Domain-General Learning of Neural Network Models to Solve Analogy Tasks – A Large-Scale Simulation

Arianna Yuan (xfyuan@stanford.edu)
Department of Psychology, Stanford University, Stanford, CA 94305 USA

Abstract

Several computational models have been proposed to explain the mental processes underlying analogical reasoning. However, previous models either lack a learning component or use limited, artificial data for simulations. To address these issues, we build a domain-general neural network model that learns to solve analogy tasks in different modalities, e.g., texts and images. Importantly, it uses word representations and image representations computed from large-scale naturalistic corpus. The model reproduces several key findings in the analogical reasoning literature, including relational shift and familiarity effect, and demonstrates domain-general learning capacity. Our model also makes interesting predictions on cross-modality transfer of analogical reasoning that could be empirically tested. Our model makes the first step towards a computational framework that is able to learn analogy tasks using naturalistic data and transfer to other modalities.

Keywords: analogical reasoning; learning; cross-modality transfer; neural network models.

Introduction

Analogy is arguably one of the most important mechanisms through which people acquire new knowledge (Gentner, Holyoak, & Kokinov, 2001). Verbal analogy task such as "PIG:BOAR::DOG:? a. WOLF b. CAT" and visual analogy task such as Raven's Progressive Matrices has been widely used in standardized test to assess students' intellectual ability (Buck et al., 1998). For a long time, there has been a heated debate over whether analogy-mapping is a domain-specific or domain-general process (Forbus, Gentner, Markman, & Ferguson, 1998). Nowadays, it has been increasingly clear that there are both a domain-general component and a domain-specific component in learning analogical reasoning.

One phenomenon related to the domain-general component is the relational shift in child development, i.e., early in development children tend to choose the item that is more associated to the third item in the analogy question A:B::C:[D1|D2] (Sternberg & Nigro, 1980), and only older kids are able to use relational matching rather than associations to perform the task. Some researchers have argued that the lack of inhibitory control, which requires a full-blown pre-frontal cortex is partially responsible for the associative response (Richland & Burchinal, 2013). The ability to inhibit associative responding and maintain the relational constraints imposed by A:B is domain-general, i.e., it is a universal prerequisite for analogy-making no matter which sensory modality or semantic domain the analogy task is built on. Hence, if one person is trained to perform analogy task in a particularly domain or modality, it is likely that the training will help them do better in other domain or modality, due to enhanced ability to suppress associative responses and to maintain contexts.

The evidence for a domain-specific component comes from the finding that children's ability to perform analogy task depends on their familiarity with the test material. Goswami and Brown (1990a, 1990b) found that when familiar concepts were used in analogy tasks, performances were much better. These findings underscore the contribution of domain-specific knowledge in analogy-making, yet change in this aspect is unlikely to boost task performance in other domain or modality.

Several computational models have been proposed to account for how people solve analogy tasks. A useful classification scheme is to group them into three types of models (French, 2002; Gentner & Forbus, 2010): symbolic models (Kuehne, Forbus, Gentner, & Quinn, 2000; Falkenhainer, Forbus, & Gentner, 1989), connectionist models (Holyoak & Thagard, 1989; Hummel & Holyoak, 1997; Kollias & McClelland, 2013) and hybrid models (Mitchell, 1993; Kokinov & Petrov, 2000). Most of symbolic models represent analogy questions using predicates and logical forms. When asked to solve an analogy task such as A:B::C:?, they use symbolic manipulations and search algorithms to find the correct answer. One of the most influential symbolic approach to analogy-mapping is the Structure Mapping Engine (SME) (Falkenhainer et al., 1989; Gentner, 1983). It represents the base and source using predicate-calculus and compares the two representations to see if there is any structural similarities between them. Once optimal matching structures are identified, the system then transfer structure in the source to the target. Later version of SME have relaxed matching criteria to allow similar, but not identical predicate to match (Brian, 1990), but whether two predicates are similar need to be explicitly computed.

