresman, D., & Wessel, D. L. (1978). *Perception of timbral analogies*. Paris: Centre Georges Pompidou.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Erdogan, G., & Jacobs, R. A. (2017). Visual shape perception as bayesian inference of 3d object-centered shape representations. *Psychological Review*, *124*(6), 740.

Everingham, M., & Winn, J. (2010). *The pascal visual object classes challenge 2010 (voc2010) development kit*. Citeseer.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. New York: Springer.

Fukushima, K., & Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and Cooperation in Neural Nets* (pp. 267–285). Springer.

Gershman, S., & Tenenbaum, J. B. (2015). Phrase similarity in humans and machines. In *In Proceedings of the 37th Annual Conference of the Cognitive Science Society.*

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (pp. 315–323).

Goldstone, R. L. (1994a). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*(2), 178.

Goldstone, R. L. (1994b). The role of similarity in categorization: Providing a groundwork. *Cognition*, *52*(2), 125–157.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative adverserial nets. In *Advances in Neural Information Processing Systems* (pp. 2672–2680).

Graves, A. (2008). *Supervised sequence labelling with recurrent neural networks* (Unpublished doctoral dissertation). PhD thesis. Ph. D. thesis, Technical University of Munich.

Graves, A., Wayne, G., & Danihelka, I. (2014). Neural turing machines. *arXiv preprint arXiv:1410.5401*.

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. In *Advances in Neural Information Processing Systems* (pp. 545–552).

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, *57*(1), 97–109.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hou, X., & Zhang, L. (2009). Dynamic visual attention: Searching for coding length increments. In *Advances in Neural Information Processing Systems* (pp. 681–688).

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Isola, P., Parikh, D., Torralba, A., & Oliva, A. (2011). Understanding the intrinsic memorability of images. In *Advances in Neural Information Processing Systems* (pp. 2429–2437).

Isola, P., Xiao, J., Parikh, D., Torralba, A., & Oliva, A. (2014). What makes a photograph memorable? *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *36*(7), 1469–1482.

Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 145–152).

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *20*(11), 1254–1259.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., … Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia* (pp. 675–678).

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology*, *8*, 1726.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2106–2113).

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014, 11). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLOS Computational Biology*, *10*(11), 1-29. doi: 10.1371/journal.pcbi.1003915

Khosla, A., Bainbridge, W. A., Torralba, A., & Oliva, A. (2013). Modifying the memorability of face photographs. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3200–3207).

Khosla, A., Raju, A. S., Torralba, A., & Oliva, A. (2015). Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE International Conference on Computer Vision*.

Khosla, A., Xiao, J., Isola, P., Torralba, A., & Oliva, A. (2012). Image memorability and visual inception. In *SIGGRAPH Asia Technical Briefs*.

Khosla, A., Xiao, J., Torralba, A., & Oliva, A. (2012). Memorability of image regions. In *Advances in Neural Information Processing Systems* (pp. 305–313).

Kim, J., Yoon, S., & Pavlovic, V. (2013). Relative spatial features for image memorability. In *Proceedings of the 21st ACM International Conference on Multimedia* (pp. 761–764).

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations*.

Koch, G., Zemel, R., & Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop* (Vol. 2).

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*(3), 558.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097–1105).

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological review*, *99*(1), 22.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016, 04). Deep neural networks as a computational model for human shape sensitivity. *PLOS Computational Biology*, *12*(4), 1-26. doi: 10.1371/journal.pcbi.1004896

Kulis, B. (2013). Metric learning: A survey. *Foundations and Trends in Machine Learning*, *5*(4), 287–364.

Lake, B. M., Lawrence, N. D., & Tenenbaum, J. B. (2018). The emergence of organizing structure in conceptual representation. *Cognitive Science*.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*.

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep neural networks predict category typicality ratings for images. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society.*

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Li, Y., Hou, X., Koch, C., Rehg, J. M., & Yuille, A. L. (2014). The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 280–287).

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., … Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Proceedings of the IEEE European Conference on Computer Vision* (pp. 740–755).

