

[文章编号] 1003-4684(2023)05-0067-08

# 基于图对比学习的长文本分类模型

刘宇昊, 高 榕, 严灵毓, 叶志伟

(湖北工业大学计算机学院, 湖北 武汉 430068)

**[摘 要]** 当前基于字符级考虑的文本分类方法在长文本分类上, 存在输入维度过大致使计算困难以及内容过长难以捕捉长距离关系, 从而导致准确度不足的问题。由此, 提出基于自适应视图生成器和负采样优化的图对比学习长文本分类模型。首先将长文本分为若干段落, 用 BERT 衍生模型对段落进行嵌入表示, 然后基于文本的高级结构将段落的嵌入表示视为节点构建图模型, 接着使用自适应视图生成器对图进行增广, 并通过图对比学习得到文本的嵌入表示, 同时在图对比学习的负采样阶段, 引入 PU Learning 知识修正负采样偏差的问题, 最后将得到的文本嵌入表示使用两层线性层进行分类。通过在两个中文数据集上的实验显示, 方法优于主流先进模型。

**[关键词]** 文本表示; 长文本分类; 图对比学习; 负采样

**[中图分类号]** TP391.1 **[文献标识码]** A

自然语言处理(natural language processing, NLP)一直是人工智能领域最重要的方向之一, 文本分类<sup>[1]</sup>则是 NLP 领域最基本的任务, 其是指通过一定的计算, 得出当前文本属于某一类别的概率的过程, 其运用于情感分析<sup>[2]</sup>、信息检索<sup>[3]</sup>、问答系统<sup>[4]</sup>、机器翻译<sup>[5]</sup>等非常具有实际应用的应用。例如日常生活中的内容推荐、搜索引擎、自动客服、翻译软件等就是以上应用的具体体现。可以说 NLP 技术正潜移默化地方便我们的生活。

文本的嵌入表示是文本分类任务的基础, 嵌入表示的好坏决定了文本分类任务的质量。传统上的嵌入表示有针对目标数据集计算的词袋模型和 n-gram 模型<sup>[6]</sup>, 但由于其数据稀疏性需要大量计算资源, 同时忽视文本的顺序和结构信息, 所以在大多数情况下对文本的语义表达并不准确。后来谷歌推出了 word2vec<sup>[7]</sup>, word2vec 考虑了词在上下文的关系, 效果<sup>[8]</sup>与通用性有显著提升。然而 word2vec 是静态的方法, 单词与向量是一一对应的关系, 无法解决一词多义问题, 但此时 word2vec 已初具预训练模型的雏形。之后因计算机硬件的发展, 研究人员着手使用大语料库训练出通用模型。2018 年, 谷歌使用双向的 Transformer<sup>[9]</sup> 在 33 亿字的无标注语料库上训练出 BERT<sup>[10]</sup>, 不仅解决了一词多义的问题, 同时使自监督模型的预训练加微调模式成了各个领域的热点。后来, 对比学习<sup>[11]</sup>的出现, 让自监督学习发展到新的高度, 由于其“轻便”的结构拥有很好

的泛化性, 在计算机视觉、NLP、多模态中都有应用。在 NLP 的背景下, 对比学习可以让模型针对特定领域进行自监督训练。例如金融、法律等专业领域有标注的数据集非常少, 如果使用人工标注则需要付出高昂的经济代价。应用对比学习不仅可以降低经济成本, 还能缓解针对特定领域准确度不高的问题。

上述工作取得了一定的成果, 然而当前工作还存在以下挑战。问题 1: 以往大多基于字符序列的工作忽略了文本的高级结构, 并且受制于文本长度。如 BERT 模型, 由于自注意力机制需要  $n \times n$  的计算矩阵( $n$  为文本长度), 所以默认只能处理 512 个字符内的文本。对于过长的文本会将不同句子化作同一句子并且截断<sup>[12]</sup>, 但这显然会导致文本的语义丢失甚至改变。问题 2: 对比学习在正负对选取上存在采样偏差问题。文本背景下, 数据增广操作可能会改变文本的语义标签, 因此增广策略需要先验知识。在负对采样<sup>[13]</sup>上, 由于自监督学习没有标签信息, 来自不同实例的增广对有一定概率具有相同的标签, 这时将其视为负对就会导致负采样偏差。

当前基于图对比学习的长文本分类模型<sup>[14]</sup>, 通常把文本随机分为两部分, 将来自同一文本的子文本视为正对, 来自不同文本的子文本视为负对, 再以句子或段落为节点, 将它们的顺序关系作为边来构建图模型, 接着进行对比学习后再分类。然而, 文本分割的比例与分割方式需要大量的实验来确定, 同时由于数据增强的方式单一, 以及负采样偏差的问

题,对比学习的提升效果有限。本文提出一种基于自适应视图生成器和负采样优化的图对比学习长文本分类模型(Graph Contrast Learning model based on Adaptive View Generator and Negative Sampling Optimization, GCL-AVGNSO),可以让段落节点自适应的选择数据增强方式,不仅增加了数据增强的手段,优化了不同文本的不同分割比例,同时也缓解了负采样偏差对图对比学习的影响。首先基于图模型构建文本,不仅可以捕捉句子的上下文关

系,也能扩展到长文本。接着利用自适应视图生成器进行数据增广,能让文本自适应地选择划分比例。然后引入PU Learning<sup>[15]</sup>的知识,在仅访问全样本分布和正样本分布的情况下,用超参数 $\pi$ 对负采样偏差进行修正。最后本文在两个公开中文数据集上证明了有效性,效果优于主流先进模型。

