Chain of Visual Thoughts

Oral Knowledge Questions

Explain each answer in your own words.

1. **What makes a continuous thought different from a text token?**

   Explain how COCONUT produces the next reasoning input without selecting a vocabulary token.

   Does each continuous thought necessarily correspond to a particular word or sentence? Why are continuous thoughts harder to inspect than textual chain-of-thought?

2. **Why does CoVT learn several types of visual thought tokens?**

   Consider these questions:

   - "How many objects are present?"
   - "Which object is closer to the camera?"
   - "What is the outline of this object?"

   For each question, identify a relevant type of visual information learned by CoVT. Why might one type of visual representation be insufficient for all three tasks?

3. **Why is learning to generate visual thought tokens not enough?**

   Suppose CoVT generates visual thought tokens that can be decoded into accurate object masks and depth maps.

   (a) Does this guarantee that the model can correctly answer "Which object is closer to the camera?" Explain the difference between representing useful visual information and using it to answer a question.

   (b) What additional ability should the **reasoning stage** teach after the model has learned to generate visual thought tokens?

   (c) In the **efficient-reasoning stage**, some groups of visual thought tokens are randomly dropped. How might this help the model avoid depending too heavily on one type of visual information?

4. **Why transfer intact-image latent thoughts to the corrupted path?**

   In CrystaL, the corrupted path receives both a degraded image and latent thoughts from the intact path.

   (a) What is the intended role of the transferred latent thoughts?

   (b) Why does degrading the image create pressure for the model to use those thoughts?

   (c) Why is this different from simply training a model to answer questions about noisy images without transferring any latent information?

5. **Does consistency guarantee correct visual reasoning?**

   Suppose CrystaL's intact and corrupted paths produce almost identical answer distributions and similar attention to latent tokens.

   (a) Does this establish that the answer is correct? Give an example in which both paths agree but are wrong.

   (b) Why does CrystaL also use answer cross-entropy losses rather than relying only on consistency between the paths?

   (c) To test whether transferred latent thoughts carry useful visual evidence, what would you expect to happen if they were removed or replaced with thoughts from an unrelated image? Explain why this is a useful experimental check.