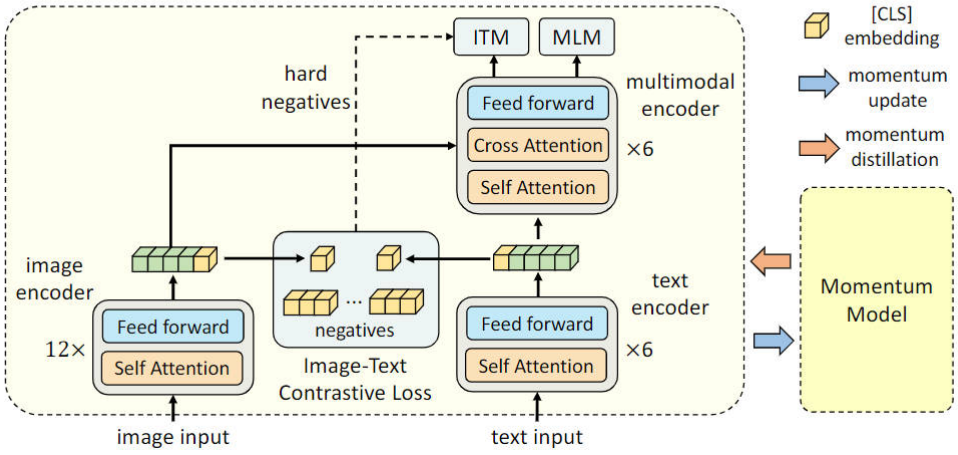
Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

---

动机以及解决：

动机/问题	解决措施
Visual Encoder给到Modality Interaction的是region-based image features，而Text Encoder给到Modality Interaction的已经是词向量word tokens了，这是不对称的。	Align Before Fuse：使用Image-Text Contrastive（ITC）实现在Fuse之前先对齐好。
Noisy Data：web image-text pairs data往往很noisy，表现在描述图片的文本只是一些关键字Hashtag。	提出Momentum Distillation，一种self-training方法（使用伪标签pseudo labels来训练模型），是采用Moco Paper中的Momentum Model来生成伪标签。
Region-Feature中需要的组件极大地延迟了Inference速度。	和CLIP与ViLT一样，全部都是用的是Transformer架构。

Architecture：



negatives sample是由Momentum Model产生的

momentum update是通过average moving实现的

Loss Function： L_{itc}、L_{mlm}和L_{itm}

目标函数：	
Image-Text Contrastive(ITC) Loss	image的CLS token为[1, 768]，downsample为[1, 256]，text的CLS token为[1, 768]，downsample为[1, 256]，这是两个正样本对；负样本储存在一个队列queue中，有65536个，这个是由Momentum Model产生的。
Image-Text Matching(ITM) Loss	在最后的representation后加上一个分类头FC，二分类任务，判断Image和Text是不是一对，这个目标函数过于简单。对负样本加上一些要求或限制，在本论文中使用的方法是，在所有负样本中选择最难的（hard negatives），即最接近正样本的那个负样本，在ITC中的queue中找，这样一来，该目标函数就变得非常challenging了。
Mask Language Modeling(MLM) Loss	将text tokens中的部分tokens mask掉，再和image tokens一起丢给multimodal encoder，要它来预测被mask掉的tokens。

在pre-training时ITC和ITM都是输入的是original text，而MLM输入的是masked text，所以在每个iteration中会有两次forward。

Momentum Distillation：

针对Data Noisy（如，负样本的文本信息可能正确地描述正样本的图像信息；正样本的文本信息可能对正样本的图像信息描述不全），one-hot labels会妨碍模型的学习。

自训练self-training，用一个momentum model生成pseudo-targets，ground-truth就不再是一个one-hot label（这个对应于Cross Entropy）了，而是一个softmax score（这个对应于KL Divergence），momentum model通过在已有的模型上做EMA来构成。

重构Loss Function，得到L_{itc}^{mod}和L_{mlm}^{mod}。

在部分downstream task的fine-tuning时也有做momentum distillation。

有scalability，在dataset层面。

Pre-training Dataset和Downstram V+L Task：

DATASET：	DESCRIPTION：
Conceptual Captions	web datasets，两个版本，CC- 3 million和CC-12million
SBU Captions	web datasets，1-million
COCO	in-domain datasets，10,000
Visual Genome	in-domain datasets，10,000

TASK：	DESCRIPTION：
Image-Text Retrieval(ITR)	包含了两个subtasks，图像到文本检索（TR）和文本到图像检索（IR）；metric，Recall（R1、R5、R10）；数据集，Flickr30K、COCO。
Visual Entailment(VE)	将其转变为三分类的问题，entailment、neutral和contradictory。
Visual Question Answering(VQA)	可以将其转变为一个multi-answer classification problem（闭集VQA），或将其转变为一个answer generation problem（开集VAQ）。
Natual Language for Visual Reasoning(NLVR)	要求模型预测一个文本能否描述一对图片，是一个二分类问题。
Visual Grounding(VG)	

Analysis：

互信息最大化角度Mutual Information Maximization Perspective进行理论分析，说明ITC、ITM、MLM和Momentum Distillation这些训练目标都是让模型从很多image-text pair中学习到不同视角的语义特征。