

潜在语义聚类与图交互驱动的少样本文本分类

王世宇, 周 刚, 卢记仓

(信息工程大学, 河南 郑州 450001)

摘要: 针对少样本长尾分类中尾部类别样本稀缺、预训练语言模型拟合不足及现有图方法依赖全局静态结构的问题, 提出一种基于潜在语义聚类与图交互的文本分类方法。首先, 利用标签描述对语言模型进行语义预热, 使文本与标签对齐到统一表示空间; 其次, 通过无监督聚类构建可学习的潜在主题节点, 作为文本与标签的语义中继, 以缓解尾部类别语义稀疏; 最后, 为每个文本实例动态构建独立的异质子图, 通过图神经网络聚合局部信息, 并以交叉注意力机制增强文本表征。在 5 个基准数据集上的实验结果表明, 该方法与 BERT、RoBERTa、DeBERTa 等 3 类基准微调模型结合后, 宏平均 $F1$ 值平均提升 2.91 个百分点、尾部类别宏平均 $F1$ 值平均提升 3.21 个百分点, 尾部性能提升更为显著, 且显存开销远低于全局图方法, 为少样本不平衡分类提供兼顾性能与效率的解决方案。

关键词: 文本分类; 语义聚类; 图神经网络; 预训练模型; 类分布不均衡

中图分类号: TP391.1

文献标识码: A

文章编号: 1671-0673(2026)03-0345-09

Latent Semantic Clustering and Graph Interaction Driven Few-Shot Text Classification

WANG Shiyu, ZHOU Gang, LU Jicang

(Information Engineering University, Zhengzhou 450001, China)

Abstract: To address the issues of scarce tail-class samples, insufficient fitting of pre-trained language models, and the reliance on global static structures in existing graph-based methods for few-shot long-tail classification, a text classification method based on latent semantic clustering and graph interaction is proposed. Firstly, the language model is semantically warmed up using label descriptions to align texts and labels into a unified representation space. Secondly, learnable latent topic nodes are constructed through unsupervised clustering to serve as semantic relays between texts and labels, thereby alleviating the semantic sparsity of tail categories. Finally, an independent heterogeneous subgraph is dynamically constructed for each text instance, local information is aggregated via a graph neural network, and text representations are enhanced through a cross-attention mechanism. Experimental results on 5 benchmark datasets demonstrate that after integrating the proposed method with 3 fine-tuned baseline models (BERT, RoBERTa, and DeBERTa), the macro $F1$ score of the models is improved by 2.91 percentage points, and the macro tail $F1$ score is improved by 3.21 percentage points, showing a more significant enhancement in tail performance. Furthermore, its memory overhead is significantly lower than that of global graph methods, offering an effective and efficient solution for few-shot imbalanced classification.

Key words: text classification; semantic clustering; graph neural network; pre-trained model; class distribution imbalance

文本分类作为自然语言处理(Natural Language Processing, NLP)的基础任务,广泛应用于新闻分类^[1]、立场检测^[2]等领域。随着深度预训练语言模型(Pre-trained Language Model, PLM)^[3-4]的兴起,该领域取得了显著进展。然而在实际应用中,模型往往面临着少样本(Few-Shot)和长尾分布(Long-Tail Distribution)的严峻挑战。当某些类别仅有极少量标注样本时,依赖海量数据驱动的PLM难以在微调过程中建立鲁棒的判别边界,极易陷入过拟合或对频繁类别的偏好。

现有少样本学习方法,如元学习^[5]、提示学习^[6]与原型网络^[7],虽部分缓解了数据稀缺问题,但仍面临本质性局限。首先,多数方法依赖单一原型表示类别,难以刻画同一标签下的细粒度语义差异。例如,“体育”类别中“高尔夫新闻”与“足球解说”的词汇分布迥异,强行压缩会丢失细粒度信息。其次,现有基于图神经网络(Graph Neural Network, GNN)的方法多采用直推式(Transductive)学习^[8],需构建包含全部测试样本的全局图,计算开销大且无法适应动态的归纳式(Inductive)场景。

针对上述问题,提出一种基于潜在语义聚类与图交互的端到端网络(Latent-Induced Subgraph Network, LIS-Net),模拟人类在面对生僻概念时的认知过程:不仅仅依赖记忆(预训练知识),而是主动检索相关的上下文线索进行辅助推理。主要贡献在于:提出一种通过无监督聚类初始化并动态更新潜在主题节点的机制,以捕捉细粒度语义信息;设计基于检索的实例级子图构建方法,实现高效归纳式推理;通过交叉注意力机制实现PLM与GNN的深度交互,使文本表示能够融合结构化知识。实验表明,其在少样本场景下性能显著优于现有基线模型,并对长尾数据具有更强鲁棒性。

