清华大学电子工程系

Show and Tell和Show, Attend and Tell两种基于递归神经网络的图像标注模型

蔡立崴 邮箱：cai_lw@126.com 学号：2014011255

2017 年 6 月 11 日

1 简介

图像标注是指自动为图像生成一段简短的文字描述，是一个需要结合计算机视觉和自然语言处理这两大人工智能领域的技术才能解决的问题。一个成功的图像标注算法不仅需要准确地识别图中的物体，还需要以简练的、符合语法的自然语言进行表达。

随着深度神经网络在人工智能各个领域的广泛使用，近年来成功的图像标注算法采用的方法大体上均为用卷积神经网络(CNN)提取图像特征，再将图像特征输入到循环神网络(RNN)中生成句子。本次任务中，我们实现了Show and Tell[1]和Show, Attend and Tell[2]两种经典的基于上述CNN-RNN架构的图像标注模型，并在一个中文图像标注数据集上比较其性能。

2 相关工作

Show and Tell的原始版本[5]是较早引入CNN-RNN架构的深度神经网络图像标注模型，也是同类模型中最简单的之一，可以看作是现今大多数深度神经网络图像标注模型的始祖。虽然结构简单，但性能并不逊色于后来提出的众多模型。本次任务实现的[1]便是该模型参加MSCOCO图像标注比赛过程中作出小幅改进之后的版本。

在Show and Tell模型中，图像的所有信息全部被压缩在一个向量中，可能会丢失较多对提高生成句子质量有帮助的次要信息，且难以得知向量中到底包含了哪些信息。针对这一情况，提出了许多从图像中获取特定语义的信息的模型。[7]是这方面较早的尝试，用RCNN直接对图像进行物体识别，将识别出的物体转化为关键词，然后由语言模型将关键词组合成语法正确的句子。

注意力机制(attention mechanism)最早由[8]提出，用于在机器翻译中生成目标句子时选择性地“关注”原句子的重点部分。类似的思想自然地被从文本到文本引申到从图像到文本，产生了许多相关研究。本次任务实现的[2]是其中较早的一个工作，注意力机制作用于简单均匀分区的图像的特征。[6]采用了更复杂的图片分区方法，先将图片用计算机视觉方法作场景分割，对各部分场景提取特征，再将注意力机制作用于其上。

3 方法介绍

3.1 问题的形式化

假设给出图像I时生成的句子为一符号序列$Y = \{y_1, y_2, \ldots, y_n\}$（“符号”可以是字符或单词），则模型的目标是最大化生成的句子的后验概率$p(Y|I)$。根据条件概率公式，$p(Y|I)$可以分解为按顺序生成每一个符号的条件概率的积的形式：

$$p(Y|I) = \prod_{t=1}^{n} p(y_t|y_1, y_2, \ldots, y_{t-1}, I) \tag{1}$$

两个模型均使用RNN直接预测$p(Y|I)$。RNN在每一步都会生成一个隐状态h_t，可以认为隐状态中编码了关于至今为止遇到的所有输入$y_1, y_2, \ldots, y_t, I$的信息。因此，用RNN预测$p(Y|I)$的数学表达式为：

$$p(Y|I) = \prod_{t=1}^{n} p(y_t|h_{t-1}), h_0 = f_0(I), h_t = f(h_{t-1}, y_t, I) \tag{2}$$

其中f_0和f是取决于网络具体形式的函数。

3.2 LSTM

近年来RNN网络的基本单元一般都采用长短期记忆单元(LSTM)，它可以有效地避免在使用反向传播算法计算梯度时梯度发生指数衰减或指数爆炸。LSTM的数学表达式如下：

$$\begin{aligned}
f_t &= \sigma(W_f y_t + U_f h_{t-1} + b_f) \\
i_t &= \sigma(W_i y_t + U_i h_{t-1} + b_i) \\
o_t &= \sigma(W_o y_t + U_o h_{t-1} + b_o) \\
c_t &= f_t \odot c_{t-1} + i_t \circ \sigma_c(W_c y_t + U_c h_{t-1} + b_c) \\
h_t &= o_t \odot \sigma_h(c_t)
\end{aligned}$$

其中$\odot$为逐元素相乘运算，y_t为输入，c_t和h_t为隐状态，h_t同时也作为输出，其他变量均为待优化的模型参数。

3.3 Show and Tell

在Show and Tell模型中，关于图像I的信息只在初始化隐状态时输入到模型中。将I输入一个已经训练好的CNN，从其全连接层中提取以向量表示的图像特征x_0，将其通过一层全连接网络用于初始化h_0。该模型的数学表达式如下：

