Catastrophic Interference in Neural Embedding Models

Prudhvi Raj Dachapally and Michael N. Jones
Cognitive Computing Laboratory
Indiana University, Bloomington
[prudacha] [jonesmn]@indiana.edu

Abstract

The semantic memory literature has recently seen the emergence of predictive neural network models that use principles of reinforcement learning to create a "neural embedding" of word meaning when trained on a language corpus. These models have taken the field by storm, partially due to the resurgence of connectionist architectures, but also due to their remarkable success at fitting human data. However, predictive embedding models also inherit the weaknesses of their ancestors. In this paper, we explore the effect of catastrophic interference (CI), long known to be a flaw with neural network models, on a modern neural embedding model of semantic representation (*word2vec*). We use homonyms as an index of bias depending on the order in which a corpus is learned. If the corpus is learned in random order, the final representation will tend towards the dominant sense of the word (*bank* → *money*) as opposed to the subordinate sense (*bank* → *river*). However, if the subordinate sense is presented to the network after learning the dominant sense, CI produces profound forgetting of the dominant sense and the final representation strongly tends towards the more recent subordinate sense. We demonstrate the impact of CI and sequence of learning on the final neural embeddings learned by word2vec in both an artificial language and in an English corpus. Embedding models show a strong CI bias that is not shared by their algebraic cousins.

Keywords: semantic models; word2vec; neural networks; catastrophic interference; statistical learning

Introduction

Distributional models of semantic memory (DSMs; e.g., Landauer & Dumais, 1997) have been hugely successful in cognitive science, explaining how humans transform first-order statistical experience with language into deep representations of word meaning. These models are all based on the distributional hypothesis from linguistics (Harris, 1970) and specify mechanisms to formalize the classic notion that "you shall know a word by the company it keeps" (Firth, 1957). There are dozens of DSMs in the cognitive literature now, with learning mechanisms inspired by different theoretical camps ranging from Hebbian learning to probabilistic inference (see Jones, Willits, & Dennis, 2015 for a review). The commonality to all these models is that they use co-occurrence counts of words across contexts in linguistic corpora, and exploit these statistical redundancies to construct semantic representations.

There has been a recent resurgence of neural network models across cognitive science and machine learning. This resurgence has included very successful predictive DSMs within connectionist architectures based on principles of error-driven learning core to theories of reinforcement learning. Earlier work had explored neural networks to learn distributed semantic representations from artificial languages using recurrent networks (e.g., Elman, 1990) and feed-forward networks (Hinton, 1986). But neural networks were essentially non-existent in the 1990s and early 2000s as models that scale up and learn from natural language corpora. Rather, the field became fixated on algebraic models based on dimensional reduction mechanisms such as singular value decomposition applied to a matrix of word counts across contexts in a corpus (e.g., classic Latent Semantic Analysis; Landauer & Dumais, 1997).

However, the recent literature has hugely reversed this trend, with the emergence of *predictive neural network models*, and the observation that they tend to outperform classic models on most tasks of semantic similarity that are commonly used in natural language processing. Rather than counting words in contexts, these models predict the word given the context, and backpropagate the error-signal through hidden layers in a neural network. Although a rather simple architecture, and essentially the same one studied by Rogers and McClelland (2004) in their seminal work, these predictive models have rapidly risen to the top of the DSM battle in their ability to account for human data across a range of tasks.

