**课程大作业要求**

**1. 任务： Visual Question Answering**

以一张图片和一个关于图片内容的自然语言形式的问题作为输入，要求输出正确答案。

参考文献：

1）VQA: Visual Question Answering (ICCV 2015)

2）Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering (CVPR 2017)

**2. 数据集**

train: 44375

validation: 21435

test: 21435

下载地址：

<https://drive.google.com/open?id=1_VvBqqxPW_5HQxE6alZ7_-SGwbEt2_zn>

**3. 评分标准**

报告评分标准（25分）

1. 按时按要求提交报告、代码，无抄袭（10）

2. 报告篇幅与排版（2）

3. 研究内容、任务中的挑战有清晰的描述（3）

4. 算法、理论知识、技术细节表述清晰，有必要图示、伪代码、公式（5）

5. 实验内容、实验步骤和评价指标表述清晰，对实验结果（方法优越性和局限性）有详细分析（5）

以个人为单位录制视频对大作业成果进行展示，限时10分钟，内容覆盖：

1. 算法：对使用的模型和算法进行介绍。

2. 主要创新点：相比于以往的工作，有哪些创新点，或是针对问题本身，做了哪些建模？

3. 实验结果： 实验结果展示与分析。

视频展示评分标准（10分）

1. 展示内容条理清晰，对算法、设计思路描述清晰（6分）

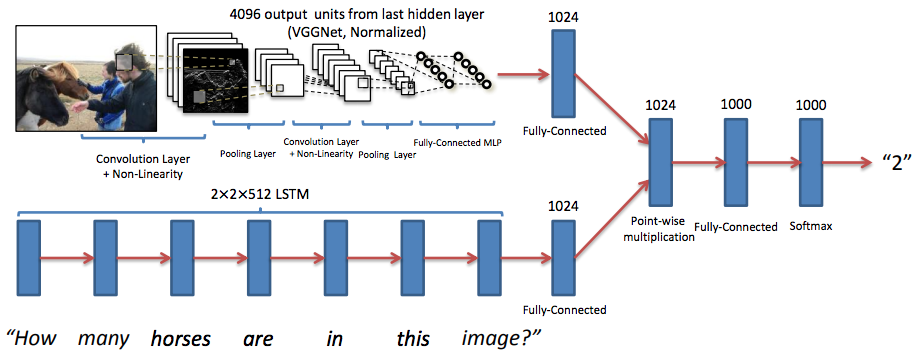
2. 创新点，针对问题本身，有做进一步分析与建模（4分）

**4. 参考方法**

4.1 2-channel vision-language model

VQA: Visual Question Answering (ICCV 2015)

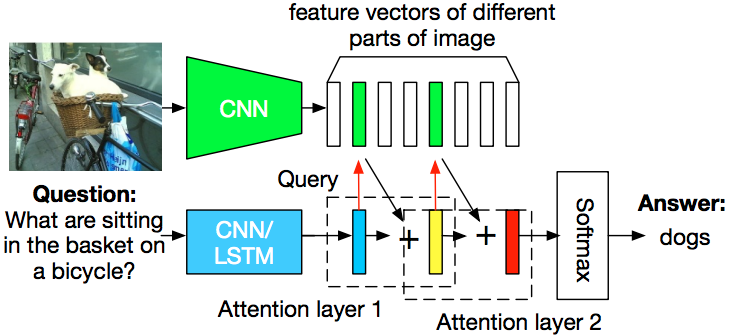
基线方法。分别采用CNN和RNN提取图像和问题的特征，再将视觉模态和文本模态的特征进行元素积，得到最终特征，经过softmax层输出正确答案概率。



4.2 Stacked Attention Networks

Stacked attention networks for image question answering (CVPR 2016)