## 1 GCL-AVGNSO 模型

GCL-AVGNSO 流程见图 1。

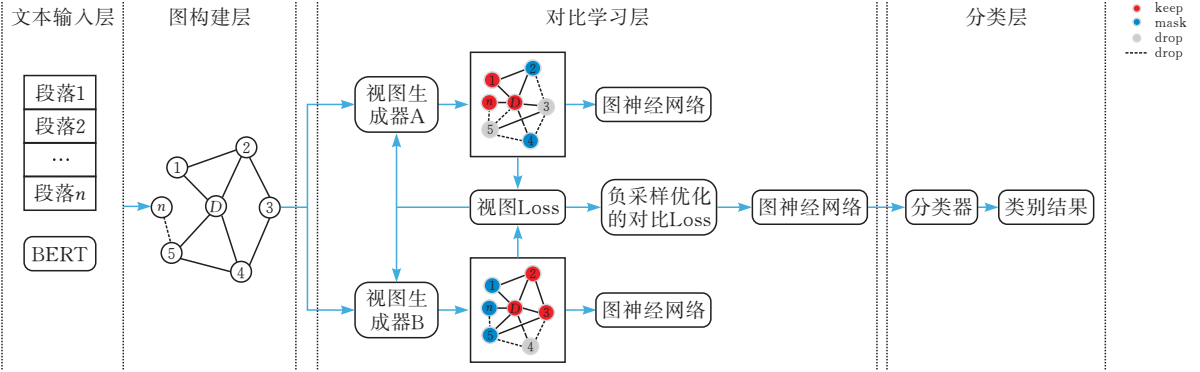


图 1 GCL-AVGNSO 模型流程

### 1.1 问题定义

将长文本  $D$  映射为图  $G$ ,再将图  $G$  映射到低维空间得到  $G$ 。设  $D = \{p_1, p_2, \dots, p_n, p_{|D|}\}$ ,  $p_i$  为文章中的段落,每个段落对应图的节点,上下文的顺序关系对应顶点间的边,所有的段落节点都与文本节点  $p_{|D|}$  有边。设图  $G = (V, E)$ ,其中  $V = \{v_1, v_2, \dots, v_N\}$ ,  $E \subseteq V \times V$  分别表示节点集和边集。 $V \in R^{N \times F}$ ,  $A \in \{0, 1\}^{N \times N}$  分别为特征矩阵和邻接矩阵。 $v_i \in R^F$  是  $v_i$  的特征向量,如果  $(v_i, v_j) \in E$  则令邻接矩阵中的  $A_{ij} = 1$ 。本文的目标是学习一个 GNN 编码器  $f(V, A) \in R^{N \times F'}$ ,其中  $F > F'$ ,  $N \times F' \geq G$ ,将一组文本  $D$  映射成  $G = \{G_1, G_2, \dots, G_M\}$  后,在没有标签信息的情况下利用 GNN 编码器将  $G$  嵌入到低维空间中得到图级别表示  $G = \{G_1, G_2, \dots, G_M\}$ ,这些图级别表示可以用线性分类器进行分类。

### 1.2 文本图构建

给定一个文本  $D = \{p_1, p_2, \dots, p_n, p_{|D|}\}$ ,定义一个无向图  $G = (V, E)$ ,其中  $V$  由  $n + 1$  个节点  $(v_{p_D}, v_{p_1}, \dots, v_{p_n})$  组成,图的边集  $E$  根据文本结构分别从段落节点和文本节点展开构造。

**1.2.1 段落节点的嵌入表示** 本文使用当前最先进的上下文语境模型来编码每个段落,由于每个段落相对较短,可采用 BERT 及其衍生模型来对其进行嵌入表示。具体来说,对于一个段落  $p_i$  由字符序列  $\{w_{i1}, w_{i2}, \dots, w_{in}\}$  组成,本文使用 Bert-WWM

进行编码,从而获得一组字符向量矩阵  $Vector_{token} = [v_{w_{i1}}, v_{w_{i2}}, \dots, v_{w_{in}}]$ ,其中  $v \in R^F$ ,  $F$  为特征维度,再对  $Vector_{token}$  做平均池化,得到段落的特征值  $v_{p_i}^{(0)}$ ,并将段落的特征值视为图的段落节点:

$$v_{p_i}^{(0)} = \text{MeanPool}(\{B(p_i); i \in [1, n]\})$$

其中  $B(p_i)$  为 Bert-WWM。

**1.2.2 文本节点初始化** 在获得所有节点表示后,使用所有段落节点的平均值作为文本节点的初始表示:

$$v_D^{(0)} = \frac{1}{n} \sum_{i=1}^n v_{p_i}^{(0)}$$

### 1.3 GNN 层

本文使用图注意力网络(GAT)<sup>[16]</sup>来捕捉节点间的关系,获得进一步的节点表示。对于第  $t + 1$  层的图注意力层,节点  $v_i^{t+1}$  通过聚合所有与之临边的  $t$  层节点信息来表示:

$$v_i^{t+1} = \text{GAT}(v_k^t \mid k \in \text{Neighbour}(j))$$

其中  $\text{Neighbour}(j)$  为所有与当前节点  $v_i$  相邻的邻居节点,最终的文本节点表示为  $v_D^T$ ,段落节点的最终表示为  $v_i^T$ 。

在获得所有节点的最终表示后,通过 READ-OUT 函数得到文本的图级别表示  $G_i$ 。

$$G_i = \text{READOUT}(\{v_j^T; j \in [1, D]\})$$

### 1.4 自适应视图生成器和负采样偏差修正

在图对比学习中,会通过数据增广(Data Augment)来扩充数据样本,对于每个图  $G$ ,有:

$$G_i \xrightarrow{\text{Aug}} G'_i, G''_i \quad (1)$$

假设一个小批次中包含有  $N$  个图,通过数据增广获得  $2N$  张增广图,则通常会将来自同一个图的  $G'_i, G''_i$  视为正对,将来自不同图的  $G'_i, G''_j$  视为负对,并由此进行正负采样来构建对比学习损失函数。受

