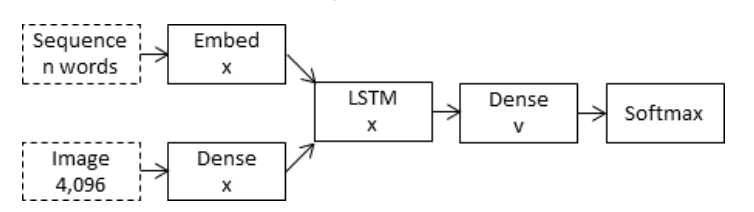
第三次提交报告 Image Caption Generator

2018202119 罗迪

**一、总述：**

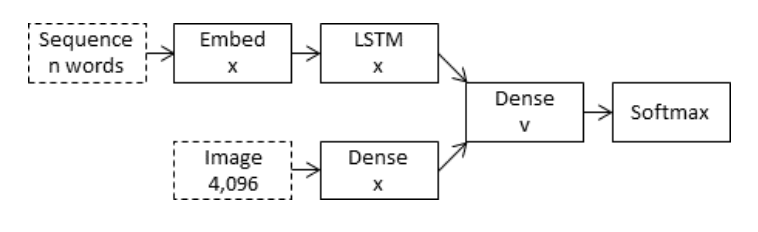
前两次报告我们主要聚焦在“Show and Tell: A Neural Image Caption Generator”这篇论文提出的NIC模型，它是一种Inject Model，使用LSTM作为文本生成模型，该模型使用图像和之前生成单词信息的序列作为输入以生成序列中的下一个单词。显著特征是把图像向量与单词同等对待，共同作为字幕前缀的一部分。在通过keras实现基础模型框架的基础上，我们对模型进行了精度调优，以达到较好的一个效果。



在中期报告之后，通过阅读和思考一些新的论文，我们开始尝试一些更好的模型架构，同时引入了attention监督机制，并且尝试了把数据集从Flickr8K换到了MSCOCO，在更广泛的图像基础上进行训练，使模型能够识别和描述出更多类别的事物。

**二、从Inject Model到Merge Model：**

Inject Model和Merge Model的灵感来自于17年的一篇论文“Where to put the Image in an Image Caption Generator”，在这篇论文中作者着重思考和实验了在Image Caption Generator中图像的重要性和输入模型的时机。最后作者提出了用Merge Model来改进之前的Inject Model，它的显著特征是仅使用循环神经网络对文本进行编码，而把图像的特征向量保留在LSTM之外，因为循环神经网络对文本特征的处理更加有效，而对图像的处理相对较弱。模型框架如下图所示，将LSTM处理后的文本向量与图像的特征向量进行相加，再放入Dense层中进行训练和输出预测结果。与此同时，作者还研究了两个向量的合并方法，结果发现用加法Add比用串联Concaten和乘法Multiply的效果要更好。