Lin, Z., Memisevic, R., & Konda, K. (2015). How far can we go without convolution: Improving fully-connected networks. *arXiv preprint arXiv:1511.02580*.

Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., & Shum, H.-Y. (2011). Learning to detect a salient object. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *33*(2), 353–367.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3431–3440).

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (p. 103-189). New York: Wiley.

Luong, T., Kayser, M., & Manning, C. D. (2015). Deep neural language models for machine translation. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning* (pp. 305–309).

Mancas, M., & Le Meur, O. (2013). Memorability of natural scenes: the role of attention. In *20th IEEE International Conference on Image Processing.*

Margolin, R., Tal, A., & Zelnik-Manor, L. (2013). What makes a patch distinct? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1139–1146).

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. new york, ny, usa: Henry holt and co. *Inc June*.

Martin, J. B., Griffiths, T. L., & Sanborn, A. N. (2012). Testing the efficiency of Markov Chain Monte Carlo with people using facial affect categories. *Cognitive Science*, *36*(1), 150–162.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., … Ostrovski, G. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529.

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human object-similarity judgments reflect and transcend the primate-IT object representation. *Frontiers in Psychology*, *4*, 128.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(3), 289.

Navarro, D. J., & Lee, M. D. (2004). Common and distinctive features in stimulus similarity: A modified version of the contrast model. *Psychonomic Bulletin & Review*, *11*(6), 961–974.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(1), 87.

Oord, A. v. d., Kalchbrenner, N., & Kavukcuoglu, K. (2016). Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*.

Perazzi, F., Krahenbuhl, P., Pritch, Y., & Hornung, A. (2012). Saliency filters: Contrast based filtering for salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 733–740).

Peterson, J. C., Abbott, J., & Griffiths, T. (2016). Adapting deep network features to capture psychological representations. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2363–2368). Austin, TX: Cognitive Science Society.

Peterson, J. C., Abbott, J., & Griffiths, T. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*.

Peterson, J. C., & Griffiths, T. L. (2017). Evidence for the size principle in semantic and perceptual domains. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Peterson, J. C., Soulos, P., Nematzadeh, A., & Griffiths, T. L. (2018). Learning hierarchical visual representations in deep neural networks using hierarchical linguistic labels. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Peterson, J. C., Suchow, J., Aghi, K., Alexander, K., & Griffiths, T. (2018). Capturing human category representations by sampling in deep feature spaces. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Plonsky, O., Erev, I., Hazan, T., & Tennenholtz, M. (2017). Psychological forest: Predicting human behavior. In *Aaai* (pp. 656–662).

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3p1), 353.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*(3), 382–407.

Rips, L. J., & Shoben, E. J. (1973). Semantic distance and the verification of semantic relations, journal of verbal learning and verbal behavior. *Journal of Verbal Learning and Verbal Behavior*, *12*(1), 1.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386.

Rumelhart, D. E., & Abrahamson, A. A. (1973). A model for analogical reasoning. *Cognitive Psychology*, *5*(1), 1–28.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. In *International Journal of Computer Vision*.

Saad, D., & Solla, S. A. (1995). On-line learning in soft committee machines. *Physical Review E*, *52*(4), 4225.

Sanborn, A. N., & Griffiths, T. L. (2007). Markov Chain Monte Carlo with people. In *Advances in Neural Information Processing Systems* (pp. 1265–1272).

Sanborn, A. N., Griffiths, T. L., & Shiffrin, R. M. (2010). Uncovering mental representations with markov chain monte carlo. *Cognitive Psychology*, *60*(2), 63–106.

Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., & Lillicrap, T. (2017). A simple neural network module for relational reasoning. In *Advances in neural information processing systems* (pp. 4967–4976).

Schwartz, R. M., & Humphreys, M. S. (1973). Similarity judgments and free recall of unrelated words. *Journal of Experimental Psychology*, *101*(1), 10.

