Vision is Language: Use LLM to understand Images

When we hear “sky is blue”, we can immediately picture a beautiful blue sky above with a few white clouds in our mind. Similarly, when we see a dog chasing after a ball, we can effortlessly describe which is doing what. Hence, intuitively, there is a large overlap between vision understanding and language understanding.

As the Large-Language Model has exhibited promising capability recently, research to leverage LLM to improve image understanding has witnessed a rapid advancement. Hence, it is a great opportunity to explore how to leverage LLM for vision understanding in order to achieve a human-like understanding and reasoning of images.

I plan to study a few recent research papers, such as “OmniVec: Learning robust representations with cross modal sharing”, “Language Models with Image Descriptors are Strong Few-Shot Video-Language Learners“, “Generating Images with Multimodal Language Models”, etc. to understand the related work and draw inspirations from.

My primary focus is the Image QA. I plan to use the https://visualqa.org/ and VisualGenome as the primary datasets. I also plan to leverage Meta’s open source LLM Llama as the frozen pretrained LLM model.

General Direction for method/Algorithm: Leverage the LLM to generate embeddings for the descriptions in the training data and use the embedding as the y value and train the vision network (TBD) with a cosine similarity as the loss function so that the vision model can produce an embedding for a given image in the same embedding space as the text.*

Evaluation: I plan to use https://visualqa.org/index.html benchmark to evaluate the performance of the model.