X-VLM

Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts

汤晨

autumn 2023

1.1 概要

针对大多是方法都需要 object detection，提出了 X-VLM，主要是多颗粒度的对齐，这个的关键是在给定相关文本的图像中定位视觉概念，同时将文本与视觉概念对齐，对齐是多颗粒度的。

1.2 当前背景与面临问题 ji

目前的对齐大约有两种方法，1）用 object detection 实时检测，但是既消耗算力，也会框出很多不重要的物体，干扰学习效果。2）用整幅图像特征做粗颗粒度匹配

而且单单只是细颗粒度对齐难以学习物体与物体的联系，粗颗粒度对齐则难从达到图中物品与文字的配对（visual grouding 等任务表现差）

1.3 意义

该模型的创新点在于使用多颗粒度的对齐，观察图 1 X-VLM 的模型结构，

1. 我们首先发现，左图给出了视觉概念的提取过程，对于一张 image，在选定的图像的边框内，通过 Vit 切分 patch，然后对所有 patch 进行池化提取全局特征，再对剩下部分提取特征，两者进行 concat。(这个图像的边框，可以指整个图像，也可以是 region 或 object，取决于训练集里输入的处理好画好边框的图像)
2. 然后是三块训练任务，1）Bounding Box Prediction ，是通过输入整张图像和不同层次的文本输入，训练模型在图像中框出目标物体，然后对比已有框的训练集计算 loss，进行训练。采用的是 GloU loss 和 L1 loss。2）Contrastive Learning 任务，该任务是基于 ALBEF 里提出的 ITC（image-text-Contrastive）的。是对不同层次的 image 和 text 的 embeding，让他们直接在特征空间内进行对比学习，匹配的拉近，不匹配的拉远。3）常用的 MLM 任务

1.4 一些细节

1. 采用的是文本描述和视觉概念相匹配，视觉概念（visual concept）可以是 object，region 或 image。

   模型预训练使用如下处理过的 data：1）描述整个图像的文本。2）区域描述与区域图像对齐（以前一般是区域与整个图像对齐。3）re-formulate 图像数据，让一张图像有多个边界框，每个框与物品描述对齐。每一个样本采用如下结构 $(I, T, \{(V^j, T^j)\}^N)$

2. Loss 采用 box regression loss，contrastive loss，masked language model loss，match loss

3. 依然基于 bert+vit 结构

4. 认为 ALBEF 仍然算粗颗粒度

1.5 一些疑问

1.6 关于细颗粒度对齐问题

1. 对于细颗粒度对齐而言，预训练任务上 BBox loss 是富有创新点的一个地方，增加了图像在细颗粒层面的理解。关于模型本身，依然是基于 Vit+Bert 的结构，然后有一个跨模态的 encoder 进行编码。这个模型也采用的 ALBEF 的 ITC，也就是在融合前对齐，然后基于此增加了 BBox，增加了融合后对齐的一个方式。

2. 该模型主要的创新点是同时进行细颗粒度和粗颗粒度的对齐，所以提高了模型的效果，于细颗粒度对齐本身而言并没有太大创新

1.7 我的思考

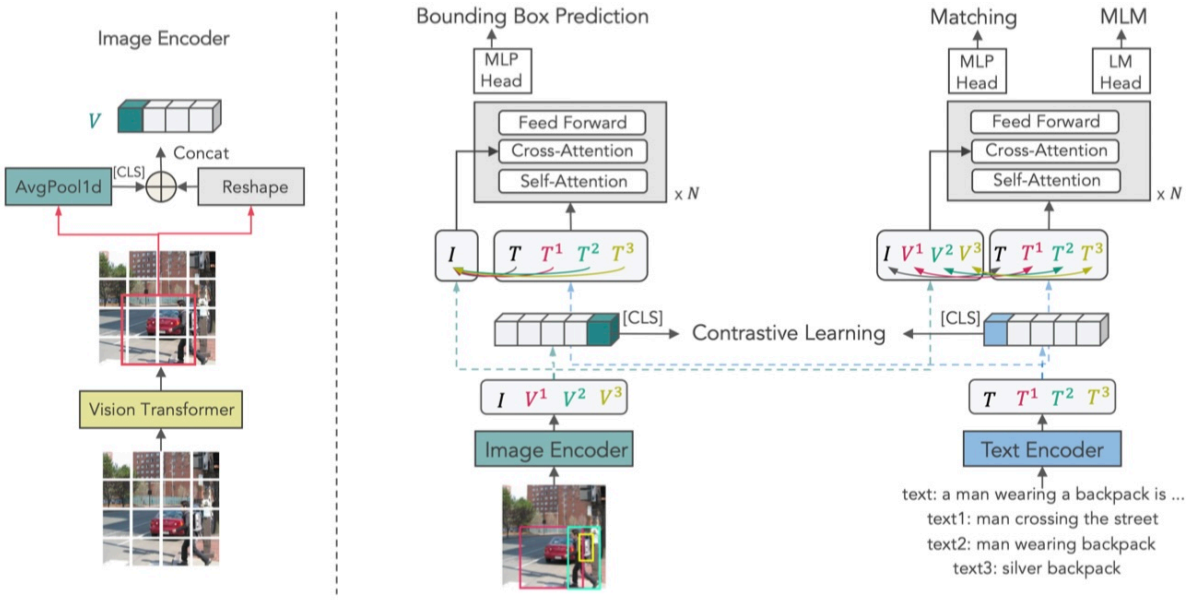

Figure 2. Pre-training model architecture and objectives of X-VLM. As shown on the left side, we extract features from the subset of patches from the vision transformer to represent images/regions/objects (I and V^{1-3}), which are then paired with corresponding text features (T and T^{1-3}) for contrastive learning, matching, and MLM. Meanwhile, the image (I) is paired with different textual descriptions (T and T^{1-3}) for bounding box prediction to locate visual concepts in the image.

图 1: Enter Caption