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge University Press.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*(4468), 390–398.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, *86*(2), 87.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., … Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Skinner, B. F. (1957). *Verbal behavior*. Copley Publishing Group.

Skinner, B. F. (1977). Why i am not a cognitive psychologist. *Behaviorism*, *5*(2), 1–10.

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Standing, L. (1973). Learning 10000 pictures. *The Quarterly Journal of Experimental Psychology*, *25*(2), 207–222.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2014). Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*.

Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319–2323.

Tseng, P.-H., Carmi, R., Cameron, I. G., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, *9*(7), 4.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327.

Vinyals, O., Blundell, C., Lillicrap, T., & Wierstra, D. (2016). Matching networks for one shot learning. In *Advances in Neural Information Processing Systems* (pp. 3630–3638).

Vondrick, C., Pirsiavash, H., Oliva, A., & Torralba, A. (2015). Learning visual biases from human imagination. In *Advances in Neural Information Processing Systems* (pp. 289–297).

Vong, W. K., Hendrickson, A., Perfors, A., & Navarro, D. (2016). Do additional features help or harm during category learning? an exploration of the curse of dimensionality in human learners. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society.*

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological Review*, *114*(2), 245.

Yang, C., Zhang, L., Lu, H., Ruan, X., & Yang, M.-H. (2013). Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3166–3173).

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, *8*(7), 32.

A
Chapter 4 Details

A.1 Experiment 1 & 2 Stimuli

Below is the total set of $1,440$ stimuli used in our experiments, grouped by domain (animals, fruits, furniture, vegetables, vehicles, and "various") and experiment (similarity judgments versus category learning). All corresponding image set pairs (e.g., fruits for similarity experiments, and fruits for categorization experiments) are perfectly stratified by subordinate class (e.g., exactly three apples in each, etc), except for the animals domain.

Figure A.1: Animal stimuli used in similarity experiments.

Figure A.2: Animal stimuli used in categorization experiments.

Figure A.3: Fruit stimuli used in similarity experiments.

Figure A.4: Fruit stimuli used in categorization experiments.

Figure A.5: Furniture stimuli used in similarity experiments.

Figure A.6: Furniture stimuli used in categorization experiments.

Figure A.7: Vegetable stimuli used in similarity experiments.

Figure A.8: Vegetable stimuli used in categorization experiments.

