ALBEF

Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

汤晨

autumn 2023

1.1 概要

作者提出新的模型 ALBEF，针对当时 VLP 模型领域存在的三个问题：1）以 CLIP，ALIGN 为代表的方法，学习单模态的 image 和 text 的 encoder，再用 contrastive loss 进行学习，让两种模态的编码都各自带上对方编码的特征，在图像文本检索任务上有很高性能。但是因为分开训练 encoder，所以他们的特征不在一个空间内，融合的时候是 coarse-grained，比较难去有效的学习融合两者特征，所以缺乏在更复杂任务上的建模和多模态交互能力。

（2）以 UNITER，OSCAR，VL-BERT 为代表的方法采用 transformers 作为 encoder 学习图像与文本交互特征，但是因为学习之前没有 align，所以需要高精度的图像特征来便于学习，所以算力开销很大。

（3）预训练的数据大多来自互联网，数据不干净，MLM 任务很容易对噪声过拟合。

针对以上问题，ALBEF 先用无 detector 的 encoder 分别对 image，text 进行编码，然后先用 ITC loss 进行对齐，再用 multimodal encoder 通过 cross-modal attention 去 fuse 两者特征。主要核心是采用动量蒸馏（MoD）的 image-text contrastive（ITC）loss ，好处是 1）可以先进行 align，方便 fuse。2）让 encoder 更好的理解语义。3）学习了一个通用低维空间来 embed 图像文本，所以可以从从 contrastive hard negative mining 中找到更多信息丰富样本

1.2 当前背景与面临问题

当时 VLP 模型领域存在的三个问题：1）以 CLIP，ALIGN 为代表的方法，学习单模态的 image 和 text 的 encoder，再用 contrastive loss 进行学习，让两种模态的编码都各自带上对方编码的特征，在图像文本检索任务上有很高性能。但是因为分开训练 encoder，所以他们的特征不在一个空间内，融合的时候是 coarse-grained，比较难去有效的学习融合两者特征，所以缺乏在更复杂任务上的建模和多模态交互能力。

（2）以 UNITER，OSCAR，VL-BERT 为代表的方法采用 transformers 作为 encoder 学习图像与文本交互特征，但是因为学习之前没有 align，所以需要高精度的图像特征来便于学习，所以算力开销很大。

（3）预训练的数据大多来自互联网，数据不干净，MLM 任务很容易对噪声过拟合。

1.3 意义

1.4 一些细节

1. ITC 的任务，是获取图像编码器的编码做为图像特征，文本编码器编码作为文本特征，计算两个特征相似度，通过训练使配对的图像文本相似度越来越高，不配对的特征相似度越来越低。同时采用动量蒸馏，将最相似的一些文本和图像也作为正标签进行训练
2. 采用了互信息最大化视角，通过计算和数学推理，证明了 ALBEF 最大程度的限制了 image-text 的 MI（mutual information）下限，ALBEF 隐含的学习了图像文本联系。

1.5 一些疑问

动量蒸馏概念不大明白，互信息最大化视角推理没怎么看明白

1.6 关于细颗粒度对齐问题

1. 本文提出的 image-text 对比损失函数（ITC）可以用于细颗粒度对齐
2. ITC，MLM，MoD 可以解释为生成一个不同视图的 image-text 对，目标是让 model 学习不随视角变化的表示，最大化 mutual information（a，b）的下边界

1.7 我的思考

掘对比来改善ITM。

2.1 Image-text对比学习(ITC)

该损失函数的目标是在融合之前更好的学习单模态表示。其会学习一个相似函数 $s = g_v(\mathbf{v}_{cls})^\top g_w(\mathbf{w}_{cls})$，使得并行的 `image-text对` 具有更高的相似分数。g_v和g_w是将 `[CLS]` 嵌入向量映射为规范化低维度表示的线性变换。受MoCo启发，维护两个队列来存储来自动量单模态编码器中最近的M个 `image-text` 表示。来自动量编码器的规范化特征表示为 $g'_v(\mathbf{v}'_{cls})$和$g'_w(\mathbf{w}'_{cls})$。定义 $s(I,T) = g_v(\mathbf{v}_{cls})^\top g'_w(\mathbf{w}'_{cls})$ 且 $s(T,I) = g_w(\mathbf{w}_{cls})^\top g'_v(\mathbf{v}'_{cls})$。

对于每个图像和文本，计算 `image-to-text` 和 `text-to-image` 的相似度为：

$$p_m^{i2t}(I) = \frac{exp(s(I,T_m)/\tau)}{\sum_{m=1}^{M} exp(s(I,T_m)\tau)}, \quad (1)$$