What We Say & What We See: Finding Alignment of Conceptual Systems in Image-Word Representations

Cindy Luo (kl3108@nyu.edu)
Center for Data Science, New York University

Bella Zhang (bz2428@nyu.edu)
Center for Data Science, New York University

Zoe Xiao (yx1750@nyu.edu)
Center for Data Science, New York University

Abstract

Deep neural networks have achieved remarkable success across various domains, including computer vision and natural language processing, through extensive training on real-world stimuli and learning representations of stimuli using thousands of features. These representations, as different sources of input, should produce similar conceptual systems, given that they are different viewpoints of the same underlying reality. This paper investigates the alignment of conceptual systems in image-word representations generated by deep neural networks. We analyze the image representations learned by a deep neural network trained for supervised classification tasks and compare them with the word embeddings of the corresponding labels. Our results demonstrate that different conceptual systems trained separately can align with each other and that the vision model trained on classification tasks has substantial power in revealing the higher-level structure of concepts. Overall, this study sheds light on the complex relationship between language and vision and provides insights into the mechanisms underlying conceptual representation and categorization.

Keywords: deep learning; neural networks; image-word representations; alignment; concept learning

Introduction

When we say the word 'dog', to what extent does this word convey the visual information of a dog? Human learning of concepts starts from labeling observed everyday objects with the use of language. During the early phase of development, supervised learning, often guided by our guardians, explicitly links distinct conceptual systems (for example, an image of a dog and the word 'dog') to help us integrate multiple sources of stimuli and form a complex understanding of the world. This remarkable ability to map both visual stimuli and language to abstract concepts sets the foundation for the learning that happens afterward (Smith & Yu, 2008; de Sa & Ballard, 1998). For this reason, various research has looked into mechanisms explaining concept learning via the lens of language or images (Lake, Salakhutdinov, & Tenenbaum, 2015; Xu & Tenenbaum, 2007; Pothos & Chater, 2002).

Although there has been a significant advancement in theory, empirical verification has been restricted mainly to controlled laboratory environments with either artificial stimuli or relatively simple representations. In addition, the relationship between language and visual information itself remained unclear in the study of concept learning, with previous research often drawing the equivalence between the abstract concept and the language used to describe the concept rather than treating language as a separate conceptual system. Moreover, the difficulty to measure and evaluate the human understanding of the abstract concept makes it nearly impossible to identify and compare mental representations formed via concept learning through different types of stimuli.

In recent years, deep neural networks (DNN) have demonstrated near-human performance in tackling challenging various visual and language tasks. This is due in part to the hierarchical architecture of these models, where each layer learns increasingly abstract representations of the input stimuli. For example, image representations observed in the hidden layers of Convolutional Neural Networks (CNN) are thought to be similar to the neural representations observed in the mammalian visual system (Rawat & Wang, 2017), suggesting that CNNs are a powerful tool for studying the underlying mechanisms of human vision. Furthermore, recent work has shown that these internal representations can be used to predict human behavioral responses, indicating that the representations learned by CNNs are meaningful and relevant for understanding human visual perception (Jha, Peterson, & Griffiths, 2023; Battleday, Peterson, & Griffiths, 2020; Kubilius, Bracci, & Op de Beeck, 2016; Lake, Zaremba, Fergus, & Gureckis, 2015).

Similar to the image representations generated by CNNs, word embeddings, a representation used in natural language processing models that maps words to vectors in a high-dimensional space, have been shown to accurately reflect the semantic meaning of words, such that words that are similar in meaning are located close to each other in the embedding space (Pennington, Socher, & Manning, 2014). In addition, researchers have shown that combining image representations with word embeddings results in a promising performance gain of language models (Lazaridou, Pham, & Baroni, 2015; Kiela & Bottou, 2014). Gaining a deeper understanding of the link between these two types of input could inform the design of model architecture, leading to better performance. Given the ability of embedding space to reflect human perceptions and the growing interest in the development of multi-modal models, it is critical to investigate the alignment of conceptual systems in image-word representations to uncover the relationship between language and vision in concept learning.

