# 计算机视觉课程报告（VQA论文介绍）

陈井爽
21821207
计算机科学与技术学院
21821207@zju.edu.cn

## 摘要（Abstract）

*视觉问答（VQA）是一门交叉学科的研究方向，是一项同时涉及了计算机视觉和自然语言处理的新任务。本文我们介绍一片发表在CVPR2018 的论文 Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering，从而由点及面来了解VQA研究中的相关方法和动向。*

### 1. 背景与意义

随着深度学习等人工智能算法的快速发展，计算机视觉和自然语言处理等领域也在不断的进步。视觉问答（VQA）是一门交叉学科的研究方向，是一项同时涉及了计算机视觉和自然语言处理的新任务，正在逐渐成为一个越来越热门的研究方向。视觉问答不仅涉及到感知层面的感知，还结合了理解层面的自然语言，而将理解语义和感知图像整合到一个统一的框架中是一个相当大的挑战。所以作为一个新兴的研究方向，视觉问答给我们带来的机遇和挑战并存，也是实现强人工智能，通过图灵测试的一个重要途径，具有非常好的发展前景。

### 1.1. 什么是VQA

简单来说，视觉问答（VQA）可以定义为：给定一张图像和与其相关的使用自然语言描述的问题，VQA的目标就是根据图像给出问题的自然语言答案[1]。举个例子，如图 1 所示，给定一张图片，使用自然语言给出一个问题：“What room are they in?”，通过VQA系统，输入这个图像及对应的问题，然后给出这个问题的答案“kitchen”。对于计算机而言，需要处理并理解图像的内容，然后理解所提的问题的语义，进行感知理解分析，从而给出问题对应的答案。计算机输出的答案不仅要合理，而且要符合自然语言的规则和习惯。由此可见，VQA的研究需要涉及计算机视觉，自然语言处理，知识表示和推理等多个领域的知识，并将这些知识合理的结合，才能设计出一个合理的VQA系统。

Question: What room are they in? Answer: kitchen

图 1 VQA示例

### 1.2. 相关数据集介绍

目前VQA研究中的主流方法是深度学习，与计算机视觉和自然语言处理的任务类似，一个特点是需要大量的数据。因此，构建好的数据集是非常关键的。在过去的研究中，VQA任务已经发布了几个大规模的数据集[2]。我们将在下面讨论这些数据集。数据集见表 1。

表 1. VQA相关数据集

| | Number of Images | Number of Questions | Avg. questions per image | Avg. question length | Avg. questiong length |
|---|---|---|---|---|---|
| **DAQUAR** | 1449 | 12468 | 8.60 | 11.5 | 1.2 |
| **Visual7W** | 47300 | 327939 | 6.93 | 6.9 | 2.0 |
| **Visual Madlibs** | 10738 | 360001 | 33.52 | 4.9 | 2.8 |
| **COCO-QA** | 117684 | 117684 | 1.00 | 9.65 | 1.0 |
| **FM-IQA** | 158392 | 316193 | 1.99 | 7.38 | 3.82 |
| **VQA（COCO)** | 204721 | 614163 | 3.00 | 6.2 | 1.1 |
| **VQA(Abstract)** | 50000 | 150000 | 3.00 | 6.2 | 1.1 |

VQA第一个重要的数据集是DAQUAR。DAQUAR包括 6794 对训练问答题，以及 5674 对测试问答题，其中的图片都来自NYU-Depth V2 数据集。但它只包括室内场景，且因为数据量较少，不适合用作复杂模型的训练和评估。

COCO-QA数据集有 123287 张来自COCO数据集的图像，其中 78736 张用于训练，38948 张用于测试。问题使用算法从COCO的图像标注中自动生成，且该数

据集只有四种问题，而且分布得不均匀。

和其他数据集相比，VQA数据集除了从COCO数据集中获取的 204721 张图片，它还包括 50000 张抽象的卡通图像。每张图有三个问题和十个答案，即总共有 76 万个问题和大约 1000 万个答案。

### 1.3. 相关方法介绍

深度学习强大的特征学习能力在计算机视觉和自然语言处理领域的得到了广泛的应用。在计算机视觉中， CNN可以端到端的学习图像特征而不依赖手工设计的特征，通过逐层的特征抽取，将图像从简单的边缘、角点等底层特征，逐层组合成为更高层次的特征。在自然语言处理中，RNN，Attention机制等取得了很大的成功。对于VQA来说，一个直观的想法是将两个领域的模型相结合，事实上，由深度CNN和LSTM网络结构组合而成的VQA模型是目前来说确实是在视觉问答中性能相对较好的模型。

Aishwarya Agrawal 等人提出Deeper LSTM Q + norm I模型，使用CNN抽取图像语义信息，使用LSTM抽取问题中包含的文本语义信息，将两者的信息融合，让模型学习到问题的含义，最后送入一个带有Softmax的全连接中产生答案输出。Mateusz Malinowski 等人提出Neural-Image-QA模型，以CNN和LSTM为基础，以一种新的使用方式，设计了一个预测结果长度可变的模型。

## 2. 论文介绍

这里我们介绍一篇论文：Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering[3]。论文发表在CVPR2018，论文的作者也是2017 年VQA Challenge的第一名。论文提出了一种自上而下和自下而上相结合，使用注意力模型方法，应用于视觉场景理解和视觉问答系统的相关问题。其中，基于自下而上的关注模型（一般使用Faster R-CNN）用于提取图像中的兴趣区域，获取对象特征；而基于自上而下的注意力模型用于学习特征所对应的权重（一般使用LSTM），以此实现对视觉图像的深入理解。

