Learning a Recurrent Visual Representation for Image Caption Generation阅读笔记

## 简介

本文探索的是图像和语言表述之间的双向映射关系。不同于将映射表示为embedding，作者考虑构建生成器，实现给定图像输出口语化语句。使用这个模型，还可以在给定图像描述的情况下重建图像的图像特征。作者使用文本递归视觉记忆的方法自动学习如何记忆长期的视觉概念，帮助语句生成和视觉特征重建。

前文提到embedding，它无法从中生成语句，或者视觉描述。作者提出的双向的表示有能力生成图像的口语化描述和图像描述的视觉表示。其中的关键便是要动态捕获所描述场景的视觉信息，因此作者使用RNN来实现。RNN一直存在无法利用到较远距离关系，作者提出在句子构建期间，把视觉表示作为已经被提及概念的长期记忆。这样网络可以自动地选取主要的概念来表达还没有说的内容。

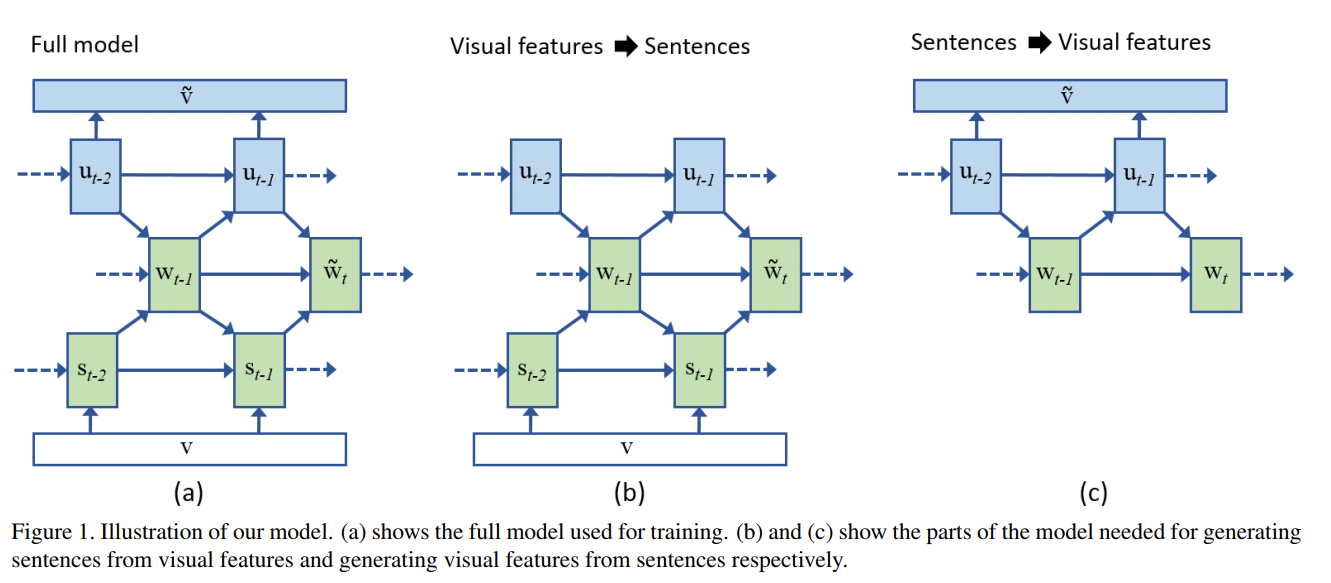
## 方法

作者的目标有二

1. 从给定的一系列视觉信息或特征中构建语句，也就是在时间通过之前生成的词和视觉特征生成使用词的概率。
2. 从给定的词序列建立视觉特征的似然估计，进而建立场景的视觉表示或者执行图像搜索。

为此，作者引入隐变量，用来编码的视觉理解，其扮演的长期视觉记忆项。作者可以使用计算和，整合这两个似然估计，可以得到全局的目标函数

#### 模型结构



在时间，表示所预测的词（one-hot向量），表示每个词的似然估计。RNN的hidden state（隐藏状态）表示在中，因为梯度消失，往往只能在小的范围内有效果。引入的是一系列视觉特征，并保持恒定，其对预测、选择有帮助。不需要建立和之间的联系，不过作者发现让连接一半的可以提升性能，因为其允许不同的单元对文字和视觉特征单独建模。

本文最大的贡献在于隐层，其通过之前的词语重建了视觉特征，换言之。也同样参与预测下一个词语。网络可以比较已经说过的记忆和当前观察的视觉，以进一步推断下一步想说什么。最开始只是视觉的先验概率表示，随着更多词汇的引入，逐渐开始反映词汇序列对视觉的理解。直观上理解，似乎图像的理解应该在模型训练完之后再继续，作者认为，当一个词汇被生成后，关于图像特征与此词汇的似然估计就会提升，那么网络就会得到正向的对抗压力，迫使在后续步骤中记住这个词汇。

因为方法的两个功能，预测图像特征和生成语句，前者未知，而后者未知。词单元将网络分割成两部分，使得和没有依赖关系，两个功能可以正常运作。否则如果和在网络中直连，那么网络则只能扮演一个自动编码器的功能。

#### 实现细节

###### 语言模型

作者的词库有约300~20000个词，如果每个词独立预测，计算成本过高。作者利用词分类的思想，将词汇预测分解成两个子式相乘

表示词的概率，表示分类的概率。词的分类标签使用无监督学习方法，将词以频率进行分类。在少量损失Perplexity的情况下大幅提升学习速度。预测词语使用标准softmax函数，每一个epoch计算Perplexity，如果没有减少则减缓学习速度。

作者为了进一步减少Perplexity，将输出和最大熵模型的输出进行组合。

作者做了两步数据预处理：将所有字母小写化，使用标准CoreNLP 工具符号化句子。

###### 学习

作者使用Backpropagation Through Time算法进行反向传播，特别的，每若干词网络会展开然后使用标准的反向传播。

一旦到达句末，网络会被重置，因此不会出现跨语句的影响。

在RNN权重和输出之间使用online learning（在线学习，及时地更新权重），而网络的其余权重则每一条语句更新一次。

激活函数使用带有clipping的sigmoid，可以抑制梯度消失现象，不带边界限制的ReLU存在数值不稳定的线性，在RNN中常会引发数值爆炸。

模型会同时学习词语和图像的表示，预测词汇的误差可以直接反向传播到图像特征中，考虑到深度CNN对数据量的要求和语句-图像数据集的图片数量，作者采用预训练的模型来避免过拟合。

## 讨论

Image caption任务不仅描述了图像中的物体，还描述了它们之间的关系。检查序列化地探索图像和它与图像描述之间的关系。许多词汇对应于空间关系，现有的模型往往难以检测到。

## 总结

作者提出了在图像特征和句子之间的双向模型，其可以口语化的图像描述和视觉特征，与使用RNN的传统方法不同的是，作者引入了长记忆的互动项，构造出了现有序列对图像的理解，规避了RNN对于较远的信息感知能力不足的缺陷。