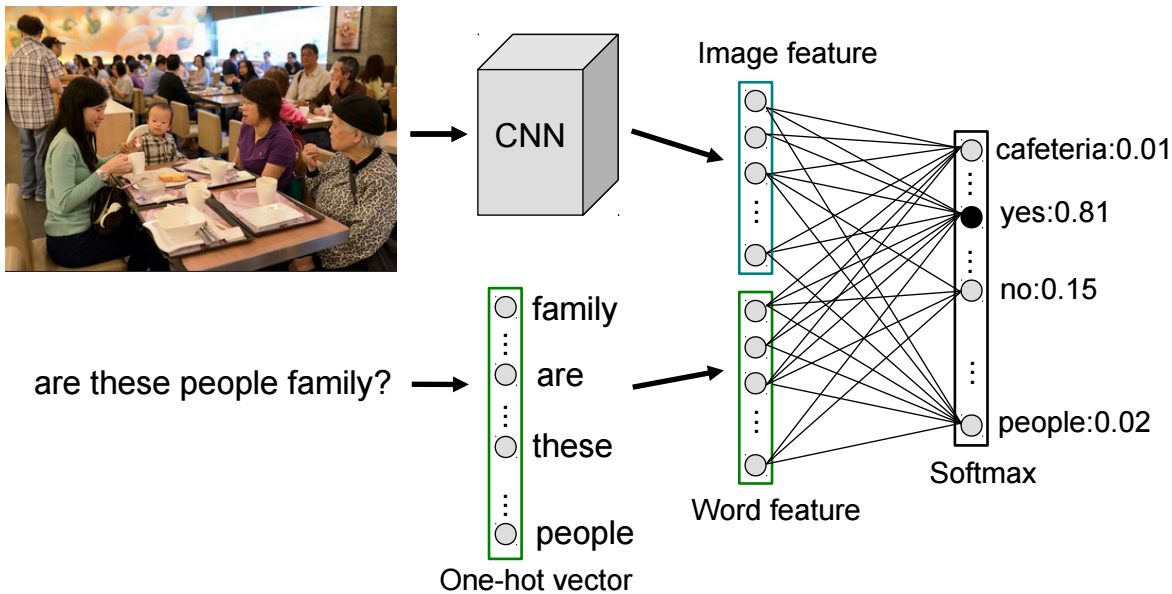
VQA任务中，机器根据给定的一张图和相关问题，回答该问题。 该任务由[Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In NIPS, 2014.]提出。和第六章中传统视觉工作不同的是，VQA需要同时识别多种物体并根据对问题的理解找到对应答案，这对机器的图像理解能力和文本理解能力要求都很高。

**1. 最简单的VQA结构：**

[B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus. Simple baseline for visual question answering.]

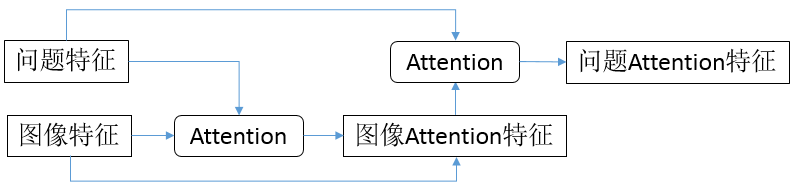


本文中用了最简单的方法实现vqa的答案预测。图像采用CNN（pretrained on ImageNet）, 问题文本采用Bag of words(BOW) 表示，两个特征连接后与输出层全连接，然后过一个softmax直接分类，最后分类的结果（即都是单个词）就是答案。简单暴力有效，调参后在VQA上达到55.88%的准确率。这其中当然可以有一些变种，比如问题文本特征可以用LSTM替代BOW，[Mengye Ren, Ryan Kiros, and Richard S. Zemel. Exploring models and data for image question answering. In NIPS, 2015.]和[Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In ICCV, 2015.]中就是这么干的。

**2. 问题指出图像该看哪里：**

把整幅图的分类特征扔进vqa真的合适么？比如问题“那个穿绿衣服的人头上的帽子是什么颜色的？”需要定位到的是帽子，但图中这个人的占比很大，类似的图在ImageNet中的分类标注为“人”。所以如果用ImageNet的分类任务做pretrain的话，得到的特征也就主要是和人相关的特征了。于是人们想，那我应该让问题来指导对图像的关注点，借助于[Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In ICML, 2015.]中提的图像中加soft attention的概念，有些文章也在vqa中加起了图像attention，即question-guided attention（往后看，第三点里有简单说明）。回忆一下soft attention的做法，大体来讲有三步：1. 将图像分块，求每块的空间特征（通常取CNN层）2. 分析文本重点 3. 获得文本与各图像块的关联程度（即attention weight）。不同的是，原文中做image caption任务的文本是caption描述，而在vqa任务中文本是给出的问题。类似的想法，[Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In CVPR, 2016.]受Memory Network（在第七章讲过）启发，用两个连续attention模块进行多步attention推断。[Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In ICML, 2016.]在动态attention网络中加入多步attention推断。[Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In CVPR, 2016.]先通过Edge box提出图片中的object proposal, 然后求各proposal的attention weight。