具体实现方面，我们的数据预处理、卷积提取图片特征的部分仍然延续之前的策略，区别在于模型的各层设计上：

在图片特征提取部分，基于VGG16模型的输出作为输入，接下来是一个Dropout层，随后是一个256的全连接层；

在文本特征提取部分，文本特征信息连接一个词嵌入层，一个256维的LSTM层；

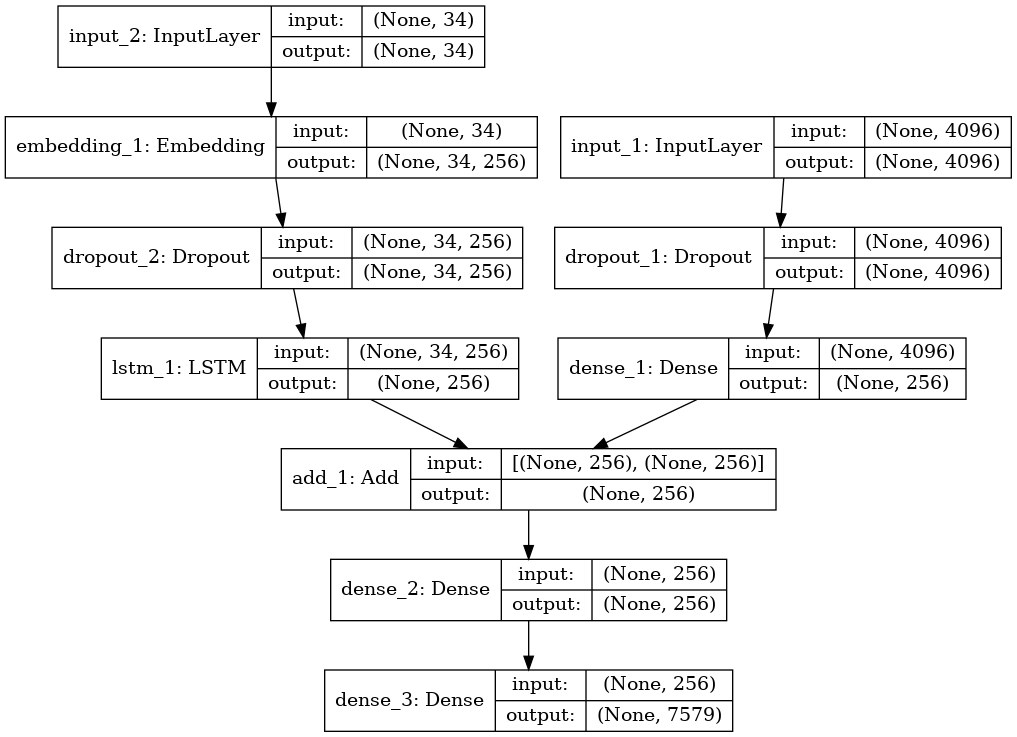
经过这两个板块的处理，图像信息和描述语句信息都被转换为了256维的向量，这时候将两者通过加法合并成一个256维的向量，再接两个全连接层，得到的输出表示序列的下一个单词在整个词汇表上的概率分布。

训练过程中使用softmax激活函数，在拟合网络时将分类交叉熵损失函数最小化。

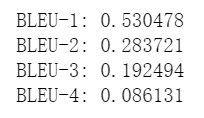
代码如下：



最终的模型可视化效果如下：

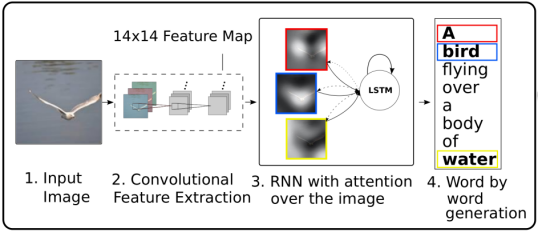


结合实际情况，我们选择对模型训练20个epoches，重复3次，来选择效果最好的模型。模型训练完成后，会对测试集的每张图片生成对应的描述语句，我们采用的是BLEU分数来评估模型的效果。通过循环多次评估模型，我们可以得到方法模式下模型的平均得分，用来衡量此方法的模型训练效果。



**三、Attention机制的引入：**

除了改变模型的架构，“Show, attend and tell: Neural image caption generation with visual attention”这篇论文把在机器翻译中广泛使用的attention机制引入到了Image Caption Generator中，整个流程被简化为两个步骤：CNN提取图像特征和带有Attention机制的RNN解码图像特征。Attention机制的核心是把图像中一些具体的位置赋予更大的权重，让机器重点关注这些部分并输出文字。