文献[16]的启发,提出对式(1)优化的自适应视图生成器和修正负采样偏差方法。

**1.4.1 自适应视图生成器** 自适应视图生成器流程见图 2。

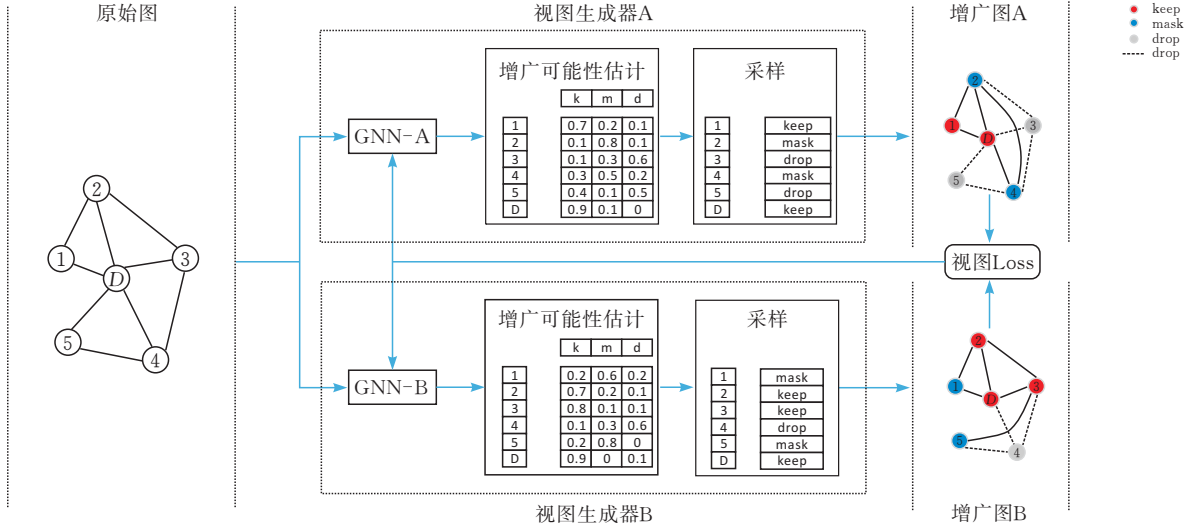


图 2 自适应视图生成器

以视图生成器 A 为例,利用节点的嵌入表示和 gumbel-softmax<sup>[17]</sup>来预测每个节点选择每种增广方法的概率。获得每个点的嵌入表示  $v_i^T$  之后,给定增广选项 keep、drop 和 mask 三种,更改最后一层 GAL,让输出的维度等于增广选项的数量,再通过 gumbel-softmax 归一化并与源节点相乘,则选择某种增广方式的概率以及最终增广后的节点表示为:

$$v_i'^T = \text{GAL}(v_k^{T-1} \mid k \in \text{Neighbour}(j))$$

$$f_v = \text{Gumbel-Softmax}(v_i'^T)$$

$$v_i = \text{Aug}(v_i, f_v)$$

$$G' = (V', E')$$

其中,  $f_v$  为与增广方法数量相同的 one-hot 向量,  $\text{Aug}(x, y)$  采用可微的计算方法,本文采用乘法。具体而言,  $f_v$  是一个形状为 (1, 3) 的 0, 1 向量, 一个由  $n$  个节点组成的图  $G$  会得到一个由  $n$  个  $f_v$  组成的 0, 1 矩阵  $A_{mat}$ , 将矩阵  $A_{mat}$  中对应的 keep 与 mask 列提取后相加得到形状为  $(n, )$  的 0, 1 向量  $Tenso r_{Aug}$ , 再将  $Tenso r_{Aug}$  与初始化后的  $V$  相乘并将 mask 对应的节点属性随机化得到  $V'$ , 同时更新边表  $E$ , 这样视图生成器的梯度就可以保留在增广节点特征中, 可以进行反向传播计算。由于本文中的边只用来指导聚合, 没有属性, 所以不参与梯度计算。增广后的图  $G' = (V', E')$ , 其中  $V' = (v'_1, v'_2, \dots, v'_{n-t}, v'_D)$  对应节点的属性矩阵  $V' = (v'_1, v'_2, \dots, v'_{n-t}, v'_D)$ ,  $t$  为被 drop 的节点数量,  $E'$  为  $E$  的子集。视图生成器 B 流程与 A 一致, 仅在随机初始化时权重值不同。由于文本分割比例存在最优值, 因此视图生成器 A/B 在经过学习后会趋于相同。所

以可以通过比较两次增广时得到的  $A_{mat}$  和  $A'_{mat}$  的相似度, 来构造视图生成器的损失函数  $\text{Loss}_{view}$ :

$$\text{sim}_{view}(x, y) = \text{mse\_loss}(x, y)$$

$$\text{Loss}_{view} = 1 - \text{sim}_{view}(A_{mat}, A'_{mat})$$

**1.4.2 负采样偏差修正** 本文利用对比学习的方式训练一个 GNN 编码器  $f(V, A) \in \mathbb{R}^{N \times F'}$ , 以下简称为  $f$ 。在得到所有图的嵌入表示  $G'_i, G''_i, G'_j$  后, 计算正对负对的相似度  $\text{sim}(G'_i, G''_i), \text{sim}(G'_i, G'_j)$ 。定义相似度函数  $\text{sim}(x, y)$  为:

$$\text{sim}(x, y) = \frac{x \cdot y}{\|x\|_2 \|y\|_2}$$

则对于所有样本  $f$  的损失期望为:

$$\text{Loss}(f) = E(-\ln \frac{e^{\text{sim}(G'_i, G''_i)}}{\sum_{j=1}^N e^{\text{sim}(G'_i, G'_j)}})$$

