题目

1.模型总体结构

给图像生成一组视觉特征以及一组编码region-grounded captions，RGC捕获对象属性及其关系。使用视觉特征和编码的RGC从输入模块构造两个图网络，GN的每个节点迭代地计算分配给它的视觉和文本信息来引导上下文表示。为了合并两个GN的信息，节点讲更新的表示形式写入外部空间存储器。

2.模型实现细节

图像进来用 Bottom-Up Attention模型，使用固定的阈值对象检测，从图像的N个不同区域中提取了2048维图像特征。用 dense captioning model为图像生成captions。每个RGC都有一个标题，一个边界框和一个置信度分数。

作者使用与字幕相同的字典对问题进行编码。 这使得模型能够将标题中的单词与问题中的单词匹配，并关注相关标题。

模型输入一组视觉特征向量、一组编码RGC和问题encoder。存储器使用注意力机制选择输入的哪些部分重点关注，它利用输入图像区域的文本和视觉信息建模成Graph Network，并利用GN进行图像文字特征进行逐对交互。

视觉GN的每个节点表示视觉特征和其关联的bounding-box；文本GN的每个节点也有bounding-box对应于检测到的图像的RGC。再两个GN中，如果两个节点的边界框的归一化中心之间的欧氏距离小于0.5则将其连接。

外部存储器是P*Q的存储网格，每个单元格都有固定位置，对应于图像中的(H/P)×(W/Q)的区域，H/W是图像的高度和宽度。如果GN的没每个边界框都覆盖了存储单元的位置，则节点会将其信息发送到存储单元。一个单元可能会从多个节点获取信息，外部存储网络负责汇总来自两个GN的信息，并消除重叠的边界框带来的冗余。

存储器聚合公式：

$$\overline{\mathbf{m}}_{p,q} = f\left(\mathbf{m}_{p-1,q}, \mathbf{m}_{p,q-1}, \mathbf{m}_{p,q+1}, \mathbf{m}_{p+1,q}\right)$$
$$\mathbf{m}'_{p,q} = \mathrm{GRU}\left(\left[\overline{\mathbf{v}}_{p,q}, \overline{\mathbf{v}}_{p,q}, \overline{\mathbf{m}}_{p,q}\right], \mathbf{m}_{p,q}\right)$$

答案预测模块也用GraphNetwork，节点表示就是是外部内存的每个存储单元的表示，每个节点间都有边，边由两两计算成独热码表示，然后更新节点属性和全局属性，最后根据全局属性来预测结果。

3.思考总结

建模方式比较新颖，但是最后的答案预测模块有些让人摸不到头脑，感觉可以在预测时加上attention。