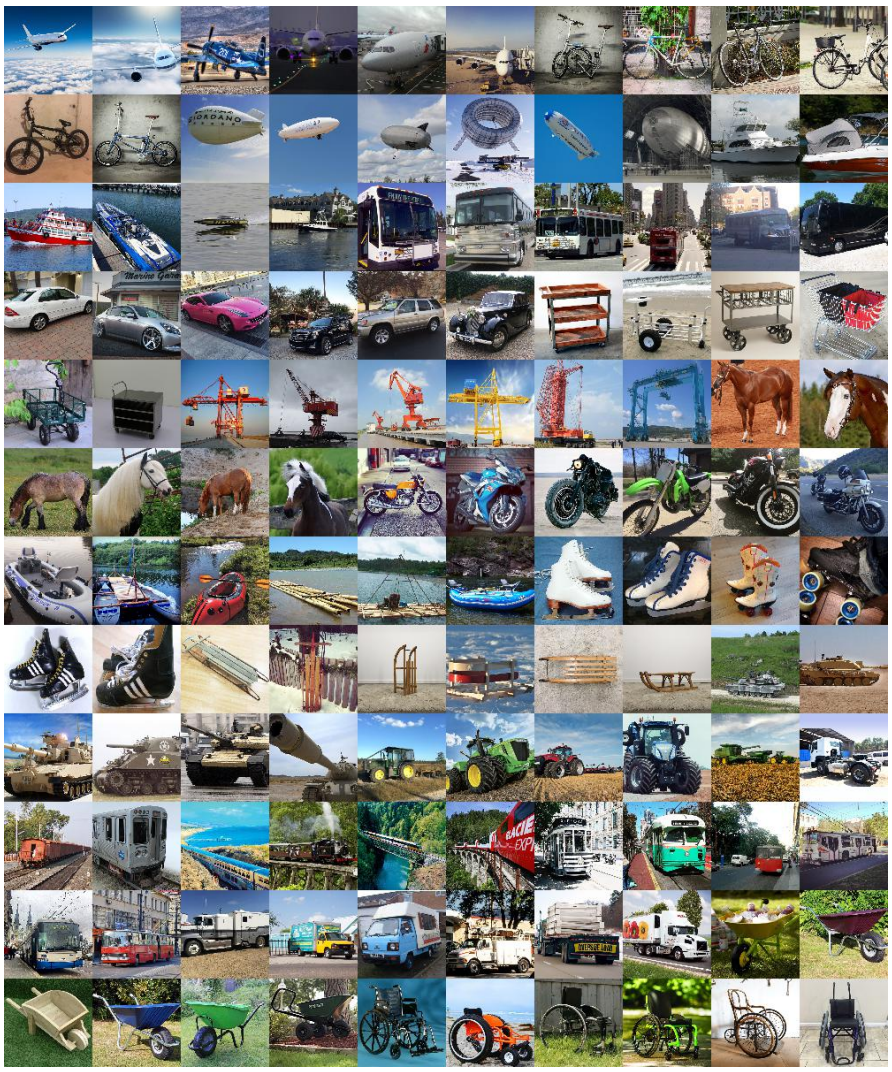

Figure A.9: Vehicle stimuli used in similarity experiments.

