**Do Sequence Length Matters for Multimodal Inputs?:**

Supervisor: Sheng Liang

Details:

The success of large-scale pretrained language models (PLMs) and pretrained image encoders has stimulated a surge of pretraining multimodal systems. Especially, recent models trained with architecture that connects PLMs with image encoders achieve human parity on visual question answering. However, multimodal inputs are often inconsistent in sequence length, specifically, when the above architecture is applied to visual question answering tasks, the length of image embeddings may be hundreds while the length of text embeddings is only tens. Intuitively, the inconsistent lengths would cause the system to be biased towards a certain modality when capturing information from multimodal inputs. On the other hand, injecting overly long sequences into PLMs is also sub-optimal. In this topic, we want to compare different ways of connecting PLMs and image encoders, e.g.linear projections or attention modules, to explore whether sequence length affects multimodal question answering systems.