CoTK-ST & CoTK-SET：图像描述复现

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

冯卓尔
2017011998
计算机系
fze17@mails.tsinghua.edu.cn

Abstract

描述图片任务将自然语言生成与计算机视觉两个领域结合在一起，是一项较为综合的任务，这个任务能够让研究者更好地了解计算机在识别图像过程中的性能，以及实现人工智能感官方面的延伸。2015年前后，谷歌研究员发布了在这个研究方向中重要的模型show-and-tell，这个模型使用了autoencoder的想法，将CNN作为encoder、LSTM作为decoder，实现了图像-语言的任务衔接。本文主要呈现我们对这个模型的复现，以及借鉴show-attend-and-tell模型，通过添加注意力机制的方法对其进行优化。相比于现有的复现模型，我们使用了CoTK工具包并采用PyTorch框架完成。

1 引言

描述图片中的内容是一项基础性的人工智能问题，它将人工智能中两个重要的领域计算机视觉与自然语言处理结合了起来。相比于区别图片中的物体，描述图片中的内容需要更多更复杂的处理：与单纯区分出图片中不同形状的物体的存在，描述图片的内容任务还需要辨别图片中的物体的主次、相互关系以及物体以外的背景等元素，同时需要基于这些观察和捕捉生成语义语境以及情感上相称的语句。一旦模型能够实现这些模型，我们能够从生成的文本中得知神经网络是如何处理这些信息的，这对我们开发计算机视觉新模型有着启发意义。这个问题很早就得到的研究，但是直到数年前才得到突破性的进展，主要阻碍着这个问题研究的困难包括如何形式化定义这个问题，使用什么方法衡量生成文本与目标图片的适配程度，以及如何将计算机视觉中图片的信息传递给文本处理单元等。要形式化定义这个问题在最初有些困难，因为缺乏衡量结果程度的指标。Ali Farhadi等人[3]试图将图片通过三个维度的特征进行表述，即（物体，动作，场所）这样一个向量，这种描述方法给出了可加工的数学模型，但是这样的简化还有些粗糙，痕量指标的意义并不明显。在2013年，数位IEEE成员发表了题为"BabyTalk: Understanding and Generating Simple Image Description"的文章[2]，使用经典的n-gram的指标，即将结果和预先由人类标注的结果进行比对，为了避免表达方式等习惯的不同，在比对过程中不再是单个字层上的比对，而是逐个取长度一定的字块，按照出现频率进行评价，以表示通过图片生成的文本与目标文本之间的重合率。通过这些方法，这个问题逐渐形成了形式化的目标，即希望目标生成文本尽可能与指定的语句相似。

> I为图片输入，最大化$p(S|I)$，$S = \{S_1, S_2, ..., \}$，其中S是所有给定单词的集合字典。

此后，这个领域众多的工作与模型都是基于这个问题而展开的。我们的主要任务是复现这个问题中的一个经典模型show-and-tell[4]，它从机器翻译的一个模型[1]变迁而来。机器翻

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

译的过程中通过autoencoder机制，使用RNN作为encoder与decoder对文本进行编码，show-and-tell将encoder的RNN换成了针对图片效率较高的CNN，形成了一个模型。最后，仿照人类地，在观察一张图片时，人类的注意力并非均匀地分布于整个图片空间内，而是集中于特定的数个点，因此2015年数位学者在show-and-tell的模型基础上增加了attention[5]，形成了当时的state-of-the-art，这两个模型便是我们工作主要复现的。我们使用我们复现的模型首字母，将我们复现的show-and-tell称为CoTK-ST，将复现的show-attend-and-tell称为CoTK-SET。我们复现的模型的主要结构与理论参照[4]论文原文。我们复现过程中使用的加载数据以及计算指标工具包为CoTK。

我们的研究过程主要分为三步。第一步，模型框架搭建，依照论文与开源代码给出的框架实现模型；第二步，模型运行与调试，这个阶段我们对模型进行调试。我们实现的模型在架构上遵循OOP原则架构，对于添加新的任务模型非常方便（例见text transfer）；第三步，测定指标与模型改动，这一阶段适用于检验我们模型实现的正确性，主要基于baseline的参数进行参照对比。

2 相关工作

n-gram

n-gram是一种重要的基于马尔科夫过程的语言模型。在自然语言处理的各种技术中，n-gram的一个经典应用是在词向量中，用于鉴别两个词在使用方式和语义上的相似程度。N-gram模型是一个概率模型，$P(x_i|x_{i-N+1},...x_{i-1})$是长度为$N$的词语序列$(x_{i-N+1},\dots x_{i-1}))$的概率分布，它预测了所有在词典中的词语作为接下来一个词语出现的概率。

