第10周汇报

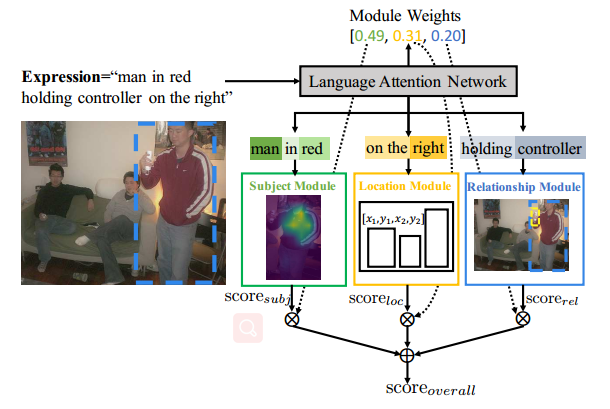
——张溢炉

1. 《MAttNet: Modular Attention Network for Referring Expression Comprehension》总结
2. 内容摘要

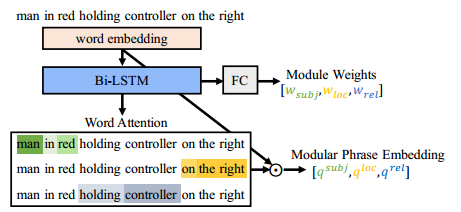
本文提出了一种多模块的模型来解决指代表述理解问题。

1. 总结及理解

* 动机：把指代表述和图片分成主体、位置、关系三个部分能够更好的捕获语言和图片中的信息
* Referring Expression Comprehension曾用模型：CNN-LSTM based frameworks、joint vision-language embedding frameworks、混合前两种模型的方法。
* 模型如下图所示，输入是图片和图中某个物体的指代表述，输出是图中物体与指代表述的匹配分数。Language Attention Network将指代表述分成subject、location、relationship三个部分，subject用于捕获主体，包括边界框内所有物体，location主要用于捕获主体位置，relationship用于捕获主体与边界框外物体间关系。Visual Modules用于提出图片中特征并得到与指代表述各个部分的匹配分数，最后分别乘module权重得到总分



* Language Attention Network：模型如下图，模型输出module weights用于最后求取匹配总分用，modular phrase embedding输出给visual module作匹配用



* 在两者关系或匹配分数不能直接算出时，可以考虑把它俩同时作为神经网络输入，通过训练来求得这种隐性关系，如文中用MLP求取匹配分数
* 物体的7种属性：类名、颜色、尺寸、绝对位置、相对位置、相关物体和通用属性

1. 《Neighbourhood Watch: Referring Expression Comprehension via Language-guided Graph Attention Networks》总结
2. 内容摘要

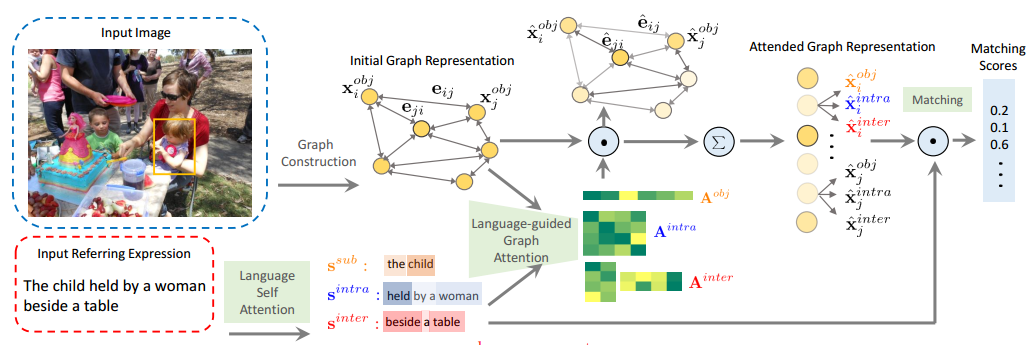
本文提出了一种用于理解指代表述的基于语言指导的图注意力模型，上一篇是本文的baseline，本文模型与上一篇模型有些相似与不同之处

