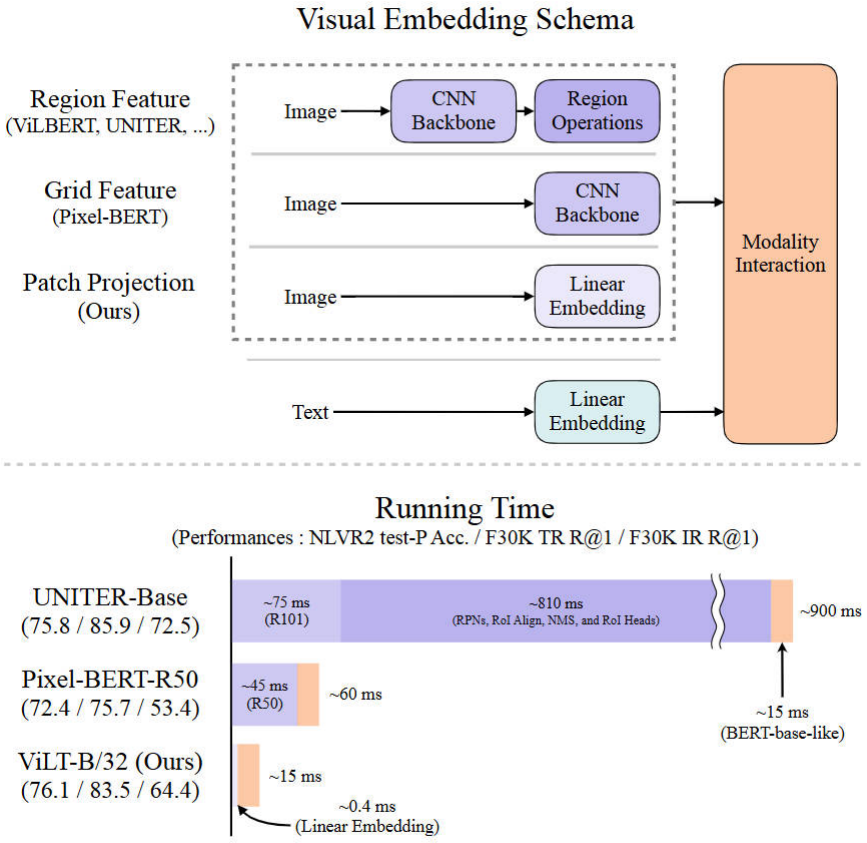
ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision

动机：



过去的Visual Embedding Schema，如Region Feature和Grid Feature，在**1)** Efficiency、**2)** Speed和**3)** Expressive Power上有限制，**1)** 和**2)** 限制了推理（Inference）的速度，**3)** 是由于未能充分使用多模态的潜力，而过多关注在Visual Embedding上——既然要做多模态，那么应当重视Modality Interaction，而不是关注单个模态的Representation多么地好。

ViLT能使得推理（Inference）加快，同时更加注重的是Modality Interaction，同时具有Competitive Performance。

Approach：

完全舍弃了**Region Feature**（如RPNs、RoI Align、NMS和RoI Heads）和**Grid Feature**（CNN Architecture）。只使用**Transformer**结构，文本转变为Tokens；图像也要转变为类似的Tokens，以便丢给Transformer，本文借鉴了ViT的思想，将一张图片打成许多Patches，经过Linear Projection Layer，将每个Patch都映射为一个Token，再丢给Transformer。

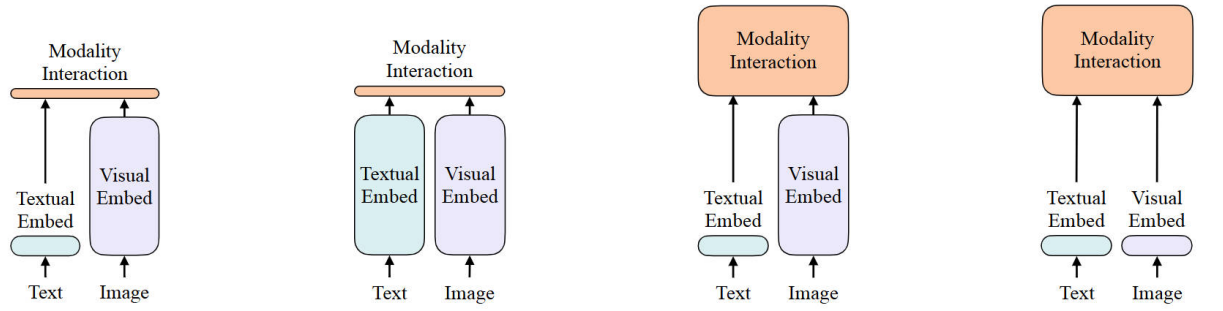
Pre-training：训练数据，Image-Caption Pairs；训练数据集，MSCOCO、VG、GCC、SBU；训练**Objectives**，Image-Text Matching、Masked Language Modeling