这里考虑更一般的情况, 将被选中的锚点样本记为  $x$ , 与  $x$  有相同标签的正样本记为  $x^+$ , 与  $x$  有不同标签的负样本记为  $x^-$ , 记整个样本空间中  $x$  的分布为  $p(x)$ 。假设当前  $x$  的标签为  $c$ , 所有标签  $c$  的样本在  $p(x)$  中的分布为  $q(c)$ , 则可得  $x$  与  $c$  的联合分布为  $p_{x,c}(x, c) = p(x \mid c)q(c)$ 。令  $h()$  是标签的映射函数, 则  $h(x) = h(x^+), h(x) \neq h(x^-)$ , 所以与锚点样本  $x$  标签相同的数据的分布可以表示为  $p_x^+(x') = p(x' \mid h(x') = h(x))$ , 与  $x$  标签不同的数据分布为  $p_x^-(x') = p(x' \mid h(x') \neq h(x))$ 。假设  $q(c) = \pi^+$  在样本空间中的分布是均匀的, 则  $\pi^- = 1 - \pi^+$  来表示取到其它不同标签样本的概率。本文规定将 “ $a \sim b$ ” 视为从分布  $b$  抽样到了样本  $a$ , 则理想情况下, 编码器  $f$  的损失期望可以写成:

$$E_{\substack{x \sim p, \\ x^+ \sim p_x^+, \\ \{x_i^-\}_{i=1}^N \sim p_x^-}} \left[ -\ln \frac{e^{\text{sim}(f(x), f(x^+))}}{e^{\text{sim}(f(x), f(x^+))} + \sum_{i=1}^N e^{\text{sim}(f(x), f(x_i^-))}} \right]$$

在实际实验中,图对比学习的负采样  $N$  通常与小批次保持一致,或者在整个内存库中取相同数量。在文献<sup>[18]</sup>中作者将负样本进行相似度排序后逐渐加大样本空间并采用均值采样,同时对采样的数量进行了研究,研究表明在顺序采样下,负采样数量和批次大小脱钩后可以取得更好的效果。由此,可以对负采样进行一定的操作,在这里引入负样本权重  $W$  作为系数,令  $W = \frac{Q}{N}$ , 其中  $\text{sim}(f(x), f(x_i^-))$  的取值在实际实验中一般不会重复,且在每次抽样过程中抽取的概率相等,于是上式可以改写为(下文简记  $e^{\text{sim}(f(x), f(x^+))}$  为  $e^{\text{sim}(+)}$ ):

$$E_{\substack{x \sim p, \\ x^+ \sim p_x^+, \\ \{x_i^-\}_{i=1}^N \sim p_x^-}} \left[ -\ln \frac{e^{\text{sim}(+)}}{e^{\text{sim}(+)} + \frac{Q}{N} \sum_{i=1}^N e^{\text{sim}(f(x), f(x_i^-))}} \right] =$$
$$E_{\substack{x \sim p, \\ x^+ \sim p_x^+}} \left[ -\ln \frac{e^{\text{sim}(+)}}{e^{\text{sim}(+)} + QE_{x^- \sim p_x^-} [e^{\text{sim}(f(x), f(x^-))}]} \right]$$

由于现实情况下,自监督的图对比学习只能获得不带标签的数据,以及由锚点样本增广而产生的增广样本,后者可视为正样本空间而前者并不能完全视为负样本空间,所以这里需要引入 PU learning 的知识。在无偏 PU learning(uPU learning)<sup>[19]</sup>中,可访问一个正的样本空间  $\mathcal{X}^+$ , 和无标注的样本空间  $\mathcal{X}$ , 即:

$$\text{Loss}_{\text{CL}} = \left[ -\ln \frac{e^{\frac{\text{sim}(G_i^+, G_i^+)}{T}}}{e^{\frac{\text{sim}(G_i^+, G_i^+)}{T}} + N \max \left\{ \frac{1}{1-\pi} \left( \frac{1}{N} \sum_{j=1}^N e^{\frac{\text{sim}(G_i^+, G_j^+)}{T}} - \pi e^{\frac{\text{sim}(G_i^+, G_i^+)}{T}} \right), e^{-\frac{1}{T}} \right\}} \right] \quad (5)$$

其中  $\pi$  为超参数,  $N$  为全样本空间的样本数量。则最终模型的损失函数为:

$$\text{Loss} = \lambda \text{Loss}_{\text{view}} + \text{Loss}_{\text{CL}}$$

其中  $\lambda$  为损失系数,默认取 1。

### 1.5 分类层

理想情况下,对比学习可以将样本的嵌入表示按照相似度大小,均匀地分布在一个超球面上<sup>[21]</sup>, 所以使用线性分类器就可以很容易地把某类与其他类分开。本文使用两层线性层来实现分类。

$$y_1 = \text{Relu}(\text{Fullyconnected}(G_i))$$
$$y_2 = \text{Softmax}(\text{Fullyconnected}(y_1))$$

## 2 实验

### 2.1 数据集和数据预处理

本文在 2 个不同的中文数据集 THUCnews<sup>[22]</sup>、

$$\mathcal{X} = \{x_j\}_{j=1}^n = p(x)$$
$$\mathcal{X}^+ = \{x_i\}_{i=1}^{n'} = p(x | y = 1) \quad (2)$$
$$p(x) = \pi p(x | y = 1) + (1 - \pi) p(x | y \neq 1)$$

