第5周汇报

——张溢炉

1. 《Multi-Target Embodied Question Answering》总结
2. Background

人工智能的一大挑战是建立智能agent，通过视觉感知周围环境，通过自然语言与人类交流，并在所处环境中行动以完成任务。Embodied Question Answering (EQA)是一个与之相关比较新的任务，它的任务是通过agent观察周围环境的第一视角来回答固定形式的问题，提问的内容一般是对象的属性，如“what color is the car?”，在这之前主要是只包含单个目标的EQA。

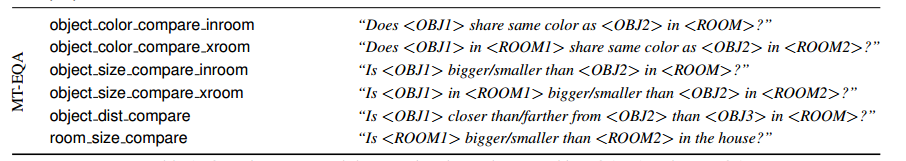
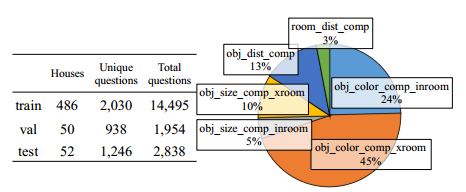
1. Motivation

之前单个目标的EQA会限制agent的能力，故本文提出了Multi-Target EQA(MT-EQA)，在其问题中包含多个目标，可以用来比较相同房间或不同房间中的两个目标的大小、颜色以及距离，如“Is the dresser in the bedroom bigger than the oven in the kitchen?”。

1. 3D environments

文中使用的仿真环境是House3D，另外还有ViZDoom，TorchCraft，DeepMind Lab，HoME，MINOS，Gibson，AI2THOR。

1. Dataset

文中所用数据集问题类型及各类型所占比例分别如下图，其中包括了6种类型的问题。数据集包含588个环境中的19287个问题，涉及8个房间中的61种对象。在设计集过程中用到一个比较重要的指标Intersection-Over Union (IOU)来选择观察一个对象的最佳位置。

1. Model

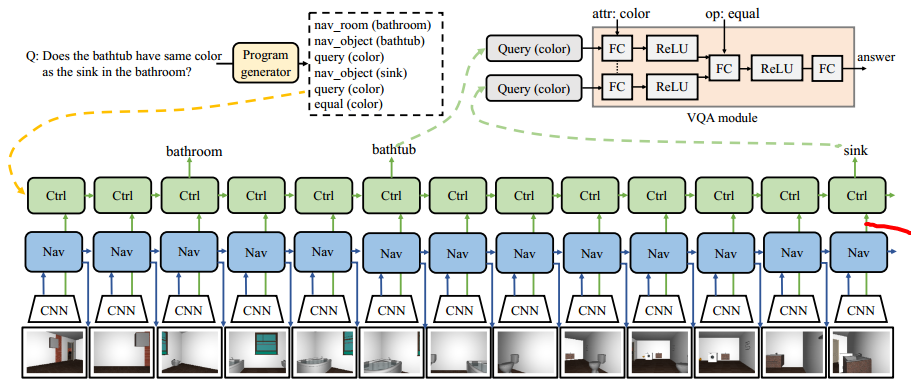
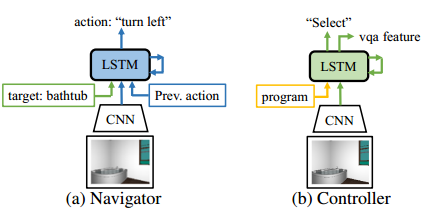
模型如下图，其采用《Visual question answering A survey of methods and datasets》

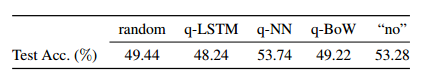
中提到的组合模型（Compositional models）的方法，包括question-to-program generator, navigator, controller, VQA module四个模块，program generator为下图的左上部分，VQA module为下图的右上部分，navigator, controller在图中为缩写形式。

