Insulating Distributional Semantic Models from Catastrophic Interference

Willa M. Mannering and Michael N. Jones
Indiana University, Bloomington
[wmanneri] [jonesmn]@indiana.edu

Abstract

Predictive neural networks, such as word2vec, have seen impressive recent popularity as an architecture to learn distributional semantics in the fields of machine learning and cognitive science. They are particularly popular because they learn continuously, making them more space efficient and cognitively plausible than classic models of semantic memory. However, a major weakness of this architecture is *catastrophic interference* (CI): The sudden and complete loss of previously learned associations when encoding new ones. CI is an issue with backpropagation; when learning sequential data, the error signal dramatically modifies the connection weights between nodes—causing rapid forgetting of previously learned information. CI is a huge problem for predictive semantic models of word meaning, because multiple word senses interfere with each other. Here, we evaluate a recently proposed solution to CI from neuroscience, elastic weight consolidation, as well as a Hebbian learning architecture from the memory literature that does not produce an error signal. Both solutions are evaluated on an artificial and natural language task in their ability to insulate a previously learned sense of a word when learning a new one.

Keywords: distributional semantic models; catastrophic interference; word2vec; random vector accumulation; elastic weight consolidation

Introduction

Distributional models of semantic memory (DSMs; e.g., Landauer & Dumais, 1997) attempt to explain how humans learn the meaning of words through statistical inference. All DSMs are based on the distributional hypothesis of language (Harris, 1970), often summarized as learning a word's meaning "by the company it keeps" (Firth, 1957). Classic DSMs use counts of co-occurrence between words in a corpus to construct semantic representations. Recently, with the development of predictive DSMs and improvements in overall computing power, the fields of cognitive science and machine learning have seen an increase in popularity of error-driven DSMs within connectionist architectures. Predictive DSMs use the backpropagation of an error signal through the network to predict context and are particularly popular because they learn continuously—making them more space efficient and more cognitively plausible than earlier DSMs.

However, a major weakness of predictive DSMs is *catastrophic interference* (CI): The sudden and complete loss of previously learned associations when encoding new ones (French, 1999). When a predictive neural network is exposed to sequential data, the introduction of completely new information causes the error signal to be very large, effectively "shocking" the model and causing it to overcorrect the weights to accommodate the new information. The problem of CI is a major issue not only for functional reasons but for implications of cognitive plausibility as well.

The standard predictive network currently discussed in the literature is Mikolov et al.'s (2013) word2vec model. Word2vec is a feedforward neural network with input and output layers that contain one node per word in the vocabulary, and a hidden layer of approximately 300 nodes. The word2vec architecture has two possible model directions. The context may be used to predict the word—which is referred to as the Continuous Bag of Words (CBOW) model—or, the word may be used to predict the context—which is referred to as the skipgram model. We will use skipgram in this paper because it maps conceptually onto most connectionist models and has been shown to perform better with smaller training corpora than the CBOW model.

Dachapally and Jones (2018) recently investigated the impact of CI on the internal representations produced by predictive DSMs when applied to sequentially learned word senses. Because of its current popularity, they used Mikolov et al.'s (2013) word2vec model to evaluate the effects of CI on the model's final semantic representations. In their study, Dachapally and Jones used homonyms to measure the effects of CI. Take for example a homonym like *bank,* with its two distinct meanings: river-bank and financial-bank. The word *bank* should have its final representation positioned equidistant to its two meanings in semantic space. Because of CI, however, if the financial sense was learned first, followed by the river sense, the final representation of *bank* would be positioned proximal to river-bank words, and the financial sense would be forgotten. This study was the first evaluation of CI in a predictive semantic model. Now that we know CI affects semantic representations produced by predictive DSMs, we can begin to propose and evaluate possible solutions for CI.

The goal of the current paper is to expand on Dachapally and Jones' (2018) work by implementing and comparing two possible solutions to CI from the cognitive and neural sciences. The first candidate solution is *elastic weight consolidation* (Kirkpatrick et al., 2017) which has been impressively successful on machine learning tasks and can be considered a "vaccination" for predictive DSMs that would prevent the effects of CI. The second candidate solution is a different architecture, *random vector accumulation* (Jones, Willits, & Dennis, 2015), which can be considered naturally "immune" to the effects of CI by way of its learning mechanism.

