


How Well Do Deep Learning Models Capture Human Concepts? The Case of the Typicality Effect

Siddhartha K. Vemuri, Raj Sanjay Shah, Sashank Varma

{svemuri8, rajsanjayshah, varma}@gatech.edu

Georgia Institute of Technology 

Abstract

How well do representations learned by ML models align with those of humans? Here, we consider concept representations learned by deep learning models and evaluate whether they show a fundamental behavioral signature of human concepts, the *typicality effect*. This is the finding that people judge some instances (e.g., robin) of a category (e.g., Bird) to be more typical than others (e.g., penguin). Recent research looking for human-like typicality effects in language and vision models has focused on models of a single modality, tested only a small number of concepts, and found only modest correlations with human typicality ratings. The current study expands this behavioral evaluation of models by considering a broader range of language ($N = 8$) and vision ($N = 10$) model architectures. It also evaluates whether the *combined* typicality predictions of vision + language model pairs, as well as a multimodal CLIP-based model, are better aligned with human typicality judgments than those of models of either modality alone. Finally, it evaluates the models across a broader range of concepts ($N = 27$) than prior studies. There were three important findings. First, language models better align with human typicality judgments than vision models. Second, combined language and vision models (e.g., AlexNet + MiniLM) better predict the human typicality data than the best-performing language model (i.e., MiniLM) or vision model (i.e., ViT-Huge) alone. Third, multimodal models (i.e., CLIP ViT) show promise for explaining human typicality judgments. These results advance the state-of-the-art in aligning the conceptual representations of ML models and humans. A methodological contribution is the creation of a new image set for testing the conceptual alignment of vision models.

Keywords: Concepts; Categorization; Typicality Effect; Machine Learning; Multimodal Models; Computational Modeling

Introduction

Categorization is a fundamental aspect of cognition. Assigning a new stimulus to a category enables humans to make inferences about its unknown or unobservable properties, facilitating taking action in the world (Murphy, 2002). A classic proposal in cognitive science is that not all the members of a category have the same status (Rosch, 1975). Rather, members vary in their *typicality*, with some members (e.g., robin) more typical of a category (e.g., Birds) than others (e.g., penguin). Moreover, people are faster to understand sentences about typical vs. atypical category members, and are quicker to give the category label (e.g., "Birds") of typical vs. atypical members when presented with images (Murphy, 2002).

Here, we investigate whether Large Language Models (LLMs) and computer vision (CV) models also show typicality gradients. Prior studies that have investigated this question have found suggestive results (Misra, Ettinger, & Rayz,

2021; Upadhyay, Mittal, & Varma, 2022). We go beyond this work to examine a larger number of LLMs and CV models, and to evaluate these models using newer human typicality data collected over a larger number of categories. We investigate for the first time whether the *combined* typicality predictions of vision + language model pairs better align with human typicality judgments than those of models of either modality alone. We also explore the potential of multi-modal models (i.e., CLIP).

The Typicality Effect

The typicality effect is that people regard some members as "better" examples of a category than others. Investigating typicality gradients requires collecting data from humans. The most common procedure is to give participants a category label (e.g., Fruits) and to have them write down as many exemplars of the category as they can in a fixed amount of time, usually 30 seconds (Battig & Montague, 1969; Castro, Curley, & Hertzog, 2021; Van Overschelde, Rawson, & Dunlosky, 2004). The typicality of a member is defined as the proportion of participants who produce it. Another approach is to provide a category label and a sequence of members and have participants rate the "goodness" of each member on a scale ranging, for example, from 1 (very typical) to 7 (very atypical) (Rosch, 1975). The typicality of a member is its average rating across participants.

Typicality in ML Models

We investigate whether ML models trained on large corpora or image sets also show the typicality effect. These models learn about the statistical structure of the cognitive environment to perform word prediction tasks or image classification tasks, respectively. Here, we evaluate whether as a consequence of this training, they become sensitive to the typicality gradients that organize the members of categories, learning them as latent representations.

