Evaluating Neural Networks as Models of Human Language-Vision Feedback

Corey Wood

Advisor: Thomas Serre
Reader: Drew Linsley

Department of Computer Science
Brown University
Providence, RI
May 2024
Acknowledgments

Thank you to my research advisor, Thomas, for his support in guiding me through my first significant research project.

Thank you to my reader, Drew, for his constant support throughout the project and for answering my countless Slack messages.

A huge thank you to my collaborators on this project, Jay and Tony. Jay’s tireless efforts to drive the project forward were essential in making this work happen, and Tony’s modeling expertise greatly improved the pace we worked at and the quality of our results.

Thank you to Professor Josiah Carberry, whose revolutionary work in pioneering the field of psychoceramics inspires me in my own research efforts.

And finally, a big thank you to my friends and family, who have supported me throughout the year in countless ways.
Abstract

The modulatory effect of language on visual perception has been demonstrated in multiple domains, but the mechanisms behind the neural circuits governing this interaction remain unclear. Computational models that exhibit similar impacts of language on visual processing as humans could help to elucidate plausible neural mechanisms underlying this effect. In this work, we evaluate how current state-of-the-art deep neural networks compare to humans on Mooney image classification, a classic psychophysics task for investigating language-vision feedback. We demonstrate that these models generally fail to match human performance and appear to use different strategies for the task. However, we find that two popular vision-language models may represent a promising avenue for building computational models of human language-vision feedback. Our work highlights the importance of developing methods that enable direct observation of the way that language modulates visual representations within vision-language models.

1 Introduction

Information derived from language has been demonstrated to impact human visual perception. Hearing or reading words appears to automatically activate visual features of the entities they describe, affecting visual perception through a top-down feedback mechanism. This effect of language on vision has been demonstrated across psychophysics experiments in multiple visual domains, particularly in noisy or ambiguous conditions. For example, Lupyan and Ward found that the presence of linguistic labels improved humans’ ability to detect the presence of an otherwise unseen object. Dils and Boroditsky demonstrated that unrelated language can alter recognition of an ambiguous visual scene. Verbal labels, even of already-known information, have also been shown to improve the efficiency of visual search. Though the impact of linguistic information on visual perception is clear, there is little known about the neural circuits that underlie these interactions.

One approach to understanding these neural circuits is to build computational models that mirror human behavior. For example, Linsley et al. developed a recurrent neural network architecture that approximates visual cortical circuits, elucidating plausible behavior of human neural circuits. If a computational model exhibits a similar impact of language on visual processing as in humans, it might represent a plausible model of the human neural mechanisms that govern this interaction. Understanding the extent to which current state-of-the-art neural networks are good models of human vision is therefore relevant for increasing our understanding of human neural circuitry.

In recent years, deep neural networks (DNNs) have become increasingly capable at a variety of language and vision tasks. Current state-of-the-art models rival and sometimes even exceed human performance on tasks such as object recognition and visual question
These increases in performance have been attributed to large increases in the number of parameters, the amount of data, and the amount of compute used to train models. Achieving human-like performance is a good first step, but to be accurate models of human vision, these DNNs should also use similar strategies as humans to solve these tasks. Unfortunately, previous work has shown that DNNs trained for image categorization reach the same performance as humans using qualitatively different strategies. This misalignment has worsened with scale. If we want to build computational models of biological vision, it’s clear that current state-of-the-art approaches won’t get us there.

Recently, increasing attention has been given to the idea of using language to ground vision models. Motivated by the desire to enable vision models to make use of information in textual data, Radford et al. developed Contrastive Language-Image Pretraining (CLIP), a model that learns unified representations of images and text. This pretraining paradigm has proved highly successful at producing performant and robust models. Spurred by this success, recent vision language models (VLMs) have combined pretrained language and vision models to improve the models’ ability to reason over visual input and enable them to be prompted with natural language questions. These models rival and sometimes surpass human performance on tasks such as visual question answering.

Does this increased performance come with increased similarity to human mechanisms of visual perception? Here, we investigate if these vision-language model architectures can drive progress in modeling – and eventually understanding – the neural feedback circuits that enable language to modulate visual perception.