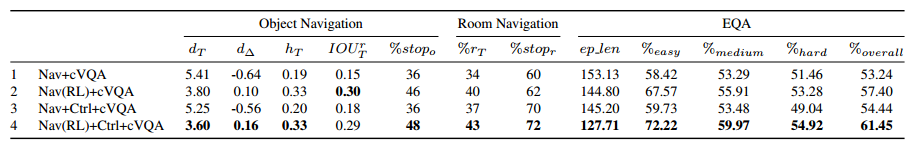
由图可以发现program generator和VQA module作为输入、输出模块相对独立，program generator将question分解为几个子程序，VQA module则用于比较controller两个输出属性产生answer。navigator, controller模块详图分别如图（a）和图（b）。controller比较当前第一视角图像（经CNN处理后）与子程序中所要求图像进行比对，判断是否到达目标房间或找到目标，逐步完成子程序所要求的任务，并给navigator模块提供当前子程序的目标。Navigator模块则根据controller输出的target、当前第一视角图像（经CNN处理后）以及上一个动作产生相应的动作，寻找controller给定target。

过程中，寻找房间和物体的过程是不一样的，房间是通过12张全景图来分辨的，而物体是通过语义信息来分辨的，CNN是预训练好的。另外，navigator的动作为左转30度、右转30度和前进。作者发现较EQA-v1转角的增大，减少的动作的次数，使得agent更容易训练。

对于整个模型的训练过程，作者没有细说，只是简单提到可以使用了模仿学习（Imitation Learning IL)或强化学习（RL）

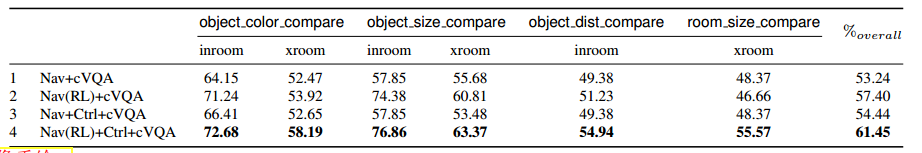


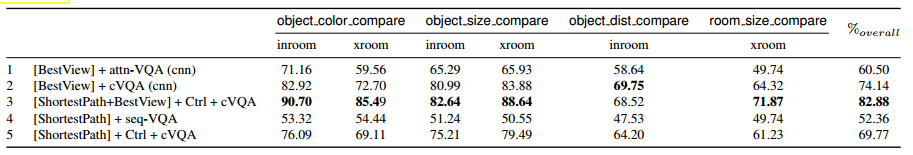
1. Experiments
2. Baselines如下图所示，它是由以下5种问答模型直接输入数据集的question得到answer的准确率
3. 针对不同模块组合的对象/房间导航和EQA准确性实验如下表所示。其中除、ep\_len以外，其他指标越大越好。由图可知，RL训练方法和Ctrl模块都有助于性能的提升。需要行动距离越短（越简单）的问题，准确率越高。



1. Oracle Comparisons如下两表所示

* 比较第二张表1、2行可以得出，本文的VQA模型优于attention-based VQA
* 比较第二张表2、3行可以得出，controller模块提取的图像特征更有助于准确率的提高
* 比较第二张表3、5行，第一张表4行与第二张表3行可以得出，controller模块作出的决定会显著影响准确率





1. Others
2. CNN的预训练是怎么进行->可阅读文献[8]
3. 怎么用IL和RL来训练网络->可日后学会IL和RL再回头思考
4. 准确率提高1-2个百分点会不会是数据集等其他因素引起而并非模型的改善？
5. 对于不同问题、不同初始位置是否都需要重新训练？
6. 各模块推导公式及优化方法，Controller和navigator模块更具体的结构、模块个数，寻找房间和对象的controller和navigator的结构是否一样->这些问题日后决定深入研究再解决
7. 这技术尝试着将图像、NLP与机器人的动作相结合，离落地还有很多问题需要解决，比如在到达目的地后拾取、放下目标物体等一系列操作；实际环境是会经常变化的需要实时三维建图；真实环境中会有很多干扰，如光照、摄像头畸变、测量误差等；机器人本体的直立行走是很难做的，目前波斯顿动力的几款机器人在这方面有了很大突破，可以过独木桥、跳跃、后空翻等，但是体型比较大，还需小型化，另外也可以采用轮式的，但不能上下楼梯等；真实运用对实时性要求也比较高，不能让主人等太久；大规模应用还得考虑成本问题；得把AI和传统技术起来，如AI用于问答和导航，拾取物品使用传统技术，目前来说机器人导航和定位主要用GPS、SLAM等。
8. 《Knowledge Acquisition for Visual Question Answering via Iterative Querying》总结
9. Background