The standard predictive network currently discussed in the literature is Mikolov et al.'s (2013) *word2vec* model[1]. Word2vec is a feedforward neural network with localist input and output layers that contain one node per word in the vocabulary, and a hidden layer of ~300 nodes that is fully connected to both input and output layers. When a linguistic context is sampled (e.g., "save money bank") the target node is activated at the input layer (+bank) and activation is forward propagated to the output layer, with the desired output being the observed context words (+save, +money). The error signal (observed output – desired output) is applied to the network with backpropagation to correct the weights and make it more likely the next time this target word is encountered that the correct output pattern will be generated. Although prediction is used to train the network, it is the final pattern of weights across the input-to-hidden layer that are exported and used as deep semantic representations of word meaning. Two words that

[1] The word2vec architecture has two possible model directions: The context may be used to predict the word (referred to as a CBOW), or the word may be used to predict the context (a skipgram). Although the theoretical claims here apply broadly to neural network models, and hence, both directions in word2vec, we will use skipgram as our demonstration example because it maps conceptually onto most connectionist models.

have similar vector patterns across these weights are predicted by similar contexts, even if they never co-occur with each other, akin to (but superior in data fit) the second-order inference vectors learned by traditional algebraic DSMs.

Word2vec has received considerable attention in the machine learning literature due to its ability to outperform all previous models (Baroni et al., 2014). To cognitive science, this success is of considerable interest as word2vec implements a potentially biologically plausible neural architecture and links to classic theories of reinforcement learning (e.g., Roscorla & Wagner, 1972), drawing theoretical connections to other areas of cognition with a unified mechanism. One feature of particular interest in neural embedding models is that they are incremental learners, in contrast to earlier algebraic DSMs which were largely batch learners. The incremental learning of neural embedding models has been taken by some as additional evidence in favor of cognitive plausibility of the models.

While the hype surrounding neural embedding DSMs is certainly warranted given their recent success at fitting human data, it is important to remember that the models also inherit the weaknesses of their predecessor neural networks. One weakness in particular that models such as word2vec are likely to have is *catastrophic interference (CI):* The tendency of neural networks to loose previously learned associations when encoding new ones. In this sense, the positive attribute of neural embedding DSMs being incremental learners is also what opens them up to a potentially serious flaw that does not exist in their batch learning algebraic counterparts. The goal of this paper is a first attempt to document the extent to which CI is affecting the semantic representations learned by word2vec in particular, although the problem of CI will apply uniformly across all neural embedding DSMs that use backpropagation as a learning mechanism.

Stability-Plasticity Dilemma in Neural Networks

The *stability-plasticity dilemma* (Grossberg, 1982; Hasselmo, 2017) refers to the problem of any learning system to learn new stimuli while preventing the new learning from distorting existing learning. We need to balance memory for individual exemplars with abstraction, recency with primacy, and it is optimal to preferentially strengthen memories that are more likely to be needed in the future. While all cognitive systems gradually forget information, biological organisms exhibit gradual forgetting of old information as new information is acquired. In contrast, artificial neural networks have long been known to forget catastrophically. Catastrophic interference (CI) is thus defined as the sudden and complete loss of previously learned associations when learning new associations (see French, 1999 for a review). CI is a consequence of using backpropagation as a learning mechanism to reuse neural connections to tune learning, and is a key flaw to all feedforward neural embedding architectures that are currently used as DSMs.

McClosky and Cohen's (1989) seminal work trained a standard multilayer network to learn single-digit "ones" arithmetic facts (e.g., 1 + 1, 9 + 1) using backpropagation until the network had perfectly learned the associations. They next trained the same network on a new set of single-digit "twos" facts (e.g., 2 + 1), until the network had been trained to respond correctly to all of them. While the network was able to correctly answer the twos facts, it had completely lost the previously learned ones facts—the associations that had been trained to zero error were now lost completely. The learning of new associations with backpropagation overwrote the previous learning. The CI pattern was duplicated in a second experiment by McClosky and Cohen by simulating standard tasks of paired associate word learning. Further, Ratcliff (1990) demonstrated that in standard sequential learning paradigms, backpropagation networks catastrophically forget previous items as new items are learned, unlike humans performing the same tasks.

