# 博客理解[[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)

<https://blog.csdn.net/amds123/article/details/65442102>

Image Caption综述

Image Caption:对计算机视觉、自然语言处理和机器学习于一体，将一段图片描述为一段文字。两方面：利用模型理解图片内容，使用自然语言去表达他们之间的关系。

主要方法：大多数方法基于encoder-decoder的模型，encoder一般为卷积神经网络，decoder一般为递归神经网络，主要用于图像描述的生成。由于普通RNN存在梯度下降的问题，RNN只能记忆之前有限的时间单元的内容，而LSTM是一种特殊的RNN架构，能够解决梯度消失等问题，并且其具有长期记忆，所以一般在decoder阶段采用LSTM.

问题描述：将问题定义为两个域的映射关系，形成二元组（I，S）的形式，原始域为图，目标S为单词序列，这里的描述是使最大似然p(S|I)取得最大值，也就是一种监督学习的形式，要有原始图像，要有标记好的目标语句，用最大似然的方法达到通过I预测S的目标，问题是自然语言很模糊，与场景、上下文都有关，怎么标记形成目标语句。

常用数据集：Flickr8k,Flick30k,MSCOCO

其中每张图像都至少有5张参考描述。为了使每张图像具有多种互相独立的描述，数据集使用了不同的语法去描述同一张图像。如示例图所示，相同图像的不同描述侧重场景的不同方面或者使用不同的语法构成。

这介绍了几篇论文，讨论的都是模型搭建上面的问题，整体结构都没有脱离CNN+LSTM的模型，但具体形式有所不同。

首先是CNN部分：1.采取不同的卷积神经网络作为提取图像特征的模块（VGG16,VGG19,GooleNet等）。2.提取图像中包含的单词，也就是说进行目标检测，但这里包含了既包含了一些实体，又包含了些概念（如拿，握）等（3）结合物体之间的空间关系，这里的想法可能是一较低的卷积层作为图像特征，同时保留空间关系，我的理解是识别处物体后要保留目标在图像中的位置（4）如何结合目标的空间信息，这里是注意力机制，这里我观察最近许多研究成果都提出了这一机制，根据我的理解实际上这里是要检测图像中的关键信息，毕竟图像中不是所有信息都有用，找到最重要的信息即可以降维，又便于后面文本生成模块的工作（5）这里还有研究图像属性对描述结果的影响，我感觉这里就是再说什么样的信息是重要的，与注意力机制有关

这部分我的感觉还是要首先提取特征，实现目标检测，因此从基础入手，提高目标分类效果🡪提高目标检测效果，然后考虑注意力机制，空间信息以及其他方面入手更适合一点。

然后是文本描述生成部分，模型上就是LSTM及其变种。（1）这里提到了LSTM梯度消失的问题（2）词嵌入

最后是两者的结合，综合考虑问题，我觉得这部分比较关键，对整个网络架构的搭建起到指导意义，（1）指导前面的CNN学习不同的特征，然后进行编码形成后面网络不同的输入，这里提到了不同的方法，如one-hot，max pooling（2）LSTM该使用什么信息，是图像的信息，还是其他，这里的理解是图像的信息实际上不能完全反应图像信息，这部分可能需要进行切换，同时在LSTM加上知识图谱（自我理解，最近听到的一个名词）指导神经网络产生一句完整的描述（3）还有针对LSTM梯度消失的问题，该如何解决，如果图像的信息没有作用了LSTM该怎么办。

这篇博客也提到这里的一切其他方面的前沿，如GAN，强化学习。

<https://zhuanlan.zhihu.com/p/25910749>

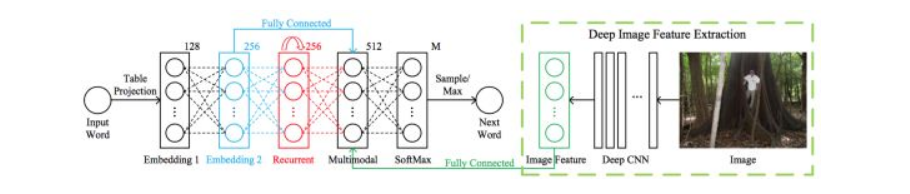
目前深度学习解决Image Caption的方法主要是Encoder-Decoder,另外attetion的机制会较明显的提高性能，进而Adaptive Attetion的机制也会较明显的提升性能。

1.encoder-decorder

这一部分的主要应用是将不等长的序列转化为等长的序列，这部分主要解决RNN对输入和输出要求等长的问题。

2.m-RNN（多模态RNN）

开创性地将深度学习用于解决Image Caption问题，是一个基准。



解释：左面是上一时刻输入的单词，经过两个词嵌入层，转化为稠密向量，RNN会对上个时刻单词向量在RNN中的激活值进行加权处理，公式如1.