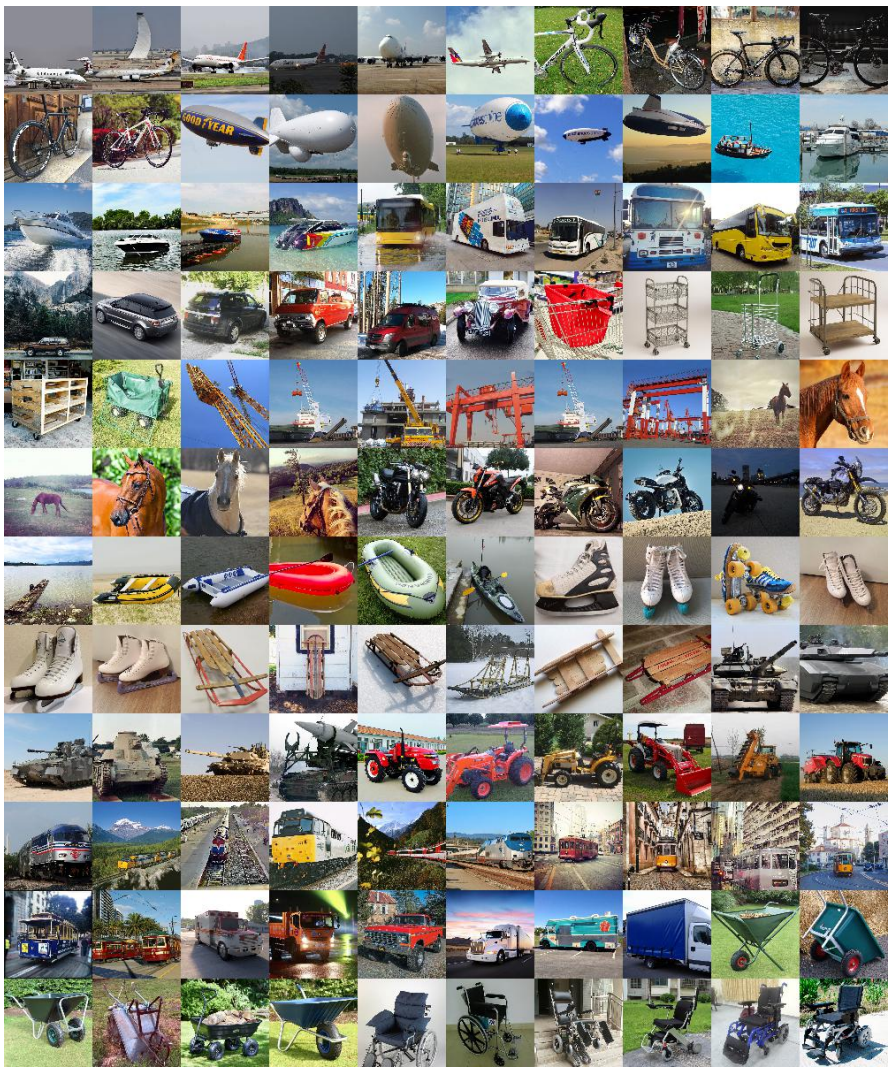

Figure A.10: Vehicle stimuli used in categorization experiments.

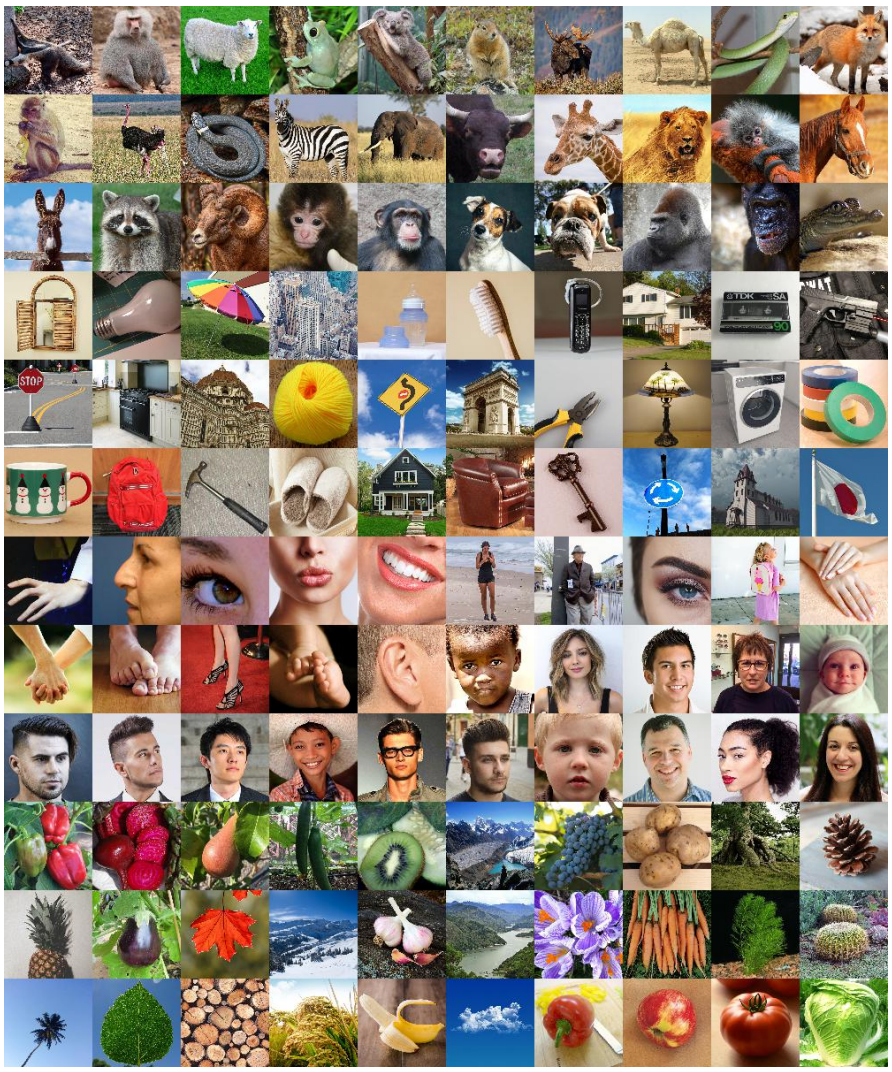

Figure A.11: "Various" stimuli used in similarity experiments.

