Grounded VLMs

Oral Knowledge Questions

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

1. **Why is Grad-CAM class-specific?**

   An image contains both a cat and a dog. Explain how Grad-CAM can produce one heatmap for "cat" and a different heatmap for "dog" from the same image.

   Which quantities remain unchanged, and which must be recomputed when the target class changes? Does generating the second heatmap require retraining the CNN?

2. **Does a heatmap prove that a model uses the right evidence?**

   A CNN correctly predicts "cow," but its Grad-CAM heatmap mainly highlights the surrounding grass.

   What possible shortcut does this suggest? Why does the heatmap alone not prove that the grass caused the prediction?

   Describe a controlled test using images in which the cow is retained but the background changes. What result would strengthen the shortcut-learning hypothesis?

3. **Why does Grad-CAM use convolutional feature maps?**

   Why does Grad-CAM need features that retain spatial information about the image?

   Explain why the last convolutional layer is commonly used rather than an early convolutional layer or a final classification vector. What is the trade-off between meaningful object features and spatial detail?

4. **How does Grounding DINO combine image and text information?**

   Describe the roles of the feature enhancer, language-guided query selection, and cross-modality decoder in processing an image and a text prompt.

   In your explanation, distinguish:

   - Interaction between image and text features.
   - Selection of promising image positions.
   - Refinement of object queries into detections.

   What does the model output, and how does the role of text differ from a detector restricted to a fixed set of trained category labels?

5. **How does deformable attention differ from standard attention?**

   Standard attention compares a query with every image position. Deformable attention instead samples a small number of locations.

   Explain this difference in your own words. Why can sampling fewer locations make attention more efficient when processing a large image feature map?