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

---

unified：“结合”understanding-base tasks和generation-base tasks。

bootstrapping：去除数据noisy。

前人工作不足：

① 模型层面：要么只适合于understanding-base tasks，要么只适合于generation-base tasks。

② 数据层面：用于pre-train的数据，部分来自web，比较noisy。

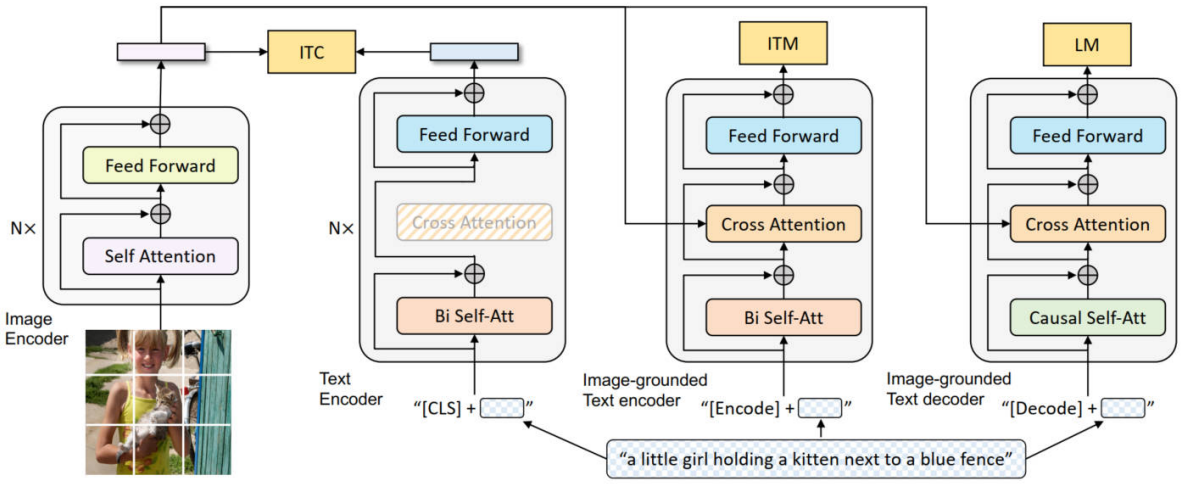
动机：

① 模型层面：用一个Framework来解决understanding-base tasks和generation-base tasks。

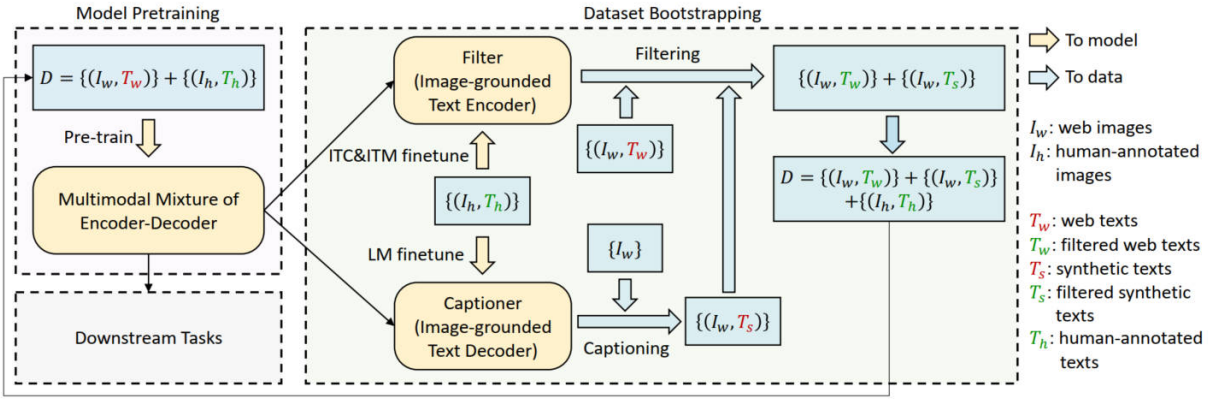
② 数据层面：去除noisy，增强数据质量。

方法：

① 模型层面：很大程度地借鉴了ALBEF（BLIP和ALBEF出自同个团队）和VLMo，对于ALBEF，如ITC Loss中的Momentum Distillation，ITM Loss中的hard negative mining strategy。对于VLMo，则借鉴它“一个Framework解决多种tasks”的思想。对于第3个objectives Loss，LM Loss和GPT类似，给定前面的一些词，要求预测剩下的词（不同于MLM，MLM类似于完形填空，在句子中间挖掉一个词，然后再预测这个词）。



② 数据层面：



流程：1) 在noisy的数据集上pre-train Multimodal Mixture of Encoder-Decoder；2) 用pre-trained的Multimodal Mixture of Encoder-Decoder组件构建Captioner和Filter，在人工标注（noisy很小）的数据集上fine tune Captioner和Filter；3) 最后，利用训练好的Captioner和Filter生成高质量的数据集；4) 用这个高质量的数据集重新pre-train Multimodal Mixture of Encoder-Decoder。

- Captioner：为Image Encoder+Text Encoder和Image Encoder +Image-grounded Text Encoder。
- Filter：为Image Encoder+Image-grounded Text Decoder。

创新点：

① 模型方面：集成了ALBEF和VLMo。

② 数据方面：！！！

这两方面导致了BLIP是分阶段训练的。