1. Introduction

本文致力于研究图像描述（Image Caption）问题，即给定输入图像，输出描述该图像的（中文）语句。我们的模型借鉴了Vinyals et al. [1]的工作，分别采用LSTM (Long Short Term Memory unit) 和GRU (Gated Recurrent Unit )作为模型的核心，以8000张图片作为训练集，各1000张图片作为验证集和测试集，实现了对图像的中文描述任务。

2. Method

**2.1 Model**

本文所实现的模型基于 Vinyals et al. [1] 的文章。

Image Caption本质上是一种时间序列预测， 为了根据输入图像和句子估计模型参数，用下式最大似然估计估计参数，

其中为输入图像，为训练标注的句子。其中，由于该问题是一个时间序列预测问题，因此同一句话的不同字之间并不独立，所以在时间的输出与之前的 都有关，因此

为了表述简洁，这里舍去了模型参数。我们的目标就是找到最大化后验概率的。

为了实现这种预测模型，我们分别采用了两种递归神经网络(RNN)：LSTM (Long Short Term Memory unit) [2]和GRU (Gated Recurrent Unit ) [3]来对句子进行预测。

**2.2 LSTM**

不同于简单的RNN，LSTM引入了记忆单元（memory cell）c，

其中是时刻的输入，为要训练的权重矩阵，是时刻的忘记门（forget gate），是输入门（input gate），是时刻的隐状态变量，表达式如下：

其中是输出门（output gate），表达式如下：

最后，由于每一个字的生成本质上是一个多分类问题，因此最终的输出概率为

**2.3 GRU**

GRU与LSTM很像，都有用来控制记忆的门，不同之处在于GRU不像LSTM那样有一个记忆单元。

GRU的时刻隐状态变量表达式如下：

其中是更新门（update gate），是候选隐状态，表达式如下：

是重置门（reset gate），用下式计算

最终输出概率为

可见，二者的结构有很多相似之处，比如GRU里的重置门类似于LSTM的忘记门。最大的不同在于计算时刻隐状态变量时，LSTM是直接将记忆单元与输出门相乘，而GRU是利用更新门求时刻隐状态变量与候选隐状态的加权平均。

**2.1 Procedure**

Image Caption经过如下流程实现句子预测：

1. 利用卷积神经网络(CNN)对输入图像编码，得到图像特征，之后通过一个全连接层，降低维数；
2. 对训练集的句子进行预处理以及单字编码，得到字典，并用其将所有数据集的句子转换为数值向量；
3. 用Embedding层对数值向量做“Word Embedding”，将字映射到向量空间，向量维数等于图像特征降维后的维数；
4. 设计LSTM/GRU，首先输入图像特征，放弃第一个输出，之后依次输入词向量，将输出导入全连接层映射到字典空间；
5. 训练网络后，采用beamsearch保留前K个最大后验概率的句子直至搜索结束，取后验概率最大的句子作为caption的句子。

该模型的结构图如图1所示。

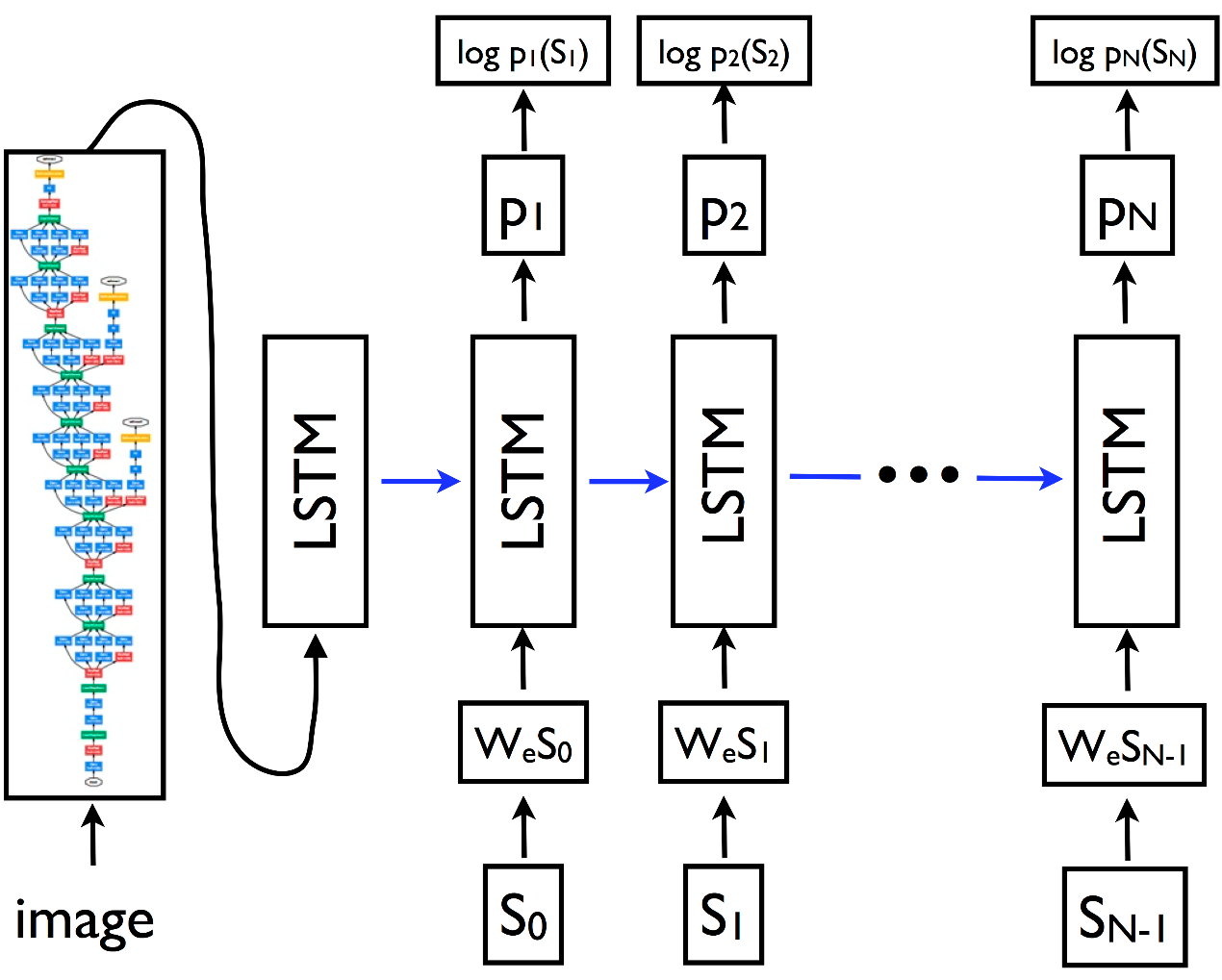


图1. 神经网络结构图[[1]](#footnote-1)

Reference

1. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3156-3164).
2. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.
3. Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

1. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3156-3164). [↑](#footnote-ref-1)