Previous research has studied the alignment of the image and the text conceptual systems under the paradigm of unsupervised learning (Roads & Love, 2020). Using three embed-

dings space (image, text, audio) obtained via unsupervised learning, the researchers have found that the alignment can be guided by a strong signal that exists between different sensory modalities. This finding suggests that there exists a unique signature for each concept within one conceptual system to be mirrored in other systems. Other studies have focused on linking embeddings generated from a single source with the abstract mental representation of concepts. Peterson, Abbott, and Griffith (2018) have found correspondence between human similarity ratings of the visual stimuli and the image representations in DNNs. In addition, semantic representations learned from language training have been shown to reflect human judgments of visual similarity, as well (Lewis, Zettersten, & Lupyan, 2019).

Here, we examine the image representations obtained via supervised learning for classification tasks and align them with the word embeddings that correspond to the image labels. First, we create a similarity matrix for each system and consider mappings between the systems by calculating the correlation. Then, to uncover the strength of alignment between the image representations and the corresponding word embeddings, we conduct hierarchical analyses on each embedding space separately to compare their structures. We also test this alignment using image representations extracted from different layers of the network and compare it across three major domains of categories. Our findings show that distinct conceptual systems, when trained independently with supervision, have the ability to align with one another and that the visual model trained for classification tasks has the ability to reveal the more abstract structure of concepts.

Methods

The primary objective of the study is to determine if the visual conceptual system can be aligned with the word embedding space when trained with supervision and examine this relationship across different layers from which the image representations were extracted and across different category domains. The secondary objectives of this study are to reveal the structure of each embedding space and provide explanations for the strengths of the alignment observed. We use two different embedding techniques. For the image representations, we use pre-trained VGG-16 to extract the activations from the hidden layers (Simonyan & Zisserman, 2015). For word embeddings, we use the results obtained via the GloVe algorithm (Pennington et al., 2014).

Image representations When feeding an image into the deep neural network, the units in each hidden layer give different activation values. We can take these activation values as the image representation generated from each layer. We chose to use the pre-trained VGG-16 model since previous research had shown that VGG-16 is predictive of human similarity judgment (Peterson et al., 2018) and can achieve near state-of-the-art performance in image classification (Simonyan & Zisserman, 2015). To create our own dataset of visual stimuli, we selected images from the ImageNet Dataset (Deng et al., 2009), a large dataset of 1.2 million images taken from 1000 object categories.

Our selection started with defining major domains of categories we were interested in. Following the approach taken by Peterson, Abbott, and Griffiths (2018), we first identified three major domains for the selection of categories: *Animals, Vehicles, and Fruits & Vegetables*. Then, we looked at the intersection between the category names in the ImageNet dataset and the words contained in the word embedding system. To correctly map between the word and the image, we filtered out categories whose names are not present in the word embedding space, leaving us with 539 categories to choose from (about 50% of the category names in the ImageNet dataset consist of more than one word, and we wanted to get an exact match between the category name and the word). We then manually labeled each category with its domain and found that, among the 539 categories, 163 were under the *Animals* domain, 31 were under the *Vehicles* domain, and 24 categories belonged to the *Fruits & Vegetables* domain. We randomly selected 20 categories for each domain.

When corresponding the word embedding with the image representation, instead of taking the image representation of a single image, we randomly selected 30 images from the same category and calculated an average image representation. For example, for the category 'koala', there is one word embedding we can use, and we generated one image representation by averaging across the image representation of 30 images of 'koala'. Notice that we also only selected images that have been correctly classified. Thus, for most categories, the image representation for that category was obtained from around 28-30 images.

As images are fed forward through the pre-trained VGG-16 model, we recorded activations in all the ReLU and maxPooled2d layers. There were 15 ReLU layers and 5 pooling layers in the model. As an example, image representations of the last ReLU layer (after the second last fully-connected layer in the classifier) for the *Animals* domain make up a 20×4096 matrix. For the activations extracted after the convolutional layer, we flattened the 2-d arrays and made them into 1-d vectors. All activations were then z-score normalized.