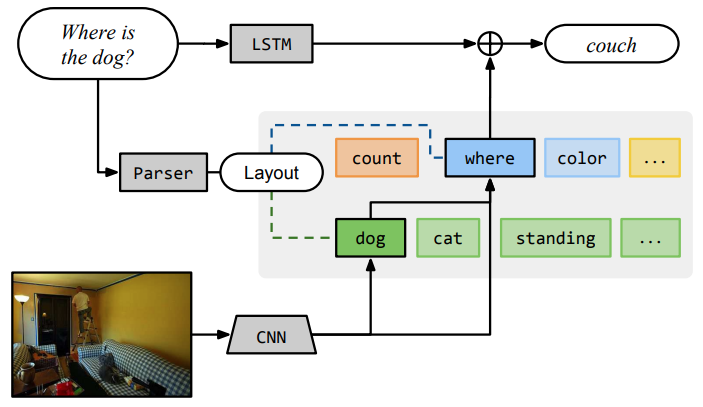
将注意力机制引入模型，由问题引导，关注图像中与问题相关的部分。



4.3 Neural Module Network

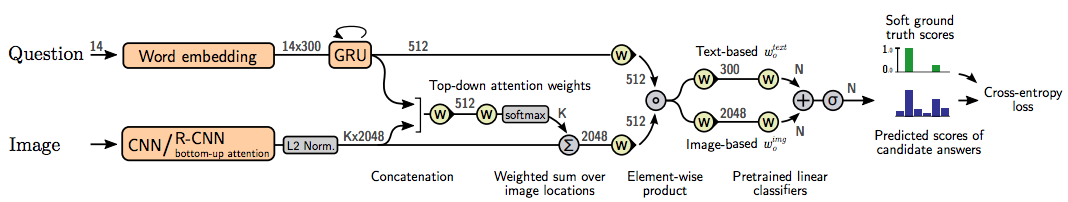
Neural Module Network (CVPR 2016)

模型由多个模块构成，根据VQA数据集中每个question类型设计特定的网络模型。



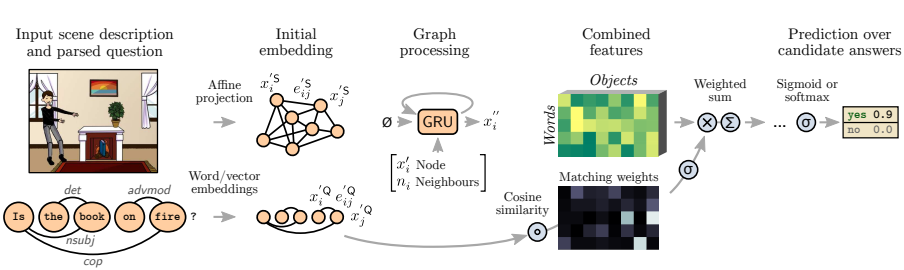
4.4 Faster-RCNN + Glove Vectors + Bottom-Up Top-Down attention

Tips and tricks for visual question answering: Learnings from the 2017 challenge （CVPR 2018）