$$h_0 = \sigma(W_0 x_0 + b_0), c_0 = 0 \tag{3}$$

$$h_t = LSTM(h_{t-1}, W_e y_t) \tag{4}$$

$$\log p(y_{t+1}|h_t) = softmax(W_r h_t + b_r) \tag{5}$$

其中W_e为可训练的词嵌入矩阵，将每个词转化为一个固定维数的向量表示。

在训练阶段，模型直接以$p(Y|I)$为优化目标。但在预测阶段，直接寻找$Y_{best} = \arg\max_Y p(Y|I)$需要指数级的时间复杂度，不可承受。这里采用一种称为限制宽度搜索(beam search)的策略来找Y_{best}的一个近似解：设定搜索宽度$k \in \mathbb{Z}^+$，每个时刻都只保留概率最大的前k个序列及其对应的陷状态，下一时刻只从这k个序列中扩展下一个符号；如果总共有n个可能的符号，则下一时刻考察所有nk个可能的新序列，从中只保留概率最大的前k个。依此不断重复，直至所有序列中都生成了句子结束标识符，此时取概率最大的序列为搜索结果。根据[1]中的研究，取$k = 3$效果最好，因此本任务中也取$k = 3$。

3.4 Show, Attend and Tell

在Show, Attend and Tell模型中，引入了在图像上的注意力机制，使得在模型生成句子时，不仅每一步都可以接收关于图像的信息，还可以根据当前的状态选择性地着重接收图像某些区域的信息。

为了保留空间信息，这里以CNN的最后一个卷积层的输出作为图像I的特征。设该卷积层有L个感受单元，每个单元产生一个特征向量，则用于表示图像的特征矩阵为$X = \{x_1, x_2, \ldots, x_L\}$。将$X$通过一层全连接网络，得到$A = \{a_1, a_2, \ldots, a_L\} = \sigma(W_X X + b_X)$作为输入RNN的图像信息。在每一步中，根据当前隐状态产生注意力向量$\alpha \in \mathbb{R}^L$表示对每个感受单元的“关注程度”，及关注度标量$\beta \in \mathbb{R}$表示对图像信息整体的“关注程度”，最终得到该时刻网络“需要”的图像信息z。上述过程即为注意力机制，具体实现方式有很多，这里采用的是与[3]类似的方法，但将“注意”的对象由“之前的输入”改为“图像信息”，其数学表达式为：

$$\alpha_t = softmax(w_\alpha^T \tanh(W_h(h_{t-1} \otimes e_L) + W_a A)) \tag{6}$$

$$\beta_t = \sigma(w_\beta^T h_{t-1} + b_\beta) \tag{7}$$

$$z_t = \beta_t(A\alpha_t) \tag{8}$$

其中e_L为L维全1向量，$\otimes$表示向量外积。

网络主体部分与Show and Tell相比，在输出的全连接层增加了“短路”(highway)连接，将当前时刻的图像信息和词向量和RNN的输出同时作用于输出层。其数学表达式为：

$$h_0 = \sigma(W_{h0}\bar{a} + b_{h0}), c_0 = \sigma(W_{c0}\bar{a} + b_{c0}) \tag{9}$$

$$h_t = LSTM(h_{t-1}, [W_e y_t, z_t]) \tag{10}$$

$$\log p(y_{t+1}|h_t) = softmax(W_r h_t + \tanh(W_{oy} W_e y_t + W_{oz} z_t + b_o) + b_r) \tag{11}$$

其中$\bar{a} = \frac{1}{L}\sum_{i=1}^{L} a_i$，$[,]$表示向量连接。

和Show and Tell一样，模型在训练阶段直接以以$p(Y|I)$为优化目标，在预测阶段用限制宽度搜索生成Y。

4 实验

4.1 数据

实验使用的数据集是来自MSCOCO的10,000张图片，其中8000张为训练集，1000张为验证集，1000张为测试集。100多名选修《模式识别》课程的学生利用课余时间参与了图像的中文标注，每张图片被标注了$1 \sim 5$句描述，平均每张图有大约4.3句。为避免样本分布不均，对于描述不足5句的图片，通过随机地重复已有的句子补全至5句。

对文本进行的预处理是：用python的jieba中文分词库进行分词，然后舍弃所有标点符号、数字、外语词汇。最终得到的语料库的词汇量约为10,000，考虑到词汇量并不太多，没有去除低频词。句子将以词序号序列的形式输入网络。

Show and Tell使用的图像特征是在其他图像分类任务上预训练好的VGG19网络的fc2全连接层的输出(4096维)，Show, Attend and Tell使用的图像特征是VGG19的block5卷积层的输出($7 \times 7 \times 512$维)。

4.2 评价方法

对验证集上的结果，在本地用BLEU-1、BLEU-2、BLEU-3、BLEU-4、ROUGE-L、CIDEr-D、METEOR共7种评分标准进行评分，以词为单位符号化(tokenize)。除METEOR直接调用[9]提供的程序外，其余评分标准均根据[4]中的公式实现。对测试集上的结果，由于测试集的标注不公开，故在评测服务器上BLEU-1、BLEU-2、BLEU-3、BLEU-4、ROUGE-L、CIDEr-D共6种评分标准进行评分，并与其他参与这项任务的学生进行排名竞争。为消除各方的分词方案不同带来的差异，统一以字为单位符号化。

