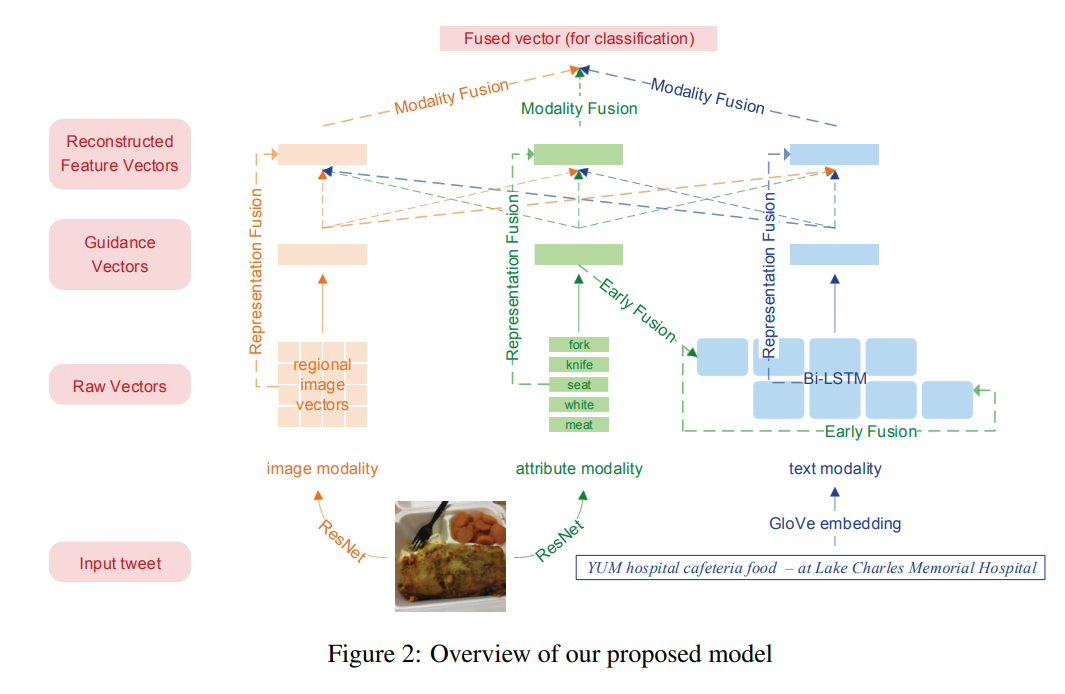
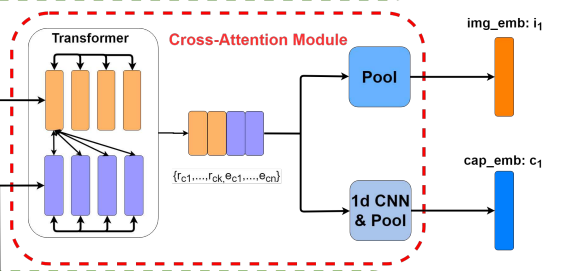
1. **主要参考文章：**

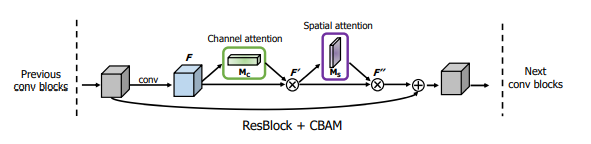
**模型框架论文：**Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model，该论文多模态讽刺检测模型框架，使用图片、图片属性、文本对twitter进行讽刺检测，该论文提出了自己的多模态twitter数据集，包含图片、图片属性、文本，直接在该数据集上训练、测试即可。