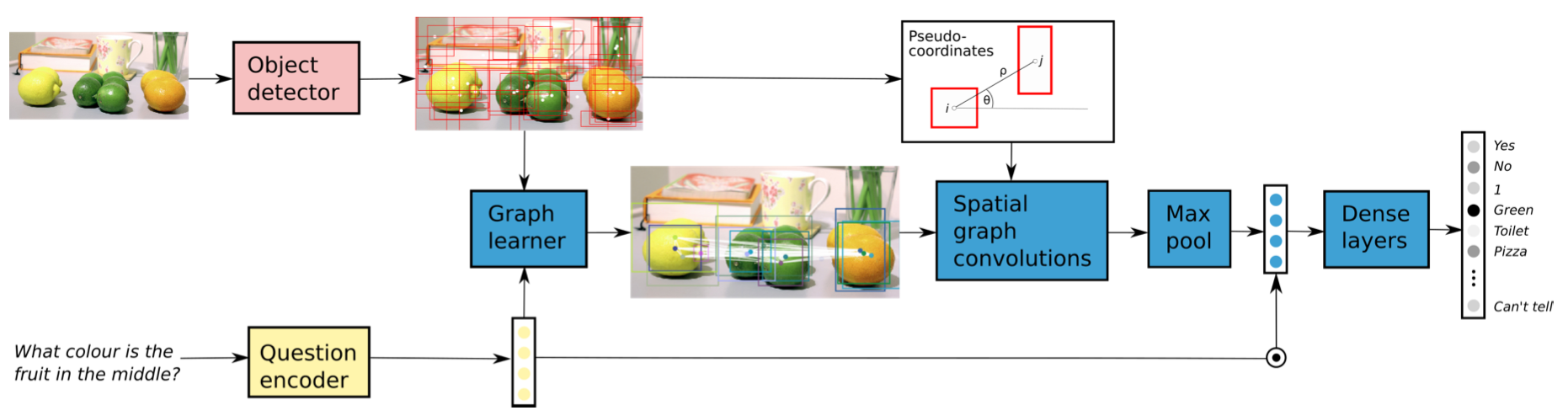


4.5 Graph representation

Graph-Structured Representations for Visual Question Answering (CVPR 2017)

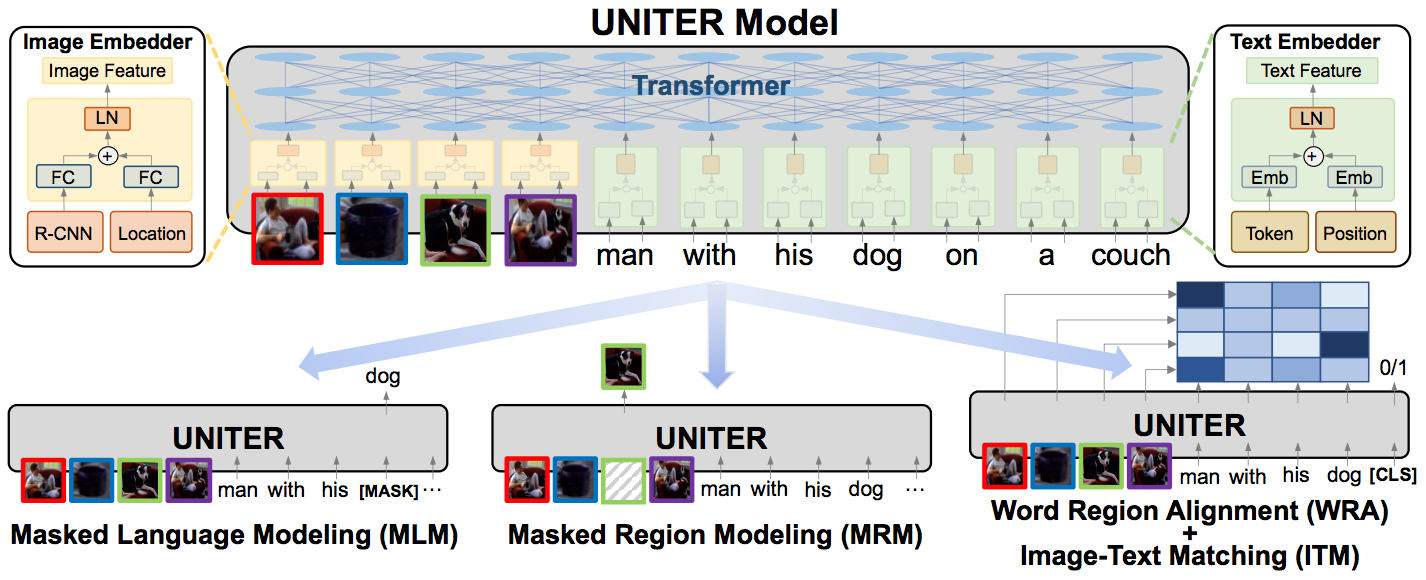


Learning Conditioned Graph Structures for Interpretable Visual Question Answering (NIPS 2018)



4.6 vision-language pre-training

UNITER: UNiversal Image-Text Representation Learning (ECCV 2020)



4.7 其他资料

1) https://github.com/KaihuaTang/VQA2.0-Recent-Approachs-2018.pytorch

2) <https://github.com/SinghJasdeep/Attention-on-Attention-for-VQA>

3) https://github.com/noagarcia/awesome-vqa-pytorch