Using Homonyms to Measure Representational Bias in Semantic Space

The standard architecture used by neural embedding models such as word2vec is susceptible to CI, but it is unclear if, or to what extent, CI would affect the final semantic representation. Word2vec uses predictive association for learning (e.g., bank → save + money) and backpropagation for error correction. But the final representations of word meanings are the contained in the vector of association weights across the input-to-hidden layer. It is reasonable to expect that if a word predicts very different contexts, such as homonyms (bank → save + money; bank → sand + river) that the final semantic representation for a word will be heavily dependent on the most recently learned sense association, potentially producing a great loss of the previously learned sense association. Hence, homonyms act similarly to classic XOR stimuli in experimental studies: The same output pattern is predicted by two (or more) orthogonal input patterns.

The goal of this paper is to explore the impact of the training sequence of context/target pairs on the final representational space in the word2vec skipgram architecture. Hence, we use homonyms as an index of movement in semantic space. For a homonym with two equally frequent distinct meaning senses, the semantic representation of the target word in word2vec should be equidistant between the two opposing meanings in semantic space. If the homonym has a dominant and subordinate sense, then the final meaning will tend towards the more frequent dominant sense in semantic space. However, if contexts containing the subordinate meaning are the most recently learned, then CI may produce a semantic representation that will erroneously tend towards recency over the more frequent meaning. Hence, we can use homonyms as an elegant measure of how a word's meaning differs from a randomly sampled corpus when we make certain sense contexts more or less recently presented to the backpropagation algorithm.

Experiment 1 uses a simple engineered language presented to word2vec. The corpus learned is the same in all conditions, but the ordering of the contexts is varied (cf. Ratcliff, 1990). Experiment 2 scales the principles up to a natural language corpus where contexts containing the dominant and subordinate senses of the target word are presented to word2vec, either in random order, or in sequentially manipulated orders.

Experiment 1: CI in an Artificial Language

As an initial evaluation of CI in word2vec, we created a simple artificial language inspired by Elman (1990). In the language, there is a single homonym, *bass*, with two distinct senses—*bass* [fish], or *bass* [guitar]. There are two actors (man/woman) who may occur in either sense context. Hence, a corpus containing the simple language was created by randomly sampling from:

```
Man/woman catch/eat trout/bass
Man/woman play/pluck acoustic/bass
```

We generated a corpus of 8,000 sentences from this grammar (e.g., "man catch bass," "woman play bass," …). The corpus was generated to ensure that *bass* occurs an equal number of times in the fish and guitar contexts, and a neural embedding model will form a semantic representation for *bass* that is an average of its two distinct sense contexts. To measure the semantic position of *bass* relative to its two distinct senses, we compute the cosine to its two sense-pure synonyms, *trout* and *acoustic*. When the corpus is sampled randomly, *bass* has an equal similarity to both *trout* and *acoustic.* However when the order of senses in the corpus is not random but favors a more recent sense, the position of *bass* is expected to tend towards the sense-pure synonym of that sense.

The word2vec skipgram architecture (Mikolov, et al., 2013) was trained on the corpus of 8,000 sentences sampled from the artificial language in three distinct orderings: random, fish context first, and guitar context first. The *exact* same set of 8,000 sentences was presented to the model in each condition, but the sequencing was varied (i.e., LSA would produce the exact same representation for all three). In the sequenced versions, ordering was randomized but with the additional constraint that one sense occurred in the first half of the corpus, and the other sense was restricted to the second half of the corpus.

Because word2vec has stochastic weight initialization, we ran the model for 200 replications on each ordering of the corpus. Each training run, the model was presented with the full corpus (8,000 training epochs) and we recorded the vector cosine of *bass* to *trout* and *acoustic*.

The skipgram model was trained by minimizing objective function predicting a target word given its surrounding context, defined as the average log probability:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j\neq 0}\log p\left(w_{t+j}|w_t\right), \qquad (1)$$

where $\log p\left(w_{t+j}|w_t\right)$ is implemented with a softmax:

$$p(w_o|w_i) = \frac{\exp\left(v'_{w_o} v_i\right)}{\sum_{w=1}^{W}\exp\left(v'_{w_o} v_i\right)} \qquad (2)$$

Results

Figure 1 shows the vector cosine of *bass* to *trout* and *acoustic* in the final semantic space for word2vec across 200 replications, while varying ordering. When the model is trained on a random ordering of the corpus, *bass* is equidistant to *trout* and *acoustic.* However, when all the fish contexts are presented first, followed by the guitar contexts, the resulting semantic spaces showed a very strong preference for the more recent sense.