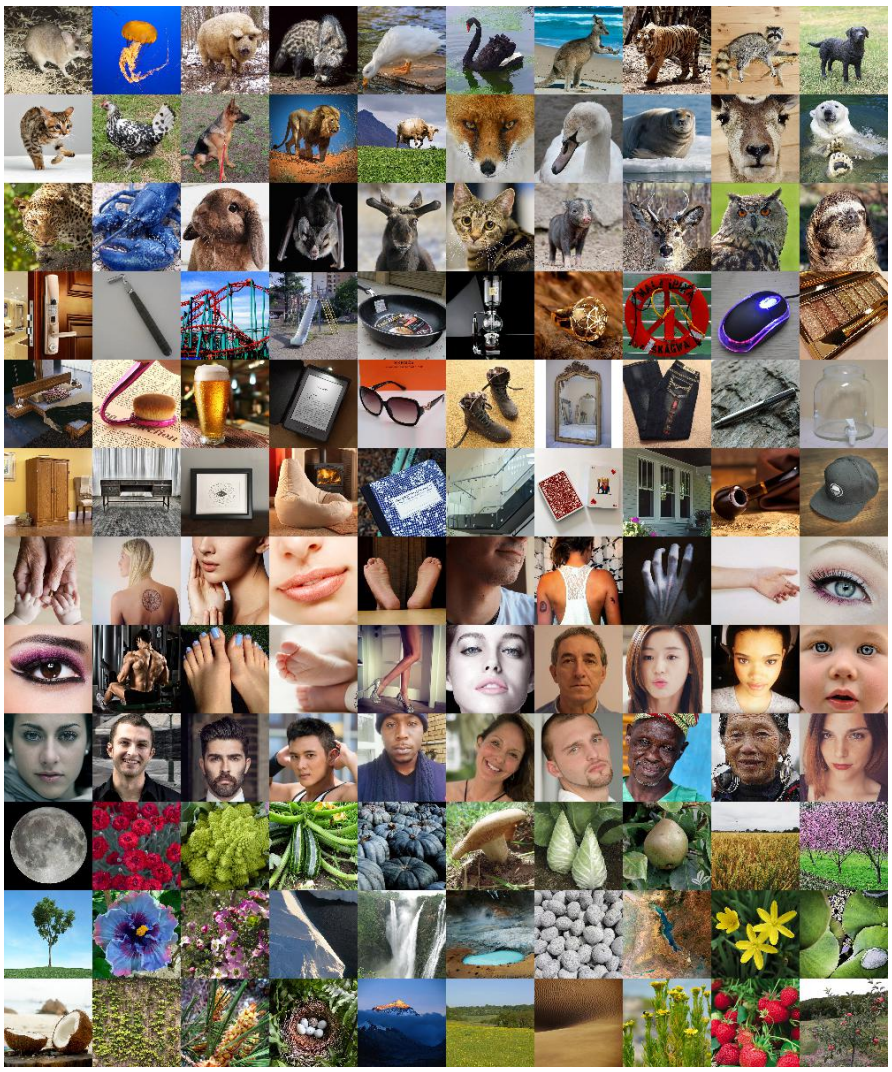

Figure A.12: "Various" stimuli used in categorization experiments.

A.2 Human & Estimated Similarity Matrices

Human (Experiment 1), deep network (VGG), and transformed similarity matrices are shown below for each of the six domains. Each domain appears to exhibit a different level of sparsity (e.g., animals versus vehicles). For most domains, the ordered alignment of the images reveals categorical clustering in the judgments that are better represented after the transformation of the deep features.

