结合BERT与图网络的传导式文本分类

2018202046 方晓坤

概述

文本分类是自然语言处理中一个非常重要的任务，给定输入文本$\mathbf{x}$，模型需要输出它的类别$y = f(\mathbf{x})$。这里的类别随任务的不同而不同，对新闻分类来说，类别可以是“娱乐”“体育”“社会”“政治”等等

一般说“文本分类”都指的是 **Inductive (归纳式)** 式的文本分类，在模型训练过程中仅使用标注数据进行训练，而测试的数据在训练的时候没有见过

与Inductive不同，**Transductive (传导式)** 文本分类则在训练的时候也提供未标注的数据，测试时的数据就是这些未标注的数据。所以，Transductive文本分类的目的在于让模型能够从**观测到**的标注数据推演到**观测到**的未标注数据，可以通过训练阶段的信息传导实现

解决Transductive文本分类的主流方法是使用图网络，如GCN，将所有的标注数据与未标注数据都构建在一个图里，图中的结点代表文档或者词，而通过结点之间的信息传递，模型就能在该异质网络中凭借已标注结点的信息推理未标注结点的特征，从而实现Transductive分类

另一方面，大规模预训练也是在学习无标注数据背后的语义信息。直觉上，如果能将大规模预训练“从纯无标注文本中学习”的能力，与图网络“从标注数据泛化到未标注数据”的能力结合，那么模型就能在Transductive文本分类上取得更好的效果

实验

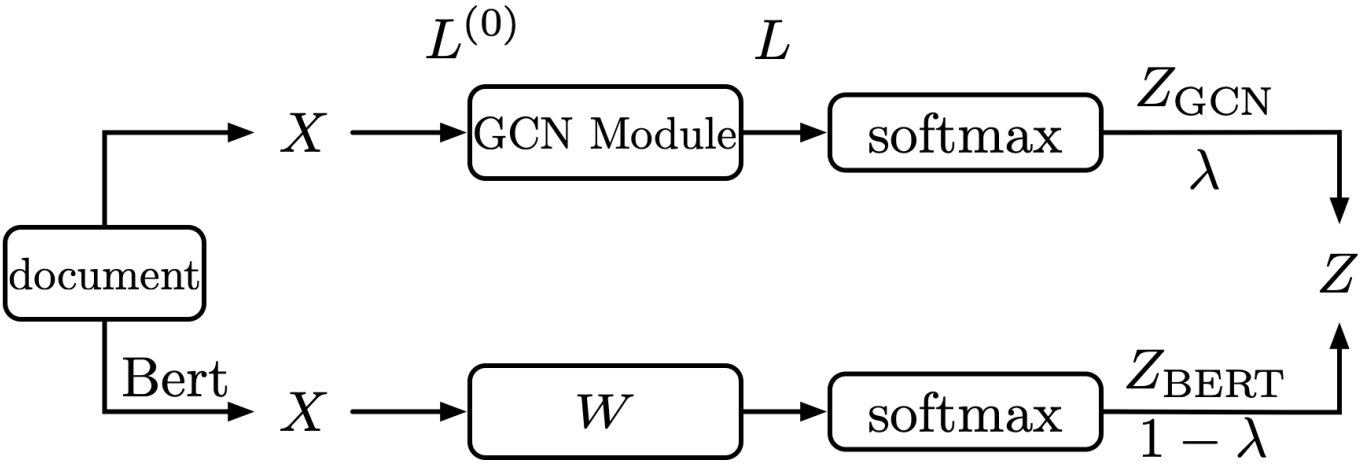
数据预处理

20newsgroups(20NG)是一个经典的文本分类数据集。该数据集合包括20个类别，共计约20000条新闻。

数据集原文本需要进行清洗，清洗包括以下几个方面：

- 仅保留字母、数字和表示语义的符号如!和?，去除空串
- 去除英语语境下的停用词
- 去除出现频率过低的词
- 最后将文本整理成(text,class)的形式

模型框架



GCN模块

首先构建一个由结点与边构成的异质图。结点分为两种：词结点与文档结点。边在词与词，词与文档之间进行连接。边的权重由TF-IDF (Term Frequency-Inverse Document Frequency）与PPMI (Positive Point-wise Mutual Information)决定：

$$A_{i,j}=\begin{cases}\mathrm{PPMI}(i,j), & i,j \text{ are words and } i\neq j\\ \text{TF-IDF}(i,j), & i \text{ is document, } j \text{ is word}\\ 1, & i=j\\ 0, & \text{otherwise}\end{cases}$$

记图中所有结点的特征组成的矩阵是 $X\in\mathbb{R}^{(n_{\mathrm{doc}}+n_{\mathrm{word}})\times d}$ ，其中n_{doc}是图中文档的数量，n_{word}是词的数量，而d则是特征向量的维度。

然后对于所有的文档结点X_{doc}用BERT进行初始化，把所有的词结点X_{word}初始化为0：

$$X=\begin{pmatrix}X_{\mathrm{doc}}\\ 0\end{pmatrix}_{(n_{\mathrm{doc}}+n_{\mathrm{word}})\times d}$$

在得到所有结点的特征向量之后，以此为基础构建一个GCN模型，将GCN最后一层输出的特征作为softmax的输入，得到关于类别的分布：

$$Z_{\mathrm{GCN}}=\mathrm{softmax}(g(X,A))$$

这里的$g()$就是图模型

BERT模块

对于BERT模块，直接将前面得到的文档的嵌入通过一个全连接层，最后经过softmax激活

$$Z_{\mathrm{BERT}}=\mathrm{softmax}(WX)$$

联合

最后以一个超参λ将两个模块得到的类别的概率分布联合在一起

$$Z = \lambda Z_{\mathrm{GCN}} + (1-\lambda) Z_{\mathrm{BERT}}$$

得到的Z是一个在所有类别上的概率分布，我们选择概率最大的类别作为预测结果。使用负对数似然损失(Negative Log-likelihood Loss)作为损失函数

贡献和尝试

- 原作使用的是BERT预训练模型，而RoBERTa是一个更强的预训练模型，用其来代替，准确度有所提升，但不明显
- 使用grid search方式找到了本模型在20NG数据集下λ的最优值
- 尝试使用$\mathrm{word2vec\ simlarity}$建立异构图新的边，但新邻接关系的引入没有带来提升
- 尝试增加GCN的层数，发现2层GCN是最优的，更多的层数没有提升反而有下降
- 使用随机游走来收集图网络上的多跳邻居的信息，但由于时间原因没有具体实现

结果

model	accuracy
bert + gcn	89.3%
roberta + gcn	89.5%

参考文献

Liang Yao, Chengsheng Mao, Yuan Luo. "Graph Convolutional Networks for Text Classification." In 33rd AAAI Conference on Artificial Intelligence (AAAI-19), 7370-7377