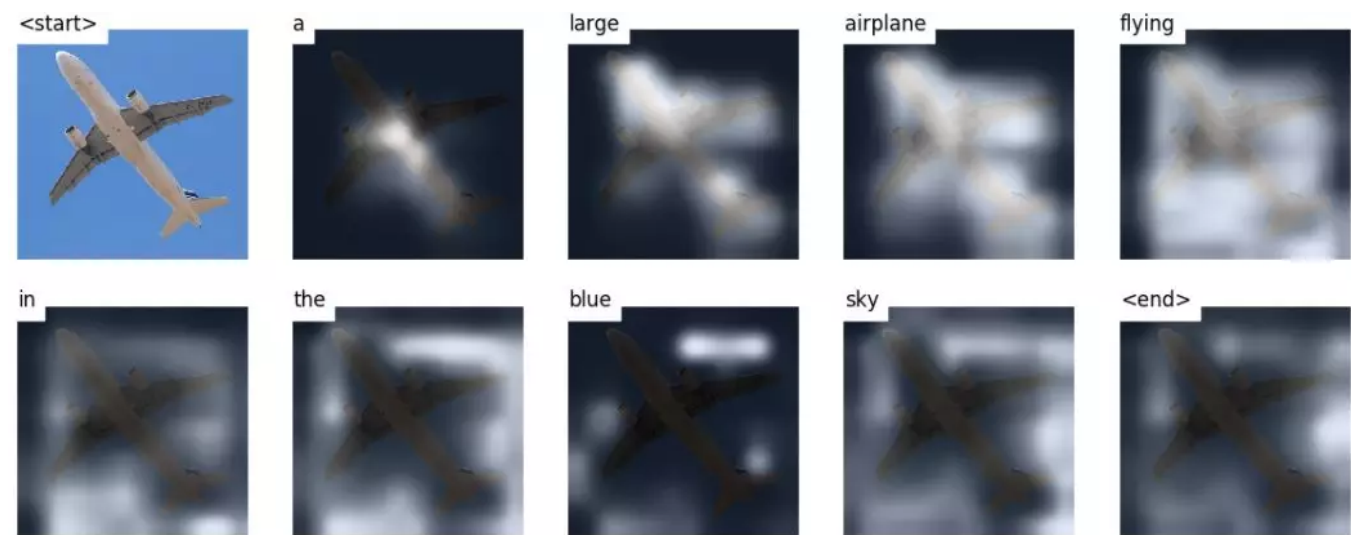


# 计算机视觉 | 图像描述与注意力机制

原创 石头 机器学习算法那些事 2019-12-28

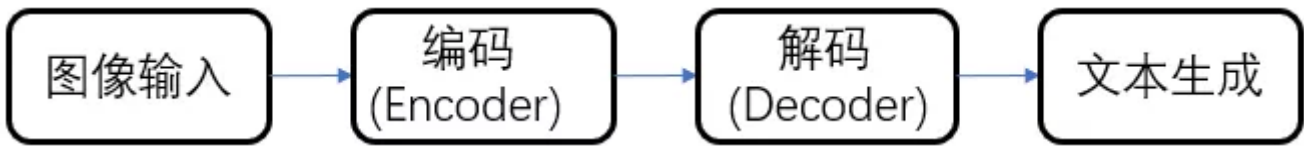
图像描述的含义是生成图像的描述，采用注意力机制生成图像标题，图像标题的每个词集中在图像中最相关的部分，并且预测下一个词。