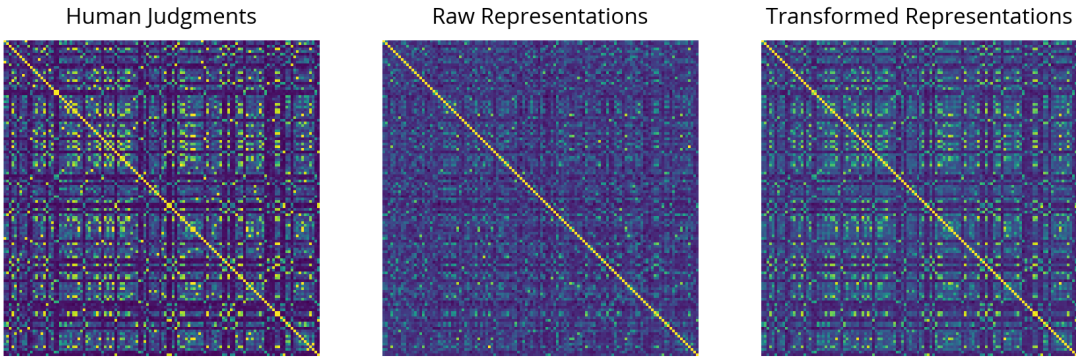

Figure A.13: Animals. Human similarity matrices, inner products from raw deep representations, and predicted similarities after transforming the deep representations.

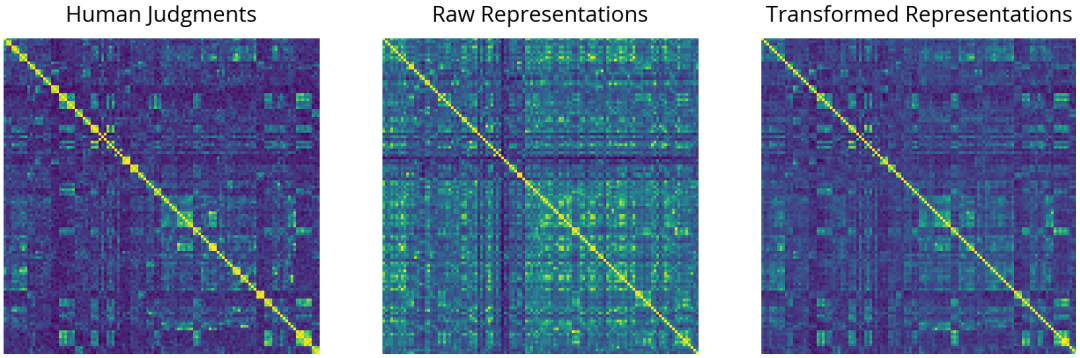

Figure A.14: Fruits. Human similarity matrices, inner products from raw deep representations, and predicted similarities after transforming the deep representations.

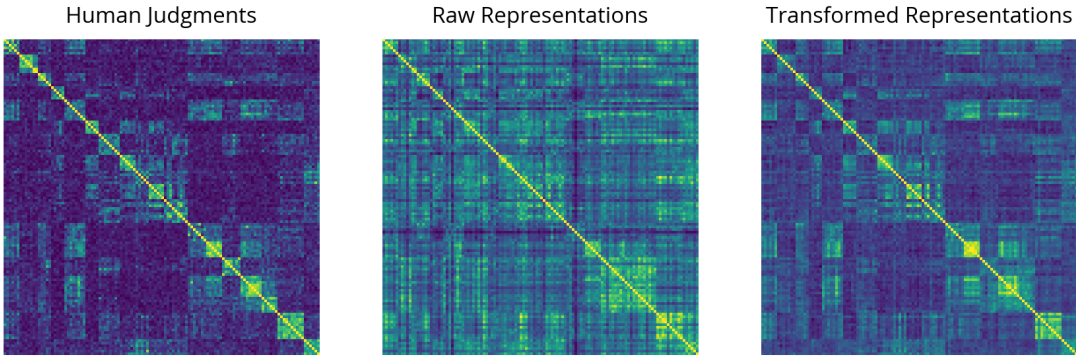

Figure A.15: Furniture. Human similarity matrices, inner products from raw deep representations, and predicted similarities after transforming the deep representations.

