AdaptVis and Entropy-Gradient Grounding

Oral Knowledge Questions

Explain each answer in your own words.

1. **Why is more image attention not necessarily better?**

   A VLM assigns most of its attention to text tokens. We increase the total attention assigned to image tokens, but its spatial-reasoning accuracy does not improve.

   (a) Why is this outcome possible?

   (b) For the question "Where is the book relative to the candle?", what should useful visual attention capture beyond simply assigning more weight to the image?

2. **How does AdaptVis decide whether to sharpen or smooth attention?**

   Explain the sequence of steps AdaptVis follows, starting with the baseline model's answer confidence and ending with a new answer.

   Why might smoothing help when the model focuses on the wrong region? Why could sharpening be harmful when the model is confidently wrong?

3. **How can a training-free method use backpropagation?**

   Entropy-Gradient Grounding backpropagates next-token entropy to visual-token embeddings while keeping the VLM frozen.

   (a) What quantity is differentiated, and with respect to what?

   (b) What does a large gradient magnitude tell us about a visual token?

   (c) Why does computing these gradients not constitute training the model?

4. **Why use multiple crops and retain the global image?**

   A question about a document requires information from two distant regions.

   (a) Why might one crop miss information needed for the answer?

   (b) How can multiple enlarged crops help?

   (c) Why should the original image remain available even when both relevant regions have been cropped?

5. **What are the two different entropies in Entropy-Gradient Grounding?**

   The method uses both **next-token entropy** and **spatial entropy**.

   (a) For next-token entropy, what outcomes form the probability distribution, and how does the method use this entropy to locate evidence?

   (b) For spatial entropy, the probabilities are the relative areas of connected active regions. What does this entropy measure, and how does it help decide when to stop refining crops?

   (c) If spatial entropy decreases, does that necessarily mean the model's answer has become more correct? Explain.