Modeling the Effects of Language on Visual Perception with Deep Learning

Jay Gopal[†] **(Jay_Gopal@Brown.edu)**
Brown University, Providence, RI, 02912, USA

Corey Wood[†] **(Corey_Wood@Brown.edu)**
Brown University, Providence, RI, 02912, USA

Drew Linsley (Drew_Linsley@Brown.edu)
Brown University, Providence, RI, 02912, USA

Pinyuan Feng (Pinyuan_Feng@Brown.edu)
Brown University, Providence, RI, 02912, USA

Thomas Serre (Thomas_Serre@Brown.edu)
Brown University, Providence, RI, 02912, USA

[†] These authors contributed equally to this work.

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. Recently, new approaches have been developed to allow deep neural networks (DNNs) to jointly learn vision and language processing. To investigate if these novel model architectures can help us understand the circuitry of language and vision, we evaluate how a zoo of DNNs compares to humans in classifying binarized Mooney images. We show that as vision-only and dual-stream language/vision feedback models have improved on ImageNet, they have become more accurate at Mooney image classification, but still fail to match human performance. However, we demonstrate that priming a single-stream vision-language DNN with language can cause it to perform similarly to humans. Our results suggest that modern vision-language DNNs represent a new opportunity to generate hypotheses on the neural feedback circuits underlying language's ability to modulate visual representations.

Keywords: multimodality; visual representations; priming; psychophysics; vision; language

Introduction

Over the past decade, the role of linguistic processing in visual perception has become clearer: the words we read and hear automatically activate visual features of the entities that they describe (Lupyan & Ward, 2013). This modulatory effect of language on vision has been demonstrated across psychophysics experiments in multiple visual domains, from the naming of viewed colors (Lupyan et al., 2020) to recognizing degraded objects (Samaha et al., 2018). Despite the undeniable impact of language on visual perception, there is still little known about the neural circuits that govern these interactions.

Deep neural networks (DNNs) have been used to successfully predict the visually-evoked behavior and neural activity of humans on many different visual tasks, including object recognition (Geirhos et al., 2020; Eberhardt et al., 2016; Svanera et al., 2019; Zeman et al., 2020). Over recent years, new approaches have been developed to allow DNNs to jointly learn vision and language processing from large-scale internet data. For example, Contrastive Language Image Pretraining (CLIP) (Radford et al., 2021; Jia et al., 2021; Sun et al., 2023) has been used to train models to learn visual representations from image caption data. This "dual-stream" approach involves using separate architectures for language and vision, and has been highly successful for training foundational models for either modality. Another approach for training DNNs to learn language and vision is Bootstrapping Language-Image Pretraining (BLIP) (Li et al., 2022, 2023). This "single-stream" approach takes images and their captions as inputs, and enables crosstalk between the two representational streams to shape the visual decisions that it makes. Here, we investigate if existing DNN architectures can drive progress in modeling — and eventually understanding — the neural feedback circuits that enable language to modulate visual perception.

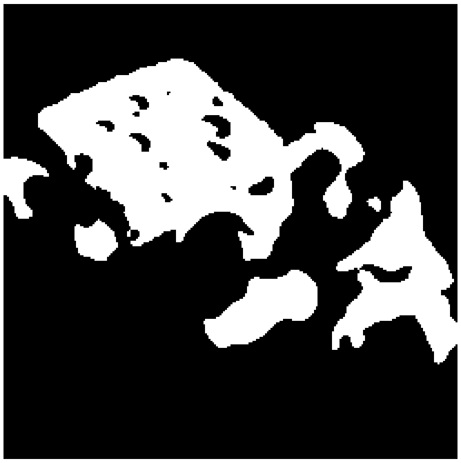

Figure 1: An example Mooney image of ”cheese” included in the psychophysics experiment. Human accuracies for free naming, basic level forced choice, and free naming when given a superordinate cue (foodstuff) were 25%, 95%, and 60%.