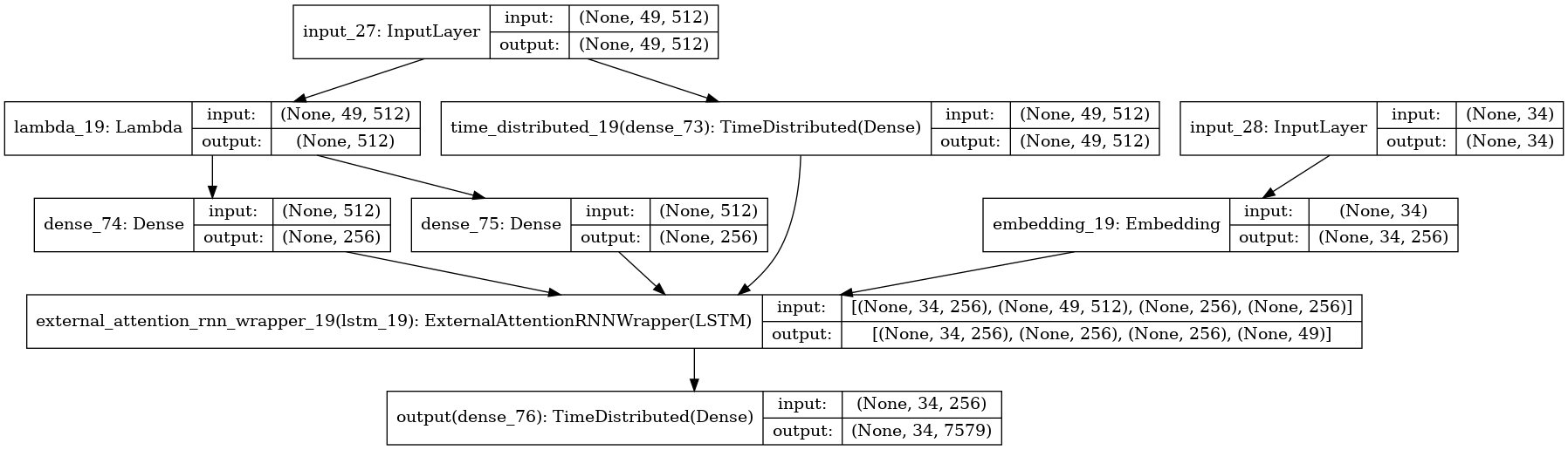


由于keras并没有实现attention层，所以在具体实现的过程中我们用到了一个基于keras实现attention的包kulc，其中的ExternalAttentionRNNWrapper即为带有attention机制的RNN解码层。训练部分图像特征经过全连接层、文本特征经过词嵌入层后，一起放入基于LSTM的attention层中，最后经过一个softmax全连接层得到输出。模型如果要输出预测语句还需要再额外训练一个inference\_model，这里不再赘述。