To evaluate the feasibility of modern DNNs as models of human language-vision feedback, we employed a psychophysics task that has previously been used to measure these effects in human behavior. For example, consider Figure 1: what object does this image depict? “Mooney” images are two-tone black and white images that make ambiguous the underlying object depicted. Previous studies have used Mooney images to demonstrate the effect of perceptual hints on humans’ ability to recognize the otherwise ambiguous images. Samaha et al. constructed a set of Mooney images by blurring and converting to black and white a collection of photographs of common objects. They found that priming participants with the superordinate category of an object boosted their recognition performance in a free-naming condition to the level found in a multiple-choice (of basic-level categories) condition. In this work we evaluate how a large zoo of DNNs compare to humans in classifying Mooney images, and investigate whether the DNNs capable of language/vision feedback exhibit human-like behavior when prompted the same way as humans.

**Contributions** We evaluate the performance of current state-of-the-art deep neural networks, including both vision-only and vision-language models, on classification of ambiguous Mooney images. These models in general fail to match human performance and appear to use different strategies than humans to solve the task. Additionally, we find that vision-language supervision during pretraining does not seem to significantly improve perfor-
mance or alignment with humans on this task compared to vision-only models. Two popular
vision-language models we evaluate exhibit a similar pattern of language impact to humans,
suggesting that they are a potentially promising avenue for building computational models
of human language-vision feedback. We highlight the importance of developing methods
that allow direct observation of how language modulates visual representations within such
models.

This work was done in collaboration with Jay Gopal and Pinyuan Feng.

2 Related Work

2.1 Similarity between vision models and humans

Geirhos et al. 11 found that modern state-of-the-art DNNs come close to and in some cases
beat human performance across a variety of datasets. However, a growing body of work has
demonstrated a fundamental misalignment between how DNNs trained for image categoriza-
tion and humans perceive the world. In cases of object recognition, DNNs that match or
surpass human performance utilize different visual features for decision-making compared to
humans.3,7,11,19,25 For example, DNNs have been shown to rely heavily on image textures for
making classification decisions, in sharp contrast to humans who rely heavily on shapes.12

Some of this misalignment with humans has worsened as performance on traditional
benchmarks has increased. Compared to previous versions, DNNs capable of matching
human-level performance are less accurate at predicting visually-evoked behaviors in humans
and neural responses in primates.24,39 DNNs with higher image classification accuracy are
also more susceptible to “adversarial attacks” which modify features in ways that are not
salient to humans. These results suggest that the increasingly large misalignment between humans and DNNs cannot be addressed by scaling the amount of data, parameters, and compute used to train models.

Because of their recency, the similarity of visual processing in VLMs to that of humans is largely uninvestigated. Recently, Gavrikov et al. demonstrated that popular VLMs can be steered with language to have higher bias toward shape information in images, making them behave more similarly to humans than un-modulated vision encoders. Our work investigates a similar question in the domain of image classification on ambiguous Mooney images.

3 Methods

3.1 Human psychophysics

We utilized experimental data made publicly available by Samaha et al., which tested human participants on Mooney images of objects. Given an image, participants were asked to identify the object it depicted in one of three different ways: (i) **Free naming** (FN): name the object category with no guidance, (ii) **Basic-level forced choice** (BLFC): choose from among 15 possible categories, or (iii) **Free naming with a superordinate cue** (SO): name the object after receiving a hint about the category it belongs to.

3.2 Deep neural network zoo

We evaluated three different types of DNNs: (i) **Vision-only**: 51 ImageNet trained models from the PyTorch Image Models (TIMM) library, including vision transformers (ViTs) and convolutional neural networks (CNNs). (ii) **Vision-language supervision**: 23 CLIP-pretrained vision-language models from the TIMM library, with either ViT or CNN base architectures. These models were trained to jointly learn representations of images and an accompanying text description, thus providing their visual representations with supervision from language. (iii) **Vision-language crosstalk**: four vision-language models: Claude 3 Sonnet, GPT-4, Gemini 1.0 Ultra, and InstructBLIP. These models make use of pretrained image encoders and large language models to enable visual reasoning modulated by language. The vision-language crosstalk models take text and images as input and output text tokens. Claude, GPT-4, and Gemini are current state-of-the-art multimodal chat models that are highly capable at a number of tasks. These models were accessed through their respective web interfaces. InstructBLIP is a significantly smaller model than the others and was chosen because it is open-source and achieves high performance on a variety of visual reasoning tasks. We used the smaller FlanT5-XL version of InstructBLIP due to hardware constraints.