The goal of elastic weight consolidation (EWC) is to allow a predictive neural network to learn two sequential tasks, Task A and then Task B, without incurring CI. To do this, Kirkpatrick et al. (2017) introduced a method to constrain the

parameters of a neural network after learning Task A so that the network can subsequently learn Task B without forgetting Task A. The new loss function they introduce is a quadratic penalty that differentially constrains parameters in the neural network depending on how important each parameter is to completing Task A. To determine which weights in the network are important for Task A they calculate the Fisher Information for each parameter—a mathematical method to measure the amount of information a variable carries about a parameter. The resulting loss function that gets minimized in elastic weight consolidation is:

$$\mathcal{L}(\theta) = \mathcal{L}_{\mathcal{B}}(\theta) + \sum_i \frac{\lambda}{2} F_i \left(\theta_i - \theta^*_{A,i}\right)^2 \qquad (1)$$

where $\mathcal{L}_{\mathcal{B}}(\theta)$ is the loss for Task B, λ controls how important Task B is compared to Task A, F is the Fisher Information calculated for each parameter, and θ represents the parameters in the network. Kirkpatrick et al. (2017) showed that EWC was able to insulate against CI when training a predictive neural network on the MNIST (LeCun et al., 1998) data set, a free data set of handwritten images. While EWC has been tested several times on categorization tasks, this paper will present the first implementation for use with distributional semantic models.

EWC has potential as a "vaccine" for predictive DSMs, that is, networks may be insulated from CI without having to implement new architectures. There is reason to suspect that EWC may have limited effectiveness when translated to the field of semantic modeling. EWC calculates the relevance of each model parameter to Task A based on the actual class of the training data. However, in the case of semantic modeling, we are not necessarily interested in the final predicted class of the training data but in the internal representations created by models as they learn. It is one goal of this paper to determine how EWC affects the internal representations of predictive DSMs.

The second candidate solution to CI that we evaluate is a different architecture: random vector accumulation (RVA; Jones et al., 2015). RVA is an alternate architecture that should theoretically be "immune" to CI by nature of the learning mechanism. RVA is the theoretical mechanism that is core to semantic models such as BEAGLE (Jones & Mewhort, 2007). Unlike predictive DSMs, which are affected by CI due to the error signal produced during learning, RVA models should be immune to CI because they utilize principles of associative learning and do not rely on an error signal to learn. These models learn via a simple Hebbian co-occurrence learning mechanism. The most basic RVAs first begin by initializing two random vectors from an arbitrary distribution and of arbitrary dimensionality for each word encountered in a corpus. One vector is unique to each word in the vocabulary, the environment vector, and the other is a summation of all context words, the memory vector. The update function for the memory vector of each word in the vocabulary is described in Equation 2:

$$m_i = e_{i-1} + e_{i+1} \qquad (2)$$

where m_i is the memory vector for an arbitrary word in a corpus, e_{i-1} is the unique environment vector for the context word before *i,* and e_{i+1} is the unique environment vector for the context vector after *i.* So, the memory vector for word *i* stores the context vectors for every other word that appears in context with word *i.*

Similar to Dachapally and Jones' (2018) study, this paper will use homonyms to measure the bias in semantic space created by CI. For each model, EWC and RVA, two conditions will be tested and compared to the performance of the original word2vec model in both an artificial and natural language. In the first condition, a target homonym will have two equally frequent senses with distinct meanings. Ideally, the target homonym should be equidistant from both of its two senses in semantic space. In the second condition, a target homonym will have two senses, one which is dominant (occurs more frequently) and one which is subordinate. In this case, the target homonym should be closer in semantic space to the dominant sense. Dachapally and Jones (2018) found that in both the artificial and natural language when word2vec was trained sequentially on equally balanced word senses, the target word was closer in semantic space to whichever sense had been trained most recently—forgetting the first sense of the word. The same effect was found when a target homonym had a dominant and subordinate sense; CI caused the target word to be more similar to the subordinate sense if the subordinate sense was trained most recently. Importantly, recency overpowered frequency, and the subordinate sense of the word became dominant if it was the most recently learned. To determine the effects of CI on a neural network equipped with EWC and on RVAs, a similar experimental structure will be used.