Typicality in Language Models Researchers have looked for typicality effects in language models. An early study investigated whether word2vec embeddings could be used to predict category typicality data (Heyman & Heyman, 2019; De Deyne et al., 2008). The mean correlation between word2vec and humans across 16 categories was only 0.29. A more recent study (Misra et al., 2021) investigated the align-

ment of more modern transformer-based models (including RoBERTa and GPT-2) to the Rosch (1975) typicality ratings for 10 categories. The 19 models tested showed a range of correlations, with the larger variants of RoBERTa and GPT-2 achieving the highest values of approximately 0.40. Bhatia and Richie (2022) developed a BERT-based model and evaluated it against 25 findings on semantic cognition, including that of typicality gradients. The model’s typicality ratings across 10 categories correlated 0.32 with the human ratings of Rosch (1975). Thus, we see that NLP models have shown modest abilities to account for the typicality effect.

Typicality in Vision Models Researchers have also investigated whether vision models align with human conceptual understanding (Battleday, Peterson, & Griffiths, 2021). An early study (Peterson, Abbott, & Griffiths, 2018) had participants rate the pairwise similarity of 120 images of exemplars from each of 6 categories. They compared these to the cosine similarities of the representations on the final fully connected layers of several CNN models, finding moderate correlations for VGG-16 (Simonyan & Zisserman, 2014) in particular. Subsequent research combined the low-level visual processing of CNNs with cognitive science models of decision-making. These hybrid models were evaluated against data collected on CIFAR-10 test images (Battleday, Peterson, & Griffiths, 2020; Singh, Peterson, Battleday, & Griffiths, 2020).

The most recent work in this area evaluated ‘stacked’ methods for approximating human similarity judgments (Marjeh, Sucholutsky, van Rijn, Jacoby, & Griffiths, 2023) This has shown the value-added of cognitive science models but has not addressed the typicality effect.

Most relevant is Upadhyay et al. (2022), who investigated whether the CNN model VGG-19 shows typicality gradients. This proof-of-concept study focused on the well-studied Bird category, finding only small (0.32) and non-significant correlations between model-predicted and human typicality ratings. Thus, it remains an open question whether vision models can account for the typicality effect observed in humans.

Research Goals Despite recent work, large gaps remain in our understanding of whether ML models trained on large data sets acquire, purely through experience, conceptual representations that resemble those of humans. The current study addressed these gaps through the lens of the typicality effect. There were four research goals:

1. To evaluate the alignment between a large number of language models of varying architectures/sizes and recently collected human typicality ratings across a large number of categories.
2. To do the same for a large number of vision models of varying architectures/sizes.
3. To evaluate, for the first time, whether combining the predictions of a language model and a vision model offers a better account of human typicality than either model alone.
4. To evaluate, for the first time, whether a multimodal model

offers better predictions of human typicality than models of a single modality (vision or language).

The current study also makes two methodological contributions. The first is to evaluate ML models across a broader range ($N = 27$) of categories than has previously been considered, using human data that has been collected in the past few years rather than decades ago. The second is to develop a new set of ‘naturalistic’ images to test the conceptual alignment of vision models.

Methods

Data Preparation

Human Typicality Ratings We used the human typicality ratings of (Castro et al., 2021). This is the most recent dataset of its kind, and it supersedes the norms (Rosch, 1975) used in many prior studies of the alignment of language and vision models to human categorization. 250 participants provided exemplars of each of the 70 categories – as many as possible within a 30-second time frame. The typicality of an exemplar was defined as the proportion of participants who produced it. We selected 27 concepts whose exemplars have concrete and distinct visual depictions, to be able to evaluate the vision models.