For example, if the contexts representing the fish sense of *bass* were presented first, the model learned that *bass* was synonymous with *trout* until the second half of the corpus when the new sense was introduced. With the introduction of a guitar sense of *bass*, word2vec then had to reuse the connection weights to learn the new context association. As a result, *bass* became synonymous with *acoustic*, and the previous prediction of fish contexts was almost completely erased.

The final similarity of *bass* after learning the corpus in the order of fish-then-guitar was maximal to the sense-pure synonym (*acoustic*), and was reduced to nearly zero for the originally learned sense-pure synonym (*trout*). In contrast, the reverse preference is seen when the ordering of senses is reversed. Note that a batch model such as LSA would produce the behavior seen in the random case across all three orderings. Very similarly to the classic McClosky and Cohen (1989) experiments, word2vec is largely overwriting the learning of the first context with training on the second context, and this pattern is reflected in the final semantic representation of the homonym.

We next varied the sense dominance of the homonym *bass*, to present a dominant and subordinate sense frequency in the artificial corpus. Rather than sampling randomly from the two senses, the subordinate sense of the homonym was sampled to be one-third as likely as the dominant sense. Hence, one sense is two-thirds more prevalent in the experienced sentences, and we test word2vec's sense preference as a function of training sequence.

The results are presented in Figure 2. In the two randomly presented conditions (two left clusters of bars), word2vec prefers the dominant sense of the word over the subordinate sense— the representation of *bass* was more similar (closer) to the dominant sense from the corpus. But a dramatic reversal can be seen in the ordered conditions (two right clusters of bars). In both of these cases, the dominant sense is presented first, followed by one-third as many sentences from the subordinate sense second. In both cases, CI shows its effect clearly in the final representations: there was a strong bias towards the more recently learned sense in the semantic space, despite the fact that it was less frequent than the dominant sense in the corpus. Even if the fish sense of *bass* was the dominant sense, recent presentations of the subordinate guitar sense almost completely overwrite that learning. Again, note that a batch model such as LSA would simply show a preference for the dominant sense across all of these presentation orders, but word2vec is influenced by CI to produce a recency bias that overrides the standard frequency bias.

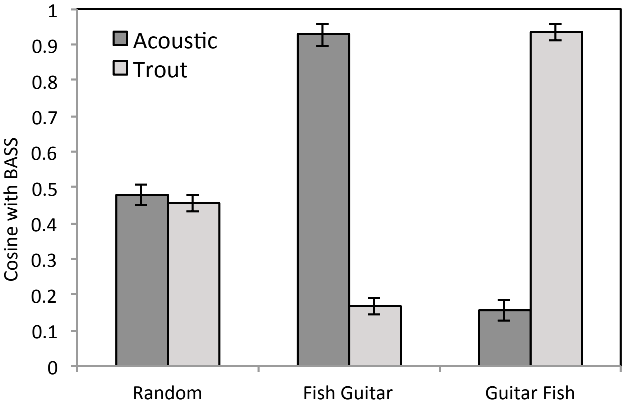

Figure 1. *Cosine similarity of the homonym bass to the fish or guitar sense as a function of sense presentation order. When the corpus is sampled randomly, bass is equally pulled between the two senses. However when one sense is presented earlier in the corpus, followed by the other sense, bass is strongly biased towards the more recently presented sense.*