Word embeddings We used pre-trained 300d word vectors from Glove Wikipedia 2014 + Gigaword 5 (Pennington et al., 2014).

Correlation Analysis

In this part, we examine the alignment by calculating the correlation between the similarity matrices generated from the GloVe300 word embeddings and the VGG-16 model layers' image representations.

First, we calculated the cosine-similarity matrix for image representations obtained from each hidden layer of VGG-16 model as well as the word embeddings respectively. The full-size matrix is 60×60 for all categories in 3 domains.

To investigate the correlation between the word embed-

dings and the image representations, we used heat maps to visualize the similarity matrices to explore whether there is a consistent pattern between the two conceptual systems. Additionally, following previous research, we calculated the Spearman correlation (Roads & Love, 2020) between the two similarity matrices to quantify the strength of the correlation and studied how it varies along the layer depth of the model.

Clustering Analysis

We then compared the internal structures of the word embeddings and the image representations by looking at the similarity matrices between objects across different domains we selected, namely *Animals*, *Vehicles*, *Fruits & Vegetables*. To achieve this, we employed two widely-used methods for clustering analysis: Hierarchical Clustering which produces a dendrogram that illustrates the hierarchical relationships among the objects, and 2D t-SNE map, which maps similarities into a spatial representation. Both methods utilize cosine similarity for clustering.

Results

There is 1 word embedding set obtained from the GloVe300 algorithm and 20 image representation sets generated from layers in the VGG-16 model. The sequence and names for layer are shown in **Table 2** in appendix.

Correlation of Similarity Matrix

In this part, we calculated the similarity matrices for three domains based on the GloVe300 word embedding set and 20 image representation sets. The within-domain matrix size is 20×20.

Extreme values in Similarity Matrices We listed the minimum and maximum (excluding 1) similarity values in matrices generated from GloVe300 word embedding and from the first 3 and the last 4 hidden layers[1] in VGG-16 in **Table 1**.

Table 1: Maximum and Minimum values for 3 domains in the similarity matrices.

Layers	Animals20		Vehicles20		Fruit&Vege20	
	Max	Min	Max	Min	Max	Min
Word	0.515	-0.043	0.491	-0.119	0.706	-0.180
Re 1	0.878	0.610	0.870	-0.533	0.884	0.144
Re 2	0.779	0.260	0.687	-0.223	0.876	0.030
Mp 1	0.783	0.330	0.770	-0.358	0.855	0.003
	...	...	...	...	...	...
Re 13	0.519	-0.009	0.696	0.001	0.690	-0.029
Mp 5	0.549	-0.012	0.747	-0.008	0.728	-0.041
Fc 1	0.475	-0.131	0.593	-0.072	0.757	-0.091
Fc 2	0.482	-0.148	0.652	-0.107	0.782	-0.063

[1]For layers in VGG-16, 'Re' represents ReLU layers, 'Mp' represents max pooling layers, and 'Fc' represents the ReLU layers after the fully-connected layer)

From this table, we can see that there is a great change, especially for the minimum similarity values, based on the first 3 layers in VGG-16 network compared to the matrix based on word embedding. This indicates that VGG-16 network extracts certain features in each domain at very first layers of the model.

For the last 4 layers, the extreme values are closed to word embedding values, suggesting that feature extraction becomes more detailed with the depth of the VGG-16 network.

Heat map for Similarity Matrices To visualize the similarity between different domains, we drew heat maps of the similarity matrices for the word embedding and the final layer 'fc2' in VGG-16 network for all 60 categories. **Figure 1** is for word embedding and **Figure 2** is for image representations extracted from the 'fc2' layer. Notice that the figure size isn't big enough, so not all categories' names are listed in pictures.

