基于半监督图神经网络的短文本分类

张斌艳 朱小飞* 肖朝晖 黄贤英 吴洁

（重庆理工大学 计算机科学与工程学院）

论文摘要

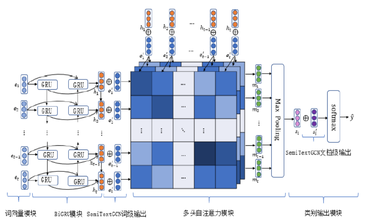
短文本分类是自然语言处理（NLP）的重要任务之一，其面临的主要挑战在于短文本数据稀疏性问题。以往短文本分类的研究工作主要关注建模文本内部词序列信息，尽管相关研究取得了较好的分类效果，但是这类方法未能有效解决短文本稀疏性问题。本文提出在短文本建模过程中引入词项与词项之间、词项与文档之间的全局结构关系来增强短文本的表示。由于有标签训练数据的缺乏，使得现有的全局结构关系建模方法，例如TextGCN无法学习到高质量的词项和文档全局结构表示，因此，我们进一步提出采用半监督学习思想来解决有标签训练数据不足的问题。在基准数据集MEDUI上，我们与现有相关模型进行对比，实验结果表明本文提出的方法比最好的基准模型在F1指标上提高了1.91%。

论文简介

动机：1、短文本数据的稀疏性； 2、有监督数据的缺乏；

方法：提出BSGNN模型，其中的SemiTextGCN模块在TextGCN构图的基础上添加部分同类型无标注样本。后将SemiTextGCN学习到的特征表示对应拼接到Base模型预先学习到的词级和文档级嵌入上，深度挖掘文本全局结构信息，学习到更为精准的词项和文档的特征表示。

系统模型



1、**词向量模块：**由word2vec进行预训练得到；

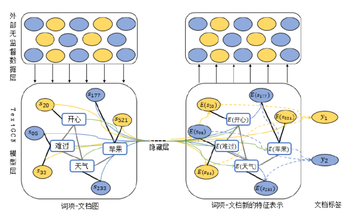
2、**BiGRU模块：**捕获得每个词语与其上下文的关联，初步建模文本的语义信息；

3、**SemiTextGCN模块：**捕捉到文本全局结构信息并得到更为精准的特征表示；

4、**多头自注意力模块：**捕捉词与词之间的交互作用，构建高质量的词的上下文依赖关系；

5、**类别输出模块：**由最大池化层捕获最重要的特征，最后用softmax激活函数输出每个类标签的概率。

图 1 BSGNN 模型总体架构图，其中e_i'是词项经过 SemiTextGCN 得到的词级嵌入，s_i'是句子经过 SemiTextGCN 后得到的文档级嵌入。



1、**TextGCN模型层：**文本图由整个语料库中的词项和文档以及外部无标注样本的词项和文档作为节点构成的，通过提取词项的全局结构信息来捕捉文档的全局依赖注意权重。

2、**外部无监督数据层：**从外部添加同类型无监督数据进行辅助训练得到文本的全局结构信息，并在训练结束后摘除这部分数据。

图 2 SemiTextGCN 模型图，其中s_i是文档节点，其他是词项节点，黑色粗体边缘为文档-词项边缘，灰色细边缘为词项-词项边缘。$E(x)$表示经过 TextGCN 之后的x的嵌入，公式中词级嵌入用e_i'表示，文档级嵌入用s_i'表示。不同的颜色代表不同的文档类别（为简洁明了，例子中仅展示两个示例类别）。