如下图的图像生成：



图像标题：<start>a large airplane flying in the blue sky <end>

图像标题生成框架：



该框架涉及的几个概念：

**图像编码 (Encoder)：** 将具有3个彩色通道的输入图像编码成具有“学习”通道的较小图像，这些编码图像包含了原始图像的信息。

**图像解码 (Encoder)：** 将编码图像逐字生成标题。

**注意力网络 (Attention)：** 编码与词相关的图像，每个标题的词集中在图像最相关的部分。

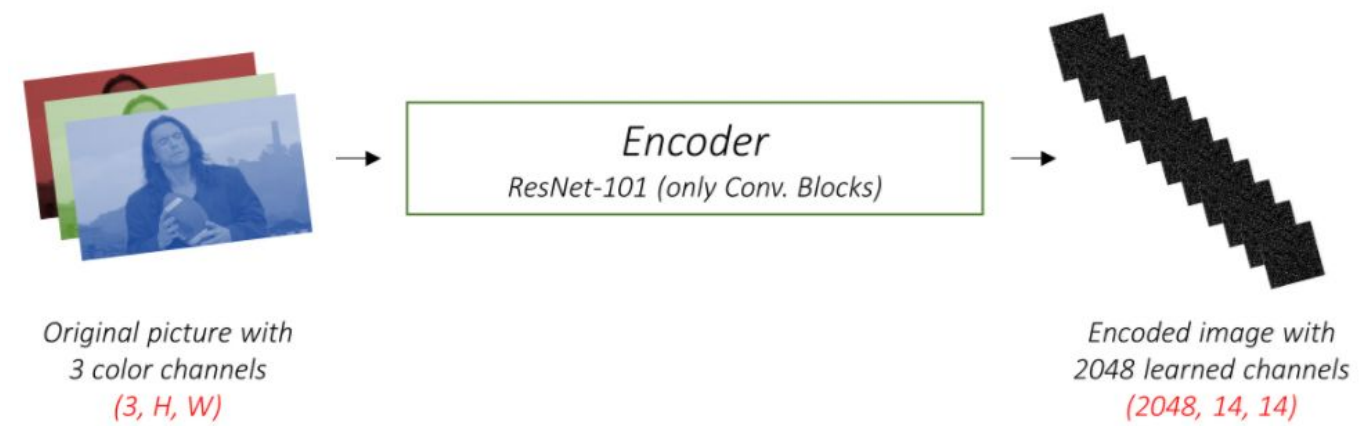
**束搜索 (Beam search)：** 解码器逐字生成的标题序列中，束搜索算法得到最优的标题序列。