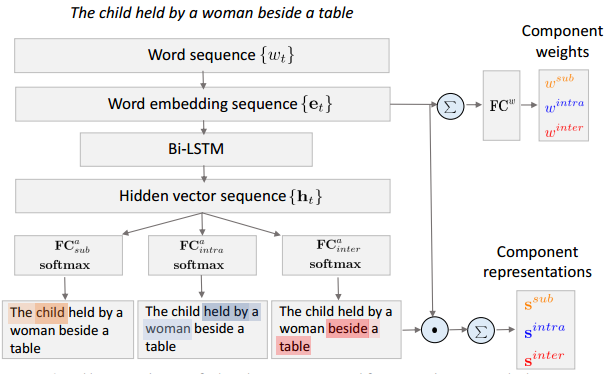
1. 总结及理解

* 动机：以前模型都未综合利用图片中物体的属性及物体之间关系，也没利用好图片与描述语言之间的对应关系。从本文模型能直观的体现出利用了这些关系
* 模型如下图所示，输入是图片和图中某个物体的指代表述，输出是图中物体与指代表述的匹配分数。图中的物体是使用预训练好的VGG16识别出来的。

模型包括语言自注意力机制、语言指导图注意力机制、匹配模块三部分。

* 语言自注意力机制将指代表述分成主体、类内关系、类间关系三部分
* 语言指导图注意力机制根据语言自注意力机制输出的三部分，使与指代表述对应的通过VGG16提取出输入图片中的物体以及类内、类间关系图部分高亮
* 匹配模块就是将经处理后的图像信息与指代表述信息使用点乘匹配，得到各个物体的匹配分数

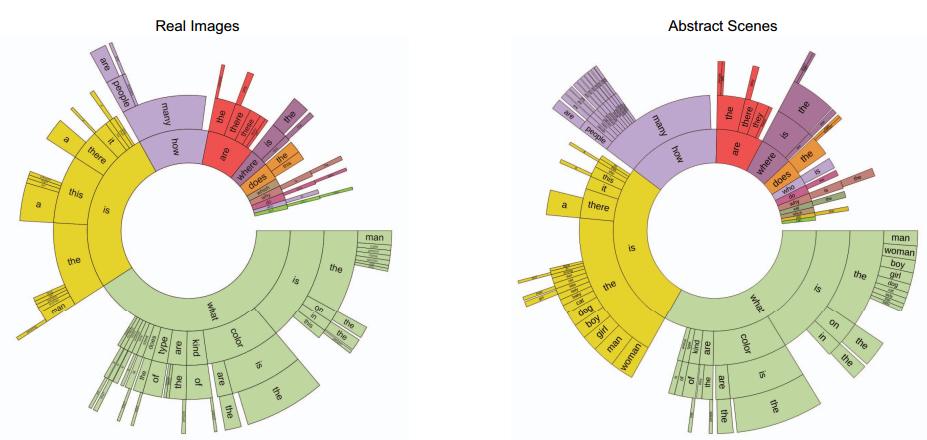


* Language Self-Attention Module：模型如下图
* 动机：为了弥补LSTM等简单编码方式未充分利用结构信息，该模块将语句信息划分成了三部分，充分利用物体属性及类内、类间关系
* 疑问：该模块是如何训练的？是否能捕获大多数指代表述中的这三种关系？
* Language-guided Graph Attention Module：先通过卷积处理的方法得到物体的属性V和位置信息l。然后使用Language Self-Attention Module输出s\_sub，s\_intra和s\_inter得到图中各部分对应注意力值。最后将原图与注意力值相乘得到加权后的图结构
* Matching Module：计算图结构与指代表述的匹配分数
* 与上一篇baseline的相同与区别
* 两篇文中都利用模块化思想来利用主体属性、位置、物间关系
* 上一篇把指代表述分为了主体、位置、关系三部分，而本文将指代表述分为了主体、类内关系、类间关系，而把位置直接作为物体的特征，与卷积网络提取的特征串联，从结果上看，本文方法优于上一篇
* 从模型上看，上一篇在visual module中，只有提取主体特征时利用了指代表述来指导提取，而本文中三个部分都利用了指代表述来指导提取

