VLMo

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

汤晨

autumn 2023

1.1 概要

当前（2022.3），在多模态大模型预训练中，大多只有单塔（fusion encoder）结构或者多塔/融合（dual encoder）结构。单塔如 VL-BERT，UNITER，ALBEF，是由一个统一的视觉-文本 encoder 同时建模文本与视觉信息，一般用 cross-attention 的形式，但是复杂度是二次的，推理时间会长很多。双塔如 CLIP，ALIGN 是分别用两个 encoder 去建模好文本图像信息，然后再用简单的对比学习去交互。但是这种交互太过浅层，难以在复杂任务上起作用。

于是作者提出了 VLMO，是一种 Mixture-of-Modality-Experts（MOME）混合多模态专家，使用的时候既可以作为单塔也可以作为双塔。

1.2 当前背景与面临问题

1.3 意义

1.4 模型结构（图 1）

模型采用 Vit 对图像编码，Bert 对文本编码，然后将两种 embedding 输入 MoME 里面。可以看到 MoME 的多头自注意力层的权重全是共享的，而在 Feed Forward Norm 里面分出来不同的权重，对应不同的任务。

1. ITC 预训练任务，分别用双塔模型做 embedding，再进行对比。
2. ITM 图文匹配任务，在分别通过各自的 FFN 层之后，通过一个统一的自注意力层和统一的 FFN，以单塔模型交互特征。
3. MLM 完型填空任务，预测 masked 的值。

除此之外，作者还用单一的图像和文本数据集做预训练，叫阶段预训练（Stagewise Pre-training）如图二：先进行视觉的预训练 Masked Image Modeling（BEIT 的方法，甚至于作者直接拿 BEIT 的权重初始化），然后再到文本预训练 Masked Language Modeling ，将 V-FFN 和 self-attention 层冻住，更新 L-FFN。最后再进行多模态的预训练（也就是上述三个任务），所有层都开放更新。