Image Collection and Processing Images were collected via the Google Image Search package (arrlo, 2022). For each exemplar of each category, we used its label as a search string and collected at least 20 images (after removing corrupted images and images with unusable file formats). We used the CarveKit Image Background Removal Tool (OPHoperHPO, 2022) to remove the backgrounds of all images where the background was distinct from the exemplar itself, e.g., removing the sky and tree branches from an image of the robin exemplar of the Bird category. Removed backgrounds were replaced with a plain white background. Background removal was not performed for the Color, Dwelling, Earth Formation, Fabric, Tree, and Weather categories because the exemplars were often inseparable from the backgrounds. We manually reviewed the outputs of automated image collection and background removal and discarded images that did not depict the intended exemplar and those with improper background removals. These images were converted to the JPG file format, resized to 224 x 224 pixels, and normalized using the ImageNet mean and standard deviation values. Our image collection and processing code is public, but please contact the authors for access to the specific image set used for the experiments discussed in this paper.¹

Model Selection

Language Models We selected several pre-trained language models spanning a range of architectures: word2vec (Mikolov, Chen, Corrado, & Dean, 2013), GloVe (Pennington, Socher, & Manning, 2014), RoBERTa-large

¹We make all our code publicly available at <https://github.com/svemuri8/cv-nlp-typicality/tree/main>

(Y. Liu et al., 2019), XLNet-base (Yang et al., 2020), MiniLM (Wang et al., 2020), MPNet (Song, Tan, Qin, Lu, & Liu, 2020), T5-large (Raffel et al., 2020), and GPT (text-embedding-ada-002) (Brown et al., 2020). All models lacked classification heads or decoders. Instead, they produced word/sentence embedding vectors. The GloVe and word2vec implementations were sourced from Gensim (Rehurek & Sojka, 2010) and the GPT embeddings from OpenAI’s API (OpenAI, 2023b). All other models are from HuggingFace’s Transformers Library (Reimers & Gurevych, 2019).

Vision Models We selected several vision models pre-trained on ImageNet1K (Deng et al., 2009) spanning different architectures: AlexNet (Krizhevsky, 2014), VGG19 (Simonyan & Zisserman, 2015), InceptionV3 (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2015), ResNet-50 (He, Zhang, Ren, & Sun, 2015), DenseNet-161 (Huang, Liu, van der Maaten, & Weinberger, 2018), MobileNetV2 (Sandler, Howard, Zhu, Zhmoginov, & Chen, 2018), EfficientNetV2-medium (Tan & Le, 2021), ViT-base-16 (Dosovitskiy et al., 2021), Swin-base (Z. Liu et al., 2021), and ConvNext-base (Z. Liu et al., 2022). We removed the classification heads, leaving only the feature extractors so that each model outputted a raw feature vector for each image passed in.

Multimodal Models We selected the pre-trained CLIP ViT-large-14 model (Radford et al., 2021) as the representative multimodal model for our investigation. We used this model to generate image and text embeddings and logit scores indicating alignment between text-image input pairs. The model implementation was sourced from HuggingFace.

Task Paradigms

Language Model Task To estimate the typicality of an exemplar of a category in a language model, we encode the exemplar name as a string, pass it through the language model, and obtain the corresponding word embedding. We then calculate the cosine similarity, as found by Bhatia and Richie (2022) to be the best metric, between this exemplar vector and that of the category prototype, with a higher value indicating that the exemplar is more typical. There are two natural methods for obtaining the prototype of a category: as the average of all exemplar vectors, and as the word embedding obtained by passing the category label to the language model. We explored both methods, finding comparable results. We adopt the former method to maintain consistency with prototype computation for the vision models and with prior work (Heyman & Heyman, 2019).

To evaluate a language model’s alignment for a given category, we compute the Spearman correlation between the cosine similarities (representing the typicality judgments of the language model) with the human typicality rankings of exemplars from Castro et al. (2021).

Vision Model Task Recall that we collected several images of each exemplar. We pass each image through the vision model, obtain an image embedding, and across these compute the *average* exemplar vector. The prototype vector for the category is defined as the mean of all of these (average) exemplar vectors. We compute the typicality of each exemplar in a category as the cosine similarity between its (average) exemplar vector and the mean prototype vector. We take these as the typicality judgments of the vision model and compute the Spearman correlation with the human typicality rankings from Castro et al. (2021).