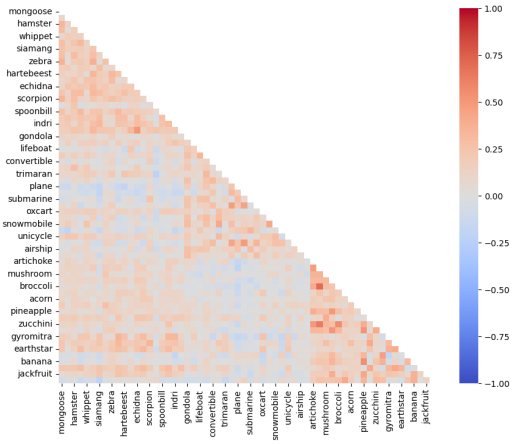

Figure 1: Heat map for similarity matrix based on word embedding.

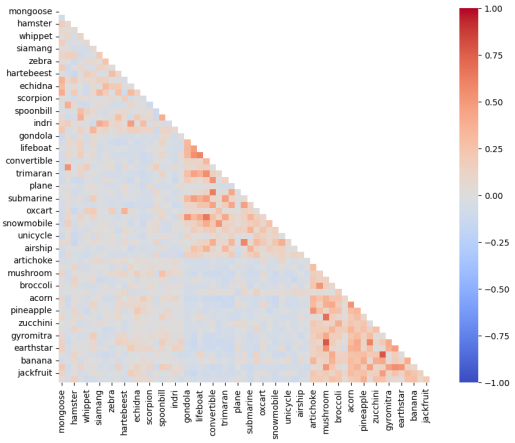

Figure 2: Heat map for similarity matrix based on the 'fc2' layer.

By comparing two figures, we can see that overall the similarity values generated from the word embeddings is higher than that from the image representations.

For all three domains, we can obviously tell that, in both two similarity matrices, in general the three triangular areas representing within-domain similarity values have darker colors than other areas, indicating higher with-in domain similarities. Additionally, the area for *Fruit & Vegetable* (the third part) has the highest similarity in both matrices, suggesting that there is higher with-in domain similarity for *Fruit & Vegetable* in both conceptual systems.

For cross-domain area, higher-level of similarity can be observed between *Fruits & Vegetables* and *Animals* domain in the word embedding matrix, compared to other cross-domain areas. As for image embedding, the same difference still exists but not that obvious. This may indicate discrepancy between natural objects and man-made objects.

In general, despite the numerical differences, both heat maps show the same pattern, which shows that the similarity matrices from word embedding and image embedding have a certain correlation.

Correlation across different layers Then, we investigated the relationship between word embeddings and layers in VGG-16 model in sequence by calculating the Spearman correlation between similarity matrices separately for three domains. **Figure 3** shows Spearman correlation for 20 layers in the order in VGG-16 model.

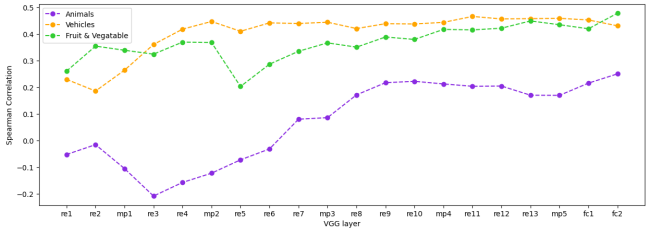

Figure 3: Spearman correlation between similarity matrices based on word embedding and on different layers in VGG-16 model.

For *Animals*, *Vehicles* and *Fruit & Vegetable* domains, the Spearman correlation values between word embedding and 'fc2' layer embedding are 0.250, 0.431 and 0.478, and the p-values are $4.98e-04$, $5.63e-10$ and $3.11e-11$, which suggests that there may exists significant correlation between similarity matrices.

Figure 3 shows that in general correlation increases as layers get deeper, which consist with the changes in extreme values in Table 1 above. And we can also notice that after about 'ReLU 9' layer, correlation values become relatively stable.

Compared three domains we picked, it's obviously that correlation for *Vehicles* and *Fruits & Vegetables* is higher than for *Animals* domain. This may because small categories under *Animals* domain contains insects, birds and mammal, which causes larger visual variance as compared to others.