1 相关工作

1.1 少样本文本分类与PLM

少样本文本分类旨在仅凭极少量标注样本识别文本类别。早期的研究主要基于度量学习^[7],它们试图在一个共享的度量空间中学习样本与类中心之间的距离。随着PLM如BERT^[3]、RoBERTa^[4]的出现,基于微调(Fine-Tuning)的方法成为主流。为进一步挖掘PLM的潜力,提示学习^[6]方法被提出,试图将分类任务重构为掩码预测或自然语言推理任务,以弥合预训练与下游任务的差异。

然而,上述方法大多存在一个共同的局限性:

语义建模的单一性。它们通常假设每个类别可以被单一的原型或提示向量所代表,忽略了复杂的类内多样性(Intra-Class Diversity)。当一个类别包含多个语义子簇时,强制聚合会导致特征空间的过平滑。本文通过引入潜在主题(Latent Topics),来建模细粒度的语义变化,从而增强PLM在少样本场景下的判别能力。

1.2 GNN与文本图模型

GNN因其强大的关系推理能力,被广泛应用于文本分类任务。经典的TextGCN^[8]在构建包含文档和单词的全局异质图上取得成功。随后,BertGCN^[9]将PLM的上下文嵌入与图卷积模型结合,取得了显著进展。尽管效果显著,但这些方法多属于直推式学习,需要在训练阶段构建一个包含所有训练和测试节点的全局静态大图。这种设定存在两大缺陷:无法对新增样本进行即时推理(需重新构图训练),且易引入不相关的长距离噪声。

相比之下,归纳式方法为每个文本构建独立的文本图,具备更好的可扩展性。例如:TextING^[10]构建单词局部共现图;HyperGAT^[11]利用超图建模主题;DADGNN^[12]引入注意力扩散机制;PaSIG^[13]将PLM文本编码器纳入归纳式学习框架等。本文提出了一种归纳式的图构建策略:不为每个样本构建包含全局信息的图,而是动态检索与当前输入最相关的主题和标签节点,形成轻量级实例子图。该方法支持对未见样本的即时推理,并能有效过滤无关噪声,聚焦于局部关键语义交互。

2 LIS-Net方法

2.1 框架概览

针对少样本长尾文本分类中存在的特征稀疏与上下文缺失问题,提出一种基于潜在语义聚类与图交互的端到端网络(LIS-Net)。该框架的核心在于联合优化特征提取与结构化推理,架构如图1所示,主要由以下4个组件构成。1)语义预热:利用自然语言标签描述对编码器进行初步对齐,构建统一的特征空间;2)潜在主题更新:通过无监督聚类算法初始化主题表示,并在训练中动态更新;3)归纳式子图构建:摒弃构建全局静态图,改成为每个输入批次动态检索最相关的潜在主题与标签节点,构建轻量级的实例级子图;4)跨模态融合:在GNN聚合子图上下文信息的基础上,通过交叉注意力机制将结构化知识融合回文本表示,实现端到端的判别性增强。