Pilot Exploration: Single vs Multiple Image Exemplar Representations Choosing to compute the exemplar vector as an *average* of multiple image vectors ensures that our results are not tuned to the choice of a specific exemplar image, and increases the chances that they generalize across images. To justify this approach, we conducted a pilot experiment using the VGG-19 vision model and the Bird category. Using our average vector approach, the Spearman correlation between the model’s typicality ratings and those of humans was 0.242. We then examined the consequences of instead using a single image for each exemplar, running 100 trials where we randomly selected one image for each exemplar and recomputed the correlation. These ranged from -0.247 to 0.469, with an average of 0.094. Thus, we conclude that using only one image of each exemplar produces unstable (and artificially low) correlations with human typicality ratings.

Combined Model Task To address the third research question, we evaluate each language + vision model pair. Specifically, for each category, we fit a linear model predicting the typicality of an exemplar from its prototype (i.e., the cosine similarity between its vector representation and that of the prototype) in the language model and its prototype in the vision model. We record the (standardized) Beta weight of each predictor variable and evaluate the respective contributions of each modality. We also capture the Spearman correlation coefficient between the predicted rank-ordering of the exemplars and the human typicality ranking from Castro et al. (2021). These values are used, respectively, to assess the respective contributions of language versus vision models in making these predictions and to determine which model pairs perform best in modeling human typicality ratings.

Multimodal Model Task The CLIP ViT model produces an embedding for each of the modalities (i.e., vision and text inputs) and outputs a logit score representing their alignment in embedding space. We use text and image-based embeddings to generate category and mean prototypes, mirroring the earlier approaches explored for the language and vision model tasks. We define the category prototype as the embedding of the category input as text and the mean prototype as the average of all representative exemplar image vectors. For the *category prototype approach*, the direct text embedding of

the exemplar is taken to be the exemplar representation. For the *mean prototype approach*, the average of all the image embeddings of an exemplar is taken to be the exemplar representation. As with the language and vision tasks, human-model alignment was computed as the Spearman correlation of human typicality ratings and the cosine similarity between exemplar vectors and the corresponding prototype vector for the category (representing model typicality).

We evaluate two additional approaches for this task. The first looks at appending the vectors of different modalities. In this *appended prototype approach*, we define an exemplar representation as the concatenated exemplar representations from the category and mean prototype approaches, producing a vector that is the concatenation of the projections of exemplar text embedding and average exemplar image embedding into joint CLIP embedding space. The typicality alignment is computed as before, with the Spearman correlation between human typicality and the cosine similarity between exemplar and prototype vectors.

In the final, *cross-modality approach* approach, we leverage the CLIP model’s computation of logit scores to represent the alignment between image and text representations. For this approach, we pass an exemplar image and the category name as an image-text pair input taken by CLIP. The produced logit score represents the alignment between modalities. It is then averaged for all exemplar images to provide the overall alignment score between the exemplar and the category. We calculate the alignment of model and human typicality for each category as the Spearman correlation between the exemplar logit scores and human typicality ratings.

Results

Language Models

Model	Mean	Stdev
all-MiniLM-L12-v2	0.429	0.153
all-mpnet-base-v2	0.424	0.185
all-roberta-large-v1	0.274	0.208
glove-twitter-200	0.421	0.186
sentence-t5-base	0.327	0.133
sentence-t5-large	0.373	0.251
sentence-t5-xl	0.402	0.279
sentence-t5-xxl	0.406	0.215
text-embedding-ada-002	0.304	0.121
word2vec-google-news-300	0.222	0.166
xlnet_base_cased	0.094	0.106

Table 1: Mean and standard deviation for Spearman correlations across all 27 categories by language model.

We first consider the alignment of the language models. Averaging the Spearman correlations across all 27 categories for each model, we observe a range across the models, with a maximum of 0.429 for MiniLM and a minimum of 0.094 for XLNet; see Table 1. The mean correlation across the models

is 0.259 ($SD = .165$). Notably, all models achieve a positive average correlation, signaling general alignment between their predicted typicalities and those of humans.

