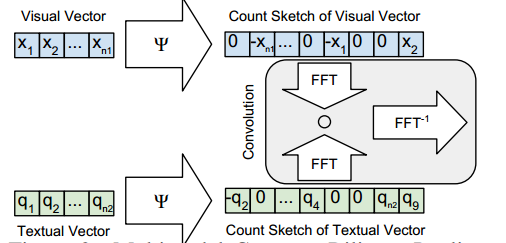
第11周汇报

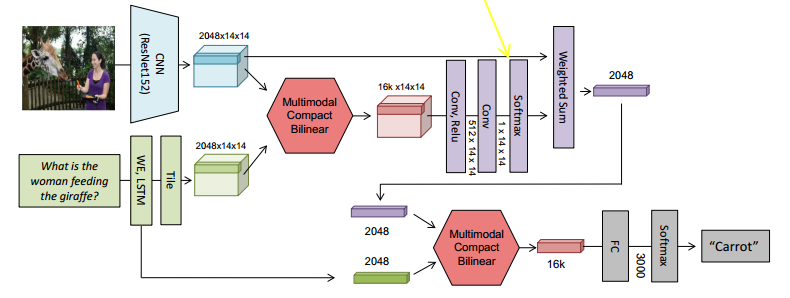
——张溢炉

1. 《Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding》总结
2. 内容摘要

本文介绍了一种能有效结合两种模态表示向量的方法：MCB

1. 总结及理解

* 动机：组合两种模态表示向量的方法：内积、和、串联、外积，本文假设外积更具表达性，但高维时计算外积不可行，参数太多，故本文提出Multimodal Compact Bilinear pooling (MCB)方法来组合多模态表示向量，此方法更有效、更具表达性
* MCB模型如下图下图所示，输入为图片表示向量和语言表示向量，输出为两者外积的count sketch，解决了外积在高维时计算量大、不可行的问题。其中FFT表示傅里叶变换，这里涉及卷积和傅里叶变换关系的公式：x0 ∗ q0=FFT−1(FFT(x0) FFT(q0))。其中Count Sketch projection function Ψ和相关证明见文章文献Charikaret al., 2002和Pham and Pagh (2013)。
* 文中提到了应用在VQA任务上的注意力机制模型，如下图所示。模型将图片划分为14\*14的区域，使用MCB模型有效利用图片和问题得到每个区域的注意值，然后通过加权和得到2048D向量。再将该向量和语言表示向量经过MCB得到16K维向量，最后通过全连接层得到输出。



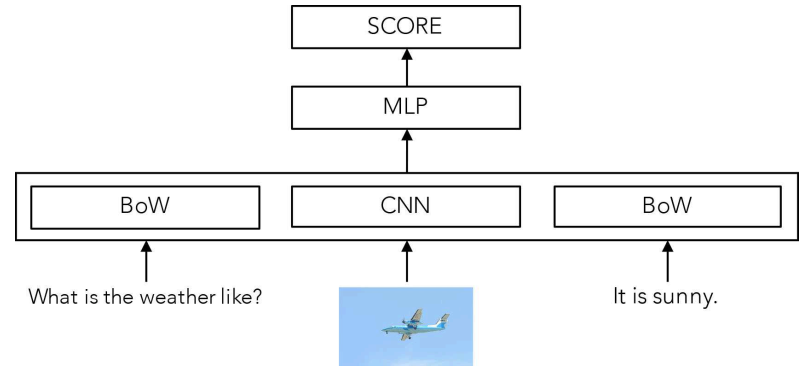
* 从实验结果来看，MCB的表现相对点积、求和串联等方式具有明显的提升，将该方法用到第4篇中能否带来性能提升？

1. 《Revisiting Visual Question Answering Baselines》总结
2. 内容摘要

本文提出了一种判断图片-问题-回答三元组是否正确的模型，用于多选项的VQA任务。文中对产生的错误结果进行了分析，发现当前VQA模型的一个关键问题在于缺乏对问题和答案中出现的概念的视觉基础。另外，当前VQA模型并没有明显优于简单利用数据集偏差的模型

1. 总结及理解

* 模型及对应公式如下图所示，是一个二分类模型，可见模型十分简单。输入为图片-问题-回答三元组，三种特征采用串联得形式组合，其中图片特征使用预训练好的卷积网络提取，问题和回答的特征分别取所有词向量和的平均值。输出为{正确，错误}，输出层采用sigmoid函数σ(x) = 1/(1 + exp(−x))来计算二分类结果