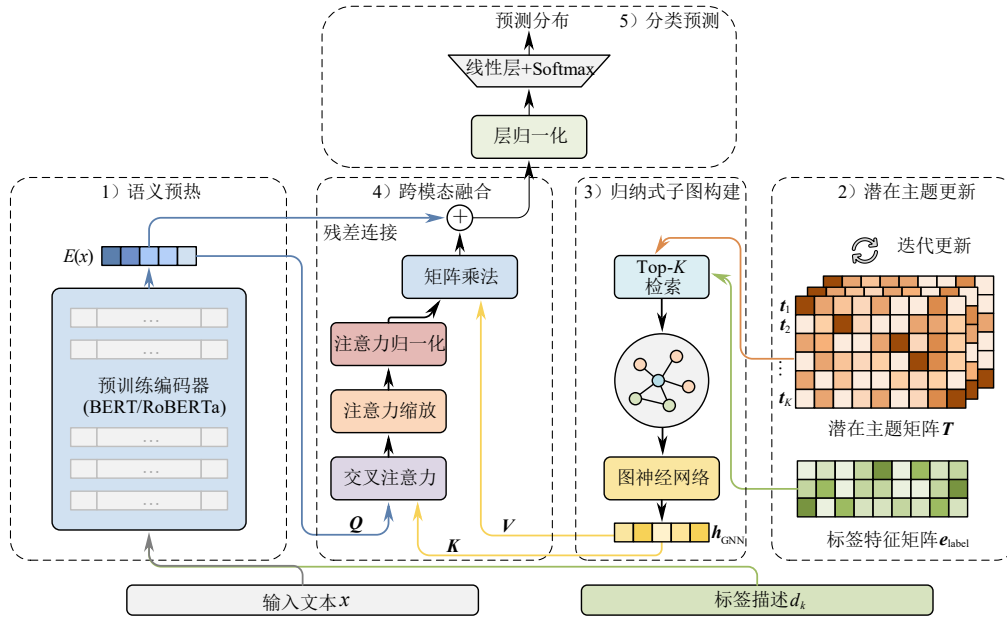


图1 LIS-Net框架示意图

2.2 语义预热

为使GNN能够在一个相对稳定的语义空间中运行,首先执行一个轻量级的预热(Warm-up)阶段。记输入文本集为 $\mathcal{X}$,标签集为 $\mathcal{Y}$ 。对于每个标签 $y_k \in \mathcal{Y}$,引入其自然语言描述 d_k (例如标签“体育”对应描述为“关于体育赛事的新闻和报道”)。将原始训练样本 (x_i, y_i) 与标签描述样本 (d_k, y_k) 混合,构成增强训练集 $\mathcal{D}_{\text{aug}}$ 。

使用PLM(如RoBERTa)作为骨干编码器 $E(\cdot)$ 。在预热阶段,优化标准的交叉熵损失:

$$\mathcal{L}_{\text{warmup}} = - \sum_{(x,y) \in \mathcal{D}_{\text{aug}}} \ln P(y|E(x)) \quad (1)$$

式中, $P(\cdot)$ 表示真实类别概率分布。

该过程有两个关键作用:一是让PLM适应特定领域的文本分布;二是强制将标签描述映射到与文本实例相同的嵌入空间中,为后续的图构建提供可靠的距离度量标准。预热结束后,将标签描述的编码向量固定,作为标签节点的初始特征 $\mathbf{e}_{\text{label}} \in \mathbb{R}^{|\mathcal{Y}| \times d}$ 。

2.3 潜在主题更新

在长尾少样本场景下,直接利用尾部类别的样本作为类中心往往存在偏差。为此,引入潜在主题概念。主题并非直接对应某个标签,而是代表数据在特征空间中的 K 个高密度聚类簇,它们捕捉了跨类别的细粒度语义模式。在预热后的所有文本嵌入上执行 K -Means聚类,得到 K 个聚类中心,初始化为主题矩阵 $\mathbf{T}^{(0)} \in \mathbb{R}^{K \times d}$ 。这些主题节点将在后续训练中作为连接文本与标签的中间桥梁,补充稀疏样本的语义信息。

2.4 归纳式子图构建

对于当前训练批次中的每一个文本实例 x_i ,其嵌入向量为 $\mathbf{h}_i = E(x_i)$,使用如下步骤为其动态构建一个以 x_i 为中心的局部异质子图 $\mathcal{S}_i = (\mathcal{V}_i, \mathcal{E}_i)$ 。

节点检索。子图的节点集合 $\mathcal{V}_i$ 包含3个部分:文本节点本身 $\{x_i\}$ 、检索到的相关主题子集 $\mathcal{T}_i \subset \mathbf{T}$ 以及相关标签子集 $\mathcal{L}_i \subset \mathcal{Y}$ 。此处采用基于余弦相似度的Top-K检索策略。1)主题检索:选取与 $\mathbf{h}_i$ 相似度最高的 K_{topic} 个潜在主题;2)标签检索:与 $\mathbf{h}_i$ 相似度最高的 K_{label} 个标签节点。这种策略确保子图仅包含与当前上下文最相关的信息,降低计算复杂度并减少无关噪声的干扰。

连边定义。在子图 $\mathcal{S}_i$ 中,定义3种类型的连边 $\mathcal{E}_i$ 。1)文本—主题边:连接 x_i 与其检索到的所有主题 $t \in \mathcal{T}_i$;2)文本—标签边:连接 x_i 与其检索到的所有标签 $l \in \mathcal{L}_i$,建立从文本到潜在标签的直接通道;3)标签—标签边:为引入标签间的先验依赖(如共现关系),预先计算一个全局的标签互近邻图。如果检索到的两个标签 $l_a, l_b \in \mathcal{L}_i$ 在全局图中存在连接,则在子图中保留该边。