Clustering Analysis

Word Embedding Clustering The dendrogram presented on the left side of Figure 4 illustrates the structure of the word embeddings using hierarchical clustering. The three clusters depicted on the dendrogram correspond well with with the three domains of *Animals, Vehicles, Fruits & Vegetables*, as indicated by the color-coded legend. Notably, one cluster exclusively contains leaf nodes representing vehicles, while some vehicles, such as "dogsled" and "oxcart" fall into the cluster dominated by animals. This finding is expected as these words contain "dog" and "ox" respectively, which are animals that may often be mentioned in conversations involving vehicles. Furthermore, we observe that different kinds of fungus, such as "stinkhorn" and "bolete", are clustered together in a cluster dominated by fruits and vegetables, but they are also located at the boundary between the *Fruits & Vegetables* cluster and the dominant *Animals* cluster.

In the t-SNE visualization of word embedding (see right side of Figure 4), we observe that the three clusters intersect in the middle. The intersection between the *Animals* cluster and the *Vehicles* cluster is exemplified by "oxcart", and a group of fungus is also located close to the *Animals* cluster. These findings are consistent with the results of the hierarchical clustering analysis of the word embedding. Additionally, the *Animals* cluster appears to be more scattered than the other two domains, which have a clearer cluster structure. This unclear cluster structure of the animal domain might partially account for the low alignment correlation.

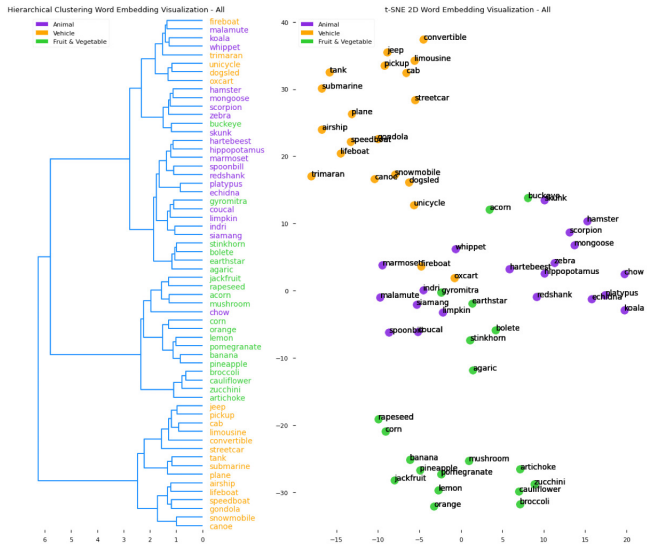

Figure 4: Word Embedding Clustering Visualization

Image Embedding Clustering For our analysis of image embedding, we utilized two layers from the VGG-16 network: the last fully connected layer and the last max pooling layer. We aimed to evaluate the network's ability to classify objects within and across domains based on these two layers.

The dendrogram obtained for the image embedding using

the ReLU layer after the second last fully connected layer of VGG-16 (see left side of Figure 5) displays a more distinct cluster structure with evenly distributed clusters and a higher level of purity which means each cluster contains objects largely from the same domain. This finding suggests that the image classification model has a greater understanding of the higher-order structure, despite the absence of any domain specification for the labels during the training process. The clear clustering of objects in the image embedding indicates that visual information is useful for higher-order categorization. This observation implies that during conceptual learning, the model might first recognize a generic shape and then add more details. Furthermore, the fact that "oxcart" is clustered into the *Animals* domain may be attributed to the co-occurrence of the object in the image during classification, which suggests that the context or background of an object is significant in the classification process.