Although MiniLM performed best among the language models, we note that GloVe, the second oldest model tested, achieved comparable performance ($\rho = 0.421$). This surprising result is consistent with the (Bhatia & Richie, 2022) study of the earlier generation language models that we also evaluated: among them, GloVe best accounted for the pairwise (exemplar-exemplar) similarity ratings made by humans.

Vision Models

Model	Mean	Stdev
alexnet	0.140	0.223
convnext_base	0.058	0.186
densenet161	0.051	0.165
efficientnet_v2_l	0.074	0.153
efficientnet_v2_m	0.053	0.215
efficientnet_v2_s	0.081	0.207
inception_v3	0.015	0.223
mobilenet_v2	0.067	0.214
resnet50	0.037	0.203
swin_b	0.058	0.226
vgg19	0.109	0.224
vit_b_16	0.069	0.230
vit_h_14	0.146	0.166
vit_l_16	0.077	0.204

Table 2: Mean and standard deviation for Spearman correlations across all categories by vision model.

We next consider the alignment of the vision models. Averaging the Spearman correlations across all 27 categories for each model, we see that the vision models show lower alignment than the language models. The average correlation is only 0.0365, ranging from a maximum of 0.1463 (ViT-Huge) to a minimum of 0.0148 (Inceptionv3); see Table 2. Like the language models, the vision models all showed positive average correlations, but their small size indicates a much weaker alignment to human typicality ratings. Paralleling what was found for the language models, there was surprising parity between newer and older models. AlexNet, the oldest vision model architecture we considered, performed nearly as well ($\rho = 0.140$) as the best-performing vision model ViT-Huge ($\rho = 0.146$), and substantially better than VGG-19, the third highest-performing model ($\rho = 0.101$).

Combined Models

We paired all language models with all vision models and, for each pair, combined the typicality predictions of each model in a linear model to predict the human typicality ratings for each category. As expected (because adding predictor variables never decreases model fit), the paired models achieved higher correlations than the modality-specific models: compare Figure 1 against Tables 1 and 2. Interestingly, the combination of the best-performing vision model (ViT-Huge) and

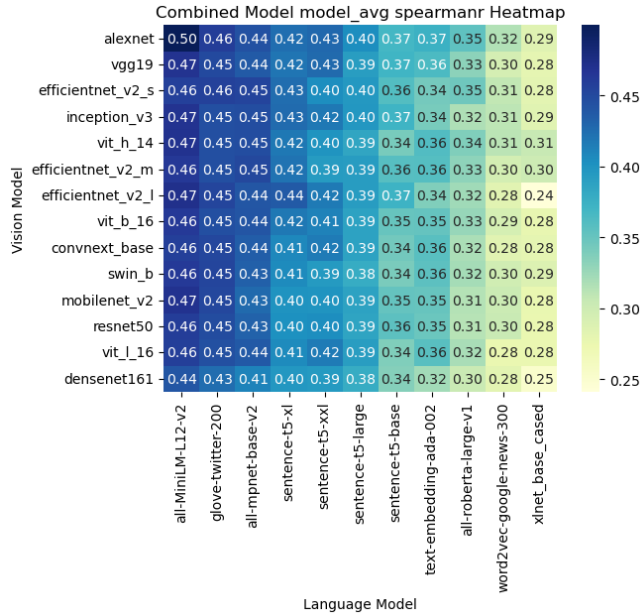


Figure 1: For each (language, vision) model combination, the Spearman correlation between its predicted typicalities and the human typicalities, averaged across all categories.

the best-performing language model (MiniLM) did not have the highest correlation with the human typicality data among all of the model pairs. Instead, the combination of MiniLM with AlexNet showed the highest correlation at 0.4995. This indicates that the two models make differential (vs. overlapping) contributions to predicting the typicality of exemplars.