Contrary to these symbolic approaches, connectionist models often use distributed representation for the items in an analogy task and encode their semantic and structural similarity in a more implicit and continuous way, e.g., Kollias and McClelland (2013). The connectionist models learn to make a correct response by adjusting connection weights so that the spreading activation within the neural networks could reveal the distributed representation of the target, thus leading to the correct answer.

Previous computational models of analogy-making have significantly deepened our understanding of the mental processes underling analogical reasoning. Many of such models claim that they provide a domain-general explanation of how people perform analogy tasks. For example, the

Structure Mapping Engine, which was originally introduced to solve analogy task in discrete semantic space, was later used to answer analogy questions in continuous visual domain (Lovett, Forbus, & Usher, 2007). Most of the connectionist models are theoretically domain-general as well, since we can easily feed the distributed representation of stimuli from different domain/modality to the input placeholders of those models. It is important for these models to be domain-general, since humans are able to make within-modality generalization, such as recognizing examples that they have never encountered before (Lake, Salakhutdinov, & Tenenbaum, 2015), or make cross-modality transfer (Hupp & Sloutsky, 2011). Specifically, it has been shown that relational knowledge is critical to the development of analogical reasoning (Goswami, 1991). If one person has received sufficient training to solve verbal analogy tasks, they would gain some experience in relational knowledge. When later asked to solve a visual analogy-making task, they do not need to learn it from scratch (Figure 1).

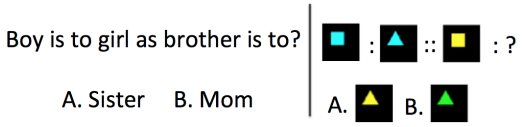


Figure 1: Stimuli used in the current study. Left: Verbal analogy task. Right: Visual analogy task.

Despite the generalization ability that previous computational models may have claimed, they are often missing some critical components. Symbolic approaches, for example, rarely address how people **learn** to make an analogy. Since the knowledge representations and the search algorithms in those models are preprogrammed, it is not clear how experience of analogy-making in one domain could facilitate analogy-making in another unfamiliar domain.

As for the connectionist models, despite their theoretically domain-general nature, none of the previous studies has directly tested cross-modality transfer. Particularly, they only demonstrate within-modality generalization, i.e., the models are trained on some examples and tested on a different set of examples in the same modality. Also, the distributed representation of stimuli are either manually defined according to the semantic features of the items, or randomly assigned to some localist codes, which are not very naturalistic.

Finally, all the previous modeling works have used small datasets, containing hundreds of examples at most. We are wondering if we could build a model that scales up to handle a very large and naturalistic dataset. Particularly, we want to understand if we use a dataset that reflects the statistical distribution of stimuli in real life, can the model still learn analogical reasoning and even make cross-modality transfer? This idea is motivated by the statistical learning account of language acquisition (Frost, Armstrong, Siegelman, & Christiansen, 2015), which proposes that language acquisition relies partially on a domain-general mechanism, which is learning and processing sensory stimuli unfolding across time and space (Saffran, Aslin, & Newport, 1996). Early since 1990s, researchers have found that if you train a recurrent neural network to predict the next word in a sentence, the word representation it learns reveals the syntactic and semantic role of the word (Elman, 1990). Inspired by these previous studies, we use distributed representations of words that reflect the statistics of word co-occurrence in everyday life, which are more naturalistic.

As for the representations of visual stimuli, we process the images (geometric figures) using a deep convolutional neural network that has been trained to perform object recognition task (Krizhevsky, Sutskever, & Hinton, 2012), and use the activation of the 7th hidden layer as the representations of the visual stimuli. Previous studies have shown that deep convolutional networks share a lot of similarities with human visual system (Yamins, Hong, Cadieu, & DiCarlo, 2013). After we obtain the representations (embeddings) of the words and the images, we build a simple, light-weighted neural network to learn the analogy tasks.

