Do VLMs Have Bad Eyes? Diagnosing Compositional Failures via Mechanistic Interpretability

Ashwath Vaithinathan Aravindan, Abha Jha, Mihir Kulkarni
{vaithina, abhajha, mkulkarn}@usc.edu

USC University of Southern California

Code & Results

Abstract:

Vision-Language Models (VLMs) have shown remarkable performance in integrating visual and textual information for tasks such as image captioning and visual question answering. However, these models struggle with compositional generalization and object binding, which limit their ability to handle novel combinations of objects and their attributes. Our work explores the root causes of these failures using mechanistic interpretability techniques. We show evidence that individual neurons in the MLP layers of CLIP's vision encoder represent multiple features, and this "*superposition*" directly hinders its compositional feature representation which consequently affects compositional reasoning and object binding capabilities. We hope this study will serve as an initial step toward uncovering the mechanistic roots of compositional failures in VLMs.

Methodology:

Visual Grounding with CLIP and Gradient-Based Localization (Grad-CAM)

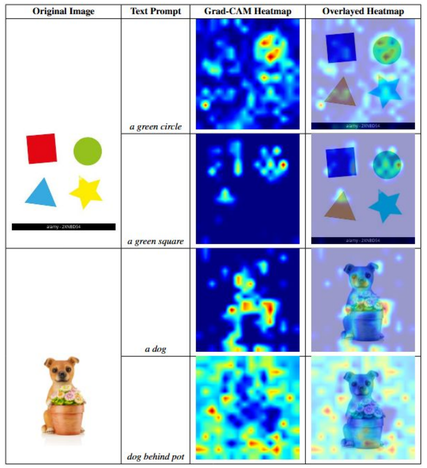
- Used the pretrained CLIP-ViT-L/14 vision encoder and registered hooks on the final MLP layer
- Captured activations and gradients with respect to text-image similarity scores
- Applied Grad-CAM to visualize spatial attention maps for different text prompts
- Compared attention localization across prompts to identify attribute–object binding failures

Neuron-Level Analysis on Toy Shapes Dataset

- Constructed a synthetic dataset of 500 images with controlled attributes: shapes, colors, and spatial positions
- Recorded activations of all 24,576 MLP output neurons across the dataset
- Computed feature affinity values for each neuron using top-k activating images
- Quantified selectivity with Shannon entropy and sorted neurons by entropy to identify highly feature-selective "feature neurons"
- Analyzed activation patterns to assess evidence of superposition

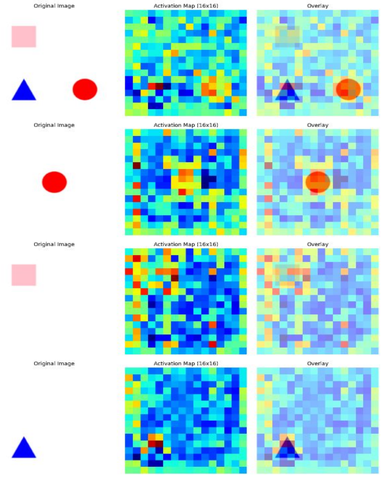
Effects of Superposition on Embedding Space

- Selected 1000 lowest-entropy neurons for superposition analysis
- For every feature pair <f1, f2>, computed a superposition score S(f1, f2) based on neuron-level affinities
- Measured embedding-level separability using two metrics:
  1. Cluster-Center Distance (D)
  2. Misclassification Rate (M)
- Correlated S with D and M to test the link between neuron-level entanglement and embedding-level binding failures



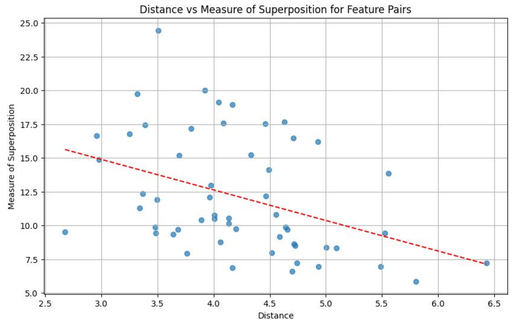
Attention analysis reveals binding failures

In the figure, each row shows how CLIP's attention shifts for various descriptions of the same image. Incorrect or partial attention localization reveals binding failures (e.g., attending to both green circle and red square for ``green square'' prompt).



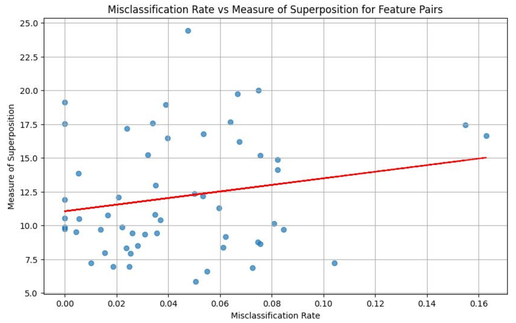
Each neuron excites to specific target features

This example shows the patch-wise activation of a neuron that has high affinity for square and circle features. The activation of the neuron is generally high (red and yellow) in images with the target features (first 3 rows) and low (blue) in the image without (last row)



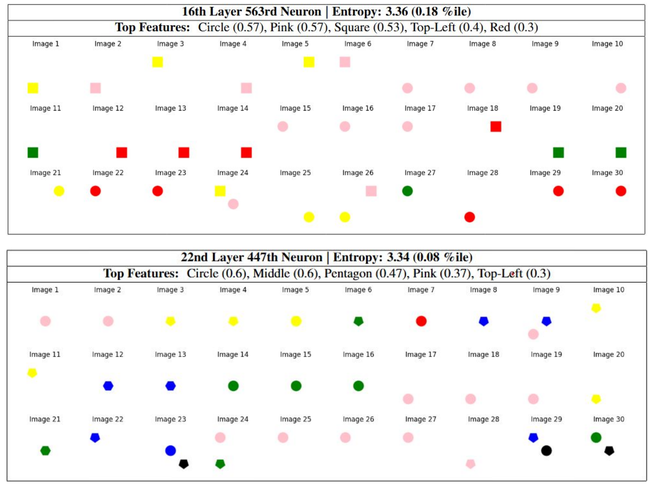
Higher degree of superposition => embeddings closer together

Feature pairs with higher combined affinity ratios in neurons tend to have embeddings that are more tightly clustered in the representation space.



Higher degree of superposition => higher rate of misclassification

Feature pairs with higher superposition scores exhibit increased misclassification rates, indicating that entangled neurons contribute to more frequent attribute–object binding errors in the embedding space.



Neurons with high feature specificity can be identified through entropy analysis

The figure shows top features and top activating images for a neuron that activates the most when handling images containing square and circle shapes and the color pink.

Conclusion:

Our study uncovers a mechanistic connection between internal feature representation and CLIP's image embeddings.

1. **Neuron-level superposition:** Feature entanglement is present not only at the embedding-level but all the way down to individual neurons, many of which may encode multiple, semantically unrelated attributes.
2. **Impact on compositionality:** The stronger this superposition, the weaker CLIP's ability to bind objects and attributes: high entanglement predicts smaller embedding-space separation and higher misclassification rates on compositional tasks.

These findings establish superposition as a key bottleneck for object–attribute compositionality and motivate future work on disentangling neuron activations to improve CLIP's feature representation.