Experiment 1: Effects of CI on EWC and RVAs in an Artificial Language

Dachapally and Jones (2018) used a simple artificial language in which there is a single homonym, *bass,* that has two distinct meanings—*bass*[fish] and *bass*[guitar]. A corpus was created from this simple language by sampling word pairs from the following Markov grammar:

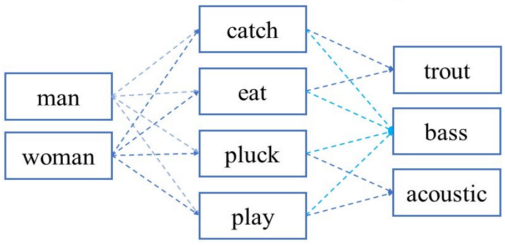

Figure 1. The artificial language used to test. Bass is the target homonym, and its position in semantic space relative to the two sense-pure synonyms (acoustic/bass) is evaluated.

In the first condition, a corpus of 8,000 sentences was generated from this grammar ("man catch bass", "woman play bass"). Each sense of the word *bass* was equally

frequent; the fish-sense made up half of the total sentences and the guitar-sense made up the other half. To measure the similarity of *bass* to its fish-sense and *bass* to its guitar-sense, the cosine similarity between the vector representation for *bass* and its two sense-pure synonyms *trout* and *acoustic* were calculated, respectively.

In the second condition, a corpus of 5,332 sentences was generated from the grammar—one sense of *bass* was dominant and the other subordinate. The dominant sense made up 4,000 of the total sentences and the subordinate sense made up 1,332. Thus, the subordinate sense was 1/3 as frequent as the dominant sense. Similar to the first condition, to determine the bias created in semantic space by CI, the similarity of *bass* to the dominant sense and *bass* to the subordinate sense was measured using the cosine similarity of the vector representations produced by each model.

The word2vec models used in this paper are both implemented using TensorFlow. Additionally, it is important to note that the implementation of the word2vec model in this experiment is different than both Mikolov et al.'s (2013) model and the model that was originally used in Dachapally & Jones' (2018) experiment. The full word2vec model as implemented by Mikolov et al. necessarily includes negative sampling and subsampling of the training data. Negative sampling is the practice of including negative information in the training data and subsampling is a method that results in less frequent words being sampled more often than frequent words. The model used by Dachapally & Jones used a different loss function called noise contrastive estimation which is common in the language modeling community because it is able to handle large input sizes. The model used in this experiment was purposely changed in order to be the most similar to the models previously used to implement EWC. This model uses cross entropy loss and does not use negative sampling or subsampling which may be responsible for the differences seen in the results of this paper.

Results

Figure 2 shows the cosine similarity of the vectors produced by word2vec, EWC, and the RVA in the case where sense 1 and sense 2 of *bass* are equally frequent. The pattern produced by word2vec is consistent with the findings in Dachapally and Jones' (2018) original experiment. When the model was trained in random order, the *bass-sense1* and *bass-sense2* similarities produced were approximately equal. When trained in sequential order, the sense which was sampled most recently ended up having a higher similarity to *bass*. The same procedure was repeated using EWC and the RVA. After exploring various parameter settings of both, we found that implementing EWC had virtually no effect on the results of the first experiment and that vector similarities produced by the RVA model were unaffected by CI.

Figure 3 shows the cosine similarity of the vectors produced by word2vec, EWC, and the RVA in the case where one sense is dominant and the other is subordinate. The pattern produced by word2vec is once again consistent with Dachapally and Jones' (2018) findings. When trained in random order, the dominant sense is more similar to *bass* than the subordinate sense. When trained in sequential order, the effects of CI reverse the frequency effects; when the subordinate sense of *bass* is trained last it becomes more similar to bass than the dominant sense. When the same procedure was performed using EWC and the RVA, we saw similar results to the first condition. The addition of EWC did not change the performance of word2vec and the RVA model was once again unaffected by CI.