Experiment

Data. We first describe the representation we use for the word. The distributed representation of words are computed using the continuous Skip-Gram model (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). It takes the current word to predict the surrounding window of context words. Hence, the estimated word embeddings capture the semantic and syntactic role of the words (Mikolov, Yih, & Zweig, 2013). We download the pre-trained word vectors from Google Word2Vec[1], which have 300 dimensions. For computational simplicity and efficiency, we reduce the dimensionality to 30 using principle component analysis (PCA) so that each word has a 30-d vector representation.

We use the same verbal analogy dataset from Mikolov, Yih, and Zweig (2013)[2], which contains 19529 examples in total with 907 unique words. We divide the dataset into three sets, a training set (percentage: 80%, 15634 examples)[3], a validation set (10%, 1955 examples) and a test set (10%, 1955 examples). We use the accuracy on validation data to tune the hyper-parameters of the model and report accuracies on the test data. We run two types of tasks. The first one is A:B::C:[D1|D2|...|Dn], in which the model is given some choices and has to select the correct one (Task 1). We simulate Task 1 with different number of choices ranging from 2 to 5. The second type of task takes the form of A:B::C:?, i.e., the model is required to find the correct D from all words in its vocabulary (Task 2).

To simulate associative responses children usually give

[1]https://code.google.com/archive/p/word2vec/

[2]http://download.tensorflow.org/data/questions-words.txt

[3]Training with fewer data (e.g., 50%) does not lead to qualitatively different results.

when first learning analogy tasks, we construct three datasets from the original analogy dataset. All of them are binary-choice questions, but the incorrect alternatives have different levels of associations with the third item C in the question. In the High Association Dataset, each questions has an incorrect response alternative (foil) that is strongly associated with the third word in the analogy question, whereas examples in the Low Association Dataset contain alternatives that is weakly associated with the third item. Finally, in the Random Association Dataset, the incorrect response alternatives are randomly selected from the vocabulary so that it does not necessarily has a strong or weak association with the third item C. We determine word associations by calculating the cosine distance between the word vectors of the two words. The smaller the distance, the stronger the association.

As for the visual stimuli, we use the Shape dataset from Reed, Zhang, Zhang, and Lee (2015). It is a dataset of 2-D colored shapes, with 8 colors, 4 shapes, 4 scales, 5 row and column positions, and 24 rotation angles. We only use one value for the rotation variable to avoid potential confusion (e.g., a square rotated 180° would be the same figure as the original figure, but it has a different label in the dataset), and vary the other 5 variables to create a dataset. An example question is showed in Figure 1, right. We generate 19080 examples in total and randomly split them into a training set (80%), a validation set (10%) and a test set (10%). Next, we use the AlexNet, a deep neural network trained to recognize objects (Krizhevsky et al., 2012), to process these images. We use the pre-trained connection weights from Caffe (Jia et al., 2014) to process each image in our dataset and use the hidden activation of the 7th layer as its embedding. We also reduce the dimensionality of the image embeddings to 30 using PCA.

Model. The model architecture is fairly simple (Figure 2). There are three layers, the input layer, the hidden layer and the output layer. The input layer contains three pools that encode the first three items (A, B and C in the analogy question). Each pool has 30 nodes, which corresponds to the dimensionality of the word/visual embeddings. The connection weights from the input pool encoding A to the hidden layer H and the ones from the pool encoding C to H are the same, denoted by W_1. The connection weights from the pool encoding B to the hidden layer H are denoted by W_2. The connection weights from the hidden layer H to the output layer O is the embedding matrix of either the choices in the current example (Task 1) or the whole vocabulary (Task 2). Mathematically, the model can be described by the following equations:

$$\begin{aligned} H &= W_1 v_A + W_2 v_B + W_1 v_C + b \\ O &= \phi(W_0 H) \end{aligned} \tag{1}$$

