Week 9 Questions

CogVLM: Visual Expert for Pretrained Language Models

Brief Summary: They send image data to a pre-trained language model so that it can accomplish tasks based on both text and image inputs. The language model receives the image data by getting embeddings from a pre-trained image embedding model. These are sent to the model in a similar way to token embeddings. Additionally, they add visual modules at each layer of the language model. The hidden embeddings for the images go through these attention matrices and MLPs instead of the regular language model attention matrices and MLPs.

1. How is their method different from previous methods that incorporate image data with a language model?
2. Give a brief description of their training data.
3. Describe the architecture of the model in more detail than the brief summary.
4. Their method gets better performance than previous methods. What are the disadvantages of their method? (Hint: there is one major disadvantage)

PaLM-E: An Embodied Multimodal Language Model

Brief Summary: they start with palm, a pretrained language model. They connect image and other data to palm with embedding models. The embeddings are input to palm similar to token embeddings for text input. The training data consists of various multimodal data, including robotic data. After finetuning with this multimodal data, their model, palm-e, can send commands to a robot to accomplish various tasks.

1. Give a brief description of their training data.
2. What types of commands does palm-e give to robots?
3. Think about the significance of figure 6.

4. Compare/contrast palm-e with cogvlm.