则可以将本文中(2)的分布重写为(3),并得到负采样的表达式(4):

$$p(x') = \pi p_x^+(x') + (1 - \pi) p_x^-(x') \quad (3)$$
$$p_x^-(x') = \frac{p(x') - \pi p_x^+(x')}{1 - \pi} \quad (4)$$

需要说明的是,此时的  $f(x)$  并不是概率密度函数,而是将抽样对象  $x$  映射到向量空间的编码器。在  $e^{\text{sim}(f(x), f(x^-))}$  中,固定锚点样本  $x$  将  $f(x)$  视为常数,则  $y = e^{\text{sim}(+)}$  可被视为一个关于负采样  $x^-$  的函数,则可将负采样期望  $E_{x^- \sim p_x^-}$  改写为:

$$E_{x^- \sim p_x^-} [e^{\text{sim}(f(x), f(x^-))}] =$$
$$\frac{E_{x^- \sim p} [e^{\text{sim}(f(x), f(x^-))}] - \pi E_{x^- \sim p_x^+} [e^{\text{sim}(f(x), f(x^-))}]}{1 - \pi}$$

此时的全样本空间可以是整个内存库也可以是一个批次内的样本,正样本空间只有增广后的一对样本,如果将其中一个视为锚点样点,那么正样本空间只有一个样本。由此可以计算对应的损失期望:

$$\frac{1}{1 - \pi} (E_{x^- \sim p} [e^{\text{sim}(f(x), f(x^-))}] - \pi E_{x^- \sim p_x^+} [e^{\text{sim}(f(x), f(x^-))}])$$
$$= \frac{1}{1 - \pi} \left( \frac{1}{N} \sum_{i=1}^{N_i} e^{\text{sim}(f(x), f(u_i))} - \pi e^{\text{sim}(+)} \right)$$

其中  $u$  为无标注的样本。由于对比学习损失函数的负对项的理论最小值为  $e^{-1}$ , 所以本文要求当上式的值小于  $e^{-1}$  时,取  $e^{-1}$ 。出于简单考虑,设  $W=1$ , 则  $Q=N$ , 则最终加入温度系数  $T$ <sup>[20]</sup> 后的修正损失函数为:

$$\text{Loss}_{\text{CL}} = \left[ -\ln \frac{e^{\frac{\text{sim}(G_i^+, G_i^+)}{T}}}{e^{\frac{\text{sim}(G_i^+, G_i^+)}{T}} + N \max \left\{ \frac{1}{1-\pi} \left( \frac{1}{N} \sum_{j=1}^N e^{\frac{\text{sim}(G_i^+, G_j^+)}{T}} - \pi e^{\frac{\text{sim}(G_i^+, G_i^+)}{T}} \right), e^{-\frac{1}{T}} \right\}} \right] \quad (5)$$

SogouCS<sup>[23]</sup>上进行了实验,验证了在长文本分类上的有效性,各数据集的相关信息见表 1。

表 1 数据集统计详情

	数据集	
	THUCNews	SogouCS
均长	883.13	727.78
数量	65 000	65 000
类目	10	10
最短类均长	437.24	690.60
最长类均长	1568.73	762.16

THUCnews 是清华大学根据新浪新闻 RSS 订阅频道 2005 到 2011 间的历史数据筛选而成,有 14 个类别共 74 万多条数据组成。SogouCS 是来自搜狐新闻 2012 年 6 月到 7 月间共 18 个频道的新闻数据,本文以频道类别作为数据标签,对其中足够数量的类目进行筛选。选取好数据后进行数据清洗,首



先依照哈工大中文停词表删除文本中大量无意义的词,再删除如网址、邮箱、电话号码等无意义但形式固定的内容。

实验里,在 THUCNews 中抽取 65 000 条长文本数据。为了兼顾一定的泛化性,本文选取文本最短长度可在 400,但总平均长度大于 600 的数据,并在对 65 000 条数据进行预训练后,以 5 : 1 : 0.5 的比例划分训练集、测试集与验证集。在 SogouCS 中抽取 65 000 条长文本数据,由于部分标签数据数量不足,本文仅抽取长度大于 300 且平均长度大于 600 的数据,并在对 65 000 条数据进行预训练后,以 5 : 1 : 0.5 的比例划分训练集、测试集与验证集。

### 2.2 评估指标与参数设置

本文基于三个评估指标精确率(Precision)、召回率(Recall)和 F1 值(F1\_Score)进行性能评估,有:

$$\begin{aligned} \text{Precision} &= \frac{TP}{TP+FP} \\ \text{Recall} &= \frac{TP}{TP+FN} \\ \text{F1\_Score} &= \frac{2PR}{P+R} \end{aligned}$$

其中,  $FP$  表示被预测为正的负样本数量,  $FN$  表示被预测为负的正样本数量,  $TN$  表示被预测正确的负样本数量,  $TP$  表示被预测正确的正样本数量,  $P$  为精准率,  $R$  为召回率。

参数设置:在对比学习阶段使用两层的 GAT 和 0.0001 的学习率,32 的 batch 和 10 的 epoch,并且设置了多组关于温度系数  $T$  和修正系数  $\pi$  的实验,最终效果最好的温度系数  $T$  和修正系数  $\pi$  分别为 0.5 和 0.12。在分类阶段,使用 0.0001 的学习率,32 的 batch 和 100 的 epoch。

### 2.3 实验设计及参数分析

本文设计了对比实验、消融实验和参数分析实验。对比实验显示本文模型优于当前主流先进模型;消融实验分析了各个模块的作用;参数分析实验得出了温度系数  $T$  和修正系数  $\pi$  的最优值。

2.3.1 对比实验 TextRCNN<sup>[24]</sup>:结合 CNN 与 RNN 的算法,通过双向的 RNN 获取上下文信息来学习包含语境信息的字符表示,再通过最大池化获取值最大的字符来代表整个文本的嵌入表示。