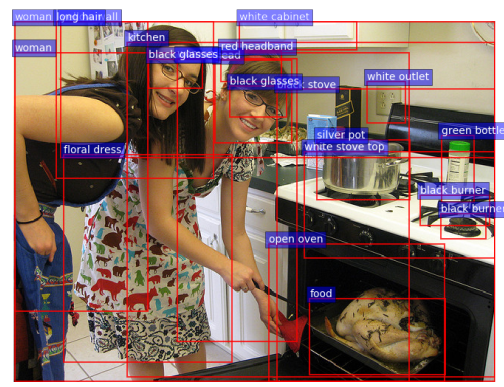
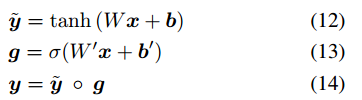


* 本文设计了相关实验对模型表现好坏的原因进行了分析，得出如下结论：
* 将问题作为输入能捕获不同回答相似性
* 不同卷积网络对结果影响较小
* 使用循环神经网络编码问题和回答在不同数据集影响不一样，可能是因为其在较小数据集上容易过拟合
* 迁移学习后微调参数能提升模型
* 文中通过对比输入为A+Q和A+Q+I时的结果，分析了图片的加入对颜色、形状、计数、空间推理、动作、因果关系六种问题影响，从结果来看，本文模型并不能很好理解图片信息
* 疑问
* 模型真的能学会推理吗?能理解where等意思吗？以后在复现论文时注意观察测试结果，观察会不会存在一些不太正常现象，如where的问题回答为数字

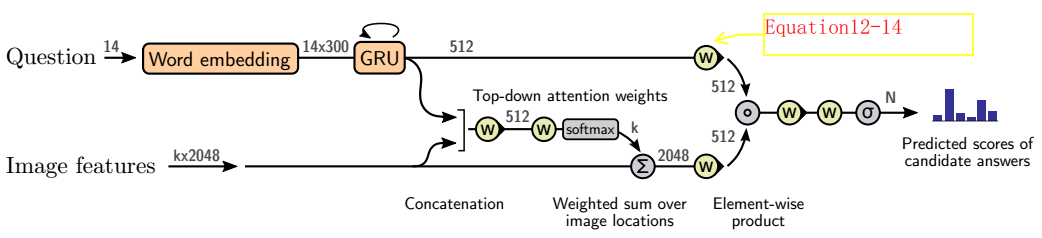
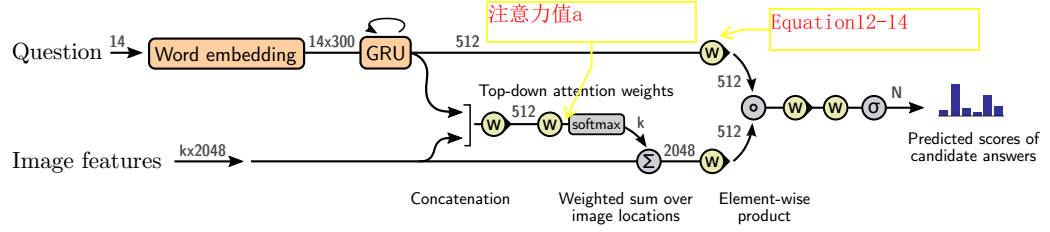
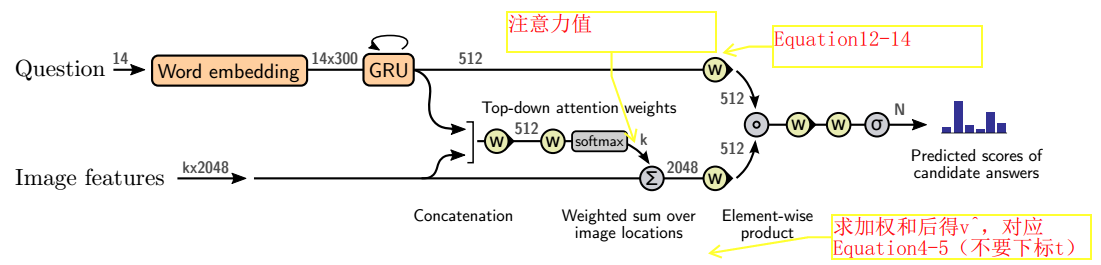
1. 《Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering》总结
2. 内容摘要

