# 自顶向下的和自下向上的注意力模型[[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

自上而下的视觉注意力机制已经广泛应用到图像字幕和视觉问答中，通过细粒度分析甚至多个推理步骤加深对图像的理解。这里作者提出了对自下而上和自上而下注意力机制的组合，使注意力机制可以在目标上和其他突出的图像区域计算。自下而上的注意力机制提出图像区域，每个区域是一个特征向量，自上向下的注意力机制决定特征权重。

## 1 介绍

自上向下的注意力机制描述为根据高层次信息决定，通过一个部分完成的图像Image Caption提供上下文信息，然后对于CNN的某一层或几个层提过空间的位置上的权重，这里无法在使用区域数量上获得一个明确的指导。

由于图像中对于一些信息的区域定义模糊，使将图像区域的绑定相应信息变的困难。

过去的论文很少考虑如何应用注意力机制到显著的图像区域中。这里借鉴了两篇基于图像区域的注意力模型。这里建立了使用Faster RCNN建立了在视觉和语言任务之间关系，直接应用目标检测上的进展。这里和在ImageNet上类似，利用了跨领域的视觉知识。

## 3 方法

给定一幅图像，输入是一个可变数量的图像特征向量，V={v1,…,vk},vi属于RD，因此每个图像特征对图像的一个显著区域进行了编码。空间特征V作为自下向上的注意力模型的输出，然后紧跟着进行标准化操作，作为一个CNN的空间输出层。这里对于自顶向下的注意力组件，多使用了简单的单注意力模型，可以通过堆叠、多头或双向注意力模型进行改进。

### 自底向上的注意力模型

空间图像特征V的定义是一般性质的。然而，在这项工作定义为空间区域的标注框使用Faster RCNN获得。Faster RCNN是一个目标检测框架将图像中目标归到特定的类和使用标注框进行定位。其他的目标检测网络也可以使用注意力机制进行训练。

在这项工作中使用Faster RCNN和ResNet 101联合使用。这里使用图像特征V是模型的最后输出，并且对于每个目标类的输出使用极大值抑制。我们然后选择了类检测概率超过一定值的所有区域。对于每个区域，vi定义为一个平均池化的卷积特征，因此图像的特征的维度D是2048维的。这里Faster RCNN高效的表现为一种硬注意力机制，只选择一部分图像区域在一系列可能区域中。

我们为了预训练自底向上的注意力模型，首先初始化了Faster RCNN使用在ImageNet初始化的ResNet101模型。我们然后在Visual Genome数据。为了获得良好的特征表示，，这里将平均池化后的卷积特征和一个已经进行嵌入的分类向量输入到附加的输出层中获得一个预定义的softmax分布在属性类上。

这里使用原有的多任务损失函数，然后加入了附加的多类损失组件训练属性预测器。

### Caption 模型

给定一系列图像特征V，我们使用了一个软的自上向下的注意力机制对于词语产生过程中的每个特征赋予权重。这里使用的是通常的软注意力机制，但是各个模块使用的最新的成果，单纯使用效果也很好。

在一个高层次上，caption模型由两个LSTM组成使用的是标准化的处理。

注意力一是选择作用，二是获取图像到固定向量的编码。视觉注意力可以被看作是一种动态的特征提取机制，它结合了随时间变化的上下文。参考的是一下概念。



### 自上向下的注意力机制LSTM

在caption模型，我们将第一层LSTM层作为一个自上向下的视觉注意力层，第二层LSTM层作为语言模型。

第一层LSTM输出一个在v的形成过程中被考虑，同时输入是过去产生单词的编码，图像均匀池化后的连接以及上一个时刻语言模型的输出。

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

We是词嵌如矩阵对于词典，是一个上一时间步产生的one-hot向量。输入为LSTM注意力提供了最大程度的上下文信息，包括语言LSTM的状态，图像的全局内容以及输出产生的部分图像Image Caption。词嵌入不经过预训练遂渐初始化进行学习。

给定attetion LSTM的输出ht1，在每个时间步t下我们产生了一个正则化的注意力权重阿尔法i，t对于图像特征vi表示如下:

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

在这里Wva属于RH\*V，Wha属于H\*M，wa属于RH是要学习的参数。经过注意力处理的向量作为语言LSTM被计算为所有输入特征的组合。

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

3D特征地图的每个2D切片对由滤波器通道引起的空间视觉响应进行编码，其中滤波器作为模式检测器执行较低层滤波器检测诸如边缘和角落的低级视觉提示，而较高级别的过滤器检测高级语义，零件和物体等模式[40]。通过堆叠层，CNN通过视觉抽象层次提取图像特征。因此，CNN图像特征基本上是空间的，通道的和多层的。

这里在使用通道注意力模型的时候使用的是多层的特征。

为什么要引入multi-layer呢？因为高层的feature map的生成是依赖低层的feature map的，比如你要预测cake，那么只有低层卷积核提取到更多cake边缘特征，高层才能更好地抽象出cake。另外如果只在最后一个卷积层做attention，其feature map的receptive field已经很大了（几乎覆盖整张图像），那么feature map之间的差异就比较小，不可避免地限制了attention的效果，所以对multi-layer的feature map做attention也是顺其自然的事。

这里将特征视为堆叠的过程，从目标反推，目标是要识别高层次的语义分类特征，然后高层次的语义特征需要良好的低层次特征支持。



## 2 相关工作

这里将工作主要分为3类:

