# 看，注意，说：带有注意力机制的神经图像标注[[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)

## 摘要

直接提出这里借鉴了机器翻译和目标检测的机制，引入注意力模型学习描述图像内容。描述了如何训练网络，展示了模型如何自动的集中与产生对应输出序列的突出目标

## 1 介绍

一方面要决定图片中哪些目标在图像中，另一方面要捕捉和解释自然语言中的关系。研究的启发，使用注意力让动态的集中与最需要关注的特征。使用前面获得的卷积网络获得的特征然后进行知识蒸馏时获得突出目标的最有效的方案，这在过去的工作中广泛的使用。这么做的缺点时损失了大量的信息（可能对于产生描述性的，丰富的句子）。

这里受人类视觉系统的影响，提出了两种用于图像描述产生的注意力机制：软注意力和硬注意力。

软注意力机制：通过标准的反向传播算法

硬注意力机制：通过最大化一个近似的变分下限或相等的强化（Williams，1992）来训练

这里作者的方法不包含显式的目标检测，允许学习一些更抽象的概念。

在前面提到三种方案中的前两种显式的包含移除无关信息的模块，但不适合主流的神经网络的方法。

## 2 带有注意力机制的图像描述产生器

### 2.1 网络细节

#### 2.1.1 编码器：卷积网络

这里卷积网络的目标是获得一系列特征向量参考标注向量。这里提取了L个D维的特征向量对应图像的一部分。这里使用低层次的卷积层特征代替全连接层特征。允许解码器集中于图像特征向量的子集来选择图像的确切部分。

#### 2.1.2 解码器：LSTM

这里作者使用了一种仿射变换表示学习到的参数。

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

这里z表示上下文信息，捕捉的是与特定输入位置有关的视觉信息。

在简化版本中，上下文信息是一个动态的表示，相对与图像输入的具体部分在时间t。这里作者用φ来表示这种机制，这种机制会计算z通过标注向量ai，i=1，…,L对应图像不同位置提取的信息。对于每个位置i，机制会产生一个正的权重阿尔法i，可以解释维位置i是正确的对应下一个单词的概率（硬机制），或是给定位置i，i在将i混合在一起中所占有的相对重要性。权重阿尔法i计算是由fatt获得，这是一个多层感知机，输入是隐藏层状态ht-1.软注意力模型机制在2014年引入。为了强调重点，我们注意到隐藏状态随着输出序列的输出顺序的变化而变化：网络所关注的下一个位置取决于已经生成的单词的顺序。

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

获得权重向量后，上下文向量z会计算通过。

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

在这里φ是给定一组标注向量，返回对应权重的机制的一个函数。

LSTM的初始记忆状态和隐藏状态是由平均的注释向量来预测的通过两层独立的MLP。

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

在这个工作中，我们使用一个深度输出层（Pascanu等人，2014）来计算输出词的概率，给定LSTM状态，上下文矢量和前一个词。

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

是可学习参数，随机初始化。

## 4 随机的硬机制和确定性的软机制

### 4.1 硬机制

用st代表位置变量，决定产生第t个词是集中的注意力。St，i是一个one-hot指示器，如果第i个位置被用来提取视觉特征则被设为1.将注意力位置作为中间的潜在变量，我们为阿尔法i假设一个多重伯努利分布，将zt视作一个随机变量。

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

我们定义了一个新的目标函数Ls，它是在边际对数-似然对数p（y j a）上的一个变分的下限，观察给定图像特征a对应的序列。模型的参数的学习算法可以通过直接优化Ls来获得。

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

方程11提出了一个基于蒙特卡罗的模型参数的梯度近似。这可以通过从方程8中定义的多ouilli分布中采样位置来完成。



采用一个移动平均基线来减少梯度的蒙特卡罗估计量的方差，遵循Weaver&Tao（2001）。类似的，但更复杂的方差减少技术以前被Mnih等人（2014）和Ba等人（2014）使用过。在看到kth的小批量时，移动平均基线被估计为前一个日志的累积和指数衰减的总和。



在每一点做出硬选择的时候，机制是一个在每一点对基于多项分布的参数阿尔法抽样的过程。

在作硬决策的时候，在每个时间点根据多元分布返回一个采样的a向量。

博客：<https://www.cnblogs.com/Determined22/p/6914926.html>

### 4.2 决策软注意力

相对于前面的硬采样，这里使用上下文向量期望值的方法返回一个上下文向量。

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

这里近似的获得了注意位置的边界似然。

#### 4.2.1 双随机注意力

由softmax的输出构建了在某一时刻个位置的权重和为1。在训练决策形式模型的时候，我们引入了一种双随机注意力机制孤立各个时刻同一位置的权重和近似为1.