EWC adds one additional parameter to the word2vec model, λ, which controls the importance of Task A compared

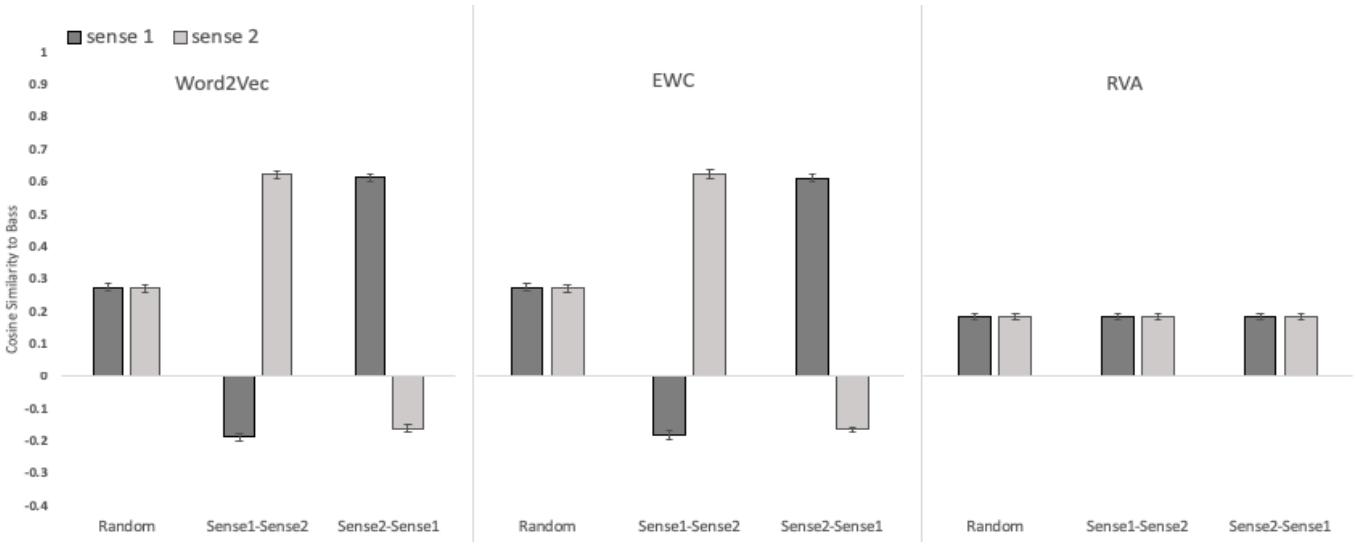

Figure 2. The y-axis represents the cosine similarity of vectors produced by word2vec, EWC, and the RVA. The x-axis represents one of three training orders: random order, sequential order with sense 1 first then sense2, and sequential order with sense 2 first then sense 1. Sense 1 and sense 2 are equally frequent in this case. CI is present in both word2vec and EWC while the RVA is unaffected by CI.