To evaluate the feasibility of today's DNNs for modeling language/vision feedback, we turned to a psychophysics paradigm that has previously been used to measure these effects behaviorally in humans. For example, consider Figure 1: what object does this image depict? Mooney images, or binarized object images, have been used since their inception to probe how feedback affects visual perception (Mooney, 1957). More recently, they were used in a psychophysics paradigm that demonstrated how priming participants with the superordinate category of an object could boost 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 test how a large zoo of DNNs compares to humans in classifying Mooney images, and if the subset of those DNNs capable of language/vision feedback act like humans do when they receive language cues for the Mooney images.

Methods and Results

Human psychophysics. We utilized experimental data made publicly available by Samaha et al. 2018, 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, (*ii*) **Basic-level forced choice** (BLFC): categorize the object as one of 15 possibilities, or (*iii*) **Free naming with a superordinate cue** (SO): name the object after receiving a hint about the category it belongs to.

Deep neural network zoo. We compared humans to three different types of DNNs: (*i*) **Vision-only:** 53 ImageNet-trained models from the PyTorch Image Models (*TIMM*) li-

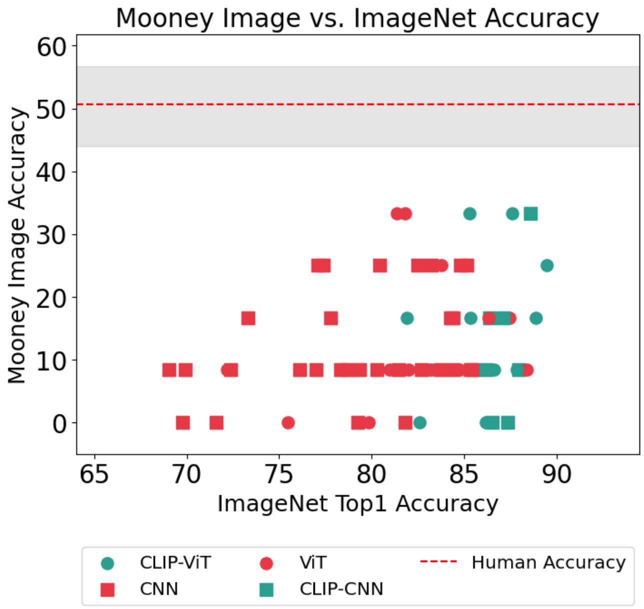

Figure 2: **DNNs tend to be more accurate on Mooney image classification as they improve on ImageNet classification.** Average human accuracy is drawn in red, and the bootstrapped interval is depicted in gray.

brary (Wightman, 2019), including vision transformers (ViTs) and convolutional neural networks (CNNs). (*ii*) **Dual-stream:** 22 CLIP-pretrained vision-language models, each with a different architecture. (*iii*) **Single-stream:** InstructBLIP, a (single-stream) vision-language model that takes text and images as input and outputs text tokens.

Human alignment. We evaluated DNNs as models of human perception of Mooney images by measuring their classification accuracy as well as the correlation of their per-image decisions with humans. Correlations were recorded as "error consistency" using Cohen's κ, following the procedure outlined by Geirhos et al. 2020. Finally, we generated bootstrapped confidence intervals for accuracy and error consistency for each model.

Image decisions were extracted from each type of model through different approaches. We linear probed **Vision-only** and **Dual-stream** DNNs by training a Random Forest classifier on ImageNet images corresponding to each Mooney object image. Decisions and classifier probabilities were taken from each image to compare to humans. In contrast, the **Single-stream** DNN was evaluated using the same procedure used for human participants.

Results. We began by evaluating our DNN zoo for classification accuracy of Mooney object images (Fig. 2). As **Vision-only** and **Dual-stream** DNNs have improved on ImageNet, they have also grown more accurate at Mooney image classification. However, a gap still remains between their performance and humans. Can the Single-stream DNN do better? Indeed, we found that after priming the Single-stream DNN with the same prompts as humans, it exceeded human accuracy and fell within the human-to-human error consistency range (Fig. 3). These results imply that the modulation of visual perception by language that takes place within the Single-stream DNN we used is a good starting point for investigating the neural feedback circuits that enable language to modulate visual perception.