下面详细介绍这几个概念。

1.图像编码

我们使用ResNet-101网络去编码图像，需要去除最后两层的线性层，因为最后两层的线性层是用于分类任务的，图像编码只需提取特征。

图像编码网络如下图：

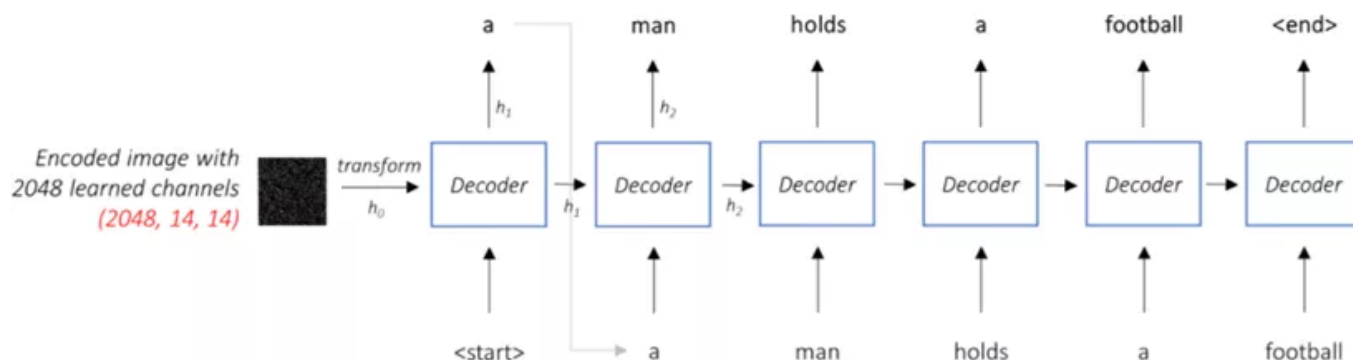


ResNet网络编码的结果是由2048个通道大小为14×14图像组成，模型参数通过迁移学习获得。

2.图像解码

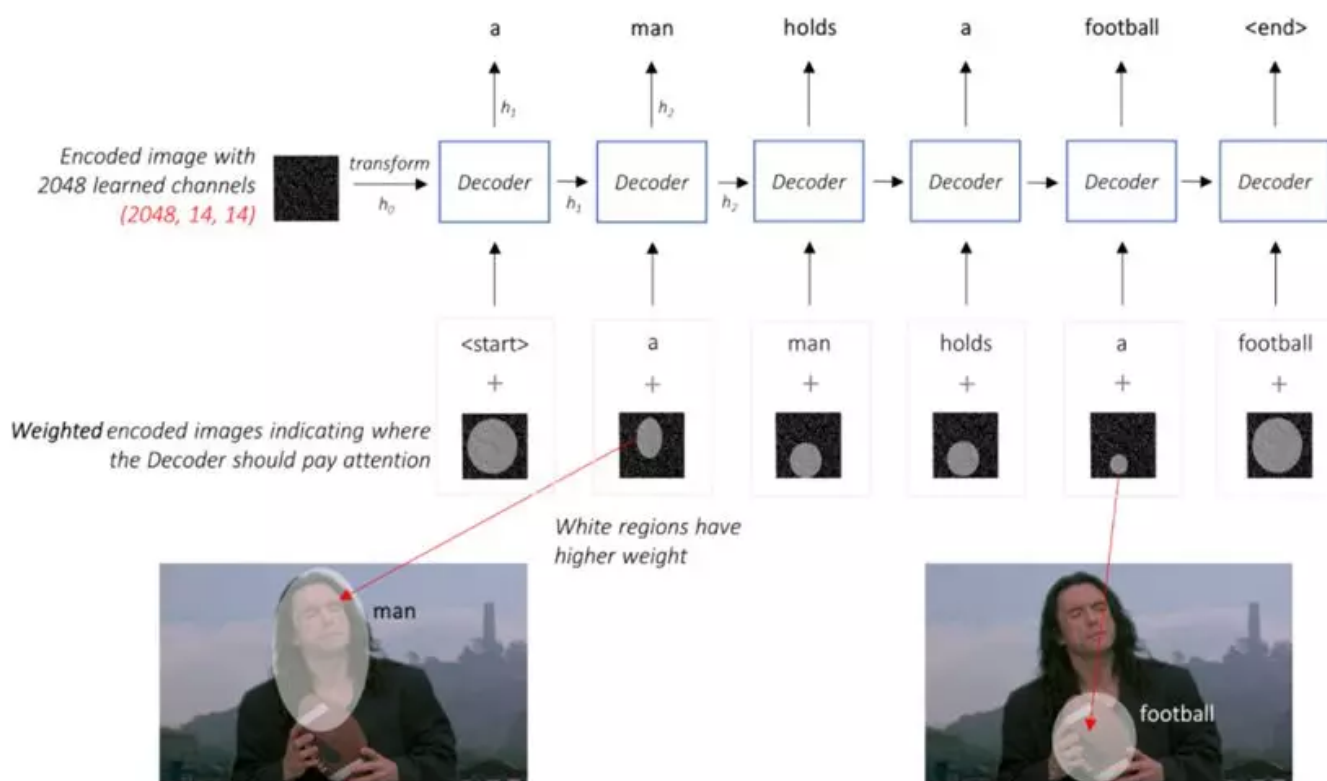
解码器是根据编码图像逐字生成标题，这里使用循环神经网络（RNN）生成标题序列，选择的RNN类型为LSTM。

若解码器不使用注意力机制，那么解码器的算法流程是：首先对编码图像所有像素进行平均，得到2048×1的向量，然后无论对该向量是否进行线性变换，都可以将其作为第一个隐藏状态输入解码器，生成第一个单词，并用该单词作为输入生成下一个单词。



若**解码器使用注意力机制**，那么解码器在生成单词时，需要考虑该单词最相关的图像部分。比如语句序列a man holds a生成单词football时，解码器需要关注图像中足球所在的区域，并给该区域较大的权重。