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

这可以被解释为鼓励模型在生成过程中对图像的每个部分给予同等的关注。在我们的实验中，我们观察到，这个惩罚对于提高整体的BLEU评分是很重要的，从定性上讲，这将导致更丰富和描述性的标题。

另外这里软注意力模型预测了一个门值贝塔上一时刻隐藏层状态ht-1中在每个时间不下。这里通过缩放因子更强调图像中的物体。

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

具体的，模型通过最小化正则化负对数似然进行端到端训练。

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

### 4.3 训练流程

我们的注意力模型的两种变体都是使用自适应学习速率算法进行随机梯度下降的训练。对于Flickr8k数据集，我们发现RMSProp（Tieleman&Hinton，2012）效果最好，而对于flickr30k/ms COCO数据集，我们使用了最近提出的Adam算法。

为了创建我们的解码器所使用的注向量释ai，我们使用了牛津VGGnet（Simonyan&Zisserman，2014）在没有finetuning的情况下对ImageNet进行了预训练。然而，原则上，任何编码函数都可以使用。此外，有了足够的数据，我们还可以用模型的其余部分从头开始（或微调）来训练编码器。在实验中我们使用了14\*14\*512的特征图（第四层卷积层未经过池化）。解码器对其进行了伸张成196\*512维的向量。由于我们的实现需要时间与每次更新最长的句子的长度成比例，我们发现在一个随机的标题组上的训练是浪费的。为了减轻这个问题，在预处理过程中，我们构建了一个字典，将一个句子的长度映射到相应的标题。然后，在训练过程中，我们随机抽取一个长度，然后检索一个长度为64的小批量。我们发现，这大大提高了收敛速度，在性能上没有明显的下降。在我们最大的数据集（MS COCO）上，我们的软注意力模型花了不到3天的时间来训练。

除了dropout（斯利瓦斯塔瓦等人，2014），我们使用的唯一一种0正规化策略是在BLEU评分中尽早停止。我们在实验的后期，观察到验证集的对数可能性和BLEU之间的相关性。由于BLEU是最常被报道的指标，所以我们在我们的验证集上使用BLEU来进行模型选择。

在我们的软注意力实验中，我们也使用了Whetlab1（Snoek等人，2012;2014年）在我们的Flickr8k实验中。我们从它所探索的超参数区域获得的一些直觉在我们的Flickr30k和COCO实验中尤其重要。

## 5 实验

我们描述了我们的实验方法和定量结果，验证了我们的模型在标题生成方面的有效性。

在所有的实验中，进行了基本的分词处理，然后我们固定单词量是10000。报告了BLEU和meteor.

### 5.2 探索机制

为了进行比较，这里采用了相同的卷积网络，使用了单一模型和相同的数据集切分。

定性的分析：通过可视化模型学到的注意力组件，我们能够为模型的输出增加一个额外的解释能力（见图1）。其他的系统依赖于对象检测系统来产生候选对齐目标（kar病变&Li，2014）。我们的方法更加灵活，因为模型可以关注非对象的突出区域。

这款19层的VGG网使用了3x3滤镜，这意味着该特征图的尺寸缩小的唯一时间是由于最大池化层。输入图像进行调整，使最短的边为256维，保持了宽宽比。卷积神经网络的输入是中心裁剪的224x224图像。因此，在第四个最大池化层中，我们得到了14x14的顶部卷积层的输出维度。因此，为了使软模型的注意力权重形象化，我们只需将权重的样本上采样到16，并应用一个高斯滤波器。我们注意到，14x14单元的接受域是高度重叠的。

正如我们在图2和图3中所看到的，模型学习了与人类直觉非常一致的排列。特别是在错误的例子中，我们看到有可能利用这种可视化来直观地理解为什么会犯这些错误。

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