题目

真实灾难帖子分类

组员

张浩楠　2006500001
蔡宇生　2006500008
吴昊原　2006500009

kaggle 链接

https://www.kaggle.com/competitions/nlp-getting-started/

科学意义

Twitter 已经成为一个重要的交流频道，在这里，我们可以交流实时的紧急事件。
人们可以发布他们实时的紧急信息。
但是，就连权威信息也会用多义的词。
我们要去建造一个机器学习模型，去预测哪些帖子是讲真实的灾难。
在这个作业中，我们用自注意力机制完成。

科学问题

运用自注意力机制分类文字。

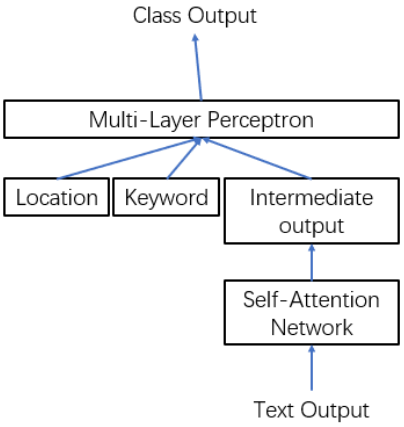
研究内容

数据

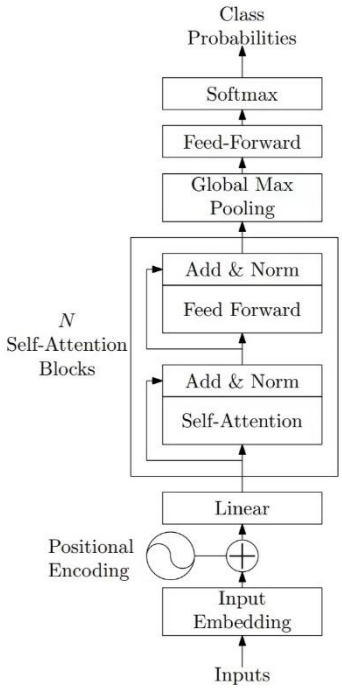
描述
数据是帖子的文字、帖子的关键字和帖子的位置。标签是这个帖子是否是一个真实的灾难。
数据增强
同义词替换，可以用 grammarly 工具，提高句子的水平。

筛选领域内数据，Xenc 工具，选出跟自己相近的语料。
数据处理
把帖子的关键字和帖子的位置独热编码，并且再除以最大的数字，使数字范围在零至一。
找一个 tokenizer 把句子 tokenize。Tokenize 把句子的每个词变成唯一的数字。

实施



我们用 Self-Attention Network，参考别人的方法。



Input Embedding

Input embedding 把 token 从一个标量的数字变成一个向量，相当于把输入变得丰富。鉴于

这个作业中，单词的数目是约三万，我们用的 embedding 维度是 30。

Positional Encoding

与循环神经网络不同，在传统自注意力机制中，每个词输入是同时的。加入 positional encoding 使得输入有一点句子顺序的资讯。

Linear

构建线性模型。

Self-Attention Blocks

Self-attention 和 feed forward 模块用于训练模型, 并利用 add&norm 归一化方法加速收敛，提高泛化性，增强网络的泛化能力和性能。

Global max pooling

用于从输出中提取特征，缩小模型的复杂度，防止过拟合，增强模型的鲁棒性。

Feed forward

将在 Global max pooling 步骤中提取到的特征向量反馈给 softmax 激活函数。

Softmax

运用 softmax 激活函数将提取到的特征向量转换为相应的概率，在本次作业中即一个帖子所陈述的内容是否是一个真实灾难的概率。

预期目标

以上通过 Self-attention network 的步骤之后，我们将得到的概率与帖子的位置、关键字相结合以得到对于每一个帖子的预测结果。希望排名在 top10%.