如下图含有注意力机制的解码器：



解码器网络的输入是前一个RNN单元的输出隐藏层，上一个单词的嵌入向量和注意力网络生成的权重图像，算法代码需要将嵌入向量和注意力网络拼接成一个向量作为输入。

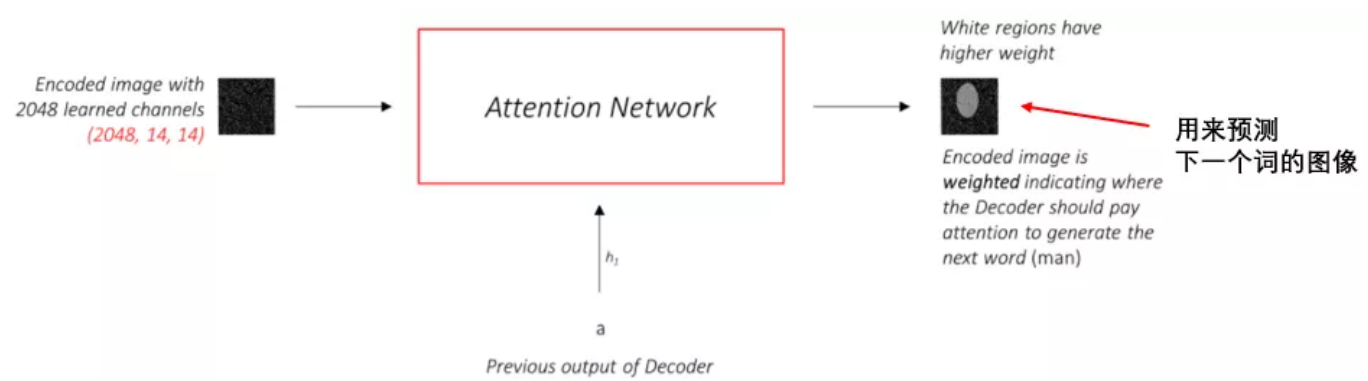
### 3.注意力网络

注意力网络计算与词相关的像素权重。

凭自己的直觉，如何估计图像某一部分的重要性？若要突出图像某一区域的重要性，那么需要提高该区域的权重。

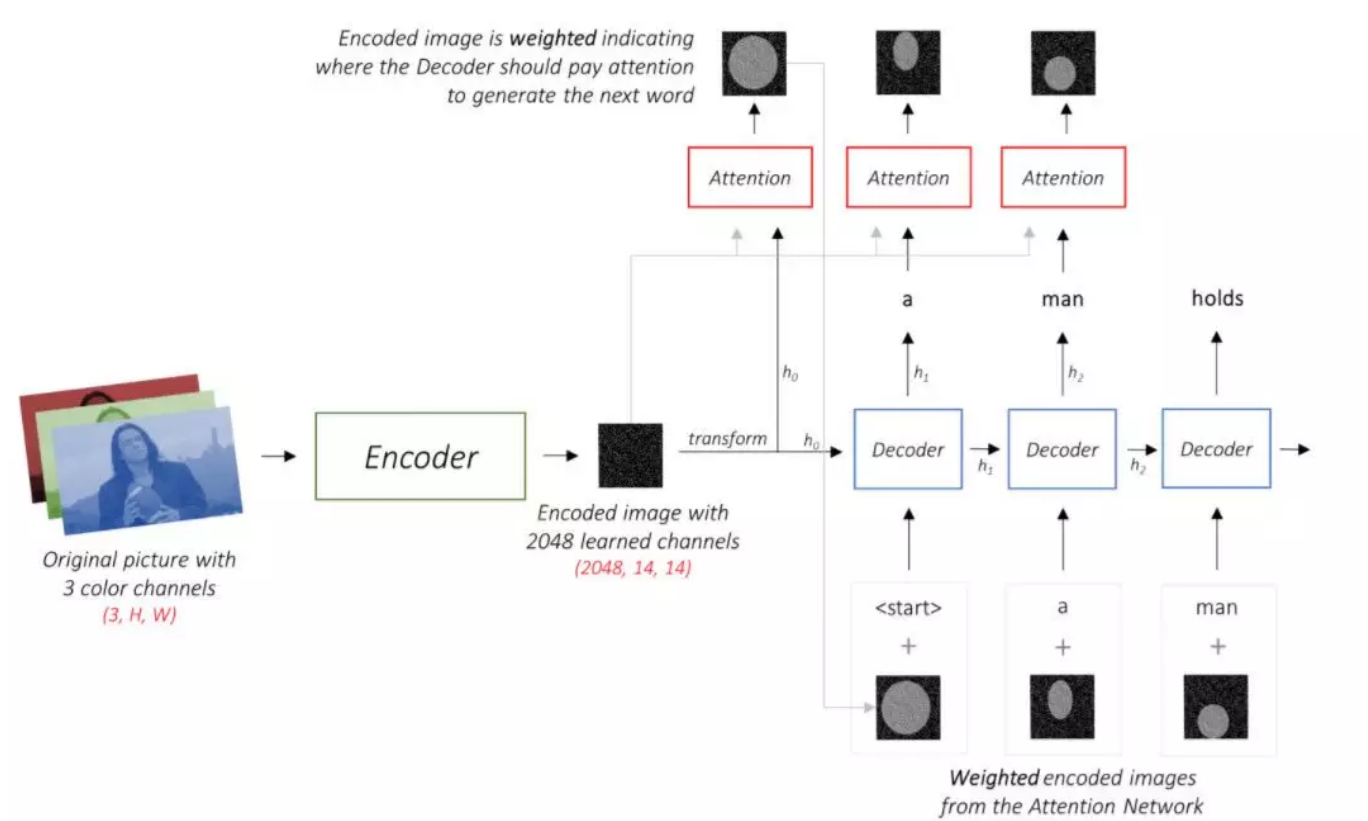
在图像描述项目中，你需要了解到目前为止生成的序列，根据注意力网络生成像素权重，并决定接下来需要描述什么。