where $v_A, v_B, v_C \in \mathbb{R}^{30}$ are the word/image embeddings for the stimuli, $W_1, W_2 \in \mathbb{R}^{30\times 30}$ are the connection weights from the input pools to the hidden layer, $b \in \mathbb{R}^{30}$ is the bias in the hidden layer, and $W_0 = [V_{D_1}; V_{D_2}; ...; V_{D_n}]^T \in \mathbb{R}^{30\times 30}$ is a matrix composed of embeddings of all the choices. We use the softmax function $\phi(\boldsymbol{x})_i = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$ to normalize the input $\boldsymbol{x}$ to the final layer O, which amounts to $W_0 H$. The ith value of the output, $\phi(\boldsymbol{x})_i$, indicates the probability of the ith choice being correct.

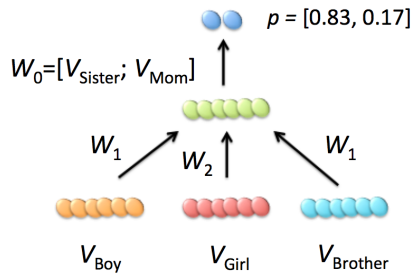


Figure 2: Model Architecture with an exemplar question "Boy:Girl::Brother:[Sister|Mom]"

Training. The model is trained by back-propagation using the TensorFlow framework (Abadi et al., 2015). We only update the weights W_1, W_2 and b. We train the neural networks with a batch size of 50 for each task. We find that the model cannot learn well in the Task 2 setting, where it needs to pick up the correct D from all the possible words. Therefore, in the following section we only report our results for Task 1. We run 2 simulations. In the first simulation, we examine whether we could reproduce the relational shift phenomenon. To this end, we train the model on a verbal analogy dataset with random association. As the training proceeds, we test it periodically on verbal High Association test set and verbal Low Association test set (within-modality), as well as the visual High Association test set and visual Low Association test set (cross-modality). In the second simulation, we look at the influences of number of choices on within-modality generalization and cross-modality transfer. Particularly, we first train the model to perform verbal analogy tasks with different numbers of choices, then test it on visual analogy tasks. We also conduct another version of the experiment in which the visual analogy tasks are learned first. We run each of the simulations 10 times with different random initialization of parameters, and in each run the model is trained for 41 epochs.

Results

Simulation 1. We first test if our model reproduces the relational shift observed during child development. Figure 3 shows the accuracy curves of four test sets: verbal High Association Dataset, verbal Low Association Dataset, visual High Association Dataset and visual Low Association Dataset.

First of all, we notice that our model clearly demonstrates the tendency of associative responding in the early stage of learning, since the accuracy on the High Association Dataset grows slowly (solid curves), compared with the accuracies

on the Low Association Dataset. As the training proceeds, the model gradually learns to inhibit associative responses for High Association Dataset (black solid curve).

Second, the test accuracies on verbal sets are consistently higher than those on the visual test sets, which is not surprising given that the model is only trained on verbal stimuli. However, we still find a decent amount of cross-modality transfer. For one thing, the tendency to give associative responses earlier in the training is carried over to the visual modality, even though no visual stimuli has been used to train the network. In addition, as the accuracy on verbal Datasets gradually increases, the model also becomes better at answering questions in visual Low Association Dataset (gray dash-dot curve). However, this improvement is not reliably transferred to the visual High Association Dataset, as the accuracy of this dataset remains near chance-level after prolonged training (gray solid curve).

Third, we find that when the accuracy on verbal Low Association test stops to grow after roughly 3 epochs, the accuracy on the corresponding visual dataset continues to improve until after 11 epochs, which then slowly decreases (gray dash-dot curve). This can be explained by the domain-general and the domain-specific component of analogical reasoning. The model first learns the domain-general component of analogical reasoning from the training in the verbal domain, but later the training becomes detrimental to cross-modality transfer since it continuously shapes the model to be specific to the verbal domain, thus reducing the accuracy in the corresponding visual domain.

