# 展示和说：一个自然语言生成模型[[1]](#footnote-1)

赵淼译

(北京交通大学电子与信息工程学院，电子科学与技术，15650761201)

## 1 摘要

描述研究的问题，本文的成果以及其先进性

## 2 介绍

这个任务的困难性，意义。描述不仅要捕捉图像中包含的对象，还必须表达这些对象如何相互关联，以及它们的属性和所涉及的活动。此外，上述语义知识必须用英语这样的自然语言表达，这意味着除了视觉理解之外还需要一种语言模型。

也就是从一个词典抽取单词描述图像。

这篇文章主要从机器翻译中获得灵感，编码RNN网络对源域句子进行编码，获得固定长度的序列，作为解码网络的初始状态，获得目标域的语句。作者用CNN代替RNN编码器获得图像编码，首先对CNN和RNN进行预训练，获得更多额外的数据训练。

作者这里采用了强大的子模块，更先进的cnn模块，rnn模块构成网络，实现了端到端的训练，并且可以使用更多的数据进行预训练，使网络生成的句子更具有灵活性，而不是单纯的模板匹配。

## 3 模型

这里通过最大化后验概率的方式给定图像，产生描述。

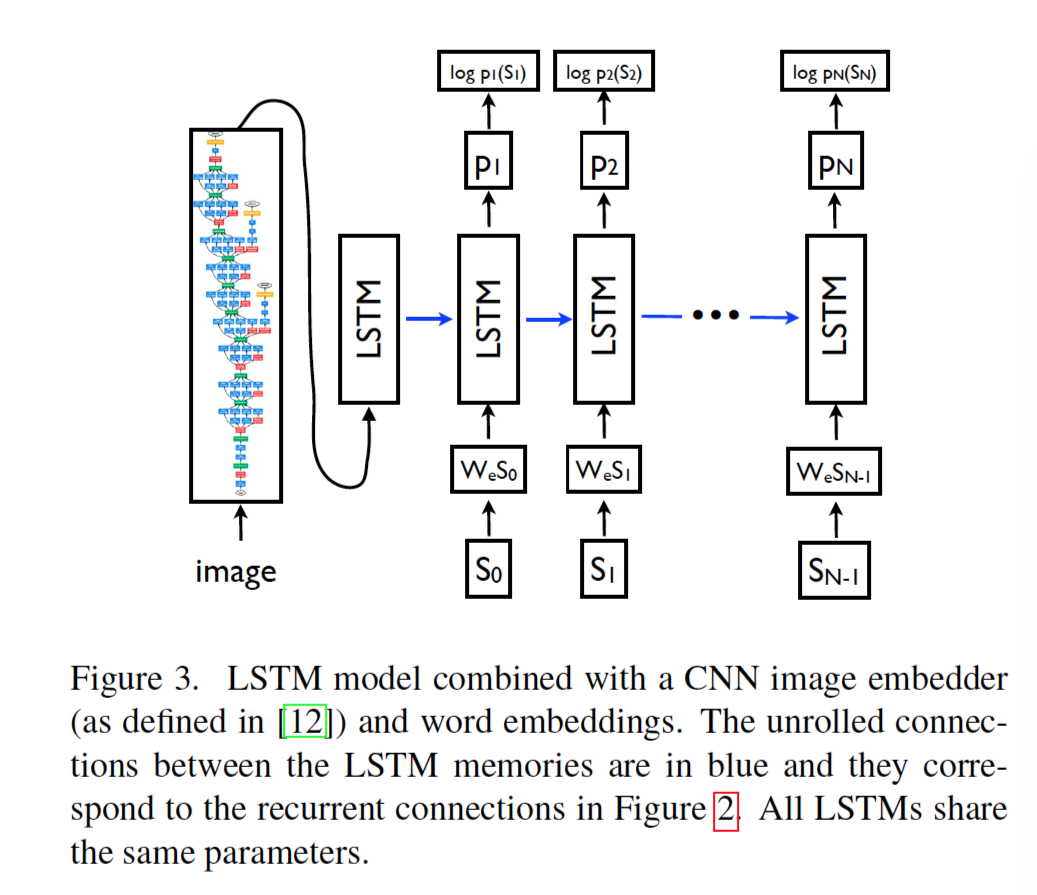
|  |  |
| --- | --- |
|  | 1 |

输入图像使正确翻译的概率最大。这个端到端的网络可以用来训练和推理，最后使用链式规则表示联合概率，这里借鉴了机器翻译的思想，将图像转化为固定大小的向量输入。

|  |  |
| --- | --- |
|  | 1 |

然后自然的用P(St|I,S0,S1,…,St-1)，采用RNN来建模，使用ht表示单词向量。为了使上面的RNN模型更具体，两个关键的问题需要确定：1）函数f的形式；2）图片和Words如何作为输入xt

