《人工神经网络》大作业中期报告

陈博涵
2018011441
计算机系计86班
cbh18@mails.tsinghua.edu.cn

冯卓尔
2017011998
计算机系计86班
fengzhuoer-thu@outlook.com

1 引言

有时，人们在Twitter, 微信等社交工具上想要发条推文/ 发个朋友圈，已经有几张精心挑选的图片，却纠结于写一段与之相关的配文。

我们试图基于已有的模型，实现一个根据**若干图片**生成相应配文的模型，并能够综合考虑图片所表达的含义，使配文尽可能接近所期望的主题，但又能有**一定程度的联想**。

我们将大致分为两个阶段。在第一阶段，我们分别处理两个子问题，借鉴现有SOTA的解决方案，得到一个baseline。在第二阶段，我们将两个模型融合，着重考虑模型连接点的处理方式，并加入对多图片输入的支持，增加文本生成的连贯性与合理性。

2 问题陈述

对于这个任务，更数学化的表述是：给定图片集S $(1 \leq |S| \leq 4)$，输出一段英文文本$\mathcal{P}$ $(\text{number of words in}\,\mathcal{P} \leq 100)$ 作为S 的配文。

Table 1: 数据集

名称	描述
Flickr30k	Flickr Image captioning dataset
COCO 2014	A large-scale object detection, segmentation, and captioning dataset

由于这个问题的输出具有一定的主观性，在完整的问题上我们暂时采取人工评价的方式，可能会在未来引入量化指标。对于其两个子问题，我们采用BLEU score作为评价指标。

我们希望对于尽可能多的输入，输出的文本能够以下要求：

- 贴近图片主题以及所表达的含义。
- 对于多图片输入，生成的文本比较连贯。
- 具有适当但不离谱的联想。

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

3 方法

3.1 起手baseline模型选取

我们首先做了一系列的调查，寻找相关的数据集、模型。对于图片处理，我们首先选用的show-attend-and-tell模型(以下简称show-and-tell)；在文本生成部分，我们采用了GPT-2。GPT-2基于预训练的BERT模型进行训练，体量庞大，我们选用了目前官方公开的精简版。

3.2 运行环境

8vCPU, 32GB RAM, 1 NVIDIA Tesla V100 GPU

3.3 参数调节

在初步调制好GPU运行环境后，我们对我们复原的模型进行了一系列的参数调节。

对于show-and-tell，我们先用默认超参数以及pretrained模型参数进行了实验，对于各种图片都能够得到比较合理的caption，但是对于一些明显超出数据集覆盖范围的图片，模型就会输出一些无关的内容，也属正常现象。后来我们更改cnn部分所选用的模型为ResNet50(原为VGG16)，重新在COCO 2014数据集上进行训练。由于完整训练时间较长，我们在训练了2个epoch后进行测试，发现模型对于不同的图片会给出一些相似的表述，且部分描述与图片无关，可能是训练不够充分所致。

对于GPT-2，影响参数较多。我们的任务对应其conditional sample generation任务。首先对于sequence的影响，我们对其马尔科夫过程考虑的seq长度进行探究，官方申明40左右的长度最佳，即top_k=40。在我们的实验中，top_k=30 50之间生成的文本都没有过大的偏离。其次是使用的数据集规模，官方发布了从124M，355M，774M，1558M四个规模的数据集，我们在云服务器上均进行了实验，结果发现1558M规模的数据集覆盖的语料面较广，对于一些奇怪的图片都能够写出较符合逻辑的后文。

3.4 数据集

我们发现了一些问题（见第五部分）后，决定自己扩充一部分的数据集来支持我们的模型。我们尝试对Twitter进行了爬取，但是我们的脚本容易触发其机器人检测，目前在这里没有获得进展。我们的计划详见第五部分。

4 初步实验结果

基于show-and-tell与GPT-2的连接，我们初步能够通过输入图片得到指定长度的配文。一下是五组实验结果，我们的设置为一张图片长度的输入，对应50字左右的文案输出。

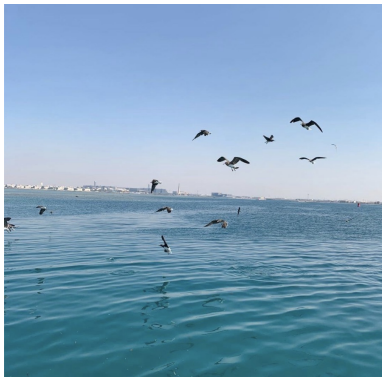
4.1 著名景点

If you climb up the clock tower. There are many places for you to go. There is a big door you can go through, and there is a long corridor where some of the rooms are. Just keep going around the tower

until you get out....

分析：该文本基本符合塔桥图片所示的场景，文本介绍了图片中场所的一些特点，并邀请读者浏览。

4.2 自然景物



A large tree.

A group of people. One man, sitting.

A group of men, standing.

A collection of different animals.

An entire forest, from top to bottom.....

分析：该文本存在着一些不足：文本中含有较多的图片中并没有出现的元素，但是可以理解为联想产物；但是这一段分享文介绍的主题已经偏离了图片的景物，而且第一句生成的文本对后文的生成有着较大的影响。

4.3 动物

Cat with its own light bulb.

Cat with its own mouse pad.

Cat trying to reach a door.

Some cats are very good at solving door problems.

The cat does not like cats.

