Learning Task-Relevant Visual Features from Large Language Model Embeddings

Michelle R. Greene[1] & Bruce C. Hansen[2]
1: Barnard College, Columbia University; 2: Colgate University

Abstract

Human perception flexibly extracts visual features depending on task demands. Here we show that sentence embeddings derived from natural language scene descriptions can guide convolutional neural networks (CNNs) to learn task-relevant visual features from images. Participants described complex real-world scenes in either a general manner or to describe the possible walking paths through the scene. Task-specific activation maps were generated from CNNs trained to predict these sentence embeddings. In a behavioral experiment, participants viewed masked images containing either high- or low-activation regions while performing categorization or navigation tasks. High-activation regions led to higher accuracy, and task-congruent regions improved performance. These findings demonstrate that natural language-derived embeddings can be used to identity task-relevant visual information, providing a novel means of linking goal-directing scene processing to diagnostic image regions.

Keywords: scene understanding; affordances; CNNs; LLMs; goal-directed vision

Introduction

A picture may be worth a thousand words, but which words depend on the task. Describing a kitchen to a blind person emphasizes different features than describing it to a contractor. Can such linguistic differences reveal task-relevant visual features?

Understanding how task goals shape perception is a core challenge in cognitive neuroscience (Kay et al., 2023). Perception is not passive; it builds representations tuned to behavioral relevance (DiCarlo & Cox, 2007). Prior work shows that tasks alter both attention (Yarbus, 1935) and neural activity (e.g., Cukur et al., 2013), but identifying relevant image regions remains difficult. Reverse correlation (Gosselin & Schyns, 2001) is data-intensive, and CNNs only partially align with human strategies (Ebrahimpour et al., 2019). Crowdsourced maps suggest importance, but imagined relevance may not match actual feature use (Kim et al., 2017).

Multimodal LLMs offer a promising alternative. These models better align with brain responses (Wang et al., 2023) and capture affordances missed by CNNs (Bartnik et al., 2024). Predicting LLM embeddings can yield more brain-like features (Doerig et al., 2022).

Here, we used sentence embeddings from task-framed scene descriptions to train CNNs and extract task-relevant visual regions via deconvolution. We show that these regions vary by task and influence human behavior in aligned ways. This approach links language to perception, offering new tools to study how cognitive goals shape visual feature selection in both brains and machines.

Methods

Stimuli: 5582 photographs from 800 locations across 260 scene categories were described by 4903 participants on CloudResearch Connect wrote either general or navigation-framed descriptions.
Sentence embeddings: We created sentence embeddings from each description using the paraphrase-mpnet-base-v2 model from the sentence embeddings Python library (Reimers & Gurevych, 2019). The resulting 768-dimensional embeddings served as input features for subsequent analyses, enabling direct comparisons between descriptions from different conditions.

CNN Training: We trained two convolutional neural networks (one for each description task) to predict the sentence embeddings for the training set. Our training pipeline builds upon a ResNet-18 architecture (He et al., 2016) pretrained on the Places-365 dataset (Zhou et al., 2017). The final fully connected layer of ResNet-18 was replaced with a 768-dimensional dense layer to map from raw pixels to the sentence embedding space. The models were trained using CosineEmbeddingLoss, which minimizes the cosine distance between predicted and true embeddings.

Visualization: We used a gradient-based visualization approach adapted from Zeiler and Fergus (2014) to visualize the information learned from the sentence embeddings. We specifically highlighted layer4.1.conv2 (the final convolutional layer) due to its proximity to the sentence embeddings.

Behavioral Validation: 60 participants performed either categorization or navigation 3AFC tasks. Stimuli revealed only the top or bottom 25% of each network's activation map. Performance was compared across activation strength and task congruence.

Results

The average cosine similarity between predicted and ground truth activations was 0.75 (test set: 0.72) for the description network, and 0.70 (test set: 0.67) for the navigation network. Both networks learned task-relevant language features because we observed lower cosine similarity between predicted and actual embeddings when we crossed tasks.

We found that layer 4 activations differed across networks. Specifically, the activations for the navigation CNN were lower in the image plane than the others, reflecting the increased importance of this region for the task. By contrast, the general description CNN contained a horizontally elongated band of high activation concentrated in the image's central region and somewhat biased to the right.

We analyzed trial-level accuracy using a generalized linear mixed-effects model (GLMM) with a binomial distribution and logit link (via the lme4 package in R). Fixed effects included Task (categorization vs. navigation), Activation (high vs. low), and Task Congruence (congruent vs. incongruent), along with all two- and three-way interactions. A random intercept for participant accounted for repeated measures.