代码如下：

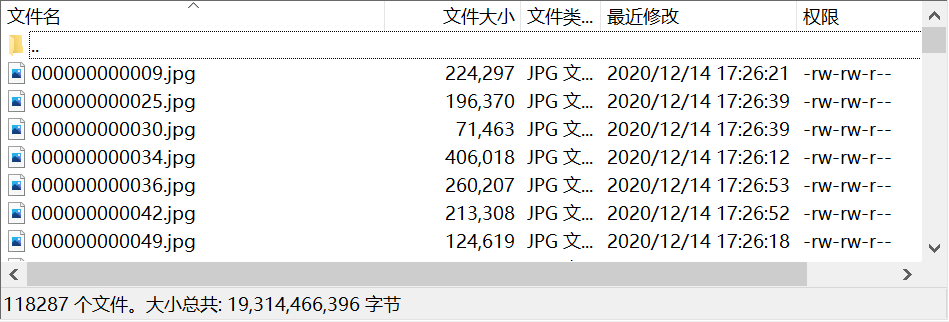


最终的模型可视化效果如下：

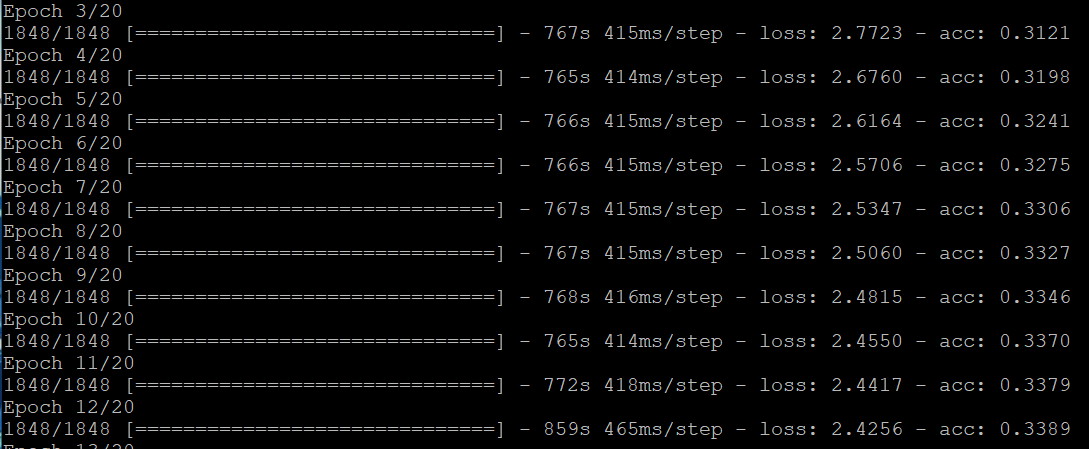


**四、更大数据集的引入：**

在上述模型的训练过程中我们还发现了一个问题，因为采用的是flickr8k数据集的8000张图片进行训练，生成的语句中只能识别出人、狗等最基本的对象，而对公交车、蔬菜、大象等不那么常见的事物效果很差，可能也是数据集有限引发的模型过拟合问题导致的。为了解决这个问题，我们尝试了更大的数据集MSCOCO，如下图所示，训练集采用了11万张图片，总大小达到19G，这样能够尽可能提高模型的泛化性，增加应用场景。



将训练集扩增了十多倍后，模型训练的时间也出现了递增，平均完成一个epoch需要十多分钟，效果自然也更好。下图为其中一个训练流程展示。



**五、各阶段模型效果展示**

挑选了不同类别、不同场景的几张图片，分别在第一阶段Inject Model、第一阶段优化后的模型、第二阶段的Merge Model和第三阶段的Attention Model下预测并生成输出，并且与数据集本身的人工标注进行比较，来验证模型更迭的结果。

① 人-滑雪



第一阶段模型输出：person is standing on the snow

第一阶段优化后模型输出：man is climbing down snowy mountain

第二阶段模型输出：man in jacket is riding down snowy mountain

第三阶段模型输出：man is skiing down a hill

数据集标注语句：man is skiing on the snow slope

可以明显看出模型认知的递进过程，从最初的人站在雪上，到人从雪山上爬下来，再到人从雪山上滑下来，再到认识到这是在滑雪。

②公交车-行驶



第一阶段模型输出：man in red shirt is standing on the beach

第一阶段优化后模型输出：man in red shirt is sitting on the street

第二阶段模型输出：a red car is driving on the street

第三阶段模型输出：a red double decker bus is driving on the street

数据集标注语句：a red bus is driving on the road

这个图片则能看出更换数据集对模型的影响，在用MSCOCO之间，模型无法识别出这是公交车，而认为是穿着红衬衫的人，而用COCO数据集训练过的attention模型甚至可以识别出这是一辆双层公交车，效果甚至好于数据集的标注。

③狗-在水上走



第一阶段模型输出：two dogs are playing in the water

第一阶段优化后模型输出：black dog swims in the water

第二阶段模型输出：a dog is running through the water

第三阶段模型输出：a black dog is walking through the water

数据集标注语句：a black and white dog is walking through the water

③人-打网球



第一阶段模型输出：two men are playing on the field of the ball

第一阶段优化后模型输出：two men are playing soccer

第二阶段模型输出：two men in red and white uniforms are playing soccer

第三阶段模型输出：a man is playing tennis

数据集标注语句：a male tennis player in white shorts is playing tennis

这张图很具有迷惑性，首先是同样出现在场景中的裁判/技术员，都误导了前三条模型的输出，误以为是两人在玩；另外是旋转的网球因为视角的原因偏大，前三条输出都误以为是足球。只有attention机制的模型成功注意到了运动员手中的网球拍而推测出这个运动员是在打网球，从而接近正确的标注语句。