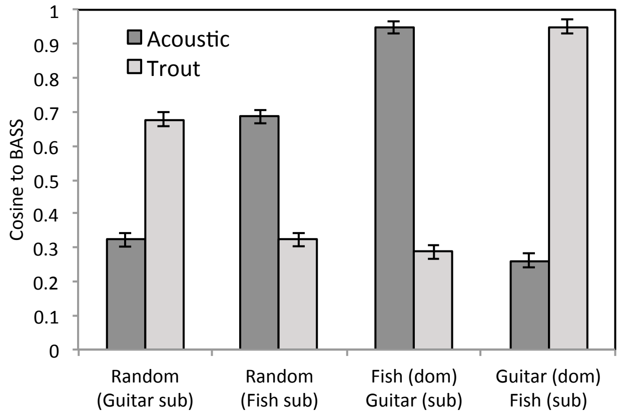

Figure 2. *Cosine similarity of the homonym bass to the fish or guitar sense as a function of order and varied sense dominance. When sampled randomly, bass tends towards the more dominant sense. However, if the subordinate sense is presented more recently, it outweighs the dominant sense, losing the association of frequency to recency.*

Experiment 2: CI in a Natural Corpus

We next test the influence of CI in the representational space learned by word2vec when trained on natural language from the classic TASA corpus (Landauer & Dumais, 1997). TASA is ideal because it contains linguistic contexts from textbooks with metadata that tags the topic of the material (e.g., Science, Language Arts, Health, etc.), which allowed us to select and track a sample of homonyms with distinct senses.

We used the homonym norms from Armstrong, Tokowicz, and Plaut (2012) to identify a sample of 14 homonyms that also exist in TASA with very distinct meaning senses as rated by human subjects. Of the sample of homonyms, half have roughly equal frequencies in TASA between the two senses, and half have a clear and dominant sense (as per frequency in TASA). An example of a sense balanced homonym is *net*—the "fabric that encloses the sides and back of the goal in various games (such as soccer or hockey)" sense occurs in an equal number of language arts contexts across TASA as the occurrence of the " a net amount, profit, weight, or price" sense in business contexts. An example of a sense-balanced homonym is *pupil*—the "part of the iris of the eye" sense occurs in an equal number of science contexts across TASA as the occurrence of the "young learner in school" sense in social studies contexts. An example of a sense-imbalanced homonym is *firm*—the "business facility" sense occurred in business contexts across TASA eight times more often than did the "having a solid structure" sense in science contexts.

Method

We applied a simple heuristic to classify whether a word had a roughly even balance between its senses, or if it had a bias towards one sense over the other based on the word's contextual uses in TASA. If the occurrence of the homonym in one sense exceeded the other sense by a factor of two, we classified it as sense biased, and classified it as sense balanced otherwise. Five of the homonyms were classified as balanced (*hamper, capital, net, slip, plane*), and the remaining nine were classified as biased (*firm, compact, hull, compound, pitch, cap, gum, bull, pupil*).

The distance of the homonym in semantic space must be measured with respect to some sense-pure synonym, analogous to the artificial grammar simulations that measured *bass* in respect to *trout* or *acoustic*. To do this, we first trained word2vec on the single sense of each homonym, e.g., training the model separately on the "student" and "eye" contexts of *pupil*. For each separate sense space, we determined the target word's most similar semantic associate. Due to the stochastic noise added by the weight initialization process, we identified each target word's closest 10 neighbors across 20 replications for sense 1 (e.g., *pupil* → *student, teacher, learn, classroom, books,* etc.) and sense 2 (e.g., *pupil* → *iris, cornea, eye, vision, etc.)*, and selected the single associate the was most often the highest ranked across the resamples. This word will be considered the target word's sense-pure semantic associate as was *trout* and *acoustic* in the artificial language, and will be used as an anchor point to measure the target word's representation as a function of corpus training order.

With the sense-pure semantic associate identified for each sense of each homonym, we then trained on the entire corpus, which mixed the sense contexts together. In the random condition, the contexts for the target word were simply sampled in random order, as in usual applications of word2vec. In the other two classes of simulations, we first trained on one sense of the target word first, and then the other sense.

