BLIP

Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

汤晨

autumn 2023

1.1 概要

针对当时(2022.2)VLP 模型中出现的两个问题:1)大多数模型只能专注于一个任务,understanding task（encoder-based）或 generation task （encoder-decoder-based），很少有两种任务都擅长的。2）大多数模型都是直接用互联网上的图文数据，这种数据集虽然数量多但是质量差，噪声多，图文难以匹配，会很影响训练效果。

针对如上两个问题，作者提出了 BLIP，1）提出 MED（multimodal mixture of encoder-decoder）以适配两种任务，预训练方法采用 ITC（用于对齐），ITM（match），以及 image-conditioned language modeling（看图生成描述语句，与原数据集中对比相似度，此时训练用的是高质量数据集）。2）提出 Captioning and Filtering（CapFilt）用来处理嘈杂数据，即用在高质量数据集微调后的 MED 的 decoder 对图片进行 caption，然后用 MED 的 encoder 判断产生的 caption 好还是原来的 caption 好，将不好的移除，然后再用这个更新过后的数据集进行再训练。

1.2 当前背景与面临问题

VLP 模型中出现的两个问题：1）大多数模型只能专注于一个任务，understanding task（encoder-based）或 generation task （encoder-decoder-based），很少有两种任务都擅长的。2）大多数模型都是直接用互联网上的图文数据，这种数据集虽然数量多但是质量差，噪声多，图文难以匹配，会很影响训练效果

也有尝试建构 unified encoder-decoder 模型的（UNVLP），但是作者说效果很差

1.3 意义

1.4 模型结构

1. 对于 MED，MED 的特点是灵活，既可以作为单模态的 encoder，也可以作为基于图像的文本 encoder，或者是基于图像的文本 decoder。

观察图 1，最左边的图像编码器是基于 Vit 的，文本是基于 Bert。而第三列是基于图像的文本编码器，作用是根据输入图像和文本做二分类，用来做 ITM 任务。而第四列是基于图像的文本解码器，采用的是 casual-attention。

需要注意的是相同颜色的层是共享权重的。

2. 对于 CapFilt，观察图 2，流程为，准备高质量小型数据集 D1，多噪声的大量数据集 D2，先用 D1 给 MED 进行微调，再输入 D2. 对 D2 的 image 用模型 decoder 生成 caption，并用模型 encoder 判断 caption 和 D2 中原有的 caption 孰优孰劣（类似于 ITM 任务），再选最好的那个留下。最后组成数据集 D3，再输入 MED 中进行微调

1.5 一些细节

1. CapFilt 可以看作是一个 KD，用来提取知识
2. 用 CapFilt 可以做 Data Augmentation ,是比以前 NLP 中的 language-only 的 data-augmentation 超越很多

1.6 一些疑问

1. 为什么以前的 encoder-decoder 模型效果不好，不能完成两个任务，该文章的模型也没有什么特殊的地方?
2. decoder 中的 causal self-attention 是什么

1.7 关于细颗粒度对齐问题

1. 对于细颗粒度对齐而言，该文章延续了 ALBEF 的 ITC 去对齐特征，最值得学习的是模型自己左手改右手，自动地筛选数据提高数据，我认为这是一种普世的思想，类似于人类的检查与反思，是很值得借鉴的