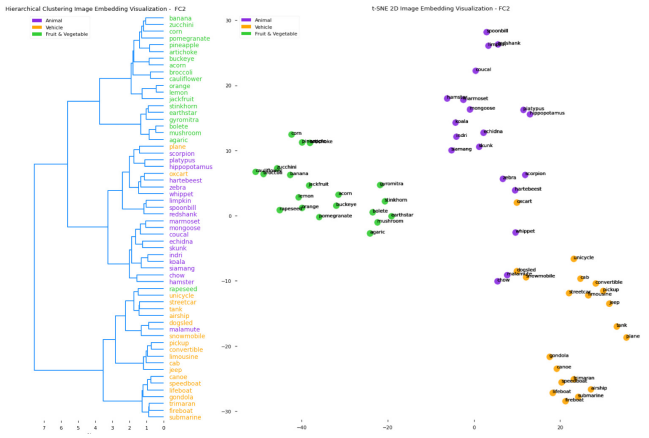

Figure 5: Image Embedding Clustering Visualization

We also generated a dendrogram for the image embedding using the last max pooling layer of VGG-16 (see appendix Figure 6), and the purity of each cluster was even higher than that of the fully connected layer. This result suggests that the fully connected layer in the classifier captures finer details and might emphasize certain aspects of the embeddings to differentiate among similar objects within one domain, thereby blurring the cluster structure. As a result, the max pooling layer may be a better layer for investigating the clustering of image embeddings.

In the t-SNE visualization of the image embedding of the last fully connected layer in VGG-16 (see right side of Figure 5), we observe a clearer cluster structure compared to the word embedding. The between-cluster distance is relatively large, indicating that the objects within each domain are separated from the objects belonging to other domains except for two special vehicles, namely "dogsled" and "oxcart", which are close to the *Animals* cluster. This observation is consistent with the result obtained in the earlier word embedding dendrogram. These findings suggest that the image classification model is effective in capturing the higher-order structure of objects, even without domain information during the training process. Notably, the group of fungus is no longer intertwined with animals, which is a difference from the word embedding analysis.

This analysis provides insightful observations about the differences in the internal structures of the two representations across domains and how these structures could explain the different alignment strengths observed in the previous analysis, highlighting the importance of considering multiple representations in studying conceptual learning.

Discussion

Our study provides evidence that visual information is closely related to the meaning of language, with much of what we see contributing to what we say. This finding has important implications for the development of multi-modal language models, which could benefit from incorporating sensory grounding (such as visual information) to improve their performance. In addition, our results suggest that applying methods from natural language processing (NLP) to vision models can provide valuable insights for cross-modal modeling, as well as enhance our understanding of computer vision.

Furthermore, our study highlights the importance of visual information in conceptual learning and categorization. Previous research has shown that vision models trained to complete high-level tasks, such as clustering images in unsupervised learning, can achieve the same level of performance as supervised models in low-level tasks like object classification. Our results provide further validation of this approach by demonstrating that learning detailed classification tasks can also inform models about clustering, indicating a bidirectional relationship between high-level and low-level tasks in model training.

Looking ahead, future research could focus on identifying the specific visual features that contribute to such clustering structures and exploring variations across different domains, such as natural objects versus man-made objects or curved shapes versus shapes with edges and corners. Another interesting direction for research would be to use categories that are not part of the training set, and examine whether the image embeddings generated by pre-trained models still align with their corresponding word embeddings and fall into the correct cluster.

While our study provides valuable insights into the relationship between visual information and language, it is important to acknowledge its limitations. First, our selection of categories was limited to single-word labels, whereas 50% of the labels consisted of two words, which may have influenced our random selection of categories for the domain. For example, the variance observed in the animal domain could have been caused by an uneven portion of Animal categories present after filtering. Second, while previous research has focused on individual images, our study averaged across images, which may not provide the most accurate image representation of the word. Although two images could be classi-

fied as the same label, they could vary a lot from each other in terms of background and lighting. Finally, we used raw activations rather than transformed image representations, which may have affected the accuracy of our results.

Overall, our findings have implications for the process of category learning, highlighting the crucial role of visual information in the meaning of language and the human understanding of concepts. The findings could also inform ways to combine different sources of stimuli and provide insights for the development of multi-modal models in the future.

Acknowledgments

This work was part of the course work for the DS-GA 1016 Computational Cognitive Modeling course in Spring 2023. It was supported by the instructors of the course, as well as the Center for Data Science at New York University.