Results

Figure 3 displays the results for the sense-balanced homonyms, analogous to Figure 1 from the artificial corpus. In the random training order, the target homonym is evenly pulled between its two senses. The homonym representation learned by the model has an equal similarity to each sense-pure associate. However, the next two sets of bars show the effect of CI that comes with sequential sense training. In both cases, the target homonym becomes more similar to the recently learned sense and less similar to the previously learned sense. Note that the forgetting is not *completely* catastrophic. However, this is a very profound effect on the final learned representation of the homonym considering that this is the exact same input corpus across all three training orders.

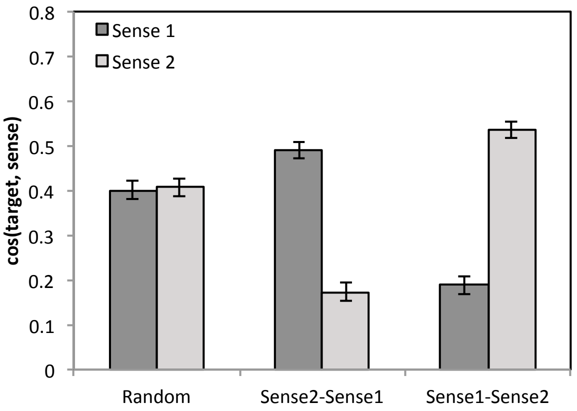

Figure 3. *Similarity of the homonym representation from TASA for the sense balanced cases, trained in random order or with sequenced senses. The CI effect is illustrated in the second two clusters of bars where the target homonym is more similar to the recently encountered sense and less similar to the previously encountered sense.*

Figure 4 plots the results for the sense-imbalanced homonyms from TASA. Each target homonym here has a bias in TASA towards one of the competing senses (e.g., *firm* is encountered in many more financial contexts in TASA than it is in science contexts). When trained in random order, the target homonym is more similar to the dominant sense in the corpus. The second two clusters of bars show the results for the sequenced training where one sense is learned before the other. If the dominant sense is also the more recently leaned sense (middle cluster of bars), then the target homonym becomes even more similar to the dominant sense and less similar to the subordinate sense. However, the effect of CI is seen prominently as a reversal in the far right cluster of bars. When the subordinate sense is learned after the dominant sense, the similarity of the target homonym favors recency over dominance. In this case, the target homonym actually becomes more similar to the recently presented subordinate sense than it is to the dominant sense. Again, the forgetting is not completely catastrophic. But the pattern is very powerful: CI produces an effect that makes recency overpower frequency. The model believes that the target homonym's meaning is more similar to the subordinate sense over the dominant sense, simply because it was encountered more recently.

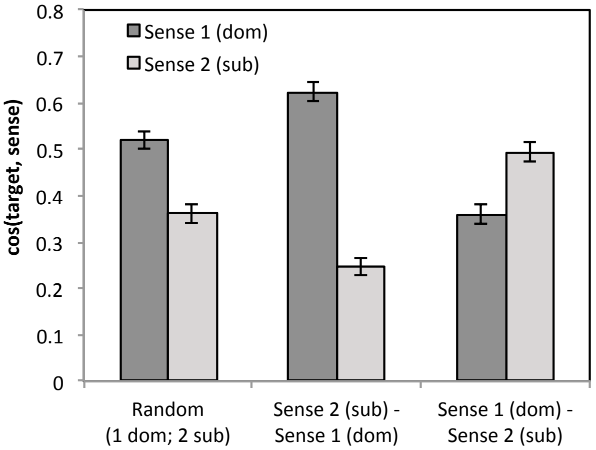

Figure 4. *Similarity of the homonym representation for the sense imbalanced cases. In random order, the target homonym is most similar to the more frequent sense (left bars). If the dominant sense is also more recent, the homonym moves closer to the dominant sense and further from the subordinate one (middle). However, CI produces an effect that reverses this trend (right bars): if the subordinate sense is more recently encountered, the homonym representation favors recency over frequency.*