通过上述过程,将原本孤立的文本样本 x_i 转化为一个包含丰富上下文信息结构化实例 $\mathcal{S}_i$ 。

2.5 动态图推理

构建好归纳式子图后,利用图结构信息来丰富文本的语义表示。不同于标准的GNN应用,LIS-Net引入动态的节点更新机制和融合策略。

1)动态主题更新。在训练过程中,潜在主题矩阵 $\mathbf{T}$ 需要适应特征空间的细微变化。简单的梯

度更新容易导致模式塌缩,即多个主题收敛到同一个语义中心。为此,设计一种基于注意力的更新机制。对于一个批次中的文本嵌入 $\mathbf{H}_{\text{batch}}$ 和当前主题 $\mathbf{T}$,首先计算分配概率矩阵 $\mathbf{P}$:

$$P_{ij} = \exp(\mathbf{h}_i \cdot \mathbf{t}_j / \tau) / \sum_k \exp(\mathbf{h}_i \cdot \mathbf{t}_k / \tau) \quad (2)$$

式中, τ 表示温度系数。主题更新通过聚合文本特征实现: $\mathbf{T}' = \mathbf{P}^T \mathbf{H}_{\text{batch}} \odot$

2) 标签关联传播。标签之间的依赖关系并非总是静态的,往往受具体上下文影响。因此,设计一个门控传播层来动态调节标签间的信息流。对于标签节点 u 和其邻居 $v (v \in \mathcal{N}(u))$, 其中 $\mathcal{N}(u)$ 表示节点 u 的邻居集合, 计算门控系数:

$$\alpha_{uv} = \sigma(\gamma \cdot \cos(\mathbf{e}_u, \mathbf{e}_v)) \quad (3)$$

式中: γ 表示缩放因子; $\sigma(\cdot)$ 表示 Sigmoid 函数。更新后的标签表示为

$$\mathbf{e}'_u = \beta \mathbf{e}_u + (1 - \beta) \sum_{v \in \mathcal{N}(u)} \alpha_{uv} (\mathbf{W}_l \mathbf{e}_v) \quad (4)$$

式中: β 表示可学习标量,用于自适应控制邻居标签信息与节点特征信息的融合比例; $\mathbf{W}$ 表示标签关联权重矩阵。软门控机制允许模型根据语义相似度自动过滤掉不相关的噪声连接,增强标签表示的鲁棒性。

3) 异质消息聚合。在子图 $\mathcal{S}_i$ 上,采用分层的异质聚合策略更新文本节点表示。在第 s 层,文本节点同时接收来自文本节点自身、邻居主题节点和邻居标签节点的消息:

$$\begin{aligned} \mathbf{h}_i^{(s+1)} &= \text{LayerNorm} \left(\text{ReLU} \left(\mathbf{W}_i \mathbf{h}_i^{(s)} + \mathbf{h}_{\text{topic}}^{\text{agg}} + \mathbf{h}_{\text{label}}^{\text{agg}} \right) \right), \\ \mathbf{h}_{\text{topic}}^{\text{agg}} &= \sum_{t \in \mathcal{T}_i} \mathbf{t}_t, \mathbf{h}_{\text{label}}^{\text{agg}} = \sum_{l \in \mathcal{L}_i} \mathbf{e}_l \end{aligned} \quad (5)$$

这种加性聚合高效地融合了多源上下文,生成包含结构化先验的图增强表示 $\mathbf{h}_{\text{GNN}}$ 。

2.6 跨模态融合与联合优化

将 GNN 输出的上下文向量 $\mathbf{h}_{\text{GNN}}$ 视为辅助知识,通过以下交叉注意力融合策略重新注入 PLM 的文本表示 $\mathbf{H}_{\text{PLM}}$ 。首先,将 $\mathbf{h}_{\text{GNN}}$ 分别投影到键 ($\mathbf{K}$) 和值 ($\mathbf{V}$) 空间。同时,将 $\mathbf{H}_{\text{PLM}}$ 投影到查询 ($\mathbf{Q}$) 空间,本文方法直接用 $\mathbf{H}_{\text{PLM}}$ 作为 $\mathbf{Q}$,以减少训练参数。

$$\begin{cases} \mathbf{Q} = \mathbf{H}_{\text{PLM}} \in \mathbb{R}^{N \times d_i}; \\ \mathbf{K} = \mathbf{h}_{\text{GNN}} \mathbf{W}_K \in \mathbb{R}^{M \times d_i}, \mathbf{W}_K \in \mathbb{R}^{d_i \times d_i}; \\ \mathbf{V} = \mathbf{h}_{\text{GNN}} \mathbf{W}_V \in \mathbb{R}^{M \times d_i}, \mathbf{W}_V \in \mathbb{R}^{d_i \times d_i}. \end{cases} \quad (6)$$

