ViLT

ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision (2021) 386

Vision-and-Language Pre-training (VLP)

Current approaches heavily rely on image feature extraction processes, most of which involve region supervision (e.g., object detection) and the convolutional architecture (e.g., ResNet).

Background

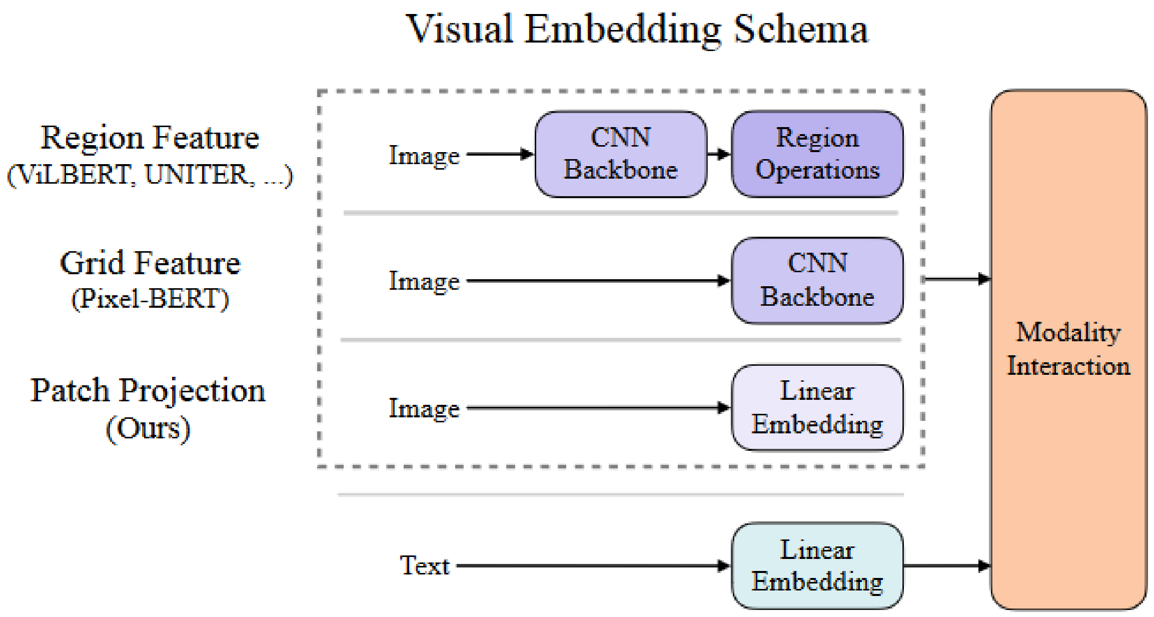
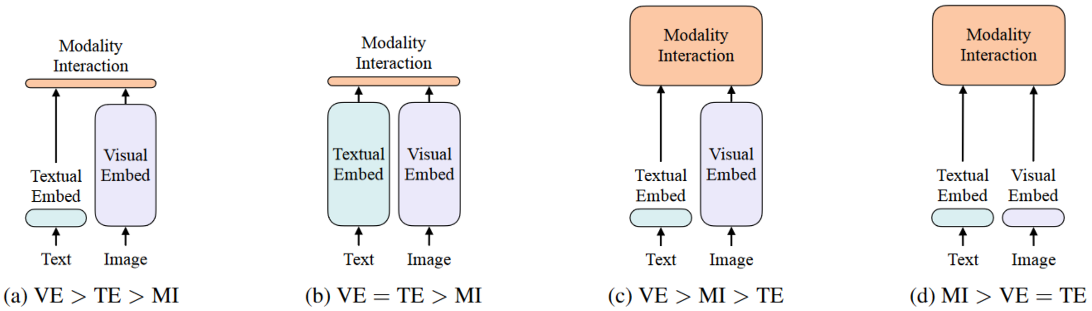


Image pixels need to be initially embedded in a dense form alongside language tokens

Most VLP models: object detectors

ViLT: simplest visual embedding scheme — linear projection that operates on image patches



(a) VE > TE > MI (b) VE = TE > MI (c) VE > MI > TE (d) MI > VE = TE

VSE CLIP ViLBERT / UNITER

ViLT

Too simple Modality Interaction for (a)(b) (e.g., dot product in CLIP).

Unbalanced dedicated parameters / computation for (c).

Model

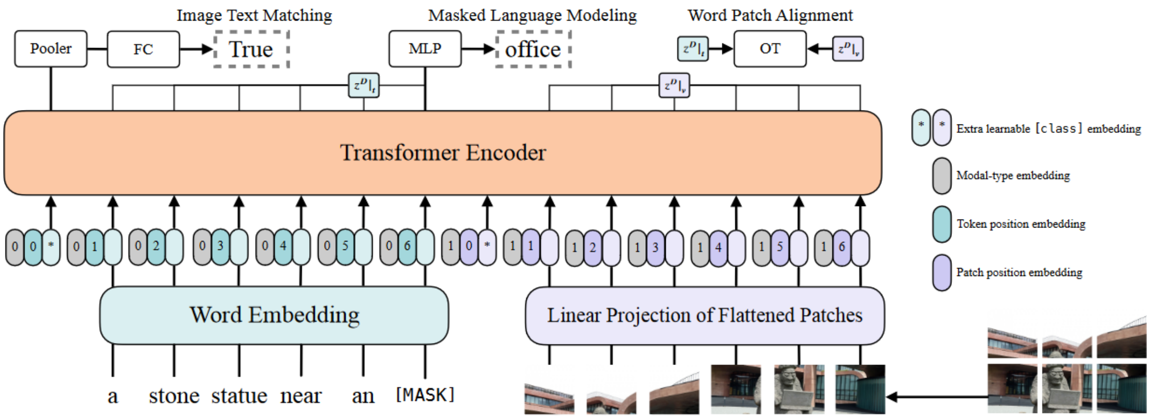


Image Text Matching loss: 大部分VIT使用的matching loss，本来图片和文字应该是配对的，如果把图片随机换成数据集中的其他图片，则图片和文字变成非配对的。通过模型得到特征，通过特征去判断图片与文本是否，看能否成功。

Word Patch Alignment loss: 附加的matching loss，把文本输出和图像输出当成概率分布，计算两个分布间的距离，希望距离越小越好。

Masked Language Modeling loss: NLP中的常见完形填空loss

Other highlights:

Whole word masking: 如果只mask单词的一部分，模型可能可以不通过image，只基于language直接猜出来这个词

Image Augmentation: 使用policies from RandAugment，除了color inversion & cutout，保证了增强后图像与文本仍然匹配