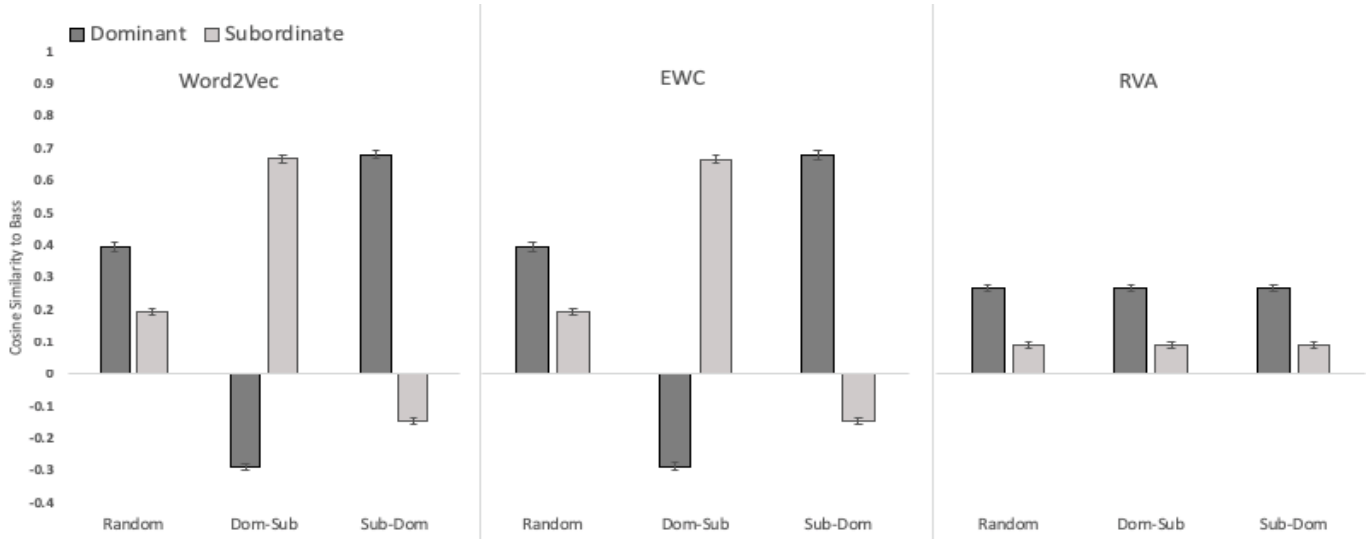

Figure 3. The y-axis represents the cosine similarity of vectors produced by word2vec, EWC, and the RVA. The x-axis represents one of three training orders: random order, sequential order with the dominant first then the subordinate sense, and sequential order with the subordinate sense first then the dominant sense. CI is present in both word2vec and EWC while the RVA is unaffected by CI.

to Task B. For both conditions in the first experiment, the value of λ had little to no effect on the final semantic representations. The results shown in both Figure 2 and Figure 3 are representative of the results obtained by any value of λ.

Experiment 2: Effects of CI on EWC and RVA in Natural Language

The texts used in this experiment are sourced from the TASA corpus (Landauer & Dumais, 1997). TASA contains language from textbooks with metadata tags which allowed us to train the models on distinct senses of a homonym without overlap. The same set of homonyms used in Dachapally and Jones (2018) was used for this experiment. They identified a sample of 14 homonyms that exist in the TASA corpus using the homonym norms from Armstrong, Tokowicz, and Plaut (2012) which determined homonyms with distinct meanings as rated by human participants.

The 14 homonyms were divided into two groups: sense-balanced and sense-imbalanced. We classified the two senses of a homonym as sense-imbalanced if one sense was at least twice as frequent in the TASA corpus, otherwise the two senses of a homonym were classified as sense-balanced. An example of a sense-imbalanced homonym is the word *slip*—the "fallen out of place" sense occurred across science contexts an equal number of times as the "shopping receipt" sense occurred across business contexts. An example of a sense-imbalanced homonym is the word *gum*—the "chewing candy" sense occurs approximately 5 times as often in language arts contexts than the "tissue surrounding teeth" sense occurs in health contexts. The sense-balanced homonyms are the counterpart to the first condition in the first experiment where the two senses of bass are equally frequent. The sense-imbalanced homonyms are the counterpart to the second condition in the first experiment where one sense of the word *bass* was dominant over the other. We then trained the word2vec model, the EWC model, and the RVA model on the entire corpus under three different order conditions. The first condition randomized the training order, the second condition was *sense1* first then *sense2* order, and the third condition was *sense2* first then *sense1* order. Cosine similarities between the target word vector and the two sense vectors were then calculated for each homonym set.

Results

The most common version of word2vec used for non-trivial training data is the model implemented within the Gensim Python library (Rehurek & Sojka, 2010). This model is optimized using C and is consequently very fast and effective. This is the model used by Dachapally & Jones in their second experiment to test for CI in natural language corpora. That model, however, is not directly compatible with the EWC implementation from our first experiment. For this reason, we did not use the Gensim model. Instead, we implemented a model in TensorFlow which is more similar to the model used in our first experiment and is compatible with EWC. However, there are some additional differences between the base models in our first experiment and the current experiment. In order to scale up to natural language, we had to include negative sampling and change the loss function to noise contrastive estimation. The model used in the previous experiment did not use negative sampling, but the model was unable to learn well from the natural language otherwise, so it was added. Additionally, while our first implementation of word2vec used a SoftMax layer to learn with a cross entropy loss function, the implementation in this experiment used noise contrastive estimation because the SoftMax method simply does not scale up well. The RVA model used in this experiment is the same model we used in the first experiment.

Figure 4 shows the results of training word2vec, EWC, and the RVA on the sense-balanced homonyms from TASA. The pattern of cosine similarities produced by word2vec and EWC are consistent with the results from the artificial language. When trained in random order the target words have approximately equal similarities to both of its senses. When trained sequentially, we see the same issue that occurred in the first experiment—the sense that was trained last becomes more similar to the target word. The RVA model shows the same pattern exhibited in Experiment 1—the similarity between the target and its two senses remain consistent no matter the training order.