到目前为止，基于深度学习的模型主导了标准的VQA基准。在这些模型中，使用CNN对图像进行编码，使用LSTM对单词进行编码是最受欢迎的方法之一。 此外，许多表现良好的模型采用了注意力机制来获得更好的结果。 最近Jabri等人提出了一种新的替代模型，即一种双层MLP，它将答案作为输入并进行二元预测。 与其他更复杂的架构相比，这个简单的网络展示出了较好的结果。

1. Motivation

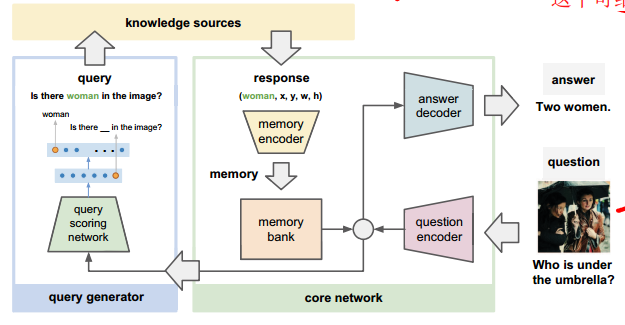
为了弥补之前方法没有充分利用知识库和基于图片的信息的缺陷，提出了利用问题产生模块通过多次迭代询问的方式从已有知识获取答案的线索的方法，本文中已有知识是指来自人类标注或算法生成的数据源的视觉信息

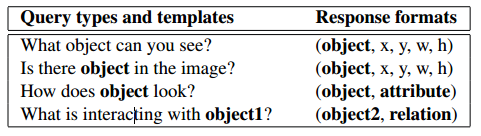
1. Dataset

本文用到的两个数据集是Visual7W、VQA Real Multiple Choice

1. Model

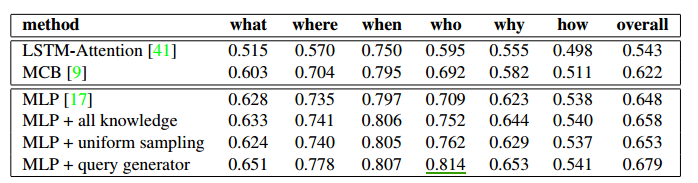
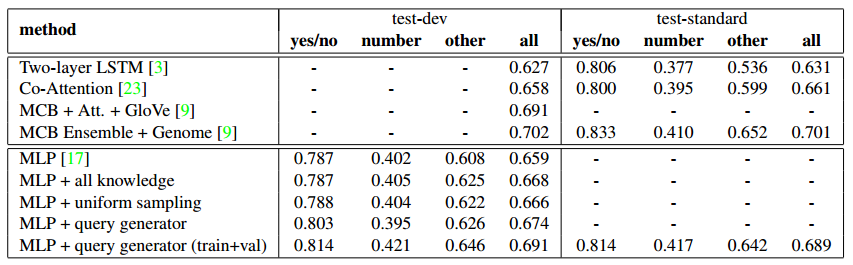
模型如下图所示，又Core Network、Query Generator、knowledge sources三个模块组成，模型输入图片和问题，输出答案。模型采用与EM相似方法来训练，将Core Network与Query Generator逐次循环训练，其中Query Generator使用Monte-Carlo方法来训练。模型的优点是结构简单，可以获取外部知识，可以展示模型的迭代推理过程。缺点是模型也可能被外部知识误导。



* Core Network
* 对图像-问题对进行编码
* 对knowledge sources信号进行编码并将其存入memory bank，其采用堆栈存储
* 由memory bank信息和image-question编码信息产生Query Generator的输入
* 由answer encoder对memory bank信息和image-question编码信息进行译码产生answer
* Query Generator：综合图像、问题和memory bank产生query到knowledge sources寻找信息，其中由两层softmax层来选择query types和object，如下图所示。
* knowledge sources：知识来自人类标注或算法生成的数据源的视觉信息，输入query返回一定形式的response，如下图所示。

1. Experiments

visual7W和VQA数据集上实验结果如下图所示，由图可以看出，在visual7W和VQA问题query generator的加入对MLP有优化的作用。在visual7W中对who问题的提升最明显，总体提升3个百分点。VQA实验结果与目前最好模型结果相当。