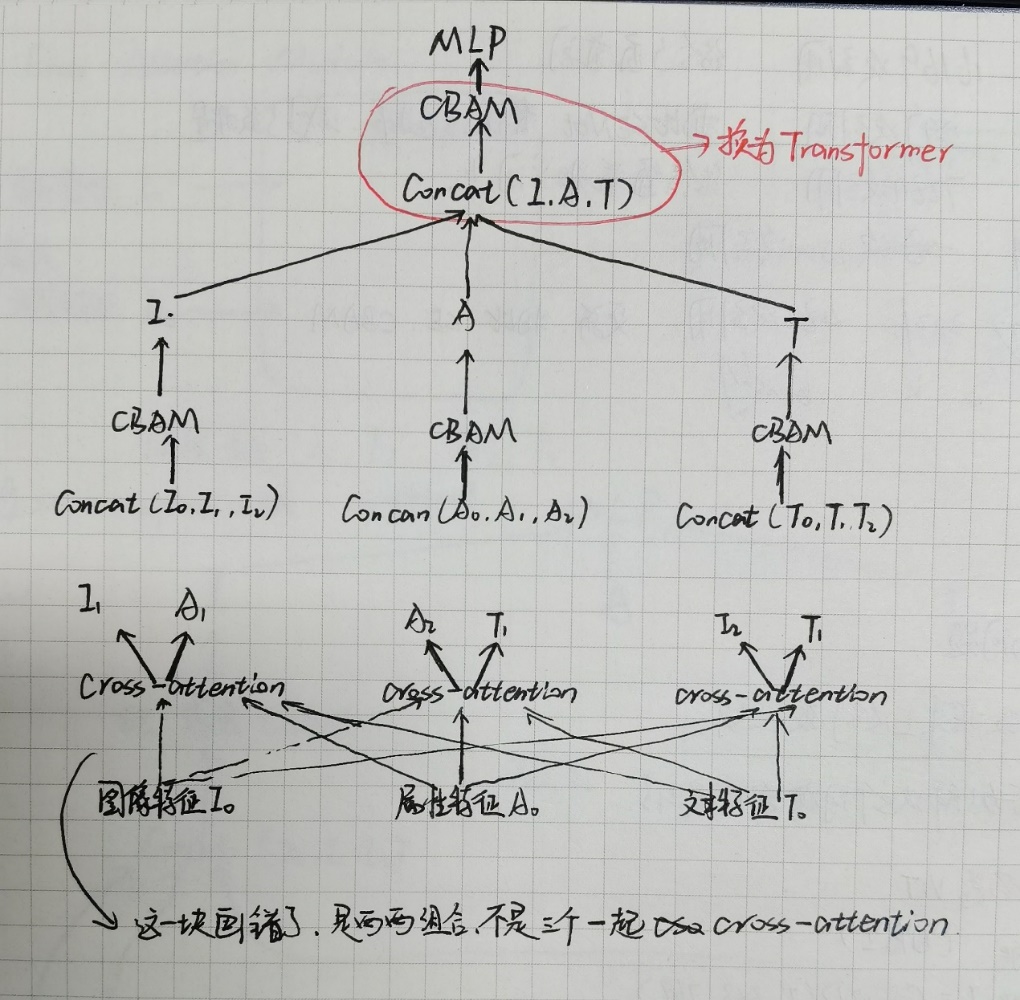
**Cross-attention:** Multi-Modality Cross Attention Network for Image and Sentence Matching，我们使用这篇论文中的Cross-attention module, 该模块输入Y为文本、图片特征的叠加，将Y传入一个TF单元中输出Y1, Y1通过一个前反馈层得到Y2，将Y2中的I、T分别传入对应的池化层得到最终的I、T。



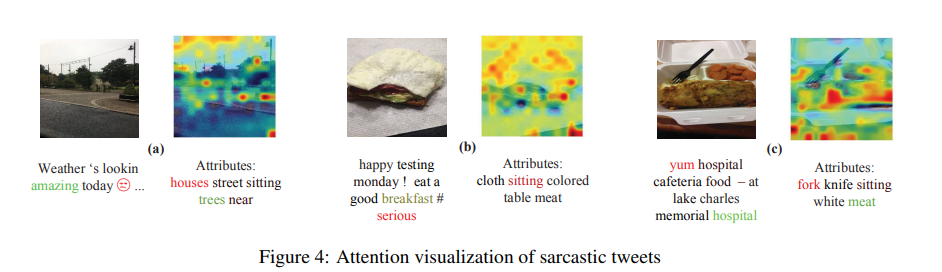
**CBAM:** CBAM: Convolutional Block Attention Module 该模块分别从通道和空间上进行attention，输入维度和输出维度相同。



1. **需求：**
2. **最开始的特征提取：**图像特征、属性特征、文本特征分别用CILP的图像编码器、文本编码器、RoBERTa提取。
3. **模态框架：**