The model revealed a significant main effect of Task ($\beta = -0.74$, SE = 0.064, $z = -11.58$, $p < .001$), with lower accuracy on navigation trials (44%) than categorization (59%). Performance was also higher with high-activation regions (55%) than low (48%; $\beta = -0.29$, SE = 0.064, $z = -4.54$, $p < .001$), indicating alignment between participant behavior and CNN-derived features. Accuracy improved when the model source matched the task ($\beta = -0.15$, SE = 0.064, $z = -2.31$, $p = .021$), and a significant Task × Congruence interaction ($\beta = 0.34$, SE = 0.089, $z = 3.79$, $p < .001$) showed this effect was stronger for categorization. No other interactions were significant, including the three-way interaction ($\beta = -0.13$, SE = 0.13, $z = -1.04$, $p = .30$). This pattern suggests that general descriptions may yield features more broadly useful across tasks.

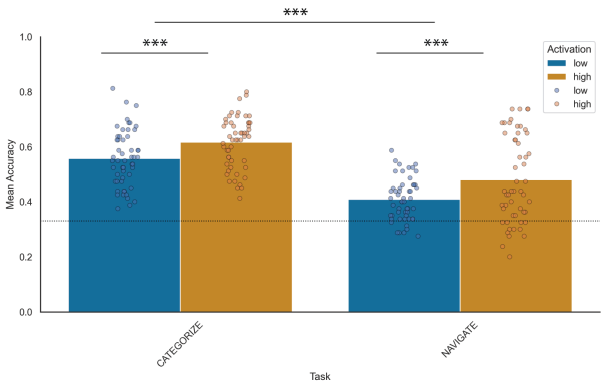

Figure: Behavioral experiment results

Discussion

CNNs trained on task-framed sentence embeddings learned distinct visual features that influenced behavior. Navigation-trained models emphasized ground-level regions, while general-description models captured broader semantic content. Each showed higher embedding alignment within its training domain.

Participants were more accurate with high-activation regions, and task-congruent features boosted performance, especially for categorization. This suggests that language-guided training yields perceptually meaningful, task-specific features.

References

Bartnik, C. G., Sartzetaki, C., Sanchez, A. P., Molenkamp, E., Bommer, S., Vukšić, N., & Groen, I. I. A. (2024). *Distinct representation of locomotive action affordances in human behavior, brains and deep neural networks* (p. 2024.05.15.594298). bioRxiv. https://doi.org/10.1101/2024.05.15.594298

Çukur, T., Huth, A. G., Nishimoto, S., & Gallant, J. L. (2016). Functional Subdomains within Scene-Selective Cortex: Parahippocampal Place Area, Retrosplenial Complex, and Occipital Place Area. *Journal of Neuroscience*, *36*(40), 10257–10273. https://doi.org/10.1523/JNEUROSCI.4033-14.2016

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333–341. https://doi.org/10.1016/j.tics.2007.06.010

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2022). *Semantic scene descriptions as an objective of human vision* (arXiv:2209.11737; Version 1). arXiv. https://doi.org/10.48550/arXiv.2209.11737

Ebrahimpour, M. K., Falandays, J. B., Spevack, S., & Noelle, D. C. (2019). Do Humans Look Where Deep Convolutional Neural Networks "Attend"? In G. Bebis, R. Boyle, B. Parvin, D. Koracin, D. Ushizima, S. Chai, S. Sueda, X. Lin, A. Lu, D. Thalmann, C. Wang, & P. Xu (Eds.), *Advances in Visual Computing* (pp. 53–65). Springer International Publishing. https://doi.org/10.1007/978-3-030-33723-0_5.

Gosselin, F., & Schyns, P. G. (2001). Bubbles: A technique to reveal the use of information in recognition tasks. *VISION RESEARCH*, *41*, 2261--2271. https://doi.org/10.1.1.24.5224

Kay, K., Bonnen, K., Denison, R. N., Arcaro, M. J., & Barack, D. L. (2023). Tasks and their role in visual neuroscience. *Neuron*, *111*(11), 1697–1713. https://doi.org/10.1016/j.neuron.2023.03.022

Kim, N. W., Bylinskii, Z., Borkin, M. A., Gajos, K. Z., Oliva, A., Durand, F., & Pfister, H. (2017). BubbleView: An Interface for Crowdsourcing Image Importance Maps and Tracking Visual Attention. *ACM Trans. Comput.-Hum. Interact.*, *24*(5), 36:1-36:40. https://doi.org/10.1145/3131275

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks* (arXiv:1908.10084). arXiv. https://doi.org/10.48550/arXiv.1908.10084

Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J., & Wehbe, L. (2023). Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, *5*(12). https://doi.org/10.1038/s42256-023-00753-y

Yarbus, A. L. (1935). *Eye Movements and Vision*. Springer.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. *Computer Vision – ECCV 2014*, 818–833. https://doi.org/10.1007/978-3-319-10590-1_53.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1. https://doi.org/10.1109/TPAMI.2017.2723009.