《人工神经网络》大作业开题报告

陈博涵
计算机科学与技术系
清华大学
cbh18@mails.tsinghua.edu.cn

冯卓尔
计算机科学与技术系
清华大学
fze17@mails.tsinghua.edu.cn

1 任务定义

考虑这样的场景:

> 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。

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$\mathscr{P}$ $(\text{the number of words in } \mathscr{P} \leq 200)$ 作为S 的配文。

2 数据集

Table 1: 数据集描述

名称	描述	大小
Twitter UK Geolocated Tweets	170K tweets from UK	47 MB
NYTimes Facebook Data	all the NYTimes facebook posts	5 MB
Flickr Personal Taxonomies	Tree dataset of personal tags	40 MB

3 挑战和基线

3.1 挑战

在这个问题中，如何将由不同配图生成的语义进行结合是一个比较有挑战性的任务。

从不同的配图可能生成出毫无关系的配文，但是现实中人们所想在推文中表达的内容一定不会是割裂的。因此如何兼顾图片的共性，令其成为一段有意义的文字，是本问题的主要挑战。

此外，两个子问题的连接可能不会非常顺利，在将Image Caption 任务的输出作为Language Generation 任务的输入时，可能会出现逻辑上的断层，导致生成结果与图片原先的意思相去甚远，这也是需要解决的问题。

3.2 基线

Show and tell (Vinyals et al, 2017) 一个能够对图片中的元素进行捕捉并且识别，生成数个标签或者一个句子来标识这张图片。

GPT-2 (Radford et al, 2019) 一个能够对文本进行阅读理解、形成摘要、翻译、续写生成的模型，相比于其前身，GPT-2的突出优势是轻量。

4 研究计划

- 第一阶段，在Milestone 提交之前先完成两个子问题的模型Baseline，即Show and tell 与GPT-2 模型的初步搭建，主要使用的工具为PyTorch与CoTK。由于资源有限，最终的Baseline 性能根据实际情况有所弱化。
- 第二阶段，在Milestone 提交之后探索将Image Caption 问题的模型结果与GPT-2 模型的输入进行对接，如果有必要，重构这两个模型之间的连接架构，使得整体组成一个系统，实现任务背景中的需求。
- 第三阶段，修改并且对原模型进行创新，两个子问题拼接而成的模型往往不是最优的模型。根据实际情况，我们将对模型进行拓展，例如增加输入图片的数量($S = 9$)，使之综合形成一个系统；或是调整Baseline 中的模型，对其进行更加细腻的调试，对整体模型的性能进行优化。

5 可行性

本任务总体上包含两个子任务：Image Caption 和Language Generation，而这两者现在都有一些相对成熟的模型和解决方案，如Image Caption 的Show and Tell, Show, Attend and Tell 以及Language Generation 的Transformer, VAE, GAN 等。

基于这些已有的模型，我们能够较快实现一个Baseline，并在此基础上进一步解决挑战和难点(3.1)，优化模型表现。

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).