References

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2020). Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature Communications*, *11*(1), 5418. doi: 10.1038/s41467-020-18946-z

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (p. 248-255). doi: 10.1109/CVPR.2009.5206848

de Sa, V. R., & Ballard, D. H. (1998). Category learning through multimodality sensing. *Neural Comput.*, *10*(5), 1097–1117. doi: 10.1162/089976698300017368

Jha, A., Peterson, J. C., & Griffiths, T. L. (2023). Extracting low-dimensional psychological representations from convolutional neural networks. *Cognitive Science*, *47*(1), e13226.

Kiela, D., & Bottou, L. (2014). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 36–45). Doha, Qatar: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D14-1005 doi: 10.3115/v1/D14-1005

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

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332-1338. doi: 10.1126/science.aab3050

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep neural networks predict category typicality ratings for images. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th annual meeting of the cognitive science society, cogsci 2015, pasadena, california, usa, july 22-25, 2015*. cognitivesciencesociety.org.

Lazaridou, A., Pham, N. T., & Baroni, M. (2015). Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 conference of the north American chapter of the association for computational linguistics: Human language technologies* (pp. 153–163). Denver, Colorado: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N15-1016 doi: 10.3115/v1/N15-1016

Lewis, M., Zettersten, M., & Lupyan, G. (2019). Distributional semantics as a source of visual knowledge. *Proceedings of the National Academy of Sciences*, *116*(39), 19237-19238. doi: 10.1073/pnas.1910148116

Pennington, J., Socher, R., & Manning, C. (2014, October). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics. doi: 10.3115/v1/D14-1162

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, *42*(8), 2648–2669.

Pothos, E., & Chater, N. (2002, 05). A simplicity principle in unsupervised human categorization. *Cognitive Science*, *26*, 303-343. doi: 10.1016/S0364-0213(02)00064-2

Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, *29*(9), 2352-2449. doi: 10.1162/neco$_{a0}$0990

Roads, B. D., & Love, B. C. (2020). Learning as the unsupervised alignment of conceptual systems. *Nature Machine Intelligence*, *2*(1), 76–82.

Simonyan, K., & Zisserman, A. (2015). *Very deep convolutional networks for large-scale image recognition.*

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568. doi: 10.1016/j.cognition.2007.06.010

Xu, F., & Tenenbaum, J. (2007). Word learning as bayesian inference. *Psychological review*, *114*, 245-72. doi: 10.1037/0033-295X.114.2.245

Appendices

Table 2: Sequence and size for layers in VGG-16 model.

Layer	Embedding Size
Re 1	3211264
Re 2	3211264
Mp 1	802816
Re 3	1605632
Re 4	1605632
Mp 2	401408
Re 5	802816
Re 6	802816
Re 7	802816
Mp 3	200704
Re 8	401408
Re 9	201408
Re 10	401408
Mp 4	100352
Re 11	100352
Re 12	100352
Re 13	100352
Mp 5	25088
Fc 1	4096
Fc 2	4096

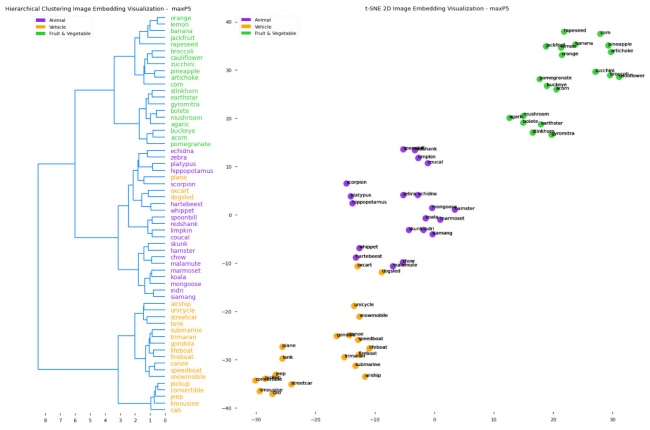

Figure 6: Image Embedding (from the 5th Max Pooling Layer of VGG-16) Clustering Visualization.

The code for this project can be found at `https://github.com/zoexiao0516/dsga1016-finalproject`.