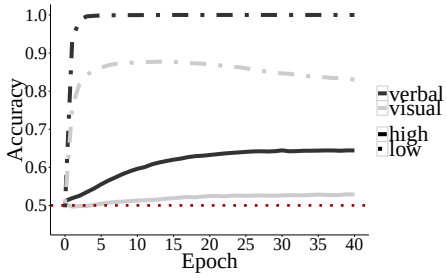


Figure 3: The effect of association on accuracy and cross-modality transfer. The dotted line indicates the chance-level.

Simulation 2. To understand the effect of number of choices on accuracy for different data type (train/test) and modality (verbal/visual), we run two linear models. In the first linear model, we look at two sets of data, which are obtained from the experiment "learning verbal analogy first" and the one "learning visual analogy first". In the "learning verbal analogy first" experiment (Figure 4, left), we find that both training and same-modality test accuracy are almost perfect, whereas the cross-modality accuracy is not. The linear model shows that for all of these three conditions, the accuracy decreases as the number of choices increases (train: $\beta = -0.005$, $t(114) = -2.57$, $p = 0.011$, same-modality test: $\beta = -0.004$, $t(114) = -2.12$, $p = 0.036$, different-modality test: $\beta = -0.1$, $t(114) = -54.63$, $p < .001$). However, we also find an interaction between test conditions and choice numbers. Particularly, the influence of choice numbers on different-modality test is much larger than the one on same-modality test, $\beta = 0.096$, $t(114) = 37.13$, $p < .001$. In the "learning visual analogy first" experiment (Figure 4, right), we find a similar effect of choice numbers on the different-modality test condition, as well as a similar interaction between test conditions and choice numbers, $\beta = 0.071$, $t(114) = 14.07$, $p < .001$.

Although both modalities demonstrate near perfect performance of within-modality generalization after sufficient training ($\sim$40 epochs), the performance of "learning verbal analogy first" is consistently better than the one of "learning visual analogy first" throughout the training. For instance, half way through the training, the same-modality test accuracy of "learning verbal analogy first" is higher than the one of "learning visual analogy first" (mean difference is 4.77 %, $t(234) = 3.279$, $p = 0.001$). This implies that the semantic space of word embeddings may have a stronger structural regularity, which makes it easier to discover relations between words than images.

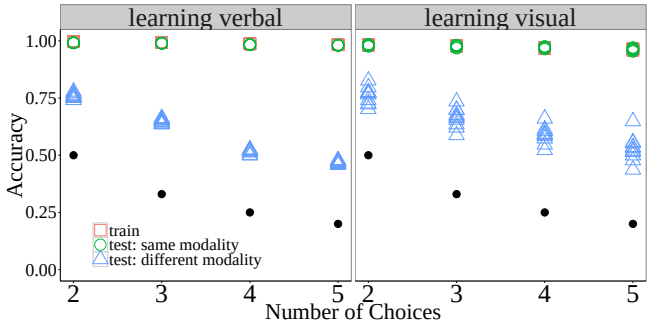
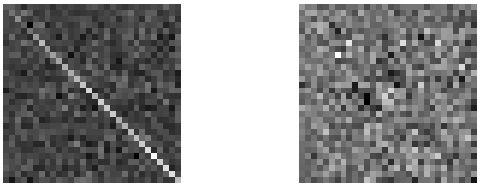


Figure 4: The influence of number of choices on accuracy. Solid dots indicate chance-levels.

Visualization of connection weights

To get a deeper understanding of what the model has learned, we visualize the weight matrix W_1 and W_2. We find that W_1 is very much like a identity matrix (Figure 5, left), whereas W_2 does not have a easily describable pattern (Figure 5, right).

We compare our model with the vector offset method, which was used by Mikolov, Yih, and Zweig (2013) to solve analogy tasks. Given the problem A:B::C:?, they found the word D such that its embedding vector had the greatest cosine similarity to $x_B - x_A + x_C$. Their method amounts to assigning an identity matrix I to W_1, $-I$ to W_2, and a zero vector to the bias b in our model. The weights of our neural network show that our model is not doing exactly the same thing as the vector offset method does, since B does not approximate the negative identity matrix. Hence, its weights are tuned to solve the current analogy task, and the same weights are also capable of solving analogy task in another modality.