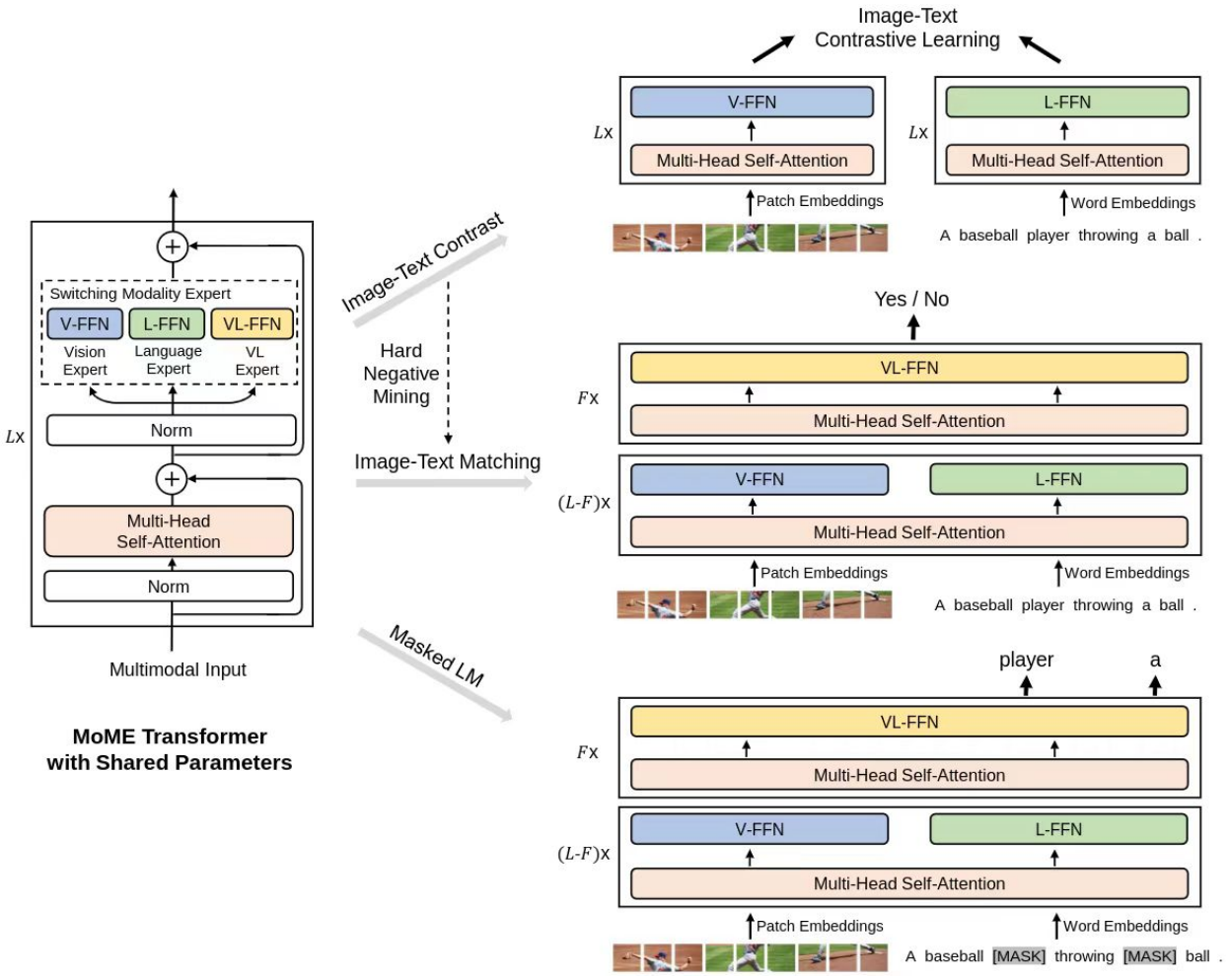

图 1: Enter Caption

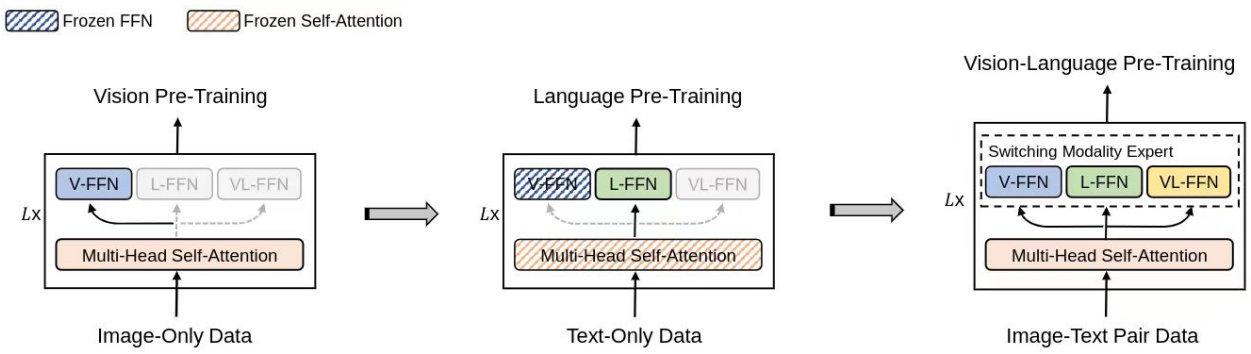

图 2: Enter Caption

因为纯图像和纯文本的数据集量是比图文数据集大很多的，所以这么训练有助于模型泛化。

1.5 一些细节

1.6 一些疑问

1. 对基本的 attention 和 transformers 掌握还不大熟悉，说实话还是觉得，现在很多论文的创新点和灵感是仅仅从直觉出发的，然后再以实验去验证。即使从直觉出发得到的创新点，能否用一种严谨的方式去证明一下它的有效性呢?
2. 从模型融合而言，这种融合的方式相比于 BLIP 等比较传统的融合方式确实是比较创新，但是效果究竟如何还有待商筹

1.7 关于细颗粒度对齐问题

1. 关于对齐问题，该模型的预训练任务有 ITC，是 ALBEF 提出的一种粗颗粒度的对齐方式。MLM 任务也只是随机 mask，没有条件掩码。从直觉上来讲对细颗粒度对齐不利

1.8 我的思考