BiLSTM-Attention<sup>[25]</sup>:通过双向的 LSTM 获得每个字符的向量表示,再通过 Attention 机制对所有向量进行加权求和从而得到文本的嵌入表示。

Capsule Network<sup>[26]</sup>:将 CNN 中的神经元合并成一个模块记为胶囊。与传统神经网络将隐藏层数据当作标量计算不同,胶囊网络的每一步计算都是向量计算。当某个低层胶囊的输出与高层胶囊的输

出方向较小甚至相反时,算法会减小这个低层胶囊对该高层胶囊的影响,在胶囊网络中这一过程被称作动态路由。输出的向量可以代表文本的特征,弥补了 CNN 不能理解语义关系的缺陷。

Longformer<sup>[27]</sup>:针对 BERT 模型仅能支持 512 个字符的问题而提出的可支持 4096 个字符的预训练模型。

BERT+NEBi-LSTM+HAN<sup>[28]</sup>:基于<sup>[29]</sup>提出的一种特征增强的非平衡 Bi-LSTM 模型(NEBi-LSTM)加上 BERT 对文本进行初步特征提取,最后用 HAN 从单词和句子两个方面对文本进行加权。

实验结果见图 3、4。

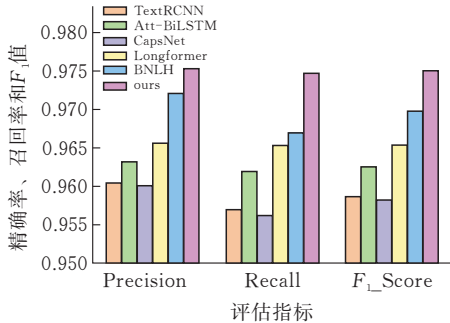


图 3 THUCnews 对比实验结果

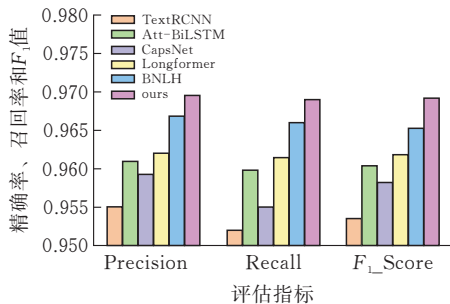


图 4 SogouCS 对比实验结果

实验分析:根据结果可以得出,本文提出的 GCL-AVGNSO 在两个数据集上均优于当前主流先进模型。原因如下:首先,文本非常适合利用高级结构来构建图模型。其次,不同的段落对于文本的贡献度不同,基于注意力机制的图模型能非常好的学习哪些段落对语义的贡献度更高。

TextRCNN:双向的 RNN 可以捕捉较长的语义信息,然而其存在越远的字符越会被重视的缺点,所以获得的上下文语境信息可能不准确。并且其是通过最大池化选取文本中最重要的含有语境信息的字符来表示整个文本,可以认为是通过字符来表示主题,再进行主题分类。关键在于文本中是否有可以代表文本主题的字符,并且通过最大池化的方法得到的字符是否就是目标字符,如果有一点不成立就会影响分类效果。

BiLSTM-Attention:BiLSTM 已经可以较好地

获取上下文的信息,并且通过注意力机制增强了文本的表示能力,实验结果也显示其好于胶囊网络。然而对于过长的文本,仍然不能很好的捕捉长距离依赖,并且通过拼接来融合前后向特征的方式不够好。对于长文本的分类任务,增加了文本结构信息的工作。

CapsNet:胶囊网络的核心路由算法与 BERT 的自注意力机制类似。自注意力机制通过  $Q$ 、 $K$ 、 $V$  三个矩阵计算出序列中其他字对当前字的权重以及加权后的向量表示,而胶囊网络通过路由权重矩阵计算所有的字符权重。相比之下,BERT 的自注意力机制比胶囊网络拥有更多的权重矩阵和更深的网络结构,并且 BERT 经过庞大的语料库进行预训练,因此基于 BERT 的分段文本工作会得到比胶囊网络更好的结果。

Longformer:Longformer 在长文本的语义表示上较传统模型已有很好提升,实验显示其效果差于基于 BERT 的分段文本工作,说明 Longformer 在远距离的语义表示上并不准确。

2.3.2 消融实验 BERT:将文本截取为 510 个字符再通过 BERT 获取文本表示。

BERT+GraphCL<sup>[30]</sup>:通过 GraphCL 的对比学习框架训练出一个 GAT,将其作为文本的特征表示器。

BERT+view+GCL:用自适应视图生成器 view 代替 GraphCL 中的数据增强模块,训练出一个 GAT 作为文本的特征表示器。

BERT+Debias+GCL:优化 GraphCL 中负采样模块,训练出一个 GAT 作为文本的特征表示器。

实验结果见图 5、6(中间三项省略 BERT 名称)。

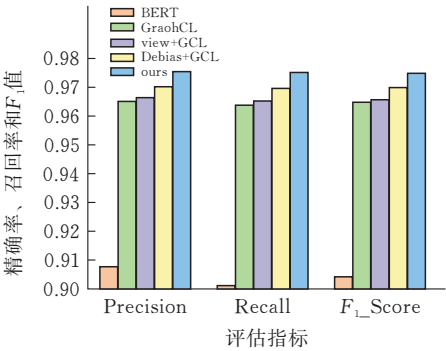


图 5 THUCnews 对比实验结果

根据结果可知,用 GraphCL 也可以得到很好的效果,但是需要大量的实验筛选出增广策略的超参数,本文并没有对其做大量的实验,所以得出中等偏上的结果是非常符合直觉的。同时用自适应视图生成器后效果略好于 GraphCL,在未做大规模实验来