Figure 5 shows the results of training word2vec, EWC, and the RVA on the sense-imbalanced homonyms from TASA. The cosine similarities produced from word2vec and EWC are consistent again with the results from the artificial language. Similarly, the cosine similarities produced by the RVA are consistent with the results from the artificial language and do not appear to be dependent on training order.

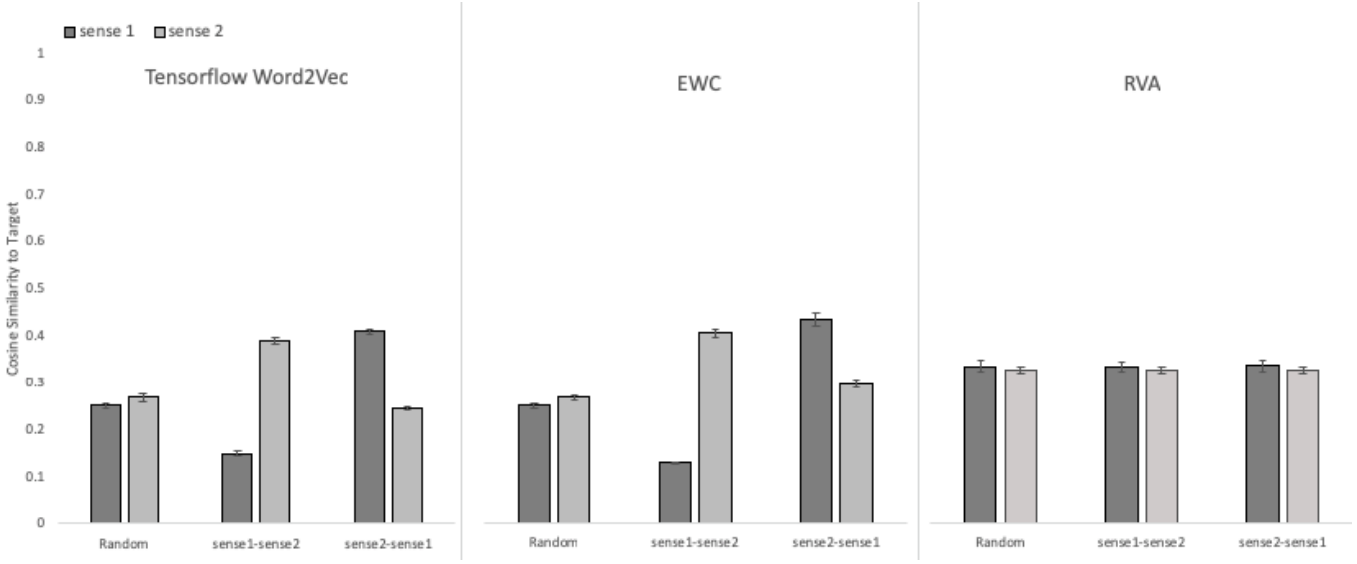

Figure 4. The y-axis represents the cosine similarity of vectors produced by word2vec, EWC, and the RVA when trained on sense-balanced homonyms from the TASA corpus. The x-axis represents one of three training orders: random order, sequential order with sense 1 first then sense2, and sequential order with sense 2 first then sense 1. CI is present in word2vec and EWC while the RVA is unaffected by CI.