### 2.1. 主要方法

#### 2.1.1 Bottom-Up Attention

论文中使用Faster R-CNN来实现基于自下而上的注意力模型。论文中的方法首先初始化Faster R-CNN结合ResNet-101 CNN在ImageNet上进行预训练，接着在Visual Genome数据上进行训练。为了生成一组用于图像描述或VQA的图像特征，论文中获取模型的最终输出，并使用IoU阈值对每个对象执行非最大抑制。然后选择所有类检测概率超过置信阈值的区域。通过设定阈值，从而允许兴趣框的重叠，Faster R-CNN有效地发挥了hard attention的作用，这样可以更有效的理解图像内容。文中对每一个感兴趣区域，不仅使用对象检测器，还使用属性分类器，这样可以获得对对象的（属性，对象）的二元描述，从而更加贴合实际应用。

#### 2.1.2 Top-Down Attention LSTM

论文中使用两层LSTM模型，一层用于实现自上而下的注意力，一层实现语言模型。对于上一层得到的特征向量，第一层LSTM用来确定特征向量的对应权重，这是一种soft attention机制，第二层LSTM则实现语言模型，对于上一层的加权特征，捕获某一时刻任一单词出现的概率，从而得到整个句子的概率分布。论文中首先使用的是最小化交叉熵进行训练，此外，还用到了SCST中的强化学习方法来对CIDEr分数进行优化。模型如图 2 所示。

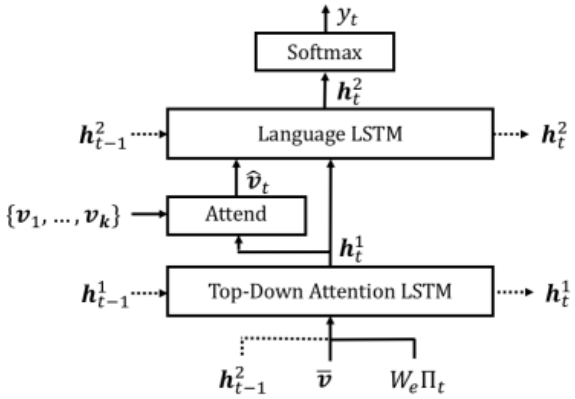


图 2 捕获模型

#### 2.1.3 VQA模型

对于给定的特征向量，论文中提出的VQA模型也使用了一种soft top-down注意力机制来确定权重。对问题和之前提取的图像特征进行联合多模式embedding，对于网络内的非线性变换，这里是用gated hyperbolic tangent activation来实现的，VQA中的问题Q被编码成GRU内的隐状态q，其中单词是以可训练的词向量来表达的，最后得到问题和图像的联合表示，来生成最终的答案。模型如图 3 所示。

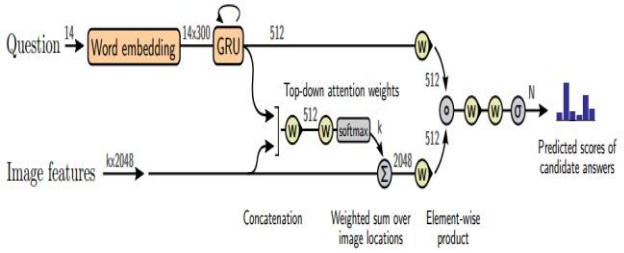


图 3 VQA模型

2.2. 实验结果

论文中的模型在VQA 2.0 数据集上进行训练和验证，如表 2 所示。在Up-Down模型中使用自底向上的注意力比各种类型的最佳ResNet baseline都有显著的改进。另外，论文中的模型比其他的集成模型也表现出了最好的效果，并取得了 2017 年VQA Challenge的第一名。除了VQA，模型也在Image Captioning上取得了很好的效果。

表 2 VQA 2.0 数据集实验结果

| | Yes/No | Number | Other | Overall |
|---|---|---|---|---|
| Ours: ResNet (1×1) | 76.0 | 36.5 | 46.8 | 56.3 |
| Ours: ResNet (14×14) | 76.6 | 36.2 | 49.5 | 57.9 |
| Ours: ResNet (7×7) | 77.6 | 37.7 | 51.5 | 59.4 |
| Ours: Up-Down | **80.3** | **42.8** | **55.8** | **63.2** |
| Relative Improvement | 3% | 14% | 8% | 6% |

| | Yes/No | Number | Other | Overall |
|---|---|---|---|---|
| d-LSTM+n-I [26, 12] | 73.46 | 35.18 | 41.83 | 54.22 |
| MCB [11, 12] | 78.82 | 38.28 | 53.36 | 62.27 |
| UPMC-LIP6 | 82.07 | 41.06 | 57.12 | 65.71 |
| Athena | 82.50 | 44.19 | 59.97 | 67.59 |
| HDU-USYD-UNCC | 84.50 | 45.39 | 59.01 | 68.09 |
| Ours: Up-Down | **86.60** | **48.64** | **61.15** | **70.34** |

2.3. 创新点

本文主要的贡献在于提出了Bottom-Up and Top-Down Attention的机制，使用Fast R-CNN检测图像中的特定目标从而可以更好理解图像语义，另外使用feature vector，包含空间信息的同时还可以对应多个单词，比如一个形容词和名词，表现力更丰富了。

3. 总结

总的来说，VQA是一个较新的研究方向，结合了计算机视觉和自然语言处理的相关技术，也是一个非常具有挑战性的方向，相关研究成果层出不穷，但是相较于计算机视觉和自然语言处理，成熟的研究成果较少，是一个值得深究的研究方向。

References


[1] Wu Q, Teney D, Wang P, et al. Visual question answering: a survey of methods and datasets[J]. Computer Vision and Image Understanding, 2017: 21-40.

[2] Gupta A K. Survey of Visual Question Answering: Datasets and Techniques[J]. arXiv: Computation and Language, 2017.

[3] Anderson P , He X , Buehler C , et al. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering[J]. 2017.