3.3 Evaluating vision-only and vision-language supervised models on BLFC

We measured the classification accuracy of the vision-only and vision-language supervised DNNs on the Mooney image BLFC task described in Samaha et al. In the original
experiment, humans were shown a set of 15 images and were asked to choose among 15 basic-level names for which object they thought was in each image.

The models were originally trained to classify 1000 classes from ImageNet. To allow them to predict over the 15 classes present in the BLFC task, we replaced the final layer of each of the models with a simple Random Forest classifier. An assumption behind the use of Mooney images is that subjects have not had an opportunity to become experienced at categorizing them. To ensure this held true for the models, we could not train the classifiers on Mooney images. We therefore constructed a mapping between classes used in the experiment and analogous ImageNet classes – for example, “guitar” maps to the ImageNet class “acoustic guitar.” Three out of 15 of the Mooney images used in the original experiment had no analogous ImageNet classes and so were excluded. We first trained the classifiers on the ImageNet images corresponding to the 12 remaining Mooney images, keeping the rest of the models frozen. We then used the classifiers to make predictions on the Mooney images.

3.4 Evaluating vision-language crosstalk models

Because the vision-language crosstalk models take text and images as input and output a sequence of text tokens, they could be prompted and evaluated using the same procedure as humans. In each task, we evaluated models on the 15 images used by Samaha et al. for the SO task. These 15 images were selected by Samaha et al. from the images that were most difficult for humans to classify, meaning our test set represented the most challenging images.

For the Free naming task, we asked the models what they saw in the image. For InstructBLIP, we used the prompt, “Question: What is the object shown in this image? Answer:”. For the other vision-language crosstalk models, we found that this prompt caused them to output abstract interpretations of the image rather than a possible object. To obtain actual guesses, we therefore used the prompt, “Give a short answer. Question: What basic level label do you see in this image? You may NOT say it is abstract, black and white, or has blobs or splatters. You MUST answer with something specific that you can discern from the image. Answer:”. This prompt more consistently produced guesses that could be evaluated for accuracy.

For the Basic-level forced choice task, we asked the models what they saw in the image and directed them to choose from one of the provided answer options. Following the questioning format of Gavrikov et al., we asked the model “Which option best describes the image?”, followed by an enumeration of all class labels in the form “A. guitar B. desk …”, followed by “Answer with the option’s letter from the given choices directly” to ensure the model produced only the chosen letter. InstructBLIP had more trouble following the directions to choose an option, likely because of its smaller size and lesser amount of instruction tuning. We found that the order the answer options were presented in had a significant impact on InstructBLIP’s performance. We therefore evaluated it over 50 trials.
with the order of answer options randomized each time and averaged the accuracy of all trials.

For the Free naming with a superordinate cue task, we prompted the models in an identical way as in Free naming, “Question: What is the object shown in this image? Answer:”, but appended a hint after the question in the form of “Hint: musical instrument.” To obtain the superordinate cues, we went two levels up from the target word in the WordNet hierarchy.8

3.5 Human alignment of vision-only and vision-language supervised models

To gain additional insight beyond accuracy, we measured the correlation of the vision-only and vision-language supervised models’ per-image decisions on the BLFC task with humans. Alignment between vision-language crosstalk models and humans was not evaluated because these models output text tokens rather than class probabilities, and so cannot be easily compared to human responses.

Correlations were computed as “error consistency” using Cohen’s $\kappa$, following the procedure outlined by Geirhos et al.10 In contrast to accuracy, which measures the number of errors, error consistency quantifies the extent to which two subjects systematically make errors on the same inputs. Error consistency therefore provides insight into how similar the perceptual strategies used by each subject are.