图像标注子任务

在2010年，几位研究人员实现了这个任务的一个子任务，即将每张图片视作一个（物体，动作，场地）的三元体，将描述图片任务转化成为三元体生成任务，通过这个三元组，可以通过便捷的生成器生成最终的语句结果。在他们的研究中，图像空间的信息被提取至三元组构成的空间中，而在句子空间中每一个句子也能够提取中这样有效的三元信息，他们的任务便是优化来自图像空间的三元组与句子空间的三元组在提取过程后结果的相似度。他们在PASCAL 2008 的数据集上进行训练，并且得到了较接近人类实验结果的生成准确程度。

在2013年发表的一篇题为"BabyTalk: Understanding and Generating Simple Image Descriptions"的文章中，研究者将生成图片的描述作为一个目标任务进行了研究。相比于前文的任务，这篇报道中研究者并不通过三元组通过固定的模板生成句子，而是通过直接生成的方式进行描述，其形式不再拘泥于"信息空间"的向量形式，形成的句子变化更多更强。

BLEU指标

在机器翻译领域, 我们需要一个指标来衡量机器翻译的结果与专业人工翻译结果的差异, 一般情况下,如果需要比较好的结果都需要专业的翻译人员介入评估模型翻译的好坏，但是这样需要大量的人力参与。BLEU 就是这样的情况下作为一个较好的且能够快速衡量翻译文本质量的指标。

在图像描述任务中，我们也可以使用BLEU来衡量机器生成的描述与人工标注之间的差异，从而对模型的能力给出大致的估计。

CoTK

CoTK 是一个用于模型构建和量化的轻量级开源框架，尤其是对于语言生成领域，其中内置了很多标准的数据集以及指标的实现。

3 方法

3.1 模型任务[4]

本模型总体任务是实现以下目标函数的优化，其中θ 是模型的参数（变量），S 为目标生成文段，I 为图片输入。

$$\theta^* = \arg\max_{\theta} \sum_{(I,S)} \log p(S|I;\theta)$$

将上式代入具体的长为N 的$S_0S_1S_2...S_N$ 组成的语句中，由上式的定义展开，得到

$$\log p(S|I;\theta) = \sum_{t=0}^{N} \log p(S_t|I, S_0, S_1, ..., S_{t-1})$$

该式为LSTM需要优化的目标。

3.2 Encoder：ResNet50

我们使用了准确度较高且模型规模适中的ResNet50对输入的图片进行encoding。对于每张图片，令$V_{\text{in}} = \text{CNN}(I)$ 为提取到的图片特征。

3.3 LSTM逻辑

在LSTM中，我们将通过CNN的图片输入I 结果记为x_{-1}，那么LSTM的初始输入为$x_{-1} = V_{\text{in}}$，根据LSTM的机制，我们将每次的output 视作概率p_t，如下图

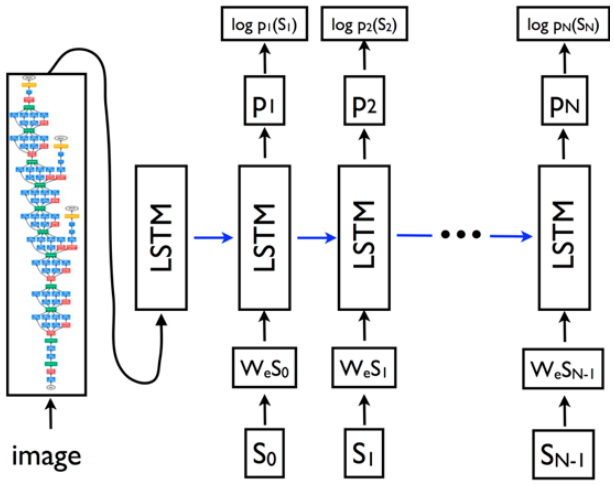

Figure 1: LSTM 结构

$$\begin{aligned} x_{-1} &= \text{CNN}(I) \\ x_t &= W_e S_t, t \in \{0, 1, ..., N-1\} \\ p_{t+1} &= \text{LSTM}(x_t), t \in \{0, 1, ..., N-1\} \end{aligned}$$