这正是注意力机制所做的——它考虑目前为止所生成的序列，并关注接下来需要描述的图像部分。如下图：



4.图像描述框架

根据前面介绍的编码器，解码器和注意力机制，图像描述框架如下图：



算法流程：

- 1) 编码器编码输入图像的信息，生成1048个通道大小为14×14的图像，编码器采用ResNet-101网络，不包括网络最后两层的线性层。
- 2) 注意力网络根据编码图像和上一层解码器的输出隐藏状态，生成与下一个单词相关的图像。
- 3) 解码器生成图像的标题序列，解码器采用LSTMcell网络。

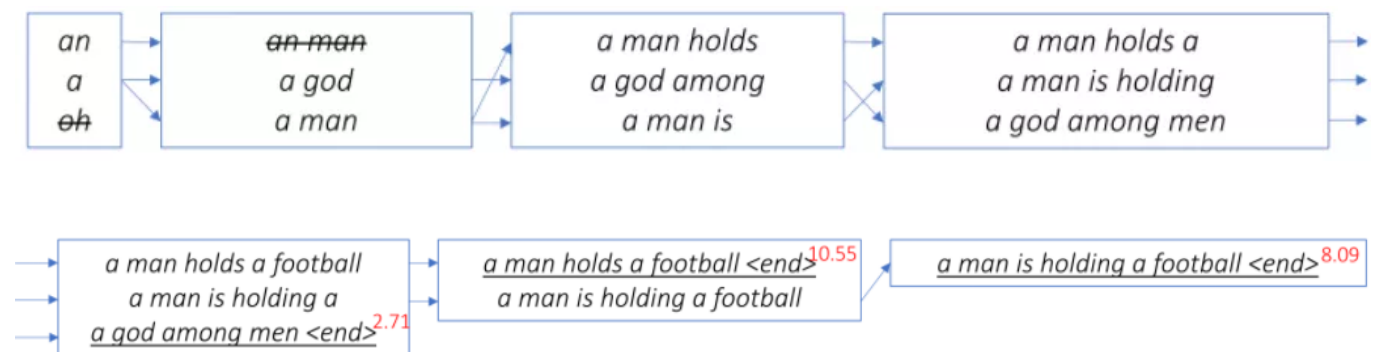
## 5. 束搜索(Beam Search)

我们使用线性层将解码器的输出转换为词汇表中每个单词的得分。

最直接和贪婪的方法是选择当前得分最高的单词来预测下一个单词，这种做法很可能生成的不是最佳序列，因为剩下的单词序列取决于你选择的第一个单词。如果第一个单词不是最好的，那么接下来的序列预测都是次优的。

**解决方法是：**每次解码器都选择最好的3个单词，比如你在第一步选择3个最好的单词，第二步根据第一步的每个单词，都生成3个最好的单词，即第二步共生成9个单词。**结合第一步第二步，选择最优的3个单词序列。**以此类推，当预测单词为<end>时，标题序列生成结束。

如下图的束搜索算法生成最优标题序列：




由上图可知，最优标题序列为：a man holds a football

参考

<https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning>

欢迎扫码添加小编微信（请备注方向），邀请你入机器学习算法群，一起交流进步：



石头 

湖南 邵阳



扫一扫上面的二维码图案，加我微信