Finally, we evaluated whether the typicalities of some categories might be driven more by vision than language. We focused on the best-performing combined model, which is AlexNet + MiniLM. Figure 2 shows the Beta weights for each of the linear models predicting the 27 categories. Interestingly, for this combined model, the vision variable contributes to predicting typicality for many of the categories. This is most strikingly the case for *Kitchen Utensil* and *Weather*.

To aid reader comprehension for Figure 2, we organized the concepts into 7 supercategories and assigned them to a designated shape: Environment (Triangle), Abstract (X), Vehicle (Diamond), Man-Made Miscellaneous (Square), Plant (Plus), Animal (Circle), Man-Made Tool (Upside Down Triangle), and Garment (Star).

Multimodal Model

Table 3 shows the results of the multimodal approaches. The mean prototype approach yields a mean Spearman correlation coefficient that is larger ($r = 0.265$) than that of the leading vision model ($r = 0.146$), showing that introducing information through alignment with text embeddings was able to align image representations more closely with human concepts.

The performance of the category prototype approach is close to the results of the best language models of similar size,

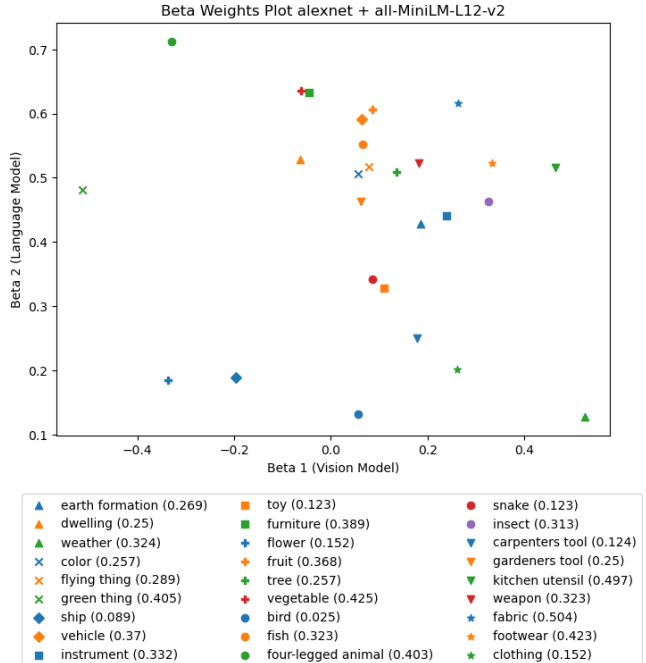


Figure 2: Beta weights of the linear models predicting the typicalities of each category for the best-performing combined model (AlexNet + MiniLM).

showing that at the very least, introducing information from the vision modality does not adversely affect performance.

The appended prototype approach does not improve alignment with human typicality judgments compared to the category prototype approach. Again, it is worth noting the silver lining here: that explicitly adding exemplar image representations did not harm the overall alignment of exemplar representations with those of humans.

Finally, the cross-modality approach, with its subpar results, suggests that there may exist a fundamentally nonhuman conceptual gap between image and text representations in the joint embedding space of the CLIP model.

Approach	Mean	Stdev
category	0.412	0.164
mean	0.265	0.169
appended	0.413	0.164
cross-modality	0.095	0.174

Table 3: Mean and standard deviation for Spearman correlation across all categories for CLIP ViT.

Discussion

Summary of Findings

The current study addressed four primary research goals. The first was to evaluate whether language models have exemplar representations that show similar typicality gradients as those