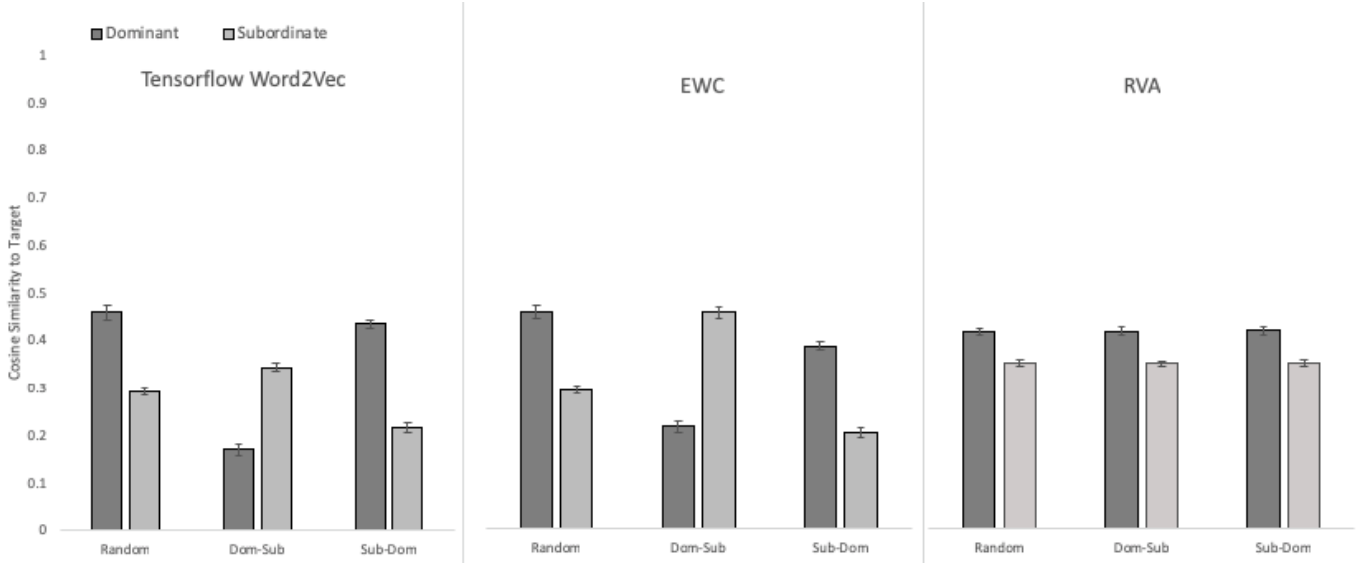

Figure 5. The y-axis represents the cosine similarity of vectors produced by word2vec and the RVA when trained on sense-imbalanced homonyms from the TASA corpus. The x-axis represents one of three training orders: random order, sequential order with the dominant first then the subordinate sense, and sequential order with the subordinate sense first then the dominant sense. CI is present in word2vec and EWC while the RVA is unaffected by CI.

Discussion

The results of this study suggest that efforts to mitigate the effects of CI need to be interdisciplinary. Within the machine learning community, insulating, or "fixing", predictive DSMs from CI is an emerging area that has seen some innovation in recent years. However, suggested solutions so far only consider the problem as it relates to strictly machine learning tasks such as categorization or image classification tasks. This study has shown that solutions from the machine learning community are not guaranteed to work when applied to tasks from different fields.

While Kirkpatrick et al. (2017) were able to show promising results from EWC on categorization tasks, the method was unable to prevent CI when applied to semantic modeling. This may be because the goal of EWC is to prevent the weights of a predictive neural network from changing based on how much information each weight carries about the *true class* of each training item. The connection between training items and their class is very straightforward in categorization tasks but is not as clear in semantic modeling tasks. When a predictive neural network learns a word representation, it is not explicitly predicting the class of a word but is attempting to predict which words belong or don't belong in context with a target word. Additionally, the window size is a variable parameter in these models which can be greater than 2, implying that a target word could have multiple "true classes" if we consider context words the class of the target word.

Additionally, EWC as it is now is not theoretically plausible for any task which requires unsupervised learning because the new loss function must be "turned on" when the network is learning a second task. This is especially cumbersome in NLP where it is impossible to supervise learning to the extent which EWC requires. Furthermore, EWC is unable to scale up well with its current implementation. Because it was designed to prevent CI in categorization tasks, it requires each training item to have a true class. This requirement prevents more efficient sampling methods which have been standardized in the DSM literature, such as noise contrastive estimation, from being used in conjunction with EWC. Similarly, calculating the Fisher Information for each node in a network becomes computationally expensive when the vocabulary and network gets large.

Introducing the RVA model as a possible solution to CI is a preliminary attempt to approach the problem of CI from the perspective of cognitive science. Within the cognitive science community, many researchers assume that the brain is primarily a predictive learner, when in reality it learns using both prediction and co-occurrence methods. Because of the tendency to favor predictive explanations of learning, predictive DSMs are still the most popular learning models in the field even though the existence of CI implies biologically implausibility. This has been documented by Ratcliff (1990) and McCloskey & Cohen (1989) who both use CI to discredit the biological plausibility of predictive DSMs. While RVAs are not a brand-new idea, they have not become as popular within the machine learning or cognitive science communities as predictive DSMs. However, they are continuous learners, can learn sequentially without incurring CI, and are computationally efficient making them a viable alternative to predictive DSMs in both the fields of cognitive science and machine learning.