算法原理

$$(1)\ \hat{y}_{semi} = \mathrm{softmax}(\overline{A'}\sigma(\overline{A'}XW_0)W_1) \quad [\text{双层GCN}]$$

$$\left.\begin{aligned}(2)\ \overrightarrow{h_i} &= \overrightarrow{GRU}(e_i,\ \overrightarrow{h_{i-1}})\\ (3)\ \overleftarrow{h_i} &= \overleftarrow{GRU}(e_i,\ \overleftarrow{h_{i+1}})\end{aligned}\right] \text{[BiGRU]}$$

$$(4)\ M = \mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad [\text{多头自注意力}]$$

$$(5)\ \hat{y} = \mathrm{softmax}\left(W_i'^T \times G_i' + b_i'\right) \quad [\text{softmax分类}]$$

$$(6)\ \mathcal{L} = -\lambda\frac{1}{|S|}\sum_{y\in S}\sum_{i=1}^{2} y_k \log(\hat{y}_i) \quad [\text{损失函数}]$$

实验仿真

表 1 模型参数设置表

参数名	值
词向量维度	200
学习率	0.01
权重正则限制	2
dropout	0.5
批处理大小	64
外部补充数据	50000

1、各个模型对比实验

表 2 不同模型在三个指标（准确率 P、召回率 R、F1）上的测试结果/%

模型	P_1	P_0	R_1	R_0	F1_1	F1_0	P_Avg	R_Avg	F1_Avg
LR	0.67	0.73	0.57	0.80	0.61	0.76	0.70	0.71	0.70
SVM	0.76	0.80	0.69	0.85	0.72	0.82	0.78	0.78	0.78
W2V+CNN	0.85	0.81	0.74	0.90	0.79	0.85	0.83	0.83	0.83
TNA	0.91	0.80	0.72	0.94	0.80	0.86	0.85	0.84	0.84
EV-CNN	0.91	0.86	0.79	0.95	0.84	0.90	0.88	0.88	0.88
UA-LSTM	0.92	0.91	0.88	0.94	0.90	0.92	0.91	0.91	0.91
HAN	88.97	96.55	97.78	83.69	93.16	89.65	92.22	91.77	91.67
NPA	90.22	96.53	93.81	85.13	93.81	90.43	92.84	92.5	92.41
BSGNN	93.73	95.28	96.76	90.98	95.22	93.07	94.38	94.34	94.32

2、参数敏感性实验

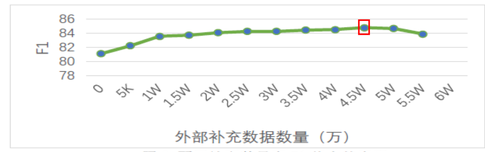


图 3 不同补充数量在 F1 值上的表现

3、模块消融实验

表 3 本文不同组件在三个指标（准确率 P、召回率 R、F1）上的测试结果/%

对比模型	指标	积极	消极	总体
Base	P	90.62	95.17	92.52
	R	96.87	86.03	92.34
	F1	**93.64**	**90.37**	**92.27**
Base+W	P	92.50	89.91	91.42
	R	92.79	89.52	91.42
	F1	**92.64**	**89.72**	**91.42**
Base+S	P	95.38	87.76	92.19
	R	90.60	93.89	91.97
	F1	**92.39**	**90.72**	**92.00**
Base+WS	P	93.29	94.09	93.63
	R	95.92	90.39	93.61
	F1	**94.59**	**92.20**	**93.59**
Base+W*	P	91.57	93.06	92.19
	R	95.30	87.77	92.15
	F1	**93.39**	**90.34**	**92.12**
Base+S*	P	93.17	91.59	92.51
	R	94.04	90.39	92.52
	F1	**93.60**	**90.99**	**92.51**
BSGNN	P	93.73	95.28	94.38
	R	96.76	90.98	94.34
	F1	**95.22**	**93.07**	**94.32**

论文结论

主要贡献：

- 将本适用于长文本分类任务的TextGCN模型进行改进，解决了在短文本分类任务中数据稀疏且没有足够上下文的问题
- 利用半监督学习的思想，从外部选取部分同类型无标注样本辅助训练并得到更加准确的词项和文档的特征表示
- 提出的方法可显著提升短文本分类任务的效果且始终优于基准方法。

西安电子科技大学 XIDIAN UNIVERSITY