documented for humans. This was the case, with MiniLM achieving the highest correlation ($\rho = 0.429$) – one higher than has been observed in prior studies that used earlier-generation language models, older human data sets, and a narrower range of categories (Misra et al., 2021). The second goal was to ask the same question of vision models. Only one prior study (Upadhyay et al., 2022) has addressed this goal, and for only one model (VGG-19) and one category (i.e., Birds). The modest correlation documented there held for the broader range of vision models and categories investigated here. The best-performing model, ViT-Huge, produced typicality predictions that correlated only modestly ($\rho = 0.1463$) with those of humans. The third goal was to examine, for the first time, whether combining language and vision models – consistent with the multimodal nature of cognition – produces even better predictions. This was indeed the case, with the AlexNet + MiniLM pairing achieving a 0.4995 correlation with the human typicality ratings. The final goal was to examine the alignment of an inherently multimodal model with human typicality judgments. We found that there was a sizable correlation between the two, with the mean approach using the vision portion of the model ($\rho = 0.2645$) showing far more promising results than vision models trained on image data alone, and the category approach using the language portion of the model ($\rho = 0.4115$) showing equal if not better results than models trained purely on text data.

Taken together, these findings advance the state of the art in aligning the conceptual representations of ML models and humans. An additional methodological contribution is the creation of a new image set for testing the conceptual alignment of vision models.

Limitations and Future Directions

The limitations of the current study lead naturally to future directions for research, so we discuss both together.

Our investigation of only one multimodal model (CLIP ViT) may not be appropriately representative of the full potential of the multimodal approach, and in more advanced multimodal models, the different modalities may complement each other even more strongly to form better conceptual representations than were observed here. For example, an image of a live chicken should lead to a textual distribution shift where the combined representation of chicken is closer to the representation of a Bird than the representation of chicken as a food in the embedding space. Further work on multimodal contextual alignment for concepts is a goal for future research.

The current study lacks a thorough investigation of large generative language models like GPT-4 (OpenAI, 2023a) and LLaMA (Touvron et al., 2023) which have shown strong performance for a large variety of tasks. Future work can investigate prompt-based cognitive modeling of the typicality effects seen in humans using these and similar models.

An important question is why the vision models show lower alignment with human typicality judgments than the language models. A possible explanation for their lower

alignment is that a more significant part of image representations are comprised of local information like texture. By contrast, humans rely more on overall shape when making categorization decisions (Kurbat, Smith, & Medin, 2019; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Conversely, the superior performance of the language models may be attributed to the text corpora on which they are trained mentioning the attributes of exemplars in the same context as their categories. This may enable them to better capture attribute frequency, which is known to be highly correlated with typicality (Rosch, 1975). Future research should explore these and other explanations for the difference performance levels of language and vision models.

Another limitation may have been our choice to process the images and replace their natural backgrounds with white backgrounds. We did so to avoid the models forming blended representations rather than representations of single exemplars. For example, consider an image of a sparrow (from the Bird category) perched on an oak tree (from the Trees category). However, it is the case that the processed images are different from the naturalistic images on which the models were trained, which may have affected the results of the experiments. In the trade-off between natural images and images with a single object, we chose to focus on the latter. Future research could examine the implications of this choice.

A final limitation concerns the human data on the typicality effect. The Castro et al. (2021) study gives typicality rankings for the exemplars (e.g., robin) of categories (e.g., Birds). To derive typicality predictions from the vision models, we sampled 6-11 images of each exemplar using Google Images and averaged together their vector representations on the final fully-connected layer. To reduce the noise in this average exemplar representation, a future study could collect human typicality ratings on the exact images provided to the vision models. This would enable less noisy evaluation of the typicality gradients of vision models and potentially increase their alignment to human typicality rankings.

References

- arrlo. (2022). *Google-images-search: [python] search for image using google custom search api and resize & crop afterwards*. <https://github.com/arrlo/Google-Images-Search>. (Accessed: 2023-08-14)
- Battig, W. F., & Montague, W. E. (1969). Category norms of verbal items in 56 categories a replication and extension of the connecticut category norms. *Journal of Experimental Psychology Monographs*, 80, 1-46.
- 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, 5418.
- Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2021). From convolutional neural networks to models of higher-level cognition (and back again). *Annals of the New York Academy of Sciences*, 1505, 55-78.