A value of $\kappa = 0$ indicates consistency at chance-level, signifying that the subjects are likely using different processing strategies. $\kappa > 0$ indicates consistency beyond chance, suggesting similar strategies, and $\kappa < 0$ indicates inconsistency beyond chance, suggesting inverse strategies. To calculate a model’s error consistency with humans, we computed its $\kappa$ score with each human subject and took the average. We also computed average human to human error consistency by computing the mean of the $\kappa$ score between each human subject. This number represents the maximum consistency with humans that a model could be expected to have.

4 Results

4.1 How do vision-only and vision-language supervised models compare to human performance on the BLFC task?

We evaluated 51 vision-only and 23 vision-language supervised models from the TIMM library on the BLFC task. Figure 2 plots their Mooney image classification accuracy against their top-1 accuracy on ImageNet. The dashed red line indicates human accuracy on the same task, with the shaded area representing the bootstrapped 95% confidence interval.

Models were not able to match human accuracy, with the best performing models reaching an accuracy of 33.33% compared to the average human accuracy of 50.69%. There is a slight positive relationship between ImageNet accuracy and Mooney image accuracy,
indicating that models improve at Mooney image classification as they get better at general image classification.

Because the vision-only and vision-language supervised models have differences in their architecture and training processes other than the presence of language, they cannot be directly compared. However, we can make a general comparison between the groups. Although vision-language supervised models tend to perform better on ImageNet, they do not perform significantly better on Mooney image classification compared to vision-only models.

4.2 How do vision-only and vision-language supervised models compare to human behavior on the BLFC task?

Beyond accuracy comparisons, we also investigated if models tend to make the same kind of errors as humans. Using the models’ predictions, we evaluated their error consistency with humans using Cohen’s $\kappa$ as detailed in 3.5. Results can be seen plotted against ImageNet accuracy in Figure 3. Bootstrapped 95% confidence intervals are plotted for each model. The purple dashed line represents human to human consistency, with the shaded area representing
Figure 3: Cohen’s \( \kappa \) score of vision-only and vision-language supervised models on the basic-level forced choice task, plotted against ImageNet accuracy. Average human to human \( \kappa \) score is represented by the purple dashed line. Vision-only models are red and vision-language supervised models are teal.

Most models had a \( \kappa \) score of close to 0, indicating chance-level consistency and suggesting the majority of the models may be using different strategies than humans. There is a slight positive trend between \( \kappa \) score and ImageNet accuracy: as models improve on ImageNet, their predictions generally align better with humans. Two ViT models, one Vision-only and one Dual-stream, were able to match human to human consistency.

We again cannot directly compare the vision-only to the vision-language supervised models due to other variations between the models. However, there generally does not appear to be an association between vision-language supervision and higher error consistency with humans.

4.3 Does language impact vision-language crosstalk models in the same way as humans?

We evaluated four vision-language crosstalk models on the FN, BLFC, and SO tasks. Results can be seen in Figure 4. The purple dashed lines represent average human accuracy for each task. Human subjects had fairly low average accuracy on the Free naming (FN) task of 11.17%. Their accuracy significantly increased on the Basic-level forced choice (BLFC) task to an average of 51.68%, indicating that choosing between 15 options enabled human
Figure 4: Performance of vision-language crosstalk models by task. Average human accuracy is represented by the purple dashed lines.

subjects to more easily detect the objects within the ambiguous images. Average human accuracy on the Free naming with a superordinate cue task was 40%, which was lower than accuracy on BLFC but significantly higher compared to accuracy on FN. This shows that the presence of the categorical hint guided visual perception of the human subjects and enabled them to name the object with significantly higher accuracy.

Two of the vision-language crosstalk models exhibited a similar pattern of the impact of language compared to humans, whereas the other two models exhibited a dissimilar pattern. Claude and GPT-4 both had low accuracies on FN, with 0% and 6.67% accuracy respectively. However, both models had a significant increase on the BLFC task, with accuracies of 13.33% and 26.67%. While there is still a gap between these accuracies and human performance on the BLFC task, Claude and GPT-4 mirrored the increase in human accuracy from FN to BLFC, indicating that they were able to make use of the provided answer options to improve their visual reasoning. The models again achieved below human accuracy on SO, with 7.14% accuracy for Claude and 20% accuracy for GPT-4, but achieved comparatively higher performance than they had on FN. This indicates that, like humans, the presence of the language hint improved the models’ ability to classify the images correctly.