While RVAs are promising, they have faced some criticism in the past. RVAs are known to have problems with metric space compression—causing most word similarities to be compressed between 0 and 1—which limits the ability of the model to discriminate between related and unrelated words (Asr & Jones, 2017). It was initially believed that predictive DSMs were able to more accurately discriminate between words because of back-propagation or the connectionist architectures they commonly use. However, recently the role of negative sampling in DSMs has been explored in more depth by Johns, Jones, & Mewhort (2019) who find that the success predictive DSMs have at discriminating between words is due to the inclusion of negative information in the training data—not the use of connectionist architecture or predictive learning method. In fact, when negative sampling information is included in the training data for other DSMs, including RVAs, their ability to discriminate words is on par with predictive DSMs.

Though this paper focused on comparing RVAs to predictive DSMs, RVAs aren't the only possible alternative architecture that could present a solution to CI. Architectures like holographic neural networks and exemplar-based models should also theoretically be immune to CI and incorporate different theoretical frameworks of learning. Holographic neural networks use convolution as an association mechanism to learn words rather than backpropagation and are able to learn complex non-linear patterns with a single layer which makes them more space efficient than predictive DSMs. Exemplar-based models, unlike other DSM models which store a semantic representation, store only episodic context. These models construct semantic meaning from the aggregation of episodic context when presented with a memory cue (Jamieson et al., 2018). Both of these models should be evaluated to determine the effect CI has on their internal semantic representations.

Up until now, the fields of machine learning and cognitive science have both been facing similar problems with predictive DSMs. Unfortunately, there has been little to no interdisciplinary communication to propose solutions. When we consider CI from a cognitive science perspective, we find that there are several possible solutions which haven't been considered yet. These solutions, which are arguably more elegant than continuously trying to "vaccinate" predictive DSMs, have the potential to introduce new mechanisms for artificial learning, assisting with new technological advances that require sequential learning and providing a framework for learning that does not exhibit the downfalls brought on by predictive DSMs.

References

Armstrong, B. C., Tokowicz, N., & Plaut, D. C. (2012). eDom: Norming software and relative meaning frequencies for 544 English homonyms. *Behavior research methods*, *44*(4), 1015-1027.

Asr, F. T., & Jones, M. N. (2017). An Artificial Language Evaluation of Distributional Semantic Models. *Proc. of the ACL Conference on Natural Language Learning*.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings Association of Computational Linguistics* (Vol. 1, pp. 238-247).

Dachapally, P. R. & Jones, M. N. (2018) Catastrophic Interference in Neural Embedding Models, *CogSci 2018*, 1566-1571.

Firth, J. R. (1957). *A synopsis of linguistic theory* (pp. 1930–1955). Oxford.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, *3*(4), 128-135.

Harris, Z. (1970). Distributional structure. In Papers in structural and transformational Linguistics (pp. 775–794).

Jamieson, R. K., Johns, B. T., Avery, J. E., & Jones, M. N. (2018). An instance theory of semantic memory. *Computational Brain & Behavior, 1(2),* 119-136.

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2019). The role of negative information in distributional semantic learning. *Cognitive Science.*

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114,* 1-37.

Jones, M. N., Willits, J. A., & Dennis, S. (2015). Models of semantic memory. In J. R. Busemeyer & J. T. Townsend (Eds.) *Oxford Handbook of Mathematical and Computational Psychology*, 232-254.

Kirkpatrick, J., et al. (2017) Overcoming Catastrophic Forgetting in Neural Networks. *Proceedings of the National Academy of Sciences*, *114(13),* 3521–3526.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211–240.

LeCun, Y., Cortes, C., & Burgess, C. J. C. (2012). The mnist database of handwritten images.

McCloskey, M. & Cohen, N. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (ed.) *The Psychology of Learning and Motivation*, *24*, 109-164.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, *97*, 285-308.

Rehurek, R. & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks,* 45-50.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, *2*, 64-99.