1. 《VQA: Visual Question Answering》总结
2. 内容摘要

本文2015一篇比较系统阐述VQA任务的文章，发布在arXiv上的原文包含更多内容。提供了VQA数据集并作了较全面的分析

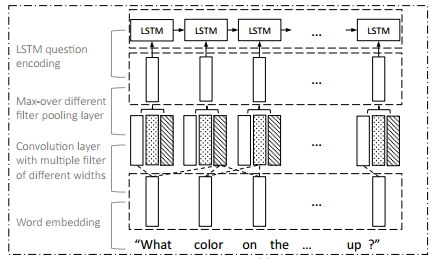
1. 总结及理解

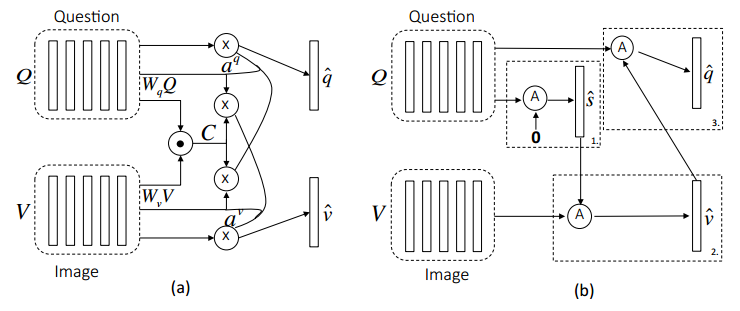
* 动机：相对于图像描述，VQA可以更好的理解图片，结合外部知识回答相关问题。VQA可以帮助视觉障碍人士和小孩更好的理解真实世界
* 数据集包括Real Images dataset（MS COCO dataset）和abstract scenes dataset，其中包括对应的开放性和多选项问题-回答对，以及对图片的描述。下图展示了问题的种类，通过前四个词来判断问题种类，从图中可看出各种问题的大概分布以及问题的多样性，而1-3个词的回答在Real Images dataset（MS COCO dataset）和abstract scenes dataset中分别为89.32%, 6.91%, 2.74%和90.51%, 5.89%, 2.49%
* 文中展示了一些简单的模型的测试结果作为baseline，准确率较低，还有很大的提升空间
* 疑问：
* 怎么让模型发现自己回答错了？用户反馈或自我判断？
* 发现之后如何调整参数？反向传播或者使用强化学习？

1. 《Hierarchical Question-Image Co-Attention for Visual Question Answering》总结
2. 内容摘要

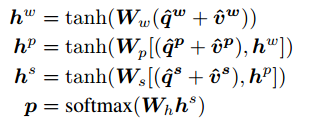
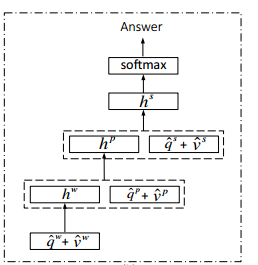
对于VQA任务，本文提出了Hierarchical Question-Image Co-Attention，对图片、问题都采用了注意力机制

1. 总结及理解

* 动机：以前模型只采用了图片注意力机制，为了更有效利用问题，本文模型不仅使用了图片注意力，还使用了问题注意力机制
* 模型输入图片和问题，输出为训练、验证集中TOP1000的回答，覆盖了86.54%的回答，可看成一个分类问题，模型可分成如下三部分
* 模型将问题分为词级、短语级、句子级三种情况，三种情况经过co-attention独立计算结果q^、v^。词级用词向量表示，短语级采用如下图1-D的CNN（unigram、bigram、trigram）模型编码而得，句子级用LSTM编码而得。短语级模型设计比较新颖，但是从消融实验结果来看，其对性能提升影响不大
* co-attention：其有两种方案（Parallel co-attention mechanism，Alternating co-attention mechanism），分别如下图所示，在本文实验中，前者表现优于后者。具体公式看原文，（b）公式中，先将q、0赋给X、g得到s^，再将v、s^赋给X、g得到v^,最后将q、v^赋给X、g得到q^



* 最后通过如下模型和公式综合计算三种问题分级情况得到分类结果



* 数据集：VQA dataset、COCO-QA
* 待解决问题：
* 学习VGG、Resnet等卷积神经网络