用多模态信息做 prompt，解锁 GPT 新玩法

原创 子龙 夕小瑶的卖萌屋 2021-10-18 12:05

文 | 子龙
编 | 小轶

自多模态大火以来，井喷式地出现了许多工作，通过改造预训练语言模型，用图像信息来增强语义信息，但主要集中在几个 NLU 任务上，在 NLG 上的研究比较少。

今天要介绍的这篇 paper **Multimodal Conditionality for Natural Language Generation** 研究的任务场景则是**以多模态信息作为条件做 conditional 的 NLG**任务。这种任务设置有许多实际的应用场景。比如，生成商品介绍文案时，仅仅基于该商品的文字标题是不够的。如果能结合商品的图片，必然能够得到更贴切的文案。

这篇工作的模型基于 GPT2，而**多模态信息则是以一种类似 prompt 的方式来使用。**虽然方法比较简单直观，但具备一定通用性，未来或许有进一步挖掘的可能。

论文题目:
Multimodal Conditionality for Natural Language Generation

论文链接:
https://arxiv.org/pdf/2109.01229.pdf

原理

作者的想法其实十分简单，一切语言模型都是为了衡量一段文字序列的概率，即：

$$p(\boldsymbol{x}) = \prod_{i=1}^{n} p(\boldsymbol{x}_i | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1})$$

而如果引入了多模态的输入，就相当于在生成时多了一个条件，即条件概率为：

$$p(\boldsymbol{x}|\boldsymbol{y}) = \prod_{i=1}^{n} p(\boldsymbol{x}_i | \boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1})$$

其中y为多模态输入序列。

以文中生成商品文案的运用场景为例：

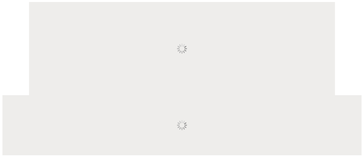
这里的Product Title和Product Images就是作为生成Product Description时的"条件"。

那么如何将多模态序列引入到自然语言生成模型呢？

本文使用了一个十分直观的方法，称作MANTIS，将作为条件的多模态序列作为前缀放置到decoder输入序列的前面，进而中解码过程中分享多模态信息。其中图片输入借助ResNet-152，将最后一层输出用线性层映射到语言模型同一个空间中。而作为条件的文本输入，即这里的product title，和生成序列一同进行编码。

效果

数据集采用FACAD，提供了商品的标题和图片，目标是生成产品描述，效果如下：



文中提出的模型在所有指标中都取得了最优结果，相比于baseline，将BLEU4提升了0.8，CIDEr提升了7.2，METEOR提升了0.8，ROUGE-L提升了1.0。同时，由于衡量生成文本质量具有主观性，作者也进行了人工评分，结果表明MANTIS依然取得了最优结果。

从生成效果来看，生成的描述成功地结合了图片信息，使得描述更加准确，而非笼统的介绍。

总结

这篇文章方法十分直观，但是结合最近火热的 Prompt，似乎又有了更多的启发。同样是生成，同样是加前缀，似乎给定条件的生成就是加上编码好的前缀？那么多模态末来能不能成为一种新的prompt呢？作者认为他们的模型可以借助各种不同的多模态条件生成，然而不得不说本文的方法对模态融合的部分做的马虎了些。本文只是单纯借助解码器进行融合，并没有在编码阶段就分享跨模态的信息。

萌屋作者：子龙(Ryan)
本科毕业于北大计算机系，曾混迹于商汤和MSRA，现在是宅在UCSD(~~Social Dead~~)的在读PhD，主要关注多模态中的NLP和data mining，也在探索更多有意思的Topic，原本只是贵公众号的吃瓜群众，被各种有意思的推送吸引就~~土下舅舅~~，希望借此沾沾小屋的灵气，paper++，早日成为有猫的程序员！

作品推荐：

1. 别再搞纯文本了！多模文档理解要被时代需要！
2. Transformer哪家强？Google爸爸辨优良！
3. 预训练语言真的是世界模型？