4.3 实验设置

实验在作者本人的笔记本电脑上进行，操作系统为Windows 10，使用一块GTX 960m显卡加速神经网络计算，程序用tensorflow实现。

两个模型的词向量维数均为512，LSTM的隐状态维数也均为512。为抑制过拟合，在两个模型的词向量嵌入层之后和最后一个全连接层之前施加了Dropout。

每个模型在测试集上训练10个世代(epoch)，每个世代结束后在验证集上计算一次评分，当验证集上的CIDEr-D评分达到最大时停止训练，以此时的模型的各项评分为该模型的最佳得分。

4.4 性能评价

两种方法在验证集上和测试集上的各项评价指标如下表所示，其中ST和SAT分别表示Show and Tell和Show, Attend and Tell。需要注意，验证集和测试集使用的评测标准很不一样（见上节），直接比较验证集和测试集的分数是没有意义的。

方法	数据集	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr-D	METEOR
ST	验证集	0.593	0.411	0.288	0.201	0.457	0.646	0.319
SAT	验证集	0.599	0.427	0.299	0.207	0.462	0.647	0.322
ST	测试集	0.657	0.516	0.401	0.310	0.496	1.067	*
SAT	测试集	0.674	0.546	0.436	0.343	0.513	1.131	*

表1 测试结果

由上表可以看出，Show, Attend and Tell的各项指标均稳定地优于Show and Tell。根据[2]中汇报的结果，两者BLEU-4分数之差在$0 \sim 0.01$之间，本实验得出的两者性能差距明显比这要大。

4.5 案例分析

主观地观察两个模型在验证集上生成的句子发现，Show, Attend and Tell似乎更容易捕捉到图中的小物体，但生成的句子不通顺的情况有所增加。

图 1: ST:一个人在海边的沙滩上放风筝 SAT:山脚下的草地上有一个白色的游艇

图 2: ST:一个女人坐在长椅上看书 SAT:一个女人抱着玩具熊

图 3: ST:一只棕熊在水中爬行 SAT:两只熊在水中嬉戏

图1、2、3显示了Show, Attend and Tell的长处。这三张图片对于图像标注都是较困难的，两个模型都没有给出令人类满意的答案，但Show, Attend and Tell给出的答案至少包含了图中的一些重要物体，而Show and Tell的答案离题万里。这可能是由于Show, Attend and Tell能接触到的图像信息维数更多，因此内容更丰富。

图 4: ST:盘子里有很多甜甜圈 SAT:桌子上放着甜甜圈和一个甜甜圈

图 5: ST:树上结了很多串香蕉 SAT:一个绿色的植物插在绿色的香蕉上

图4、5、6显示了Show, Attend and Tell的短处。Show, Attend and Tell生成的句子时常有啰嗦重复的现象。这可能是由于Show, Attend and Tell在每一步都接触到同样的图像信息，和LSTM中“不能重复描述同一物体”的逻辑发生矛盾，两者的竞争中若前者占上风则会输出啰嗦重复的句子。

5 结论

本次任务实现了只有LSTM语言模型的Show and Tell和结合了合图像上的注意力机制的Show, Attend and Tell两种图像标注模型。在中文图像标注数据集上的实验结果表明，Show, Attend and Tell能够取得更好的结果。

本次任务的一大遗憾是，使用的数据集是专门为本次任务而准备的，此前没有人在同样的数据集上

图 6: ST:一个有马桶和洗手池的卫生间 SAT:卫生间里有镜子和镜子

做过测试，因此也无法将本次实验的结果与前人的工作比较。相比之下，英文的图像标注早已有了MSCOCO和Flickr等几乎成为业界标准的数据集，已有大量工作在这些数据集上完成。希望本次实验的组织方面能够将本次实验使用的数据集开源，为中文图像标注的研究工作设立一个标准数据集。

参考文献

[1] Vinyals, Oriol, et al. "Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge." IEEE transactions on pattern analysis and machine intelligence 39.4 (2017): 652-663.

[2] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In International Conference on Machine Learning (pp. 2048-2057).

[3] Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kočiský, T., & Blunsom, P. (2016). Reasoning about entailment with neural attention. In International Conference on Learning Representations.

[4] Chen, X., Fang, H., Lin, T. Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft COCO captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325.

[5] 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).

[6] Jin, J., Fu, K., Cui, R., Sha, F., & Zhang, C. (2015). Aligning where to see and what to tell: image caption with region-based attention and scene factorization. arXiv preprint arXiv:1506.06272.

[7] Karpathy, A., & Li, F.. (2015). Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3128-3137).

[8] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In International Conference on Learning Representations.

[9] Lavie, A., & Agarwal, A. (2007, June). METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In Proceedings of the Second Workshop on Statistical Machine Translation (pp. 228-231). Association for Computational Linguistics.