我们将模型的loss 定义为每一步output 概率p 的负倒数之和。

$$L(I,S) = -\sum_{t=1}^{n} \log p_t(S_t)$$

3.4 Attention

在show-and-tell模型基础上，我们在decoder阶段为LSTM添加了attention。

Attention作用于LSTM，主要对每一处input调整权重，其公式为

$$e_{ti} = f_{\text{att}}(\mathbf{a}_i, \mathbf{h}_{t-1})$$
$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})}$$
$$\mathbb{E}_p(s_t|a)[\hat{\mathbf{z}}_t] = \sum_{i=1}^{L} \alpha_{t,i}\mathbf{a}_i$$

$$\begin{aligned}\text{NWGM}[p(y_t = k|a)] &= \frac{\prod_i \exp(n_{t,k,i})^{p(s_{t,i}=1|a)}}{\sum_j \prod_i \exp(n_{t,j,i})^{p(s_{t,i}=1|a)}} \\ &= \frac{\exp(\mathbb{E}_{p(s_t|a)[n_{t,k}]})}{\sum_j \exp(\mathbb{E}_{p(s_t|a)[n_{t,j}]})}\end{aligned}$$

4 实验

4.1 数据集

在训练和测试中，我们选用了MSCOCO2014 (Lin et al., 2014) 数据集。该数据集在此任务上是benchmark数据集，相对而言具有典型的参照价值。在CoTK中，已经实现了MSCOCO2014的dataloader，但是没有图像描述任务所需的图片。我们在阅读代码后扩展了该dataloader，使其能够同时支持图像加载。

4.2 参数设置

以在MSCOCO2014数据集的训练为例，我们的模型参数设置如下。

首先，针对一组数据（以一组数据为例，实际数据增加一个batch的维度），图片缩放至$224 * 224$ 的大小作为输入经过ResNet50，输出维度为$9 * 2048$，该向量将经过attention 作为LSTM 的输入。标注数据通过embedding_layer 进行embedding，词向量维度embedding_size=300。LSTM 的隐藏层大小hidden_size=500。在LSTM 的输出和下一层的连接处我们均加了0.5 的Dropout。

我们采用Adam 优化器，其中关键参数取值为：learning_rate=1e-4, beta1=0.9, beta2=0.999。

训练过程中我们使用了Tensorboard X 实时监测Loss 的变化趋势（附图见下一页）。

在Eval 阶段，我们实现了beam_search 生成具体的文本，在实际测试过程中取beam_size=3，最后我们能够获得可能性前3 大的输出。

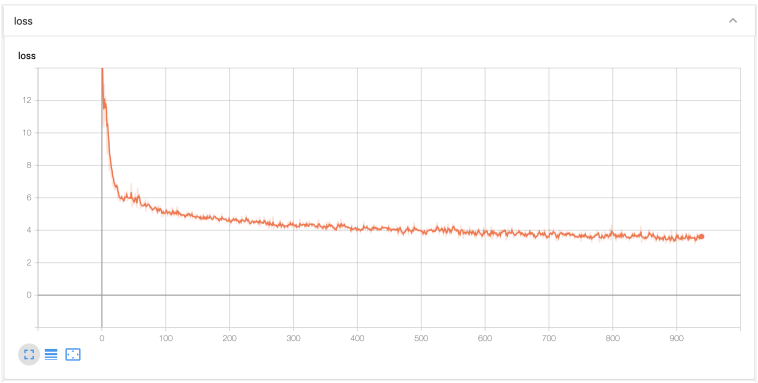

Figure 2: Loss 曲线

4.3 Ablation Test

4.3.1 Show-and-tell

在论文中，show-and-tell使用了autoencoder的方法，其灵感来自于machine translation的autoencoder模型，后者的模型结构为LSTM作为Encoder + LSMT作为Decoder，前者的模型为CNN作为Encoder，LSTM作为Decoder。因为图片本身并非序列，使用RNN对图片信息进行encoding进行ablation test过于没有意义，因而我们忽略这个ablation test，认为这是一个阳性指标。（原因：RNN模型适用于序列任务）

4.3.2 Show-attend-and-tell

这个模型与show-and-tell模型的区别是attention的使用，因此在原论文的结果和我们的实验结果中可以看出，attention在模型表现中的重要作用，它提升了所有指标结果。

4.4 定量结果

Table 1: 数据集

数据集	模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4
MSCOCO2014	Google NIC	66.6	46.1	32.9	24.6
	COTK-ST	56.7	37.2	24.9	19.1
	NITK Soft attention	70.7	49.2	34.4	24.3
	COTK-SET	59.2	39.7	26.5	19.6

4.5 案例分析

4.5.1 成功案例

a. Model: a man riding a snowboard down a snow covered slope.
   Human: a male snowboarder traveling down a ski slope.

b. Model: a box filled with lots of donuts.
   Human: a pack of donuts with different flavors.

c. Model: a man is playing tennis on a court.
   Human: a tennis player is standing on the court.