- Bhatia, S., & Richie, R. (2022, 10 27). Transformer networks of human conceptual knowledge. *Psychological Review*.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.
- Castro, N., Curley, T., & Hertzog, C. (2021). Category norms with a cross-sectional sample of adults in the united states: Consideration of cohort, age, and historical effects on semantic categories. *Behavior research methods*, 53(2), 898–917. doi: 10.3758/s13428-020-01454-9
- De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other dutch normative data for semantic concepts. *Behavior research methods*, 40, 1030-1048.
- 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* (pp. 248–255).
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Hounsby, N. (2021). *An image is worth 16x16 words: Transformers for image recognition at scale*.
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep residual learning for image recognition*.
- Heyman, T., & Heyman, G. (2019). Can prediction-based distributional semantic models predict typicality? *Quarterly Journal of Experimental Psychology*, 72, 2084-2109. Retrieved from <https://doi.org/10.1177/1747021819830949> doi: 10.1177/1747021819830949
- Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2018). *Densely connected convolutional networks*.
- Krizhevsky, A. (2014). *One weird trick for parallelizing convolutional neural networks*.
- Kurbat, M. A., Smith, E. E., & Medin, D. L. (2019). Categorization, typicality, and shape similarity. In *Proceedings of the sixteenth annual conference of the cognitive science society* (pp. 520–524).
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). *Roberta: A robustly optimized bert pretraining approach*.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... Guo, B. (2021). *Swin transformer: Hierarchical vision transformer using shifted windows*.
- Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). *A convnet for the 2020s*.
- Marjeh, R., Sucholutsky, I., van Rijn, P., Jacoby, N., & Griffiths, T. L. (2023). Large language models predict human sensory judgments across six modalities. *arXiv preprint arXiv:2302.01308*.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*.
- Misra, K., Ettinger, A., & Rayz, J. T. (2021). Do language models learn typicality judgments from text? *arXiv preprint arXiv:2105.02987*.
- Murphy, G. (2002). *The big book of concepts*. MIT press.
- OpenAI. (2023a). *Gpt-4 technical report*.
- OpenAI. (2023b). *New and improved embedding model*. <https://openai.com/blog/new-and-improved-embedding-model>. (Accessed: 2023-08-14)
- OPHoperHPO. (2022). *image-background-remove-tool: A tool for removing the background from images and video*. <https://github.com/OPHoperHPO/image-background-remove-tool>. (Accessed: 2023-08-14)
- 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. Retrieved from <https://aclanthology.org/D14-1162> 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, 2648-2669.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140), 1–67.
- Rehurek, R., & Sojka, P. (2010). *Software framework for topic modelling with large corpora*. <https://radimrehurek.com/gensim/index.html>. (Accessed: 2023-08-14)
- Reimers, N., & Gurevych, I. (2019). *Sentence transformers: Multilingual sentence embeddings using bert / roberta / xlm-roberta & co. with pytorch*. Retrieved from <https://huggingface.co/sentence-transformers> (Accessed: 2023-08-14)
- Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3), 192.
- Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8, 382-439.
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510–4520).
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Simonyan, K., & Zisserman, A. (2015). *Very deep convolutional networks for large-scale image recognition*.
- Singh, P., Peterson, J. C., Battleday, R. M., & Griffiths, T. L. (2020). End-to-end deep prototype and exemplar

- models for predicting human behavior. *arXiv preprint arXiv:2007.08723*.
- Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). *Mpnet: Masked and permuted pre-training for language understanding*.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). *Rethinking the inception architecture for computer vision*.
- Tan, M., & Le, Q. (2021). Efficientnetv2: Smaller models and faster training. In *International conference on machine learning* (pp. 10096–10106).
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... Lample, G. (2023). *Llama: Open and efficient foundation language models*.
- Upadhyay, N., Mittal, K., & Varma, S. (2022). Typicality gradients in computer vision models. *Proceedings of the Annual Meeting of the Cognitive Science Society, 44*.
- Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the norms. *Journal of Memory and Language, 50*, 289-335.
- Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). *Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers*.
- Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2020). *Xlnet: Generalized autoregressive pretraining for language understanding*.