

CLIP: Oral Knowledge Questions

Explain each answer in your own words and justify the underlying mechanism.

1. CLIP's two encoders and shared space.

Describe how CLIP converts an image and a caption into comparable vectors. What are the roles of the image encoder, text encoder, linear projections, and L2 normalization? Do the image and text interact through cross-attention during encoding?

2. Why CLIP's text encoder does not generate text.

CLIP's text encoder uses causal Transformer blocks, which are also used in autoregressive language models. Why does this not make CLIP a caption generator? Explain what CLIP extracts from the end-of-sequence position, how that output is trained, and how it is used.

3. Positive pairs, negatives, and collapse.

In a batch of matched images and captions, what counts as a positive pair and what counts as a negative pair? Why could simply making matched embeddings similar lead to every input receiving the same vector? Explain why that representation would perform poorly on CLIP's contrastive matching task.

4. **Zero-shot classification versus training a classifier.**

How can CLIP classify images into a new set of categories using only the category names and text prompts? Contrast this with training a linear classifier on frozen image features. Which parameters change in each approach, and why can prompt wording affect zero-shot predictions?

5. Recognizing objects versus understanding relationships.

Consider the captions “a dog under a table” and “a dog on a table.” Why might CLIP recognize the objects yet struggle to distinguish these captions? Explain how training with captions that contain the same objects but incorrect relationships could encourage better relational understanding.