生成图像描述

Paper: Deep Visual-Semantic Alignments for Generating Image Description(2015 CVPR)

任务：为图像和区域生成文字描述。

方法：

- 利用图像和文字描述信息，建立视觉和语言之间的多模态联系，也叫做“Alignment”。图像经过CNN得到视觉特征的Embedding，文本描述通过BRNN得到语言特征的Embedding。然后通过一个“结构化的损失函数”来建立两个模态特征之间的“Alignment”。
-

相关工作：

“Dense Image Annotations”

- 建立多模态联系[2],[48];
- 整体场景理解；

相关工作的缺陷：只在有限的集合里面标注场景、区域、物体。本文工作则关注更丰富的、更高语义层(更复杂)的区域描述。

“Generating Description”

- 基于检索任务的
- 基于生成语法或者图像内容进行固定模板填充的。

现有工作的缺陷：限制了生成描述的多样性；他们使用固定窗口大小的上下文信息，而本文使用RNN，使用所有前面单词的上下文信息。也有其他文章使用RNN的，但是本文使用的RNN比大部分都简单，但同时也存在效果的瓶颈。

“Grounding natural language in images”

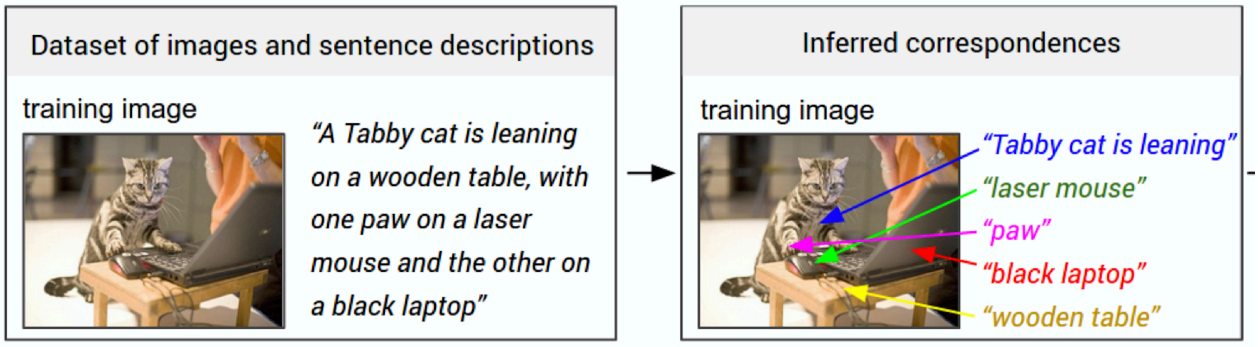
这些文字和本文十分相关，但是本文做的是为图像区域关联连续的句子片段，这样的结果更有意义、更有可解释性、而且长度也不固定。

模型

本文的最终目的是为图像生成描述信息，但是模型的训练分成两个阶段：先训练visual embedding和sentence embedding之间的“Alignment”,然后将训练好的这种"Alignment"作为训练数据输入到为图像生成描述信息的模型中。

Align Visual and Language Embedding

描述图像的句子，往往一些连续的单词都在描述图像中某个区域，这是一种潜在的联系，如下：



步骤：

1. 物体检测（RCNN），在ImageNet上预训练过,又在ImageNet检测数据集上的200个类进行了微调；每张图像保留20个候选区域（包括整张图像在内，为了生成整张图像的描述？）
2. 然后使用一个全连接层，将每个候选区域的embedding转换到1000~1600维之间：

   $$v = W_m(CNN_\theta(I_b)) + b_m$$

   其中I_b是RCNN中，分类器之前的全连接层输出的4096维向量。RCNN接近有6亿参数。W_m是$(h, 4096)$维；
3. 用BRNN得到句子的embedding，也是h维的，BRNN的每个隐藏状态对应于句子中一个单词的Embedding，BRNN每个timestep的输入是每个单词经过word2vec转化的embedding；还可以有一些其他的方法可以用来获取这个embedding，比如利用BOW、word bigrams、dependency tree relation[24].（注：这个地位用word2vec先得到了单词的表示，再输入RNN中，得到单词的新的表示。我觉得这里主要是为了编码句子的全局信息，虽然word2vec也有上下文信息，但是针对这种上下文信息可能是更通用的上下文信息，而不是针对某些领域的上下文信息）
4. 计算整张图像k和整个句子l的相似性，图像子区域i和句子单词t之间的相似性为两者Embedding的内积:

   $$similarity_{it} = v_i^T s_t$$

   那么整个图像和句子的相似性分数就为：

   $$S_{kl} = \sum_{t=1}^{N}\sum_{i=1}^{K} max(1, v_i^T s_t)$$

   N个单词，K个图像子区域，本文将此相似度简化为：

   $$S_{kl} = \sum_{t=1}^{N} max_{i\in[1,K]}(v_i^T s_t)$$

   它的含义就是说对每个单词t,$t \in [1, N]$)，在所有图像区域中找一个最相关的区域i。
5. 损失函数

   $$C(\theta) = \sum_k [\sum_l max(0, S_{kl} - S_{kk} + 1) + \sum_l max(0, S_{lk} - S_{kk} + 1)$$