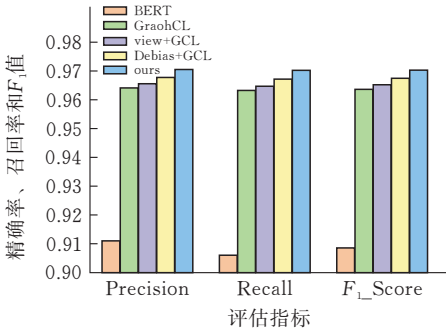


图 6 SogouCS 对比实验结果

选择数据增广策略时,用自适应视图生成器就可以得到不错的效果。接着修正负采样偏差后,效果好于 GraphCL,可以认为负采样偏差确实会影响对比学习的效果。最后 GCL-AVGNSO 的效果达到最优,说明本文提出的两项工作在文本分类的背景下,对图对比学习有加强作用。

2.3.3 参数分析实验 固定温度系数  $T=0.5$ ,再设置不同的修正系数  $\pi$ ,来选择对当前数据集修正负采样偏差最好的  $\pi$ 。固定修正系数  $\pi=0.1$ ,再设置不同的温度系数  $T$  来选择最好的温度系数。THUCnews 数据集结果见图 7、9,SogouCS 数据集结果见图 7-10。

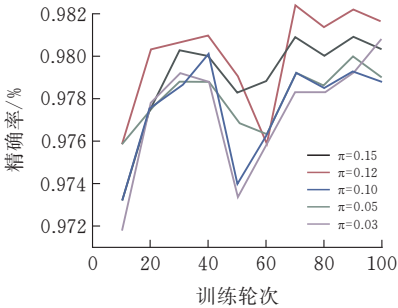


图 7 THUCnews  $\pi$  实验结果

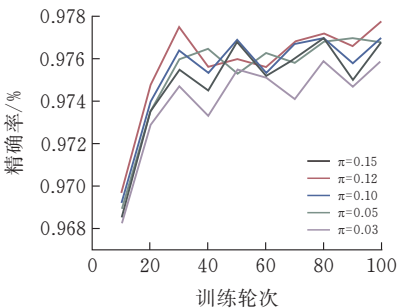


图 8 SogouCS  $\pi$  实验结果

从式(5)的推导显示,当所有类目中的实例数量相等时, $\pi$  的理论取值为某类目的实例总数除以总数据数。对于当前 THUCnews 和 SogouCS 数据集,理论最优值均为 0.1,但实验结果显示 0.12 为最优值,说明数据集中某些类在相似度上较为靠近。

$T$  的取值不是越小越好,当  $T=0.5$  时能最优地区分正负样本。

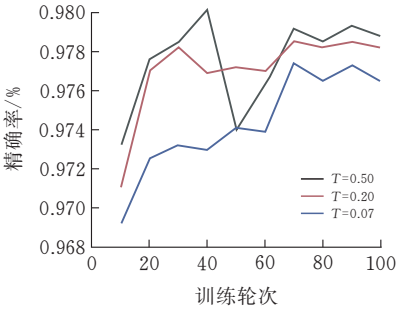


图 9 THUCnews  $T$  实验结果

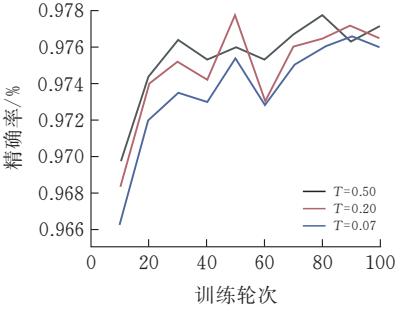


图 10 SogouCS  $T$  实验结果

3 结束语与展望

本文利用文本的高级结构构建图模型,通过对比学习的方法学习一个图神经网络来获得文本的嵌入表示,并在细粒度上适配任何基于 Transformer 的文本预训练模型。在对比学习的数据增广阶段,基于传统 NLP 领域数据增广中将一篇文本随机划分为两个文本的方法,引入一个自适应视图生成器,利用段落本身的属性,能在微观上让每个段落都自发地选择增广方式,同时在宏观上也实现了文本可以自发的选择划分比例。在负采样阶段,通过引入修正系数  $\pi$ ,实现在仅访问正样本分布和全样本分布下对负采样进行修正。在两个数据集上进行实验对比,结果显示本文方法好于主流先进模型。

本文是通过引入图节点的属性来实现自适应视图生成器,相比于 GraphCL 而言少了对边的利用。由此在构造图的时候,如何利用段落间的关系来对边进行赋值便成了很直观的问题。如果有很好的对边赋值的方法,那么就可以让文本的图结构更加多样化,数据增广的策略也会相应变多,或许可以得到更好的效果。同时,对比学习如果训练过多会导致数据间的距离被拉得过开,如何设置停止机制也将是未来研究的重点。

[ 参 考 文 献 ]

[1] KOWSARI K, JAFARI MEIMANDI K, HEIDAR-YSAFA M, et al. Text classification algorithms: A survey[J]. Information, 2019, 10(04): 150.

[2] MEDHAT W, HASSAN A, KORASHY H. Sentiment analysis algorithms and applications: A survey[J]. Ain Shams engineering journal, 2014, 5(04): 1093-1113.

[3] YATES A, NOGUEIRA R, LIN J. Pretrained transformers for text ranking: BERT and beyond[C] // Proceedings of the 14th ACM International Conference on Web Search and Data Mining. 2021: 1154-1156.

[4] MA X, ZHU Q, ZHOU Y, et al. Improving question generation with sentence-level semantic matching and answer position inferring [C] // Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(05): 8464-8471.