空间注意力

语义注意力

多层次注意力

这里对空间注意力模型进行了总结，认为在最后一层的特征图上使用注意力机制，使感受野范围过大，忽略了细节的信息，不同区域之间的特征图差异也较小，难以达到注意力几种的目的。

语义注意力：这里提到使用图像与标题之间的相关性作为全局语义信息来指导LSTM生成句子。然而，这些模型需要外部资源来训练这些语义属性。在SCA-CNN，一个卷积层服务器的每个过滤内核作为语义探测器40。因此，SCA-CNN的频道式的关注类似于语义上的关注。

多层注意力：这里解决的即是上面提到的问题。

仔细体会，可以看出channel wise attention是在回答“是什么”，而spatial attention是在回答“在哪儿”，二者是不一样的。

## 3 空间和通道注意力

CNN+LSTM整体架构。

SCA-CNN通过通道的注意力和多个层的空间关注力使最初的CNN多层特征图适应了句子的上下文。

这里假设获得image caption的第t个单词，这里已经获得LSTM的记忆ht-1，d维向量，在第l层，空间和通道注意力权重是ht-1的权重和CNN特征Vl的函数。因此，SCA-CNN将Vl以一种重复的、多层的方式来调节Vl。表达式如下所示。

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

这里Xl是调整后的特征。机制是空间和通道注意力函数。Vl是卷积层输出的特征图。F是一个线性权重函数应用于调整后的CNN特征和权重。这里不再对视觉特征进行加和，而是应用元素级乘法。

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

这里L是总体卷积层数，pt是一个概率特征和D是一个预定义的词典包括所有caption words。

注意到这里是与Vl和Xl具有相同的维度（Wl\*Hl\*Cl），这样要求一个WlHlClk的空间作计算，对于GPU内存代价太高，因此这里分别获得空间注意注意力权重和通道注意力权重。

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

通过这里的运算将复杂度分解为WlHlk和Clk。

### 空间注意力

这里作者将V=[v1,v2,…,vm]重新进行变形，将宽和高展平，这里vi是C维向量，m=W\*H。考虑vi作为第i个位置的视觉特征。给定LSTM过去的时间步隐层状态ht-1，使用一个单层的神经网络紧跟着softmax函数产生注意力分布阿尔法。下面是空间注意力模型的定义;

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

这里Ws是k\*C维矩阵，Whs是k\*d维矩阵，Wi是k维矩阵。作为转换矩阵将图像视觉特征映射到与隐藏层状态同样的维度上去。这里包含矩阵和向量之间的加法，通过矩阵和向量之间的加法是通过把矩阵的每一列加上向量来完成的

### 通道注意力

这里注意前面计算空间注意力权重时需要视觉特征V的参与，但是视觉特征V不是基于注意力机制的。因此，我们引入通道注意力机制来关注特征V。这将不同的卷积核代表的是不同模式的匹配，因此可以将获得的不同通道的特征图是做对语义信息的属性选择。

这里首先将特征图从V缩放到U，每个通道用ui表示，然后对每一个通道应用平均池化获得通道特征v：

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

获得向量v后，通道注意力模型可以如下定义。

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

这里第一个等式中是一个外积，Wc（k\*1），Whc（k\*d），Wi（k\*1），bc（k\*1），bi（1）是偏置向量。

根据两种注意力模型实施的顺序可以分成两种类型。

通道-空间：在空间注意力前增加通道注意力。如下图所示。



具体公式如下所示；

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

空间到通道

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

两种方案都是先获得一种注意力权重，对信息进行一次过滤，再使用过滤后的信息获得另一种注意力模型的权重，在两种权重的综合作用下的到使用的特征图，从信息的引导上这两种模型并不相同。

## 实验

实验主要验证1。通道注意力的重要性2.多层注意力的重要性3.整体的性能

这里使用了VGG19和ResNet-152作为基础的SCA-CNN架构。词嵌入模型设置为100维，LSTM隐藏层状态为1000维，两种注意力权重为512维。

使用了流搜索的算法，流大小为5.

这里主要对比了S，C，硬注意力模型，S-C和C-S。

### 通道注意力

在S中，VGG网络表现更好，因为VGG网络保留了空间信息，ResNet进行均匀池化后破坏了空间信息。在C中，ResNet表现更好，显然ResNet有更多的通道。在ResNet中，C-S和S-C均获得了比S更好的表现。这证明了本文方法的有效性当通道数量比较大时。在两种网络下，C-S和S-C效果接近，通常，C-S效果更好。

### 多层注意力探索

这里实验了1，2，3层注意力机制下的效果，当层数增加时，两种模型均取得更好的效果。太多层增加时，效果会下降，过拟合，同时由于Flicke数据集更小，也更易过拟合。

### 整体效果

作者将这个模型的成功归结于多种注意力机制的考虑。

这里通过可视化的方法分析各各层个通道的分布。

我们在图3中提供了一些定性的例子，以便更好地理解我们的模型。为简单起见，我们只在一个词预测步骤中可视化结果。例如，在第一个示例中，当SCA-CNN模型试图预测单词伞时，我们的通道将会根据像雨伞、棒状和圆形的形状等语义，在由过滤器生成的功能地图通道上分配更多的权重。每个层的直方图表示所有通道的概率分布。上面的地图是空间注意力地图，白色表示模型大致上的空间区域。对于每一层我们选择了两个通道具有最高通道的注意力概率。

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