1. 后续因为要写论文，还需要做消融实验证明框架各部分的有效性。
2. 同时，最好能通过直观的热力图显示出模型的优势，比如下面这样

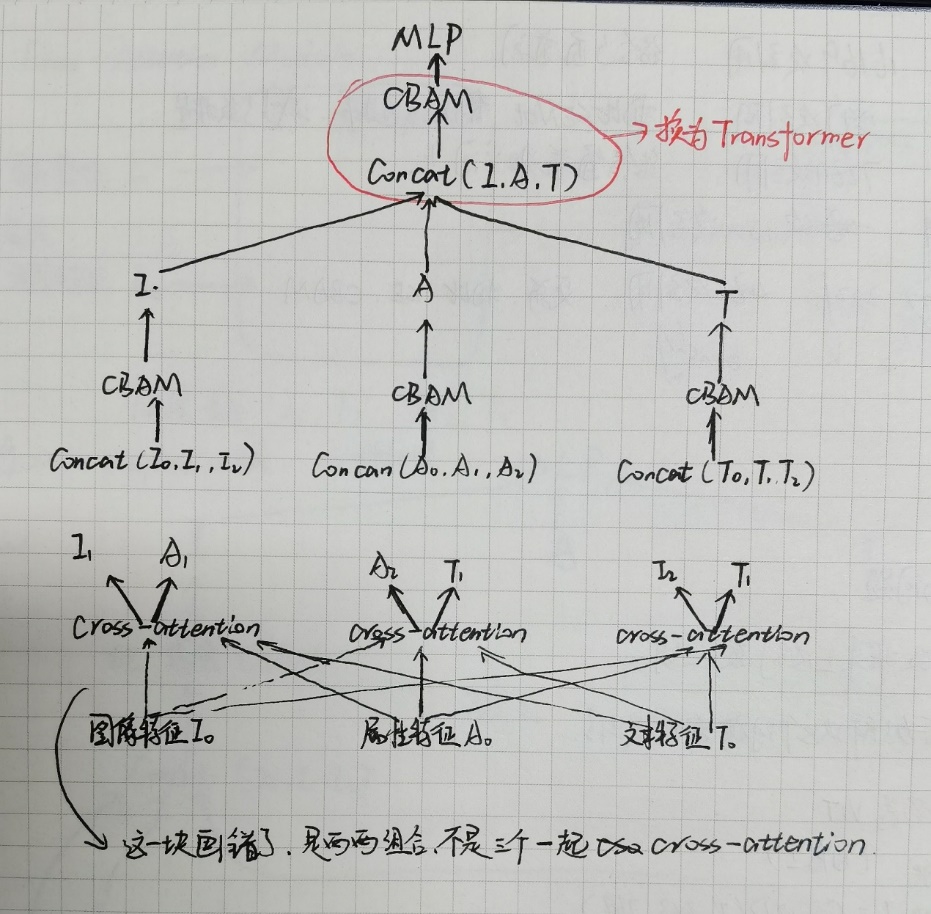


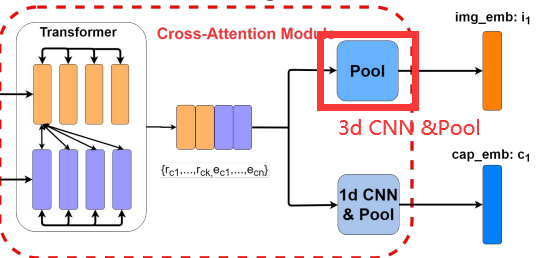
1. 我是本科生，这方面基础很一般，后续可能还需要进一步帮助。

**三、 步骤：**

**不必完全按照我的想法进行修改，假如有更合适的操作的话，给我说一下就行。**

1. 模型框架论文Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model的代码复现、数据都在<https://github.com/ZLJ2015106/pytorch-multimodal_sarcasm_detection>，可以以它们的代码为基础进行改进，首先更换特征提取模型，分别用CILP的图像编码器、文本编码器、RoBERTa提取图像特征、属性特征、文本特征。
2. Cross-attention、CBAM的代码网上有很多，我没有仔细看。根据如下模型框架图进行改进。Cross-attention可以进行一些改进，针对不同的模态采取不同的操作，对图片特征处理时增加一个三维CNN，如下图所示。最后通过transformer将I、A、T融合，这部分我不清楚该怎么做，老师这样告诉我的。





1. 代码改完在模型框架论文的数据集中进行训练、测试。
2. 效果理想的话再在其他一到两个数据集上进行测试。
3. 后续还需要做一些消融实验、对比实验之类的，画一下热力图。