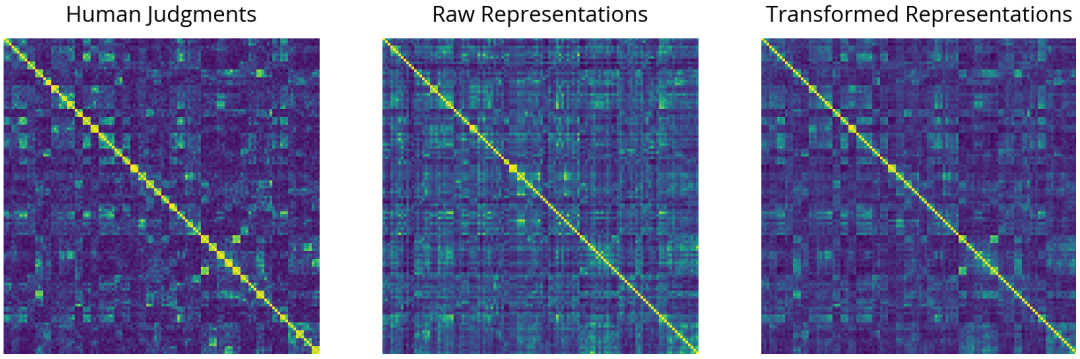

Figure A.16: Vegetables. Human similarity matrices, inner products from raw deep representations, and predicted similarities after transforming the deep representations.

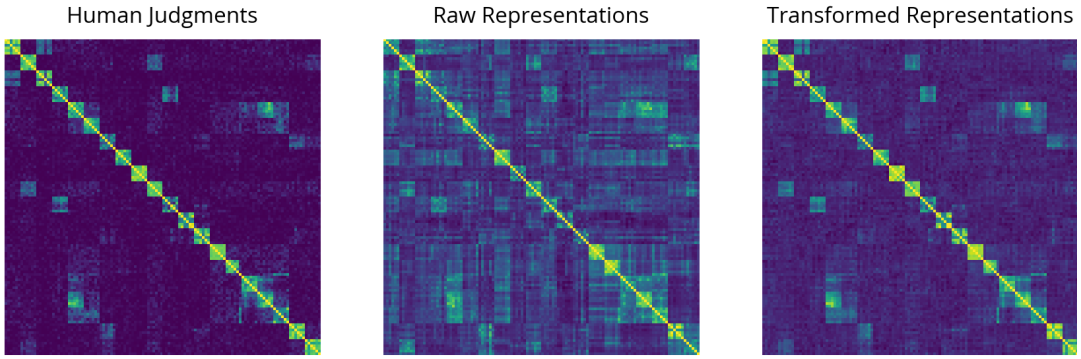

Figure A.17: Vehicles. Human similarity matrices, inner products from raw deep representations, and predicted similarities after transforming the deep representations.

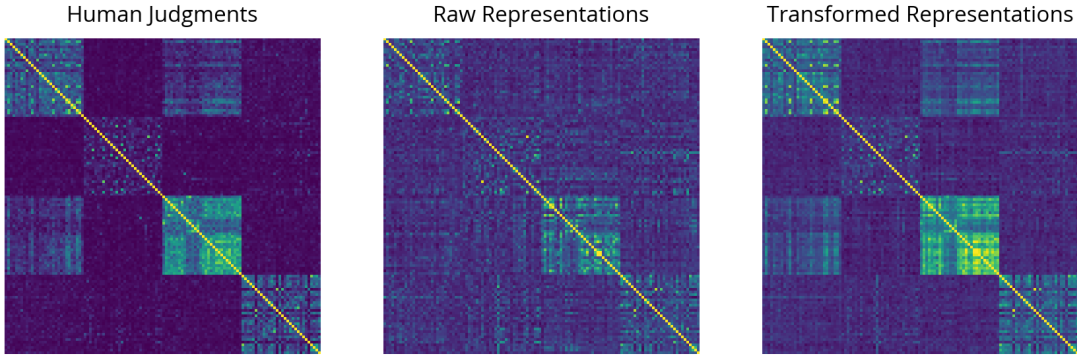

Figure A.18: Various. Human similarity matrices, inner products from raw deep representations, and predicted similarities after transforming the deep representations.