[5] WANG Z, LIU X, YANG P, et al. Cross-lingual text classification with heterogeneous graph neural network [C] // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, (ACL/IJCNLP) 2021, (Volume 2: Short Papers), 2021: 612-620

[6] CHAFFAR S, INKPEN D. Using a heterogeneous dataset for emotion analysis in text[C] // Canadian conference on artificial intelligence. Springer, Berlin, Heidelberg, 2011: 62-67.

[7] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[C] // 1st International Conference on Learning Representations, (ICLR 2013), Scottsdale, Arizona, USA, 2013. <http://arxiv.org/abs/1301.3781>.

[8] LILLEBERG J, ZHU Y, ZHANG Y. Support vector machines and word2vec for text classification with semantic features [C] // 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI \* CC). IEEE, 2015: 136-140.

[9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C] // Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. 2017: 5998-6008.

[10] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C] // Proceedings of NAACL-HLT. 2019: 4171-4186.

[11] VAN DEN OORD A, LI Y, VINYALS O. Representation learning with contrastive predictive coding[C]. CoRR, 2018, abs/1807. 03748. <http://arxiv.org/abs/1807.03748>.

[12] SUN C, QIU X, XU Y, et al. How to fine-tune bert for text classification? [J] // China national conference on Chinese computational linguistics. Springer, Cham, 2019: 194-206.

[13] MOHANTY I, GOYAL A, DOTTERWEICH A. Emotions are subtle: learning sentiment based text representations using contrastive learning[J/OL]. [2021-04-

- 15].CoRR, 2021, abs/2112.01054. <https://arxiv.org/abs/2112.01054>.
- [14] XU P, CHEN X, MA X, et al. Contrastive Document Representation Learning with Graph Attention Networks[C]// Findings of the Association for Computational Linguistics; EMNLP 2021; 3874 - 3884.
- [15] DU PLESSIS M, NIU G, SUGIYAMA M. Convex formulation for learning from positive and unlabeled data [C] // International conference on machine learning. PMLR, 2015; 1386-1394.
- [16] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[J]. stat, 2017, 1050; 20.
- [17] JANG E, GU S, POOLE B. Categorical reparameterization with gumbel-softmax[C]// 5TH International conference on learning representations, (ICLR 2017), Toulouse, France, 2018. <https://openreview.net/forum?id=rke3y85ee>.
- [18] CHU G, WANG X, SHI C, et al. CuCo: Graph representation with curriculum contrastive learning[C]// Proc. IJCAI, 2021; 2300-2306.
- [19] KIRYO R, NIU G, DU PLESSIS M C, et al. Positive-unlabeled learning with non-negative risk estimator[C]// Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. 2017; 1675-1685.
- [20] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J/OL]. [2022-04-15]. CoRR, 2015, abs/1503.02531. <http://arxiv.org/abs/1503.02531>.
- [21] WANG F, LIU H. Understanding the behaviour of contrastive loss[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021; 2495-2504.
- [22] 李景阳, 孙茂松. Non-independent term selection for Chinese text categorization[J]. Tsinghua Science and Technology, 2009(01); 115-122.
- [23] WANG C, ZHANG M, MA S, et al. Automatic online news issue construction in web environment[C]// Proceedings of the 17th International Conference on World Wide Web, WWW '08, pages 457-466, New York, NY, USA, 2008. ACM.
- [24] LAI S, XU L, LIU K, et al. Recurrent convolutional neural networks for text classification[C]// Twenty-ninth AAAI conference on artificial intelligence. 2015.
- [25] ZHOU P, SHI W, TIAN J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C] // Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers). 2016; 207-212.
- [26] KIM J, JANG S, PARK E, et al. Text classification using capsules[J]. Neurocomputing, 2020, 376; 214-221.
- [27] BELTAGY I, PETERS M E, COHAN A. Longformer: The long-document transformer[J/OL]. [2022-04-15]. CoRR, 2020, abs/2004.05150. <https://arxiv.org/abs/2004.05150>.
- [28] HUAN H, YAN J, XIE Y, et al. Feature-enhanced non-equilibrium bidirectional long short-term memory model for Chinese text classification [J]. IEEE Access, 2020, 8; 199629-199637.
- [29] YANG Z, YANG D, DYER C, et al. Hierarchical attention networks for document classification[C] // Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. 2016; 1480-1489.
- [30] YOU Y, CHEN T, SUI Y, et al. Graph contrastive learning with augmentations[J]. Advances in Neural Information Processing Systems, 2020, 33; 5812-5823.

## A Long Text Classification Model Based on Graph Contrast Learning

LIU Yuhao, GAO Rong, YAN Lingyu, YE Zhiwei

(School of Computer Science, Hubei Univ. of Tech., Wuhan 430068, China)

**Abstract:** The current text classification methods based on character-level consideration have the problems of computational difficulty due to the large input dimension and the difficulty of capturing the long-distance relationship due to the long content, which leads to a lack of accuracy in long text classification. Thus, the proposed graph contrast learning long text classification model is based on an adaptive view generator and negative sampling optimization. Specifically, the long text is first divided into several paragraphs, and the paragraphs are embedded with the BERT-derived model, then the graph model is constructed based on the high-level structure of the text by considering the embedded representation of the paragraphs as nodes, then the graph is augmented using the adaptive view generator, and the embedded representation of the text is obtained by graph contrast learning, while PU learning knowledge is introduced to alleviate the problem of negative sampling bias in the negative sampling phase of graph contrast learning, and finally the obtained embedded representation of the text is classified using two linear layers. Experiments on two Chinese datasets show that the method outperforms mainstream advanced models.

**Keywords:** text representation; long text classification; graph contrastive learning; negative sampling

[责任编辑: 张岩芳]