6. 前面都是得到的单词和图像region之间的“Alignment”。为了得到句子片段和区域的对齐，本文使用了MRF，此处不做详述。

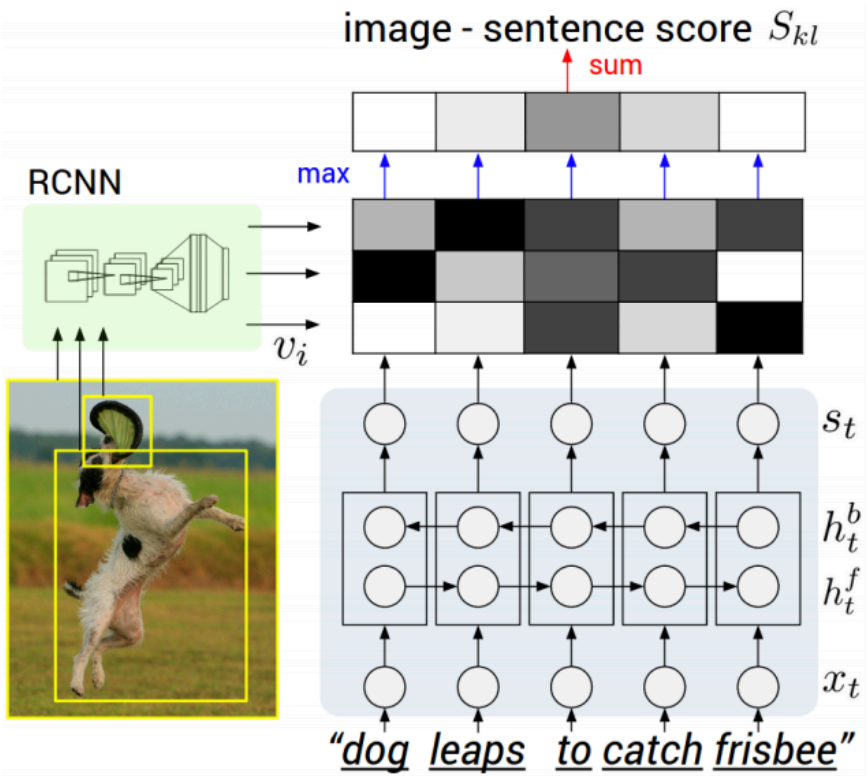


Figure 3. Diagram for evaluating the image-sentence score S_{kl}. Object regions are embedded with a CNN (left). Words (enriched by their context) are embedded in the same multimodal space with a BRNN (right). Pairwise similarities are computed with inner products (magnitudes shown in grayscale) and finally reduced to image-sentence score with Equation 8.

Generating Descriptions

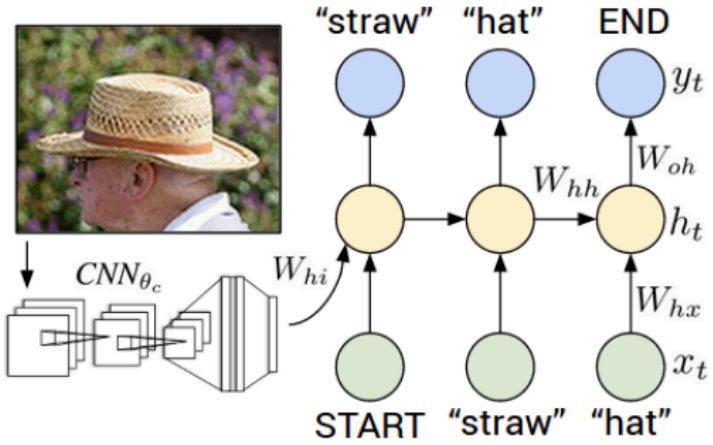


Figure 4. Diagram of our multimodal Recurrent Neural Network generative model. The RNN takes a word, the context from previous time steps and defines a distribution over the next word in the sentence. The RNN is conditioned on the image information at the first time step. START and END are special tokens.

这一部分怎么利用前面的模型呢？理解了其实很简单，前面的模型其实为此阶段的描述生成做了这样的事情：获得region(物体检测得到)和sentence 片段(先对齐单词，再通过MRF得到句子片段)的训练集，作为本阶段的训练数据。

那么次阶段是怎么做的呢?

1. 训练时，输入的是区域和句子片段，图像信息只在第一个timestep作为输入，而句子片段的每个单词作为每个timestep的输入。
2. 测试时，只将图像作为输入，第一个timestep，h0=0，输入单词为“START”，然后之后的timestep的输入都是h_{t-1}和上一个timestep预测出来的单词的embedding。

注意这个模型的embedding都是word2vec.