VLMO: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts

---

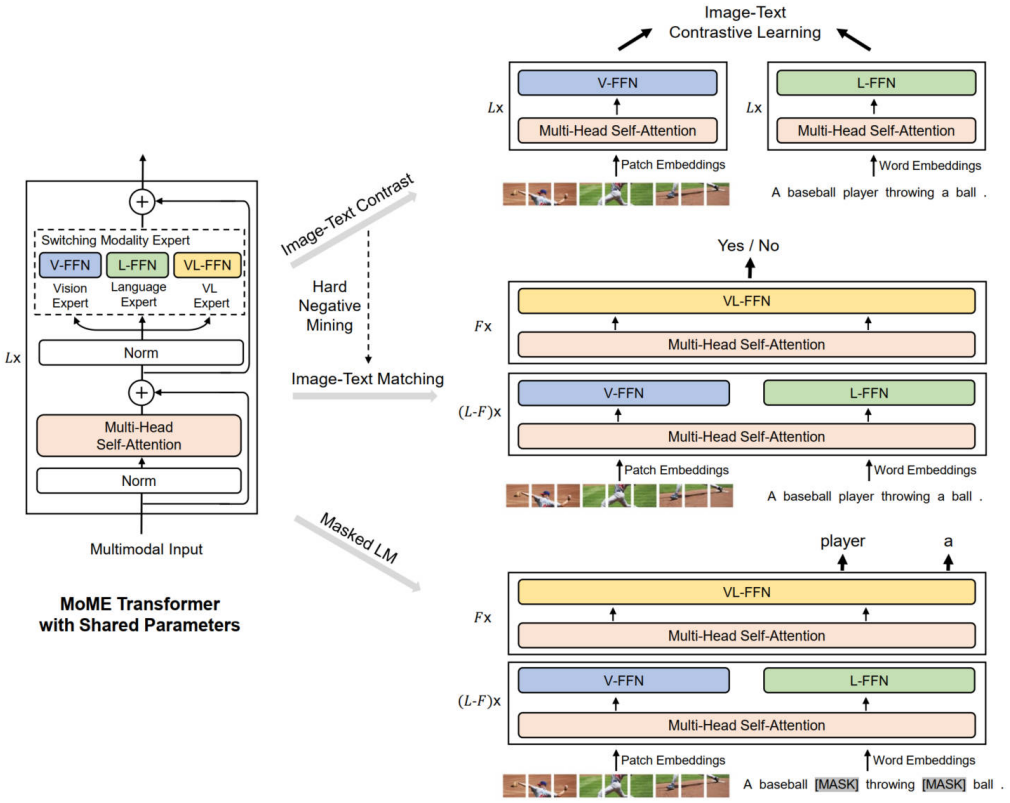
动机：

前人工作的不足，有两种主流的架构，① dual encoder + cosine similarity（模态之间的交互太shallow了），如CLIP，ALIGN，适合于VL retrieval tasks，不擅长VL classification tasks（VQA，VR，VE）；② fusion encoder，适合于VL classification，但是在做VL retrieval推理时，用时较长。**想结合两种主流框架的优点。**

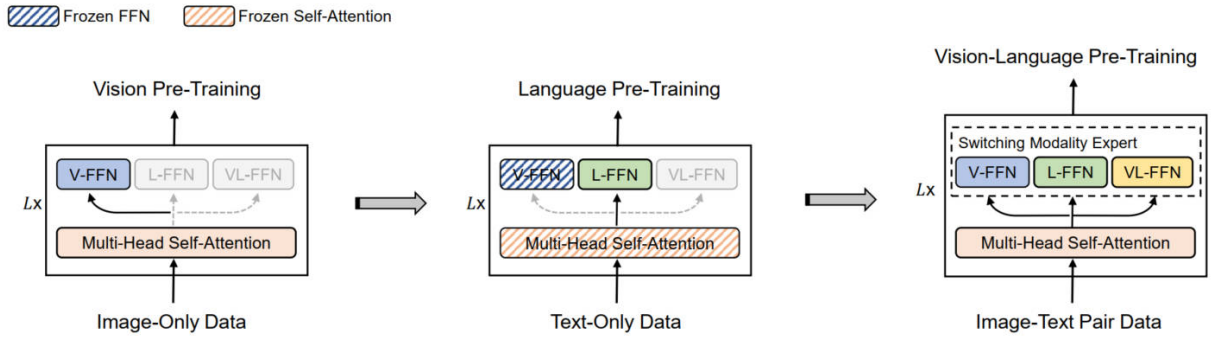
想在多模态模型上看到scalability，需要用到更多的数据。

方法：

针对第1个动机，提出来MoME（Mixture-of-Modality Experts） Transformer。Multi-Head Self-Attention是共享学习权重的，将Transformer的Encoder组件中的feed-forward network改成了Switching Modality Expert子层，它不共享学习权重。



针对第2个动机，提出来stagewise pre-training strategy。 在Image-Only Data阶段中，训练任务是Mask Image Modeling；在Text-Only Data阶段中，训练任务是Mask Language Modeling；在Image-Text Pair Data阶段中，训练目标是上图的3个pre-training tasks，分别是Image-Text Contrastive Learning、Image-Text Matching Learning和Mask Language Modeling。



创新点：

MoME，这是在模型架构上进行改动。

Stagewise pre-training strategy，这是在训练方式上进行改动。