In contrast, the presence of language information appeared to impact Gemini and InstructBLIP in different ways than humans. Gemini and InstructBLIP each had higher accuracies on FN than humans, with 26.27% and 13.33% accuracy respectively. This suggests that these models were able to better detect the objects in the Mooney images compared to humans when no language information was provided. However, Gemini and InstructBLIP both saw a decrease in accuracy on the BLFC task, to 13.33% and 0.53%. This result is surprising as
choosing from 15 options should be an easier task compared to naming an arbitrary class. Additionally, Gemini and InstructBLIP also had lower accuracy on SO compared to FN, with SO accuracies of 20% and 6.8%. Compared to humans, who benefitted from the additional language information in the BLFC and SO tasks, Gemini and InstructBLIP saw decreased performance on these tasks compared to FN, indicating that these models were not impacted by language in the same way.

5 Discussion

The modulatory effect of language on visual processing in humans has been demonstrated across various domains. The presence of linguistic information activates certain visual features and aids humans in recognizing ambiguous images. However, there is little known about the neural circuits that govern these interactions. One avenue towards better understanding this effect is to build computational models that mirror human behavior, which could provide insight into plausible human neural mechanisms.

Modern DNNs are capable of matching or exceeding human performance at a variety of language and vision tasks. Performance gains in recent years have largely come from increases in the number of parameters, the amount of data, and the amount of compute used to train models. However, DNNs trained for image categorization have been demonstrated to use different strategies than humans, making them poor models of biological vision.

The last few years have seen an increased interest in grounding vision models with language. Vision-language supervised models like CLIP and vision-language crosstalk models like InstructBLIP have shrunk the gap between human and model performance and enabled a wide range of new applications. However, it has remained unclear whether these DNNs are good models of language-vision feedback in humans.

To address this question, our work evaluated how current state-of-the-art DNNs, across a range of different architectures, perform on Mooney image classification compared to humans. We first investigated how vision-only and vision-language supervised models compared to human performance on the BLFC task. Our results demonstrated that as models have improved on ImageNet, they also have become better at recognizing the ambiguous Mooney images. However, there is still a gap between model and human performance. Additionally, there did not appear to be a significant difference in performance between vision-only and vision-language supervised models, although a direct comparison cannot be made due to other architectural variations between the groups.

To determine how similar the strategies used by the models were to humans, we also evaluated whether models made similar errors to humans. We found that most models we tested had low error consistency with humans, indicating a lack of similarity in strategies. Vision-language supervised models did not appear to be significantly more consistent with humans, although due to the additional variations between models mentioned previously, this is not a conclusive result.
We also evaluated if language impacts vision-language crosstalk models in the same way as humans. Two of the models, Claude and GPT-4, exhibited a similar pattern of language impact compared to humans. These models saw increased accuracy on the BLFC and SO tasks compared to the FN task, indicating that additional language information had a similar effect on them as it did on humans. Conversely, Gemini and InstructBLIP exhibited dissimilar behavior compared to humans with decreased accuracy on the BLFC and SO tasks compared to FN.

This work explored how current state-of-the-art vision models perform on a psychophysics task commonly used to investigate the effect of language on human visual processing. Our results begin to illuminate how vision-language models compare to humans, and we demonstrate the viability of Mooney image classification as a task to evaluate the impact of language on vision models. Future work evaluating the similarity between vision-language models and humans on other psychophysics tasks would be valuable for investigating the impact of language on models in different domains. Additionally, directly comparing models that are identical other than the presence of language supervision during training would provide a clearer picture of the impact of language on these models.

Vision-language crosstalk models appear to be a potentially promising avenue for building models that mirror the ways in which language modulates visual perception in humans. Two of the vision-language crosstalk models we evaluated exhibited a similar pattern of language impact to humans, suggesting that they might capture some of the mechanisms underlying human vision. However, it’s challenging to directly observe the way that language modulates visual representations within these models. Developing more interpretable methods of incorporating language into vision models would make it easier to leverage these models in understanding the neural circuits that underlie human language-vision feedback.
References


