CoCa: Contrastive Captioners are Image-Text Foundation Models

---

Co：Contrastive Loss。

Ca：Captioning Loss，即LM Loss，也就是GPT中的Loss Function。

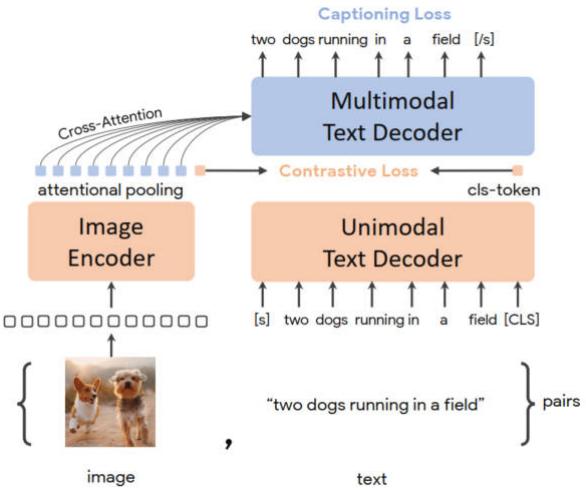
前人工作不足：

（Vision-Language） Pre-trained Model的发展历程：① single-encoder models，Loss Function是cross-entropy loss，不适用于多模态任务；② dual-encoder models，Loss Function是contrastive loss，不适用于较复杂的多模态任务，如VQA、VR、VE、generation tasks；③ encoder-decoder models，Loss Function是LM loss，不适用于简单的多模态任务，如retrieval。

动机：

对于前人工作的不足之处，作者想要实现One Framework solve multi type VL tasks。

方法：



attentional pooling参数可学的pooling 层。

Unimodal Text Decoder使用的Transformer的Decoder，解决了分阶段训练的问题，提高了效率，即1次iteration，forward一次image和text。

创新点：

集成single-encoder models、dual-encoder models和encoder-decoder models，并能很好的transfer到众多downstream tasks。

避免的分阶段训练，提高了训练效率。

效果抢眼！

效果：

在2.1billion数据上进行预训练，得到如下的“多边形”效果图。