Medical Visual Question Answering Notes

General VQA model:

1. 多模态特征提取:
    1. 视觉特征由中间层的特征图(Map),一般是CNN(Faster-RCNN)
    2. 问题的语义特征由RNN提取,LSTM或者是GRU
2. 注意力机制的特征融合,塑造视觉和文本的特征相互作用的关系来融合
3. 通过一个classifier预测

本篇论文创新点:针对不同类型的Med-VQA任务来学习task-adaptive推理能力
相比传统的用神经注意力机制作为一个简单的推理模块,我们的模型还学习了一个QC的调制*modulation*
这个词我查出来是调制分类?
还有一个把closed-ended和open-ended分开的模块

QCR 有条件的问题的推理模块

QC modulation component:

1. 和人类的推理类似,对不同的任务要求特定的能力
2. 从问题中提取任务信息来引导modulation

题外话,word embedding(词嵌入)
是一种文本表示的方法(独热编码,整数编码),两种主流的是word2vec和GloVe
本文用的就是GloVe

步骤:

1. 有l个单词的str q输入CloVe输出一个l维的Qemb
2. Qemb输入一个dG维度的GRU来获取问题的词嵌入,输出一个dG`*`l维度de
Qfeat = η
3. 现在Qfeat是一个对不同词语没有重点的向量组了,我们通过一个注意力机制来对不同的词语施加不同的权重
4. ~Q是一个(dG + dW) x l的向量组,即Qemb和Qfeat并起来.
5. Y是(~Q)乘一个可训练的权值W1,被tanh激活-1, 1.~Y同理,被sigmod激活0, 1.
都是dG x l维的 ***不懂这个~Y的控制作用***
6. G = Y Hadamard ~Y(对应的每个元素相乘) 这样我们就把GRU和GloVe的
优点集合了起来
7. 注意力向量α l x 1维,是Wa x G转置来的在经过一层softmax,
Wa是1 x dG
8. 最终结果Q att = Q $feat$ α
9. QCR(q) = MLP(q att) MLP是一个多层感知器

正常VQA的映射函数f\theat由两部分组成,A θ m D θ c
之后又是一个Hadamard乘把QCR(q)和A θ m (v, q)搞一起喂给上述
的D函数算分数***这里A和D函数都没有明确说明,不知到是啥.***
而且A都翻译不出来-什么多模特征联合.D是分类

TCR Type-Conditioned Reasoning 问题分类

我们要将问题分成cloesd-ended和open-ended来增加回答的准确性
过程:

1. 经过了和上面相同的步骤,我们得到了问题嵌入和合并好的map ϕ
m D(A(v, q)元素乘QCR(q))

```
也就是到这里只讲了文本特征提取上的创新,一个对问题内容,一个对问题类型
文章中说的是text和visual都被一个共享的特征提取之后,吧联合的特征喂给
对立模块,这里对问题进行分类.之后QCR输出的问题特征向量和
(多模特征联合)元素乘喂给MLP classfiier打出分数.
*那其实看到这里我还是挺糊涂的,这就介绍完了?下面就开始案例分析了?
明明觉得还有最重要的answer部分没说*
```

```
他之后说出彩的地方是用了两次bilinear-attention network
算A的时候用了V和Q两类feature,通过一个BAN.然后拿VAQ算联合特征f又用了
一次BAN
对于视觉表现(编码)采用的是pre-trained初始化MAML模型和CDAE模型
对于语义文本特征,用GloVe来词嵌入和一个1024维的LSTM从问题中提取
QCR和TCR中所有的GRU都有1024维隐藏层
Adam optimizer 是一个优化算法 针对学习率的
```