之后,计算注意力权重,并应用于 $\mathbf{V}$:

$$\mathbf{A} = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax} \left(\frac{\mathbf{Q} \mathbf{K}^T}{\sqrt{d_k}} \right) \mathbf{V} \quad (7)$$

式中, $\sqrt{d_k}$ 表示缩放因子,防止内积过大导致梯度消失。最后,将注意力输出与原始 $\mathbf{H}_{\text{PLM}}$ 相加(残差连接),然后进行层归一化,得到最终表征:

$$\mathbf{Z}_{\text{final}} = \text{LayerNorm}(\mathbf{H}_{\text{PLM}} + \mathbf{A}) \quad (8)$$

模型训练采用交叉熵损失函数进行优化,将表示 $\mathbf{Z}_{\text{final}}$ 输入一个线性分类器,得到每个类别的 logits:

$$\mathbf{S} = \mathbf{Z}_{\text{final}} \mathbf{W}_{\text{CLS}} + \mathbf{b}_{\text{CLS}} \quad (9)$$

式中, $\mathbf{b}_{\text{CLS}}$ 表示用于分类预测的偏置向量。

随后对每个样本的 logits 进行 Softmax 操作,将其转换为概率分布:

$$\hat{y}_{ic} = \exp(s_{ic}) / \sum_{j=1}^C \exp(s_{ij}) \quad (10)$$

对每个样本计算预测概率分布与真实标签之间的交叉熵,并对所有样本取平均:

$$\mathcal{L}_{\text{CLS}} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C (y_{ic} \ln(\hat{y}_{ic})) \quad (11)$$

通过这种端到端训练机制,PLM 的特征提取器、GNN 和潜在主题矩阵同步更新,实现了语义理解与结构化推理的深度耦合。

3 实验

3.1 数据集选取与划分

在 Ohsumed、20NG、R8、R52、AGNews 这 5 个基准数据集上进行实验。为贴近实际应用场景,在划分训练集时遵循少样本设置,为每个类别采样至多 10 个训练样本;验证集和训练集设置较大的样本数并保持规模一致。为验证模型在类别不平衡场景下的泛化能力,采用一种基于幂律分布的长尾划分策略来构建训练集。具体流程如下。

1) 类别频数统计。给定包含 N 个样本的数据集,其标签序列记为 $\{y_i\}_{i=1}^N$ 。首先统计每个类别的原始频数,并按照频数从高到低对类别进行排序,得到类别排名 $\text{rank}(c)$ 。

2) 长尾分布生成。对于排名为 $\text{rank}(c)$ 的类别,其标记训练样本数量 n_c 由以下幂律关系决定: $n_c \propto \text{rank}(c)^{-\alpha}$ 。其中, α 为分布参数,控制长尾的倾斜程度。 n_c 被归一化到预设区间 $[K_{\min}, K_{\max}]$ 内,即

$$n_c = \max \left(K_{\min}, \min \left(K_{\max}, \left\lfloor K_{\min} + (K_{\max} - K_{\min}) \cdot \text{rank}(c)^{-\alpha} \right\rfloor \right) \right) \quad (12)$$

式中: $K_{\min}$ 与 $K_{\max}$ 分别表示每个类别最少与最多的训练样本数; $\lfloor \cdot \rfloor$ 表示向下取整。确保高频类别获得较多标记样本,而低频类别样本量显著减少,形

成长尾分布。

3)数据集划分。对于每个类别,随机选取 n_c 个样本作为训练集,其余样本进入候选池。候选池进一步按比例随机划分为验证集与测试集,同时保持各类别在数据中的原始分布,以反映真实评估场景。

4)参数设置。设 $\alpha = 1.0$ 、 $K_{\min} = 1$ 、 $K_{\max} = 10$,验证集与测试集占原始语料比例均为10%,该划分方法使训练集呈现明显的类别不均衡,旨在考察模型在少量样本类别上的学习能力,同时通过验证集与测试集保持原始分布,以评估模型在均衡评价下的泛化性能。表1展示了采用长尾划分后各子集的样本数量统计。