1. Others

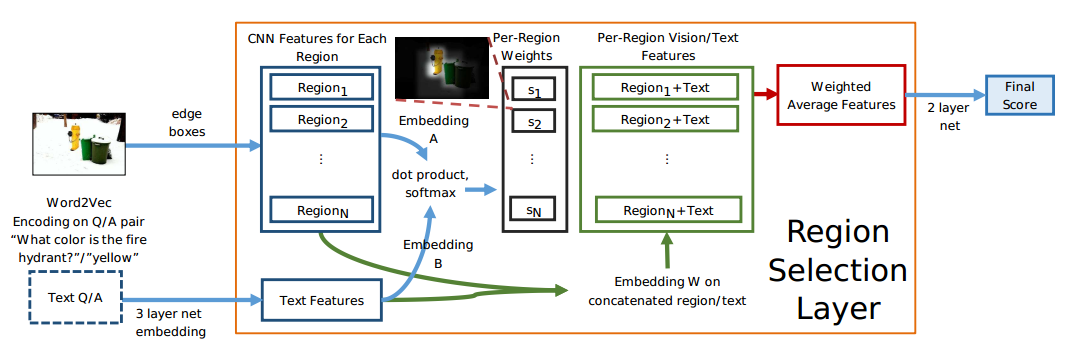
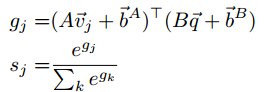
* 可不可以尝试把图像输入knowledge sources模块，将knowledge sources模块改成另一个问答模块，相当于由两个人来完成同一个问题
* 有空阅读文献3、9、17、23了解其他VQA模型
* 在训练模型时能否将语义部分打印出来，方便更好理解模型
* 模型的设计有点类似于人类的思考方式，如本文模型通过借助外部知识来提升性能

1. 《Where To Look: Focus Regions for Visual Question Answering》总结
2. Motivation

本文主要是为了使模型能通过问题、回答和图像来选择图像中对于回答问题比较重要的部分

1. Model

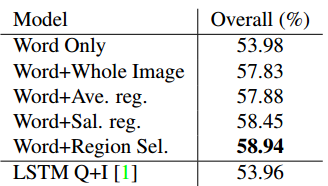
模型如下图所示，输入图像、问题和答案，输出是分数。其中，图像通过ImageNet预训练好的CNN网络来提取特征，使用Stanford Parser方法表示问题，再与答案的表示组成15000的向量，这种方法可以将不一样长度的句子变换成固定长度。训练时发现，低初始值可以防止选择用的softmax过早飙升以及高维视觉向量过早占主导地位。



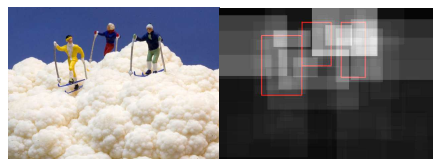
将两个向量串联得W

1. Experiments

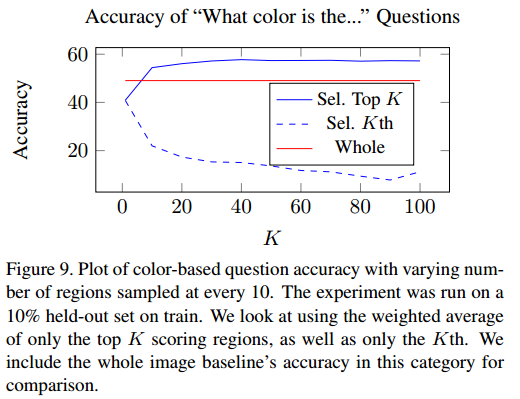
* 实验结果如下表，比较第一行和第五行可以发现，相对纯文本问答，图像信息的加入会有5个百分点的提升，提升较大。第二行为简单将图像特征与问题-答案特征串联起来替代选择环节的方法，比较第二行和第五行发现作者方法只有1个百分点的提升，说明选择环节对性能改善效果并不是很明显。



* 作者后面用图像展示模型根据问题-答案选择图像对应部分的能力，如下图，其中问题是“Are the people real?”，红框是经过训练后权重较大部分。综合文中其它几个图片可以发现，该模型能将问题-答案与图像对应起来，有点类似于文本问答的attention机制。



1. Others

* 文章的图9说明了权重越大图像区域对准确率影响越大，也说明了权重排在前K的图像区域对准确率具有决定性作用，而且选择合适K值准确率可能高于选择全部图像区域
* 组合2种类型向量时可以考虑用内积或串联的方法