# 对于视觉转语言学问题明确的高层次概念意味着什么[[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

在主流的CNN-RNN架构中没有明确的表示高层次的语义概念年，而是间接的进行图像特征到文本特征的转换，这里探索高层次语义信息的表示的重要性，并且这种表示可以对图像理解给出了提高。这里展示了可以引入外来的语义信息。

## 1 介绍

尽管一幅图像可能等于千言万语，但是在每种数据形式下信息的表示形式几乎没什么不同。语言是人类之间交流的过程，但一幅图像中包含许多我们无法控制的物理过程信息的集合。

这篇文章的重点探究的问题就是标题给出的问题。我们调查了加入高层次语义信息对V2L的影响。这里插入了明显的场景属性进入模型中。

这里的关注点是高层语义注意力，如在LSTM第一个时间步加入语义信息或是以软注意力机制加入语义信息，这种方式是受限制的。

这篇文章提出通过对一些单个的语言标记进行组合以获得对图像整体含义的描述。这里使用CNN获得图像特征，然后对LSTM进行了改造，SCN对对传统LSTM进行扩展。

这里探索了这种模型，并证明了其有效性，同时又进行了细节的分析，并应用到了视频理解中。

## 2 相关工作

LSTM在语义表示上的有效性。LSTM和CNN是先进的方法，但是不清楚绕过高层次语义性信息的影响，LSTM和CNN没有直接对于单词的映射。

## 3 一个基于属性的V2L模型

整个模型包括两个部分：图像分析部分和一个语言产生部分。图像分析部分，基于监督学习训练了一些属性探测器。（一个多标签分类器）获得一个特定属性的概率向量。语言产生部分将产生的语义向量输入到LSTM中。



这里将最近基于深度学习的图像理解分为两类，一类是编解码的结构，另外一类是高层语义分析。

这里重点是使用了语义概念向量，使用LSTM对所有标签得到一个概率值，同时语义的概念使模型的可解释性更好。

### 属性探测器

这里获得了最常用的单词组成的单词集，忽略了时态，单复数等，获得了256个属性。

从一个更大的属性单词集产生句子是另一个探索。

这里使用一个CNN进行属性检测，使用的是基于CNN的方法，参考了基于区域的属性检测器。

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

### 语言模型

LSTM，表示如下：

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



## 5 实验

实验显示使用属性向量效果更突出

### 使用WordNet扩展

这里选择了与原来属性有联系的WordNet词作为扩展词典，然后将原来属性预测器的预测向量通过一个转移矩阵获得新的向量。可以表示为全连接和最大池化层的结合。

这里显著提高了性能，提供了额外的信息。

## 3 SCN

### 语义概念检测

这里首先要选择一批标签作为语义概念，使用了K个最常用词作为标签，包含了最常见的名词，动词或是形容词。

为了预测给定测试图像的语义概念，这里将其处理为一个多分类的标签问题。这里对图像打上标签为y=[yi1,…yik],每个维度代表一个语义单词，为0时代表不存在这个单词，为1相反，这时可以通过对损失函数的最小化：

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

这里s是一个K维的向量，首先对视觉向量进行全连接操作，然后使用sigmoid函数获得一个对应所有语义标签的概率值。

### SCN-RNN

这里SCN对传统的RNN进行了扩展，形成了一个标签独立的权重矩阵。特定的，SCN-RNN计算隐藏层状态如下：

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

这里z=1\*Cv，当时刻t为1时，v输入。W和U是一组标签独立的矩阵，根据语义概念矢量对图像中出现的语义标签存在的概率进行主观判断。

给定s属于Rk，我们定义两组权重向量Wt属于Rnh\*nx\*K，Ut属于Rnh\*nh\*k，nh是隐藏层单元数目，nx是词嵌入的维度。这样可以得到下面表达式。

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

Sk是s的第k个元素，Wt和Ut是一个k索引的二维切片，相对地。第k个语义元素和RNN权重Wt[k]和Ut[k]相联系，显式指定了k个RNN。结果，训练这样的模型可以使用k个个RNN的组合进行联合训练。

这里明显参数数量过大，因此按照下面的形式对Ws和Us进行了分解。

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

在这里Wa 2 Rnh\*nf ,Wb 2 Rnf\*K andWc 2 Rnf\*nx

将上面两个式融合到前式中，的到下面的公式。

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

Wa和Wc在所有caption中共享，高效的捕捉语义模式。相同的分析同样适用于U。在这个分解模型中，RNN的权重矩阵对应每个语义概念分享结构。



为了提供进一步的对分解的观察，这里用wbk代表Wb的第k列。

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

对Us也有一个相似的分解。上面方括号中的部分可以解释为一个权重张量的第k个切片。每个切片对应对应k个语义概念中的一个，每个是nh乘nx维的。因此，通过分解表达式可以提高K个RNN参数学习的效率。这种组合由于分享矩阵参数变得高效。这里张量的第k维的权重取决于各自的概率sk，将相对应的语义概念与图像i相关联。

### SCN-LSTM

这里将前面的模型泛化到了LSTM单元。特定的ht=（xt-1，ht-1，v，s）。





总结：这里对于视觉信息和语义概念进行了结合和利用。V送入LSTM中在初始步，给图像信息有一个概览。LSTM在全局的视觉概念v进行初始化后，使用一个K个LSTM集合的组合进行解码，以权重概念向量s为权重产生描述。

训练函数：



这里介绍了在视频领域的扩展。

## 实验

训练流程：使用ResNet-152的2048维池化层。预训练在ImageNet数据集上。为COCO和Flicker30k选择了1000，2000的语义概念词典。所有的参数是[-0.01, 0.01]平均分布初始化。所有偏移项初始化为0.词嵌入使用公开可得的word2vec向量。在预先训练好的词集中不存在的词是随机初始化的。SCN-LSTM隐藏层单元和元素隔宿都是512，mini-batch的大小是64.三个数据集上运行epoch的最大数量都是20.梯度在范数过大时会被截断。没有使用特定的数据调整和正则化除了dropout和early stop。优化算法Adam的学习率时2\*10-4.所有实验在Thenao上进行。测试时使用beam size为5.

### 探索

这里比较了视觉特征，语义特征，语义特征超出了视觉特征的能力，结合之后有更大的提高，直到使用本文的方法进行。

这里进行了模型的组合使用取得了更好的效果。

### 定性分析

可以对语句进行替换，形成不同的句子，也可以通过合适的功能词选择为通过的caption形成合理的描述。

感觉推理能力的重要性，如何通过图片信息判断图片中事物的合理状态。

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