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

右侧卷积网络产生特征向量I,I，RNN输出与上一个单词的词嵌入向量一起并入到multimodel中，再做加权处理，再激活后输出，形成概率矩阵输出当前时刻应该输出的单词

|  |  |
| --- | --- |
|  | 2 |

代价函数是似然函数，是当前图像向量下输出的各单词的概率与实际的该输出单词的交叉熵，这里每个时刻，图像特征都参与了计算，这里是一个问题。

NIC模型

相对于m-RNN

1这里使用了LSTM

2使用了效果更好的卷积网络

3CNN提取特征只输入一次。

这里还认为目前的评价指标难以客观地对模型的caption做评价，我们从实践经验上证实如果在每一个时间点都输入图像数据，将会导致较差的结果。网络可能会放大图像数据中的噪音，并且更容易过拟合这里来引出了每个时刻都应该关注图像的不同区域，即注意力模型。

Attention机制

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

这里介绍attention就是考虑不同时刻的情况下，神经网络应该关注的区域，encoder部分不变，decoder加入注意力机制，即对不同时刻，不同区域设置一个权值，然后学习这个权值，公式如下。

|  |  |
| --- | --- |
|  | 3 |
|  | 4 |
|  | 5 |

这里考虑每个通道的权重相同，即空间attetion，同时也有论文考虑这方面。

SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning

Adaptive Attention

这里考虑一些单词实际上与图片没什么关系，可以直接通过文本信息推出，这里就设置了一个过滤器来过滤该时刻是否要用到图片信息来预测产生的词语，具体架构如图所示。

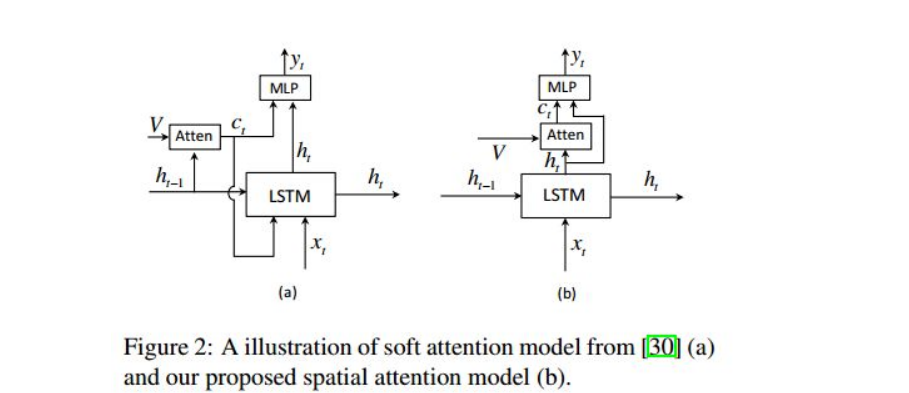
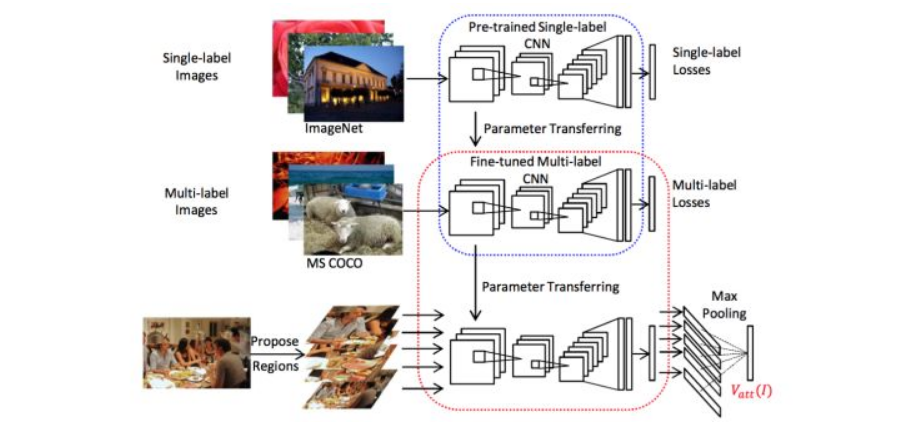


图 soft注意力模型的改进

理解：这里LSTM用来保存历史信息，x是这一时刻的输入，h（t-1）是上一时刻的历史信息，原本的注意力机制，是使用h（t-1）也就是上一时刻的隐藏层状态和当前的视觉信息输入进行注意力判断，现在以当前隐藏层状态和视觉信息，进行注意力判断，直接判断这两个信息的重要性，分配哪一个信息更重要，注意力信息不再影响LSTM隐藏层状态。

What Value Do Explicit High Level Concepts Have in Vision to Language Problems?

这篇文章以在图像的语义信息表达部分作为研究点，以ImageNet的但标签模型作为与训练模型，到MS COCO上进行多标签任务的fine\_tune,然后进行迁移学习提取图像的语义特征，这里再次强调了评价指标的不足。



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