# 为了产生图像描述的深度视觉语义联合[[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

作者提出的是一个生成式模型，用于生成图像和区域的描述。通过CNN获取图像的特征，双向RNN获得句子特征，然后联合通过多模态嵌入的方式联合两种模态形成结构化目标。然后使用多模态RNN价格普使用推断比对的方式学会产生新颖的图像区域描述。作者的方法在一个新的图像和区域级标注数据集中产生超出检索基准的效果。

## 2 介绍

作者首先承认了其他视觉任务的成就，然后提出进行图像的描述更有挑战性，更有意义，也更困难。

这里说了一些其他人的工作，又指出他们方法的不足，这些方法多是套用模板，而且局限于一个句子来描述图像，这里作者认为这是不适宜的。

这里作者的任务是产生对图像密集的描述。作者认为

此次实验所使用的数据集实现这一目标的主要挑战是设计一个足够丰富的模型，以同时推理图像的内容及其在自然语言领域的表现。此外，该模型应该没有关于具体的硬编码模板，规则或类别的假设，而是依赖于从训练数据中学习。还有现在的数据集只有生成的句子而没有将句子中的实体对应出来。

作者的核心观点是，我们可以通过将句子视为弱标签来利用这些大型的图-句子数据集，句子的每个连续片段对应图像中的位置，作者的方法是推断他们的对齐，然后学会描述产生模型。

作者使用深度学习开发了一个深层神经网络学习图像区域与句子之间潜在的对齐关系，通过多模态嵌入空间和结构化目标将两种模态联合起来。

## 3 相关工作

图像标注：相对于简单的目标、场景、区域，这里集中于更丰富和多层次的区域描述。

生成描述：作者使用更简单的RNN，同时模型表现上也会受到影响。作者通过实验对比了这种影响。

将自然语言放到图像中：作者通过延续了语义的方法和通过片段进行内部对齐。而我们的模型根据了依赖关系树，使用了了连续片段，更加有意义，更具有可解释性，在长度上也不固定。

视觉和语言领域的神经网络

## 4 我们的模型

和上一篇文章相比，这里首先是做片段对齐，然后在片段对齐的基础上进行获得一个多模态的循环神经网络。整体思路是从图像描述的检索到生成。

### 4.1 学会视觉和语言数据的对齐

这里建立在前人的基础上，使用双向循环神经网络计算句子的单词表示，并获得单词和他们在上下文中无限的关系。这里对嵌入表示进行了描述，实际上就是找到一个空间使语义相似的图像与句子的距离更短一些，理解为有监督的聚类（聚类与监督分类的效果到底是一致h，可以无监督是监督的一种扩展）。

#### 4.1.1 图像表示

这里卷积网路与上篇论文介绍的类似（RCNN）,公式如下：

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

这里Wm是4096乘以h的矩阵，将卷积网络获得较高内聚的特征表示经过转换映射到h维的共同的多模态空间中。

#### 4.1.2 句子表示

这里是对依赖树关系的扩展，为了实现上下文捕捉窗口的任意大小，使用双向循环网络结构获得词语的表示。BRNN采用单词在一个词典中的编码和将这种编码形式转换到h维向量。每个单词都是根据在其周围可变大小的上下文内容来丰富。用索引1…N表示句子中的位置。

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

刚开始输入是一个表示单词在词典中位置的向量，通过公式1转换为300维的word2vec（考虑到过拟合的影响）。这里作者发现最后的效果与不会发生改变，即使是随机初始化的情况下。公式3和公式4表示两个独立的处理流程。一个从左到右，一个从右向左。最后获得第t个单词的最终h维表示st是句子中该单词和其周围上下文的函数。W和b是学习的参数，ReLU做激活。

#### 4.1.3 对齐目标

首先对齐目标函数来自于上一个论文中提到的对齐函数，然后加上多实例目标，让点积为正的即可获得正例，然后进行了简化。

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

鼓励单词与一个最佳的图像区域对齐。

#### 4.1.4 将文本片段对齐解码为图像

在训练时，将Skl解释维描述任意边界框的第t个单词的非归一化对数概率。这里要生成的是文本片段，因此需要将单词序列与对应的边界框对齐。作者使用了马尔科夫随机场将对齐关系作为了潜在的变量。相邻单词的二元作用推动了对同一个区域的对齐。