(a) (b)

Figure 5: (a): Connections from A to the hidden layer. (b): Connections from B to the hidden layer. Lighter areas represent larger weights.

Discussion

In this article, we build the first neural network model that can learn to solve analogy tasks and make cross-modality transfer. It uses word representations and image representations estimated from large-scale naturalistic corpora. The model demonstrates both the domain-general and the domain-specific component of analogical reasoning.

Specifically, we see that the accuracy on the same-modality test set is consistently higher than the accuracy on the cross-modality test set. This is aligned with the empirical finding that domain-specific knowledge boosts performance in analogical reasoning (Goswami & Brown, 1990b). On the other hand, the model demonstrates the domain-general property of analogy-making by showing cross-modality transfer. This is relevant to a broader topic in cognitive science, the zero-shot learning. Zero-shot learning refers to the ability to solve a task despite not having received any training examples of that task. As human beings, we do zero-shot learning all the time. Only recently did researchers begin to simulate zero-shot learning using neural network models. For instance, in Socher, Ganjoo, Manning, and Ng (2013), they showed that learning the distributions of words in texts as a semantic space helps the model understand the visual appearances of objects, and enables the model to recognize objects even if no training data is available for that category. Our model contribute to the zero-shot learning literature by showing that zero-learning is possible for analogy-making task as well. It also makes some interesting predictions that can be empirically tested. The success of our model suggests the possibility that there might be some similar structural regularities in the word embeddings extracted from naturalistic corpus and in the image embeddings extracted from object recognition models. This similarity explains why our model makes cross-modal transfer.

Our results are also relevant to another line of research, the one-shot learning. One-shot learning refers to the problem of learning from one or very few examples. Classic deep learning neural networks could not perform one-shot learning, which is a common criticism of neural networks being plausible cognitive models of human learning (Lake, Ullman, Tenenbaum, & Gershman, 2016). However, recently Vinyals, Blundell, Lillicrap, Wierstra, et al. (2016) showed that if you match the training task with the test task, neural networks are able to learn from few examples. In their paper, they trained a network to map a query example to one of the four candidates example so that both of them belong to the same category. The model learned the task very well. Our results lend further support to their approach. We find that our model only learns efficiently under the Task 1 setting, where it chooses among a few choices rather than the whole vocabulary. Our work extends Vinyals and colleagues' results by showing that our network model can make an inference not only on unseen stimuli, but also on unseen stimuli from a completely different modality.

There are some limitations of the current work. First, the model is a simplification of the actual mental processes underlying analogy-making. A lot of previous computational model have given very insightful explanations of those mental processes (Gentner, 1983; Morrison et al., 2004; Kollias & McClelland, 2013; Gergel' & Farkaš, 2015), and our goal is not to argue against those models or to provide a better model. Instead, our goal is to demonstrate the possibility that domain-general neural network models can learn from large-scale, realistic datasets to solve analogy tasks. Second, we have not directly compared our model performance with human performance. It would be interesting to see how human would respond to the analogy questions in the current study and whether our model predictions align with human data in the future.

Acknowledgments

A.Y. is grateful to Jay McClelland for helpful discussions about this work, and the whole Parallel Distributed Processing Lab at Stanford for their useful comments.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., . . . Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems.* (Software available from tensorflow.org)

Brian, F. (1990). Analogical interpretation in context. In *Proceedings of the 12th Annual Meeting of the Cognitive Science Society.*

Buck, G., VanEssen, T., Tatsuoka, K., Kostin, I., Lutz, D., & Phelps, M. (1998). Development, selection and validation of a set of cognitive and linguistic attributes for the SAT I Verbal: Analogy section. *ETS Research Report Series*, *1998*(1).