(a)

(b)

(c)

(d)

Figure 3: 成功案例

d. Model: a train traveling down train tracks next to a forest.

   Human: a train on a train track near many trees.

4.5.2 失败案例

a. Model: a bird is flying over a body of water.

   Human: a dog jumps to catch a thrown frizbee.

   分析：物体从狗识别成了鸟。

b. Model: a window with a door open window.

   Human: a horse looking out a window in a brick building.

   分析：没有识别出“马”。

c. Model: a man sitting at a table with a banana.

   Human: a man making a goofy face while sitting near a cake.

   分析：由于图片的截取，蛋糕被识别成了香蕉。

d. Model: a view of a building with a window.

   Human: a tall clock tower being reflected in a window.

   分析：正确识别出了窗户的镜像，但是无法识别出是一个钟塔，因为图片中的变形有些严重。

(a)

(b)

(c)

(d)

Figure 4: 失败案例

4.5.3 模型定性分析总结

在定性分析中，我们不难看出，对于特征明显、符合事物一般规律的图片，我们的模型确实能够得出令人满意的结果。但是对于一些异常的图片，例如非正常视角看到的事物、残缺的事物、镜像等等，我们的模型很容易判断失败。不过，如我们研究背景中提到的，在这个问题的研究过程中，我们能够发现神经网络捕捉到的图像特征、偏差所在。

5 结论

5.1 参数与结果分析

通过BLEU指标我们可以发现，show-and-tell模型相比于目标模型的差距是9.9；show-attend-and-tell模型相比于目标模型的差距是11.5。虽然存在着这个差距，但是从具体的文本性能上来看，结果与目标模型的结果相近。

产生这些指标上的差距的原因比较复杂。

目标模型是基于TensorFlow进行实现的，我们将模型用PyTorch进行重构改写，在这个过程中存在着一些不足或是本身运行环境上的差异，因此难免产生这些误差。

另外一个重要原因是训练时间，由于原来的选题不够恰当，我们的工作是临时换题的，这带给了我们莫大的精神压力。在训练模型的过程中，目标模型需要约20 epochs训练稳定的模型，我们在Tensorboard显示的loss几乎不下降的阶段便停止训练，平均约为3∼4 epochs。由于无法预测接下来的拟合情况，我们只能在时间和指标优秀率之间做一定的取舍。

5.2 贡献

本次作业中，我们尝试了作业中没有实现过的autoencoder架构，实现了CNN实现的encoder和基于LSTM的decoder。在实现过程中，我们掌握了更多的工具的使用方法，如tqdm等，使得模型运行界面清晰明朗。

在本次大作业中，我们使用了CoTK库提供的dataloader以及metric，大量的代码阅读与实践增强了我们实现人工神经网络框架的能力。我们实现的模型能够为CoTK库提供一个example，为这一语言处理工具包的推广提供了样品。在CoTK工具下，数据的加载使用了CoTK给出的统一规范，能够兼容更多类型的数据，这个模型验证了CoTK的简洁性与强扩展性。

5.3 总结

我们的工作在训练约束条件接受范围内能够复现给定的两个模型的结果。其中，show-attend-and-tell模型通过增添注意力机制对模型进行了优化，从BLEU的各项指标上来看，能够看出Attention的优化可见，但是没有那么明显。

5.4 展望

Show-and-tell模型启发于机器翻译的autoencoder机制，由于机器翻译是seq2seq模型，因此它的encoder与decoder均为RNN。Show-and-tell提供了一种通过修改encoder的机制来实现图像-序列间的变换。另外一方面，如果将decoder换成CNN，将encoder保持为RNN模型，或许可以实现文字-图像的转换。事实上已经有相关的工作与研究（wordseye.com），但是seq2image还可以通过对seq使用文法解析的方式进行encode，这一个方向的研究或许有不同于神经网络方法的手法。

5.5 分工

陈博涵:

模型框架设计

代码阅读(CoTK，NLTK)

代码编写与调试

对CoTK库进行拓展

实验数据运行

报告撰写

冯卓尔:

文献阅读

代码阅读(CoTK，Texar)

代码编写

实验数据运行

报告撰写

参考文献

[1] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

[2] Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., ... & Berg, T. L. (2013). Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12), 2891-2903.

[3] Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010, September). Every picture tells a story: Generating sentences from images. In *European conference on computer vision* (pp. 15-29). Springer, Berlin, Heidelberg.

[4] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the *IEEE conference on computer vision and pattern recognition* (pp. 3156-3164).

[5] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning* (pp. 2048-2057).