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

β控制对齐的长度，β为1对一个单词进行对齐，当β很大时，整个句子被对齐到一个单一得分最大的区域，输入获得的是一段使用文本注释的图像。公式2考虑的是每个单词与各个区域之间的对应关系，式3考虑相邻单词与图像之间的对应关系是否相同，当贝塔很大是意味着所相邻单词被映射到同一个区域，如果贝塔减小，这种限制减弱。这里一个二元组的形式将整个句子内单词的对应关系，避免设置超参数对齐长度，改变公式形式。

### 4.2 用于产生描述的多模态循环神经网络

这部分，假设一组输入图像及其文本描述。这些可以是完整的图像和他们的句子描述，或区域和文本片段，如前一节所述。这里的目标是设计一个给定图像的可变尺寸输出序列。这里对RNN进行简单的扩展。输入是图像和一系列的输入向量x。计算获得一系列隐藏层向量和输出通过迭代。

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

式1输入是卷积层的最后一层，整体输出字典中每个单词的概率和附加停止表示符。图像特征只最开始时输入网络。隐藏层是512个神经元。前面获得对应关系可以作为这部分的训练数据，也就是学习生成的句子片段与图像之间是否对齐。

RNN训练：RNN训练是在结合一个词和h（t-1）的基础上，预测下一个词的概率。刚开始设h0为0向量，x1为一个特殊的Start向量，y1为一个初始生成单词的标签。最后生成最后一个词目标是产生结束标志向量。

测试时计算每一个通过网络获得一个词，然后嵌入生成下一个输入。直到最后结束标志产生。

## 4 实验

对数据集预处理：小写化，去除非字母数字。

### 4.1 图像，句子对齐实验

作者的BRNN模型表现突出。

简单的损失函数提升了性能。

相对于以前的依赖树关系有较大的提升，BRNN学习到了超过两个词的上下文关系。

这一模型发现了可解释的视觉 - 语义对应，即使对于小型或相对罕见的物体，如“手风琴”，也是如此。这些可能会被那些只有完整图像的原因的模型所遗漏。

我们模型的一个重要特征是它学习调整区域的大小和字嵌入。由于内在的相互作用，我们观察到，像“皮划艇，南瓜”这样具有视觉辨别力的单词的表示具有较的嵌入向量，这反过来又转化为对图像句子分数的较高影响。相反，停止诸如“现在，简单地，实际上，但是”这样的词被映射到原点附近，这减少了它们的影响。

### 4.2 产生叙述：全帧评估

利用完整的图像进行句子和图像间的映射。这部分使用了VGGNet图像特征。

定性描述：一般而言可以在训练集中找到生成句子中的较大的一部分，当这一比例变小时错误率上升。

与检索基准相比。我们的方法表现更好。

与相似工作相比，他们使用了复杂的LSTM和AlexNet，可能是由于AlexNet更简单的影响，他们的表现较差一些。

产生描述：区域性探索

在这种方法中，与全图像标题的统计信息不同，我们的区域性表述更全面，包含一些在全局标注中不被突出的部分。

定性的：模型必须学会以字符串为基础，学会生成它。

区域模型比全帧模型和排序基准表现更好。与全帧模型相似，这里将预测任务定义为二维的像素矩阵到一个字符序列，用BLEU分数评价。这里检索的基准时每个图像区域与BRNN判断匹配的字串相匹配。这里相对于全帧模型，使用了更小的图像区域作为决定其表现。

### 4.4 限制

只能产生一个固定分辨率大小的图像的描述。更好的方法也许是对图像多次扫描获得所有实体，相互作用以及上下文关系。RNN只是使用一个偏置接受图像信息比复杂的多重相互作用更差。最后，这里是两个独立的模型，不能从图像句子数据集到区域级注释。

## 5 总结

这篇文章实际包含两部分内容：一部分多模态的视觉和语言模态对齐的嵌入进行检索。另一部分是分别基于全局图像和局部图像的多模态的RNN结构

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