Discussion

The resurgence of predictive neural network models has already led to reconceptualization of the mechanisms underlying cognitive phenomena. In addition, the utility of these models to machine learning demonstrates the importance of basic cognitive science to solving applied problems. However, it is crucial to remember our history: CI is a problem that was never fully addressed in the original connectionist models of the 90s. We have made a clear first demonstration here that CI can have a strong impact on the semantic representations learned by neural embedding models such as word2vec. This is particularly important because the potential source of error goes beyond our basic science and into the many applications of word2vec. McCloskey and Cohen (1989) and Ratcliff (1990) used CI to demonstrate an inherent flaw with neural networks as theoretical models of cognition. But beyond that, neural embedding models such as word2vec are being used in a massive number of NLP and knowledge mining applications[2], bringing the errors that come with CI along for the ride.

This does beg the question: How big of a problem is CI really likely to be in the regular application of neural embedding models? Firstly, CI is still very much a theoretical flaw with these architectures as theories of cognition, in that they show forgetting and interference patterns that are extremely unlike biological systems. But the version of word2vec that we ran here was essentially stripped down to a core predictive neural network. The full word2vec algorithm includes negative sampling, frequency subsampling, and other machine learning tricks to speed learning and scale up to large amounts of text. Frequency subsampling adjusts the sampling of contexts of words inversely proportionate to their normative frequency. The idea is to sample word tokens roughly an even amount, and so the model needs to account for the Zipfian law of word frequencies or it will be spending far too much time on already well-learned frequent words, and not enough on rarer words further out on the Zipfian tail.

However, the addition of frequency subsampling can potentially allow CI to do even more damage to the final semantic space than without it. The practice means that fewer samples are taken from high-frequency words, meaning that there is a greater likelihood that a single rare sense of the word lemma could be sampled most frequently, undoing previous learning for the more common sense of the word. Frequency subsampling assumes a single sense for a word, but does not take into account the fact that there is also a Zipfian distribution of senses within homonyms and polysemes, which constitute half the lexicon. Clearly, more research is needed to evaluate how much variance there is in word2vec's estimate of a word's position in semantic space across multiple replications, and in different sequential orders, of a training corpus.

With the renewed interest in neural networks, the field has begun to discuss architectures to insulate networks against CI. The most promising approaches take their organizations from the anatomical and functional organization of the brain. For example, complimentary systems theory (McClelland, McNaughton, & O'Reilly, 1995) posits that a slow neocortical processing system and faster hippocampal encoding system in the human brain allows deep processing of predictive information while avoiding problems that come with CI. Other architectures, such as holographic semantic models (Jones & Mewhort, 2007) have already been shown to have a near immunity to CI and are promising continuous learning candidates to pursue.

More recently, new algorithms that capitalize on Elastic Weight Consolidation (EWC) have shown great promise at learning new associations while insulating previously learned associations against forgetting (e.g., Kirkpatrick et al., 2017). EWC constrains the weight space of a deep learning network within the optimal parameter space of a previously learned task, essentially having the effect of a spring (or elastic) on already learned weights that are important to a previously learned task. Kirkpatrick et al. demonstrated that EWC networks could learn new strategies and associative patterns with minimal loss to previously learned but orthogonal associative patterns. However, EWC and its relatives have not yet been implemented in semantic neural embedding models.

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.

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).

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

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.

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review, 87*, 1-51.

Hasselmo, M. E. (2017). Avoiding catastrophic forgetting. *Trends in Cog Sci*, 21.

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

Hinton, G., E. (1989). Learning distributed representations of concepts. In. R. G. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology*, 46-61. Oxford: Clarendon Press.

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

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, et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proc Natl Acad. Sci.* U.S.A., 114, 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.

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.

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.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press.

[2] The original word2vec papers, published in the Proceedings of NIPS in 2013, have already been cited over 10,000 times.