本文是2017年VQA比赛中获得冠军的模型，其结合了Bottom-Up Attention和Top-Down Attention，前者通过预训练好CNN（本文用Faster R-CNN）提取图片中的物体和显著区域，分别用相同维度特征向量表示，后者用于确定各个特征向量的权重。它们在图像描述和VQA任务中都有很好表现

1. 总结及理解

* Faster R-CNN bottom-up attention模型包括两部分：Region Proposal Network (RPN)和region of interest (RoI)池化层。第一层提取object proposals，然后通过一个intersection-over-union (IoU)阈值得到RoI输出给第二层，第二层通过池化得到固定尺寸的特征。模型的结构和实现还需阅读文中提到的相关文献进一步理解，模型的输出结果如下图所示，输出包含了物体类别和属性
* VQA模型如下图所示，输出的回答是在训练集中出现过8次以上的。模型多次使用了‘gated tanh’ layers，对应图中带尾巴的W。具体公式如下

另外，将回答作为简单分类，不能很好理解答案语义，能否将回答的词向量以合适的方式加入到模型中提升性能？



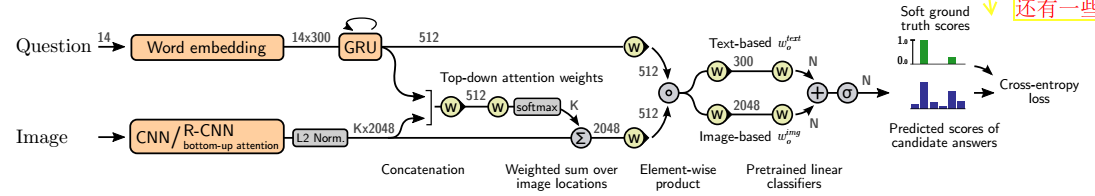
* 默认一般是列向量，加T转置变成行向量。外积：将两向量所有元素分别相乘得矩阵，注意和几何叉积的区别；element-wise product：两张量对应元素相乘；串联各个向量；累加：将所有向量对应元素相加；element-wise sum：两向量对应元素相加

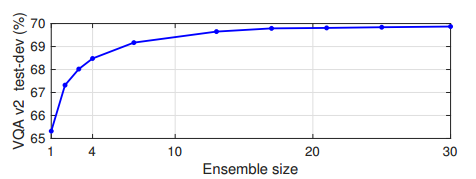
1. 《Tips and Tricks for Visual Question Answering:Learnings from the 2017 Challenge》总结
2. 内容摘要

本文讲述的是2017年VQA比赛的冠军模型，里面提到了很多模型设计、训练测试的一些重要细节，作了大量的对比实验并作了分析，值得以后自己在实验时回过头来细读

1. 总结及理解

* 作者在选择模型和超参数时花费了3000GPU-hours，这说明VQA研究需要花费大量时间和精力，从文后实验可见作者做了大量的对比实验来论证自己观点
* 模型如下图所示，模型中的top-down attention也叫image attention，是典型的question-guided attention，自己对question-guided attention的理解是：question特征和image特征通过一定计算后得到注意力值，然后使用这个注意力值作用于image 特征，因而称为question-guided image attention，同样地，可以将此理解拓展到其他注意力机制的命名。从图可以看出，本文模型与上一篇模型大体一致，只是在模型最后变成两个text-based和image-based分支对应元素相加的形式。整体模型看起来比较简单，但是在比赛中取得的效果是最好的，作者把它归功于模型结构和超参数的选择。包括以下几个方面：
* 用sigmoid output替代常用的softmax output
* 对参考答案选择了柔性化评分，不是一个问题只有一个正确的答案，而是一个问题可以对应1个或多个评分为[0,1]的答案，这也像现实生活中的一些问题，不同的人有不同的解答
* ‘gated tanh’激活函数，从实验结果来看，相对于ReLU、tanh和gated ReLU有0.6-1.5个百分点的提升
* 用bottom-up attention提取图像特征，相对ResNet-200 7×7 features，有3.8个百分点的提升
* 使用预训练好的参数赋给，具体方式还需结合代码进一步理解
* 选择了较大的小批量训练数据，本文选择了一种比较有目的性的打乱方式：让相同问题对应的不同图片和不同答案不要出现在同一个小批量训练数据中。作者目的是为了避免在一次迭代中梯度下降往两个不同方向跑



* 从下图可以看出集成（ensemble）对性能提升的显著影响，本文使用了相同模型，只是采用了不同的初始化参数，最后将各个模型预测分数相加作为预测依据