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

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, *41*(1), 1–63.

Forbus, K. D., Gentner, D., Markman, A. B., & Ferguson, R. W. (1998). Analogy just looks like high level perception: Why a domain-general approach to analogical mapping is right. *Journal of Experimental & Theoretical Artificial Intelligence*, *10*(2), 231–257.

French, R. M. (2002). The computational modeling of analogy-making. *Trends in Cognitive Sciences*, *6*(5), 200-205.

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: the paradox of statistical learning. *Trends in Cognitive Sciences*, *19*(3), 117-125.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170.

Gentner, D., & Forbus, K. D. (2010). Computational models of analogy. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(3), 266–276.

Gentner, D., Holyoak, K. J., & Kokinov, B. N. (2001). *The analogical mind: Perspectives from cognitive science*. MIT press.

Gergel', P., & Farkaš, I. (2015). Connectionist modeling of part–whole analogy learning. In *Proceedings of the EuroAsianPacific Joint Conference on Cognitive Science.*

Goswami, U. (1991). Analogical reasoning: What develops? a review of research and theory. *Child Development*, *62*(1), 1–22.

Goswami, U., & Brown, A. L. (1990a). Higher-order structure and relational reasoning: Contrasting analogical and thematic relations. *Cognition*, *36*(3), 207–226.

Goswami, U., & Brown, A. L. (1990b). Melting chocolate and melting snowmen: Analogical reasoning and causal relations. *Cognition*, *35*(1), 69–95.

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science*, *13*(3), 295–355.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*(3), 427–466.

Hupp, J. M., & Sloutsky, V. M. (2011). Learning to learn: From within-modality to cross-modality transfer during infancy. *Journal of Experimental Child Psychology*, *110*(3), 408-421.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., … Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

Kokinov, B., & Petrov, A. (2000). Dynamic extension of episode representation in analogy-making in AMBR. In *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society.*

Kollias, P., & McClelland, J. (2013). Context, cortex, and associations: a connectionist developmental approach to verbal analogies. *Frontiers in Psychology*, *4*, 857.

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).

Kuehne, S., Forbus, K., Gentner, D., & Quinn, B. (2000). SEQL: Category learning as progressive abstraction using structure mapping. In *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society.*

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. (2016). Building machines that learn and think like people. *arXiv preprint arXiv:1604.00289*.

Lovett, A., Forbus, K., & Usher, J. (2007). Analogy with qualitative spatial representations can simulate solving Raven's Progressive Matrices. In *Proceedings of the 29th Annual Meeting of the Cognitive Science Society.*

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).

Mikolov, T., Yih, S. W.-t., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013).* Association for Computational Linguistics.

Mitchell, M. (1993). *Analogy-making as perception: A computer model*. MIT Press.

Morrison, R. G., Krawczyk, D. C., Holyoak, K. J., Hummel, J. E., Chow, T. W., Miller, B. L., & Knowlton, B. J. (2004). A neurocomputational model of analogical reasoning and its breakdown in frontotemporal lobar degeneration. *Journal of Cognitive Neuroscience*, *16*(2), 260–271.

Reed, S. E., Zhang, Y., Zhang, Y., & Lee, H. (2015). Deep visual analogy-making. In *Advances in Neural Information Processing Systems* (pp. 1252–1260).

Richland, L. E., & Burchinal, M. R. (2013). Early executive function predicts reasoning development. *Psychological Science*, *24*(1), 87–92.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Socher, R., Ganjoo, M., Manning, C. D., & Ng, A. (2013). Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems* (pp. 935–943).

Sternberg, R. J., & Nigro, G. (1980). Developmental patterns in the solution of verbal analogies. *Child Development*, 27–38.

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

Yamins, D. L., Hong, H., Cadieu, C., & DiCarlo, J. J. (2013). Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream. In *Advances in Neural Information Processing Systems* (pp. 3093–3101).