Dataset	# Images	# Captions	Caption Length
MSCOCO	113K	567K	11.81 ± 2.81
VG	108K	5.41M	5.53 ± 1.76
GCC†	3.01M	3.01M	10.66 ± 4.93
SBU†	867K	867K	15.0 ± 7.74

Downstream Tasks：方式，Fine-tuning or Zero-shot；任务，Classification Task（VQAv2、NLVR2）、Retrieval Task（Flickr30k、MSCOCO）

Related Works：

VLP（Vision-Language Pre-training）Models的分类：**(a)** 如VSE；**(b)** 如CLIP；**(c)** 如ViLBERT、UNITER；**(d)** 如ViLT；

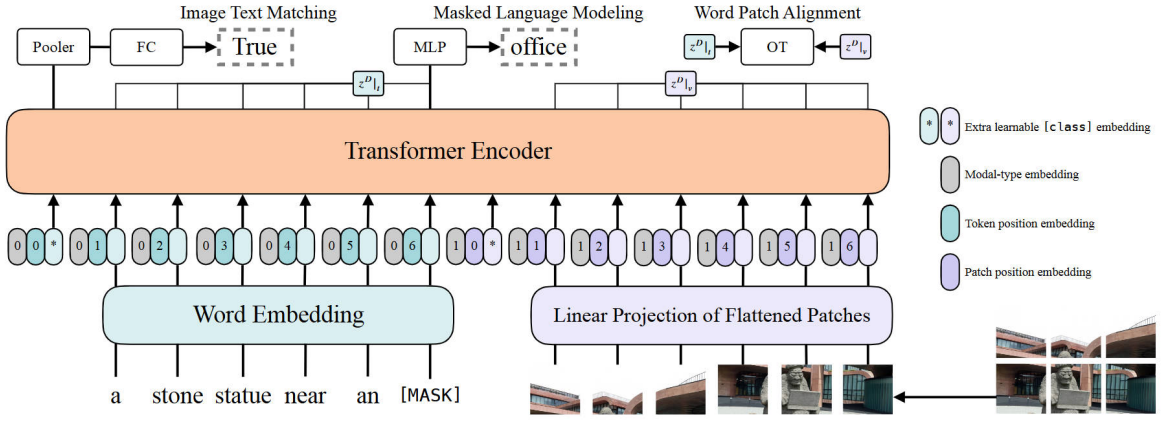


(a) VE > TE > MI (b) VE = TE > MI (c) VE > MI > TE (d) MI > VE = TE

Modality Interaction Schema：**1)** Single-Stream Approaches，Concatenate；**2)** Dual-Stream Approaches

Architecture：

图示：为Single-Stream Approaches；**Transformer Encoder**输入Tensor为[(N+1)+(L+1), H]；**Pooler**学习矩阵为[H, H]，它的输入Tensor为[1, H]，通过**FC**实现**Image Text Matching** Objective训练（二元分类），即Image-Text匹配或不匹配；**Masked Language Modeling** Objective训练与BERT的完全一致，即做完形填空，又有特殊的地方，使用了**Whole Word Masking**（15%的占比），这要求要完全依赖于图像的信息来做文本的完形填空；**Word Patch Alignment** Objective训练，利用最优运输理论，通过降低文本分布和图像分布间的距离来训练；图像数据增强，**Rand Augmentation**，RA有很多的Policy，但有两个不用，Color Inversion、Cutout。



Future Works：

Scalability：Parameters、Datasets

图像的完形填空：

Image Augmentation Strategies：