**3. 图像&文本一起加关注点：**



图像和文本一起加attention的主要思路如上图所示。首先读完整句问题，得到问题的关注点，指导模型获得图像各区域的attention weight，进而得到图像的attention特征（这是左边这个attention做的事）；类似地，再用带有attention的图像特征指导问题，获得更有针对性的问题文本attention特征。整个过程叫做co-Attention，其结果比单独的图像部分attention要好很多，基于co-Attention的思路, Berkeley和Sony合作的文章[Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding]在VQA数据集上达到66.9%的准确率，[Jiasen Lu, Jianwei Yang, Dhruv Batra, Devi Parikh. Hierarchical Question-Image Co-Attention for visual question answering]达到62.23%的准确率。具体文章里怎么做的细节大家感兴趣自己回去看，很有意思的两篇paper，这里我们只讲基于co-Attention的最简单实现。对于每个Attention模块，有两个输入：X和g，g指导X获得X的attention。以上图中左边的attention为例，X为图像特征，g为问题特征，

Question-Answer Embedding feature:

Attention weights：

Question-guided Image feateure:

假设每幅图像分成了N个块，每块的CNN层特征有D维（如ResNet中取最后一层卷积层“res5c”的特征），即。文本的输入特征q也映射到每个词D维，共T个词，通过时序模块（比如过一个RNN）得到固定维度D的文本特征即g 。上面公式中的三步

**第一步**，通过对x和g做非线性变换得到H，其中, , 之所以对文本特征又乘了一个**1**向量是为了保证文本和图像映射到统一维度,其中。

**第二步**，Attention weight 的计算方法和[show, attend and tell]那篇文章中的做法一样，将隐层通过一个非线性变换映射到一个N维向量, 对应到图像的N块。

**第三步**，由question引导的图像attention特征为N个块的加权和，（或者也可以通过连接concat将这些加权特征连接起来，这样）。

具体实现时， 由于是每隔x的D个位置做一次D维按位相乘，共乘N次，所以实际可以用卷积层实现，用K个卷积核，每个卷积核的width = stride = D。类似地，也可以由卷积操作实现。中与**1**相乘的部分可以用外积实现。

需要注意的是，图像和文本的co-Attention也并不是其突出效果的唯一原因。在机器翻译文章[Effective Approaches to Attention-based Neural Machine Translation]中比较了soft attention中embedding feature的两种获得方法，dot () 比concatenate的方法 () 强。PS: concatenate顾名思义就是把x和g拼起来，其实就是x和g分别与下一层连接，输出是两个连接起来的加和。类似地，[Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding]中也采用dot代替了concatenate，并通过把bilinear变换看作多项式核简化了复杂度。[Hierarchical Question-Image Co-Attention for visual question answering]中除了co-attention还采用了层级的文本特征，分别在文本的word-level, phrase-level, sentence-level分别计算和图像之间的alignment。而同样是采用了co-Attention, [Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering]在VQA数据集上的准确率却只有57.99%，其在attention实现细节（比如文本特征的组织，x与g进行dot后的降维处理）上与上面两篇文章都有所不同，可见细节的重要性。

**4. 让问题决定看几步**

[Training Recurrent Answering Units with Joint Loss Minimization for VQA]中说，每个问题的难度都不同，比如“苹果是什么颜色的？”和“桌上的苹果是什么颜色的”需要查找的步骤就不同，那我能不能通过一个RNN让机器自己决定要走几步呢？ 于是文中给思考过程加了一个RNN，RNN中每个timestep都留一个自带attention的答案单元，这些单元共享模型参数，最后的loss是所有timestep的loss的和。由于在分析每步给出的answer时发现，步数越多越容易在简单问题上过拟合，所以只要第一个timestep的答案单元饱和了，就停止学习。

此外，还有一些非主流方法，也达到了不错的效果。[Neural Module Networks]人工定义了几个问题中的关注点（如寻找类、数数类等），并给每种关注点人工定义了子网络模块，整个句子通过一个language parser(语言分析器)选择整个网络的结构，其在VQA数据集上的准确率达到59.43%。[Dynamic memory networks for visual and textual question answering] 为动态记忆网络引入了一个新的输入融合模块, 该模块从周围patches提取图像空间信息，然后从一个attention GRU模块中寻找答案。

**不得不看：**

<Co-attention>

1. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding
2. Hierarchical Question-Image Co-Attention for visual question answering
3. Show, attend and tell: Neural image caption generation with visual attention