本文中，f采用LSTM网络（已经在序列任务例如翻译中取得state-of-art效果）。对于图片的表示，我们采用CNN模型。我们使用的一个特别之处是采用了比较新颖的BacthNorm[12]。Words采用一个embedding model来表示。



图像的向量只在开始的时候输入LSTM作为输入，然后每个时刻LSTM的输入均是上一时刻的词向量，输出对于下一刻单词的预测，是标准的LSTM架构。

|  |  |
| --- | --- |
|  | 1 |
|  |  |
|  |  |

单词为词典中的one-hot向量，然后和图像一起被映射到相同的空间。S0与Sn是一个被特殊设计作为起始和结尾向量。

最后损失函数为-log似然。

|  |  |
| --- | --- |
|  | 1 |

这里有多种方法可以产生句子。

一种是抽样，考虑p1对第一单词获得，然后提供嵌入向量获得第二个样本，直到特殊的句末标记或是最大最大最大长度。第二种是考虑直到直到时间t的k个最佳句子的集合作为t+1时刻句子的候选者，保留最好的结果。接近是S= maxS0 p(S0jI)。

## 4 评估

人工评估

BLEU：这里作者认为虽有缺点，但仍具有参考价值。

困惑度：预测值逆概率的几何平均值，这里几乎不会报告，因为以BLEU为评估机制。

另外这里还可以用于对图像描述进行排序。另一方面，将图像描述生成任务转换为排序任务是不令人满意的，因为伴随着图像复杂度的提升，所需要的句子数量将指数级增长，这是不现实的。

在实验中，作者探索了数据集规模的影响，如何使用迁移学习，如何处理弱监督标签问题。

### 4.3.1 训练细节

作者主要探讨了处理过拟合的问题，cnn使用ImageNet上训练的权重，这个很有帮助，另一方面尝试了在大型新闻数据集上处理词嵌入模型，效果不显著，因此使用了随机初始化。作者也尝试了dropout和模型组合的方式提升性能。

使用了不带动量和学习率的随机梯度下降方法。LSTM存储量为512维。描述经过了基本分

词处理，保持所有词出现至少5次。

### 4.3.2 生成结果

作者认为评估BLEU4指标是有意义的，但与人类评估效果有较大的差距。

### 4.3.3 迁移学习，数据规模和标记质量

作者训练了大量模型，并探索对模型碱性迁移是否可行，领域之间的不匹配可以被更高质量的标记和数据集不长。

迁移学习和数据大小最明显的情况是在Flickr30k和Flickr8k之间，两个数据集的标签类似，都是由同一组创建的。事实上，当使用Flickr30k进行训练（大约有4倍以上的训练数据）时，得到的结果更好的。因为整个过程是数据驱动和过度拟合的倾向。 MSCOCO甚至更大（比Flickr30k多5倍的训练数据），但是由于收集过程是不同的，因此词汇可能会有更多的差异，并且会有更大的不匹配。事实上，所有的BLEU分数都降低了10分。尽管如此，这些描述仍然是合理的。

### 4.3.4 生成多样性

这里探究是否可以产生新颖的描述同时又有比较高的质量。，在样本中的句子是多样性的，描述了图像的不同方面。这里发现产生数据集较小时，最好的句子中百分之80是训练集中句子，如果是在评分最好的15个句子中挑选，有1半是新颖的句子

### 4.3.7 嵌入分析

嵌入分析一方面是端到端的神经网络，可以直接被训练，而且向量的规模可以与字典规模相独立，并且学会词语之间的语义关系，并且引导CNN学习一些关联的特征。

## 5 结论

进一步，作者对无监督学习如何促进图像域和文本域的学习很感兴趣。

1. 赵淼 电信研1706班 电子科学与技术 17120036 15650761201 [↑](#footnote-ref-1)