分析：该文本基本符合预期，文本的主要内容是围绕着猫展开的，并没有明显的偏差，同时也起到的鼓励分享的作用。

4.4 交通工具

What I'm seeing is a little like flying. I know this because when I was seven and riding in a red school bus through Manhattan, I felt like I was on a different planet. Every street in the city looked different. The bus.....

分析：该文本基本符合预期：明显，双层巴士是伦敦的，文本中出现曼哈顿的内容是联想得到的，并没有逻辑上的错误，同时故事内容具有吸引力。

4.5 人物

The man walks around the building, looks in a window, and then walks to a back door, opens it, locks it, and walks out.

The man then drives away.

The next evening, the man appears again, this.....

分析：该文本存在着一些不足：文本介绍了图片中主要的内容为穿西装的人，可能因为图片的清晰度，在识别的过程中识别的动作为驾驶；但是和图片本身的意义存在一些差距。

5 目前的困难

5.1 GPT-2模型的修改

目前，我们下载了官方发布的精简版的GPT-2类型。通过环境调试能够实现从图片读入（224 × 224）进行特征提取，将提取后的特征带入baseline的GPT-2进一步处理，最后生成给定长度（例如50）的文本。这个模型很大程度上依赖于原初的模型，特别是的训练的语料，我们试图对原有的参数进行调试，目前认为对于GPT-2，top_k（文本生成时，模型考虑sequence后的最大概率出现文字的时候，其前缀单词的考虑数量）为40-50时，产出的文本相关度最高，并且不至于生成长文本的时候不随着长度增长导致注意力转移。

5.2 show-and-tell模型机制修改，Transformer的应用

show-and-tell我们使用的baseline模型采用的是先将输入图片通过CNN，提取出a string of characters，此后将这串特征加入之后的LSTM模型，将LSTM的long term output给出的结果进行编码，输出结果。然而，目前我们的baseline是通过将show-and-tell中出输出的结果作为text-token输入GPT-2中的Transformer，这个过程存在着较大的信息丢失和曲解。我们在处理show-and-tell输出部件的时候，希望能够将LSTM在运行过程中提取到的特征信息（raw seq）取出，并且将其处理成为按照重要性降序排列的list。但是这一个过程操作的验证较为复杂：我们能够快速取出这些信息，但是这些信息是否确实是图片的真实特征、这些信息是否能够支持GPT-2进行进一步的加工，这两点在操作上会耗费较大的时间人力和精力。

5.3 多张图像的特征提取

目前我们使用的show-and-tell模型能够支持单张图片输入后，给出特征提取的内容。但是对于多张图片的输入，我们需要修改原有模型，使其能将其编码成合适的特征传入下一层。最直观的方式是对每张图片分别用LSTM提取时序特征，然后加入分隔符后首尾相接作为下一层的输入，但这样可能会使得本来无序的图片输入产生人为安排的顺序。另一个可能的方式是对每张图片分别用Transformer提取特征(包含位置信息)，然后加一层Attention将多张图片的特征信息综合，这样在生成文本时模型也许就能同时处理多个图片的信息，使得产生的文本在语义上更为连贯。我们需要进一步实验来验证以上方法的效果。

5.4 GPT-2的拓展

GPT-2基于Transformer实现unconditional和conditional的文本生成，在conditional的文本生成中，Transformer依赖于自编码器auto-encoder实现了语义的切分和输出文本的整合。但是，如果要对精简版的GPT-2进行优化，特别是使用LSTM的方法对其Transformer优化、使得其生成的文本与后文的相关程度更高，需要修改的内容略大。

在问题的定义过程中，我们没有考虑到“生成的分享文案”可以通过怎样的标准定义：我们缺少一个可量化的指标对模型生成的文案与人工写的文案进行比较，从而没有办法对模型进行监督式的干预。目前模型生成的文本主要通过其训练预料进行调整，所幸GPT-2原模型使用的语料体量非常大，能够覆盖非常多的话题和文本风格，但是我们更加希望能够使用Facebook或者Twitter来源的数据对其进行finetuning。具体的操作方法，需要我们实现一个网页脚本，对Twitter或Facebook中配图的英文动态进行爬取，爬去的过程本身就建立了图片到seq的一个标注；然而，对这样的数据集进行后一步的处理（即加入模型训练）或许需要对目前模型的pretrained模型——即BERT进行重训，这些操作都需要比较大的算力，以及数据标注中的人力。

如上文提到的，我们最终的结果也缺少一个硬性的指标来衡量。评判机器生成文案的优劣可以从语法、表达、口语化、情绪等方面进行评价，我们试图引入这些指标来评价。目前两个试探性的解决方案如下。一、引入语法、表达、口语化、情绪等参数评价，对于机器生成的样本，我们制作训练集或者人工观察输出，对于某一个输入图片，其分享配文的情绪、口语化程度会有一个expected的结果，根据这一结果（TorF）的比对，我们可以计算模型的所谓“生成命中率”。二、对于同一张图片或者一系列图片，雇佣人力对其进行配文撰写，随后邀请实验者（一般为同学，或者微信朋友圈发布）对机器生成与人力撰写的配文进行图灵测试，测试准确率可以在统计推断的意义上认为是模型准确率。

参考文献

[1] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2017). Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4), 652-663.

[2] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., & Salakhutdinov, R., et al. (2015). Show, attend and tell: neural image caption generation with visual attention. *Computer Science*, 2048-2057.

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances *in neural information processing systems* (pp. 5998-6008).

[4] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).