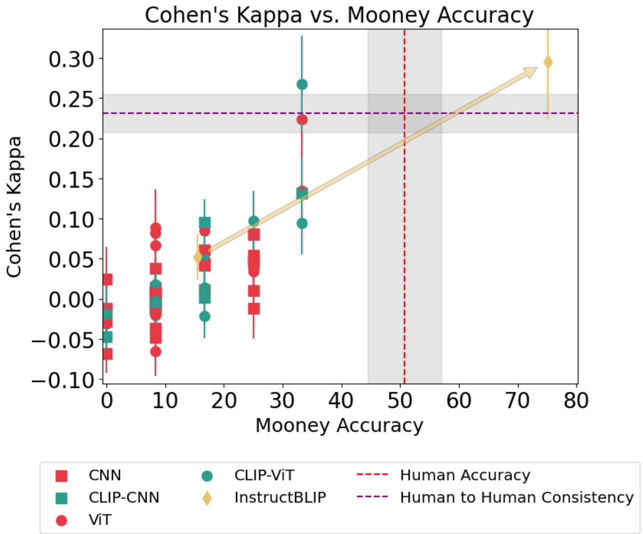

Figure 3: **Single-stream vision-language models can explain human accuracy and decision-making on Mooney images**. The yellow arrow depicts the change in performance and human decision correlation of a single-stream DNN after it is prompted like humans are on Mooney object classification.

Conclusion

The interplay between language and vision is extremely challenging to explore. In this work, we show that today's DNNs trained to model language and vision represent a new opportunity to generate hypotheses on the circuitry of language and vision and how it drives different visual behaviors. Our work represents a first step towards this goal, and we believe additional progress will require the development of architectures that are more interpretable and more easily mapped to known neural systems than the **Single-stream** and **Dual-stream** models we investigate here.

Acknowledgments

This work was supported by ONR grant N00014-24-1-2026. Computing hardware was supported by NIH Office of the Director grant S10OD025181 through the Center for Computation and Visualization at Brown University. JG was supported by Brown's Undergraduate Teaching and Research Awards (UTRAs).

References

Eberhardt, S., Cader, J. G., & Serre, T. (2016). How deep is the feature analysis underlying rapid visual categorization? In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29). Curran Associates, Inc.

Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, *33*, 13890–13902.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., ... Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning* (pp. 4904–4916).

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning* (pp. 19730–19742).

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning* (pp. 12888–12900).

Lupyan, G., Abdel Rahman, R., Boroditsky, L., & Clark, A. (2020, nov). Effects of language on visual perception. *Trends Cogn. Sci.*, *24*(11), 930–944.

Lupyan, G., & Ward, E. J. (2013). Language can boost otherwise unseen objects into visual awareness. *Proceedings of the National Academy of Sciences*, *110*(35), 14196–14201.

Mooney, C. M. (1957). Age in the development of closure ability in children. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *11*(4), 219.

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).

Samaha, J., Boutonnet, B., Postle, B. R., & Lupyan, G. (2018, Apr). Effects of meaningfulness on perception: Alpha-band oscillations carry perceptual expectations and influence early visual responses. *Scientific Reports*, *8*(1). doi: 10.1038/s41598-018-25093-5

Sun, Q., Fang, Y., Wu, L., Wang, X., & Cao, Y. (2023). Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Svanera, M., Savardi, M., Benini, S., Signoroni, A., Raz, G., Hendler, T., ... Valente, G. (2019). Transfer learning of deep neural network representations for fmri decoding. *Journal of neuroscience methods*, *328*, 108319.

Wightman, R. (2019). *Pytorch image models.* GitHub. doi: 10.5281/zenodo.4414861

Zeman, A. A., Ritchie, J. B., Bracci, S., & Op de Beeck, H. (2020). Orthogonal representations of object shape and category in deep convolutional neural networks and human visual cortex. *Scientific reports*, *10*(1), 2453.