Aligned Dual Channel Graph Convolutional Network for Visual Question Answering

关键词：多模态；VQA；图卷积

Keywords: Multi-Modal, VQA, Graph Convolutional Network

简述

本文感觉东西不新啊，把各种技术拼凑到一起就中了ACL...

简单总结就是VQA中首先对图像建图，再对文本建图，跑两个GCN后，把特征融合一下（说是Align但其实只是Attention），做Answer Prediction。

有一说一这不是Visual Grounding里用烂的技术么...

方法

贺老板告诉我们要从文章中挖掘闪光点，那就从文章的方法描述中寻找一些可以借鉴的思路和技巧吧。

视觉建图

作者首先用Faster-RCNN抽取regions以及对应的feature，将regoin作为node，根据regions之间是否有overlap决定是否存在一条edge，边权则通过MLP读入region的feature输出。之后就是常规的Attention GCN得到视觉表示 $H_v = h_{vi}^{(l+1)}$ 。

文本建图

文本方面，作者基于Standford生成文本的依存树，基于此树得到以词为节点，以依存关系为边的带权图。同样跑一遍Attention GCN得到 $H_q = h_{qi}^{(l+1)}$ 。

基于注意力的对齐

以为做了Visual Grounding其实并没有，就是跑了个Transformer的QKV，首先文本自己做self-attention(Q=K=V=q)得到对齐的文本表示 $\tilde{H}_q$ ，再用 $\tilde{H}_q$ 和 H_v 做Attention(Q=q, K=V=v)，得到 $\tilde{H}_v$ 作为对齐的视觉表示。

答案预测

拼接特征-MLP-Softmax。

总结

趁着GCN还能水，多灌几篇吧。