BLIP2

Bootstrapping Language-Image Pre-training with Frozen Image Encoders 安定 Large Language Models

汤晨

autumn 2023

1.1 概要

当前（2023.1）多模态大模型存在一些普遍问题，1）因为预训练的时候常常要将图像编码器和文本编码器重新再预训练一次，所以会需要消耗非常大的计算资源。2）并不是没有人想用预训练好的单模态模型，但是因为图像和文本编码器分开进行预训练，所以两者信息难以对齐。而已有的 image-to-text generation loss 也难以弥补差距。

针对这些问题，作者提出了 BLIP2，1）直接使用现成的图像文本编码器，并 frozen 参数，2）同时使用一个 Q-Former（Querying Transformers）作为一个 information bottleneck 存在于 image encoder 和 LLM 之间。有两个阶段的训练，表征训练和生成训练。

1.2 当前背景与面临问题

当前（2023.1）多模态大模型存在一些普遍问题，1）因为预训练的时候常常要将图像编码器和文本编码器重新再预训练一次，所以会需要消耗非常大的计算资源。2）并不是没有人想用预训练好的单模态模型，但是因为图像和文本编码器分开进行预训练，所以两者信息难以对齐。而已有的 image-to-text generation loss 也难以弥补差距。

1.3 意义

1.4 模型结构（图 1）

训练分为俩阶段，首先是表征训练，然后是生成训练。

1. 表征训练阶段：Q-former 图像部分采用的基础模型是 Bert，输入可学习的 query（随机初始化）经过 self-attention（该层与文本部分共享权重）先学习自己相关的特征。观察图 2 右侧的 mask 表示，Query 和 Text 是一起输入 self-attention 的，所以能互相学习。然后再通过 cross-attention 与图像 encoder 输出的 embedding 进行混合，再输出到后面两个 loss，Image-Text Matching 和 Image-Text Contrastive Learning。

ITM 任务中 text 的 token 和 Query 不 mask，学习好彼此的特征之后，在 ITM 任务中找到负样本。

CL 任务中，Query 和 text 互相的注意力被 masked，以学习在特征空间进行对齐。

generation 任务中，让 attention 中允许 Text 看到 Query 和之前的 Text token，不允许 Query 看到 Text 信息。

2. 生成训练阶段：作者尝试了两种语言模型，分别是 decoder 和 encoder-decoder（图 3）。训练一个全连接层将 query 转换成 LLM 可以接受的信息。

1.5 一些细节

1. 多年以来出现了很多 VLP 预训练任务，经过时间考验的有如下：1）image-text contrastive learning 2）image-text matching 3）(masked）language modeling
2. 历年也有尝试在使用预训练好的 encoder 情况下去做对齐，方法有在 LLM 中插入 cross-attention，并用十亿级的 image-text pair 去预训练；或者是微调一个 image encoder，让输出直接作为 LLM 的 soft prompts

1.6 一些疑问

1. 对 attention 的了解不够，可学习的 query 本质是什么（attention 不够了解）
2. 表征训练中 attention 的过程还是不够清晰（transformers 不够了解）
3. 生成训练中的具体过程还是不大清晰（LLM 大模型不够了解）

1.7 关于细颗粒度对齐问题

1. 这篇文章给了一个非常好的思路，对齐不一定要对语言 encoder 和视觉 encoder 做预训练，可以使用现成的很好的大模型，然后做一个可训练的轻量级的传递信息的 bottleneck。这个思路的可发展性非常强，我觉得未来多模态大模型的思路，要不然就是在一个统一的模型下，从一开始就进行信息的融合对齐，要不然就是像本文一样，用现成的视觉和语言模型做一个桥梁。

虽然从最终发展而言，还是前者比较重要，但是后者不失为一种发展对照思路。

1.8 我的思考