1.8 我的思考

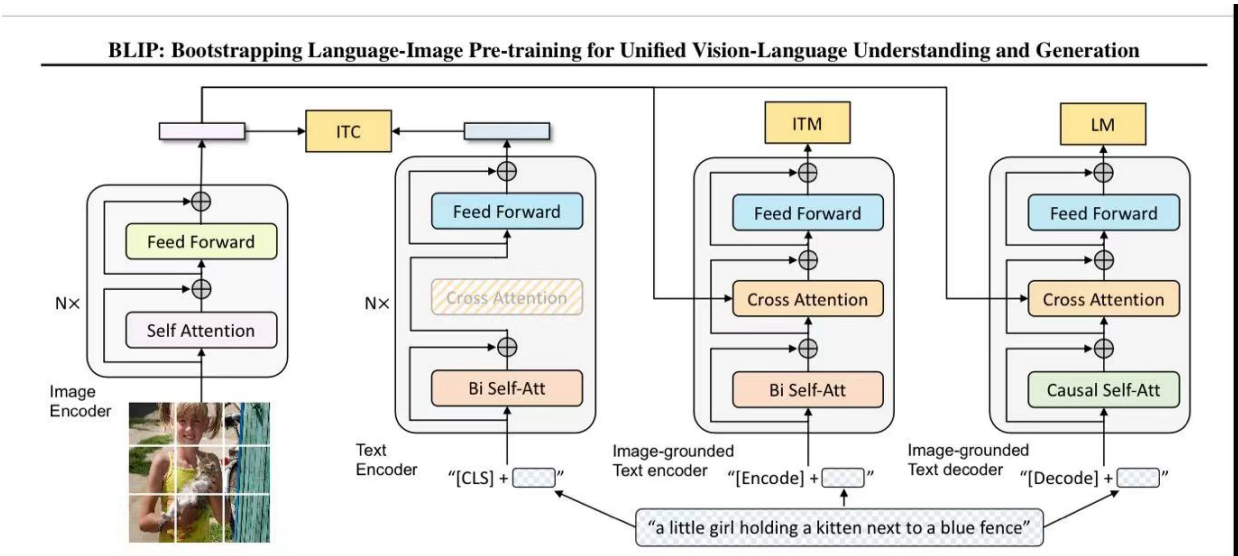

Figure 2. Pre-training model architecture and objectives of BLIP (same parameters have the same color). We propose multimodal mixture of encoder-decoder, a unified vision-language model which can operate in one of the three functionalities: (1) Unimodal encoder is trained with an image-text contrastive (ITC) loss to align the vision and language representations. (2) Image-grounded text encoder uses additional cross-attention layers to model vision-language interactions, and is trained with a image-text matching (ITM) loss to distinguish between positive and negative image-text pairs. (3) Image-grounded text decoder replaces the bi-directional self-attention layers with causal self-attention layers, and shares the same cross-attention layers and feed forward networks as the encoder. The decoder is trained with a language modeling (LM) loss to generate captions given images.

图 1: Enter Caption

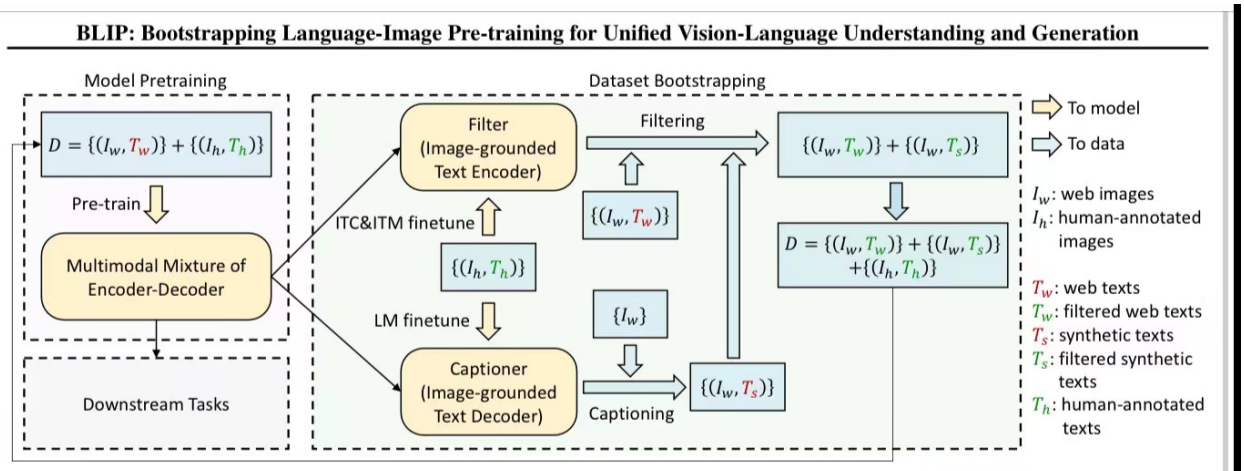

Figure 3. Learning framework of BLIP. We introduce a captioner to produce synthetic captions for web images, and a filter to remove noisy image-text pairs. The captioner and filter are initialized from the same pre-trained model and finetuned individually on a small-scale human-annotated dataset. The bootstrapped dataset is used to pre-train a new model.

图 2: Enter Caption