表1 不均衡划分下的数据集统计信息

数据集	训练集 样本数	验证集 样本数	测试集 样本数	类别 个数	文本 平均长度
Ohsumed	46	735	735	23	187.1
20NG	43	1 880	1 880	20	275.7
R8	30	764	764	8	102.4
R52	75	902	902	52	109.6
AGNews	22	897	897	4	38.4

3.2 基线方法

为全面评估 LIS-Net 的性能,选取3类基线方法进行对比。

1)直推式:TextGCN^[8]、BertGCN^[9]。

2)归纳式:TextING^[10]、HyperGAT^[11]、DADGNN^[12]、TextSSL^[14]、TextFCG^[15]与 PaSIG^[13]。LIS-Net 面向归纳式场景,所以重点与此类方法比较。

3)PLM: BERT^[3]、RoBERTa^[4]、DeBERTa^[16]。此组用于分析 LIS-Net 相对于基础 PLM 的性能增益。为确保公平,主要对比对象为同样基于 PLM 的 BertGCN 与 PaSIG。

3.3 实验设置

为构建异构图,设置主题节点与标签节点的数目相同,文本—主题边的最近邻数 $K_{\text{topic}} = 2$,文本—标签边 $K_{\text{label}} = 2$,标签—标签边的连接数为3,以捕捉不同节点类型间关系。训练分为两个阶段:预热阶段使用较大批量 $B_1 = 8$ 进行10轮微调;联合训练阶段则采用较小批量 $B_2 = 4$ 进行10轮训练,并使用两层 GNN 进行信息传播。预热阶段中,PLM 与分类器的学习率均设为 2×10^{-5} ;联合训练阶段中,PLM 学习率保持为 2×10^{-5} ,而 GNN、分类器及主题—标签映射模块的学习率调整为 10^{-3} ,以适配不同模块的训练动态。两个阶段均采用 10^{-2} 权重

衰减以抑制过拟合。

本实验对所有基线模型均进行了严格复现。所有基线模型均采用官方开源代码实现,超参数设置遵循官方默认配置,且统一采用与本文方法一致的数据划分策略。对基于 PLM 的 BertGCN 和 PaSIG,二者均选用 bert-base-uncased 作为基础 PLM 架构。为确保实验结果的可靠性与稳定性,每组模型训练过程均采用不同的随机数种子独立运行,最终实验结果为5组独立实验结果的平均值。所有实验均在 NVIDIA Tesla V100S GPU 硬件平台上完成。

3.4 评价指标

为全面评估模型在长尾分布下的性能,采用一套综合的评价体系,重点关注模型对整体及尾部类别的识别能力。整体性能采用宏平均 $F1$ 值进行评估,该指标计算每个类别 $F1$ 值的算术平均,对少数类别性能敏感,能够平等反映模型在所有类别上的表现;尾部类别性能则通过尾部类别宏平均 $F1$ 值专门衡量。具体而言,首先将所有类别按其训练样本数量从多到少排序,并均等划分为头部、中部和尾部3个区段。本文指标仅计算尾部区段(即样本最少的1/3类别)的宏平均 $F1$ 值,从而直接反映模型对稀疏、困难类别的分类能力。通过结合整体宏平均 $F1$ 值与尾部类别宏平均 $F1$ 值,该体系能够多维度、公平地比较不同方法,尤其凸显模型在提升长尾数据中尾部类别性能的有效性。

3.5 性能比较

为全面评估所提 LIS-Net 方法的有效性,与基线方法在长尾类别划分下进行了系统性比较,如表2所示。其中: $F1$ 表示宏平均 $F1$ 值;Tail $F1$ 表示尾部类别宏平均 $F1$ 值。加粗数字表示当前指标的最优结果。分析表明,长尾分布对所有模型的性能均构成严峻挑战,而 LIS-Net 框架展现出较强的改善能力。在多个 PLM 基线上,引入 LIS-Net 后,Tail $F1$ 获得显著提升。如以 RoBERTa 为基座时,在 20NG 和 R52 数据集上,Tail $F1$ 分别由 27.33%、48.94% 提升至 32.57%、54.02%。这说明 LIS-Net 通过构建“文本—主题—标签”异构图,为样本稀疏的尾部类别建立了更丰富的语义关联,促进知识从头部向尾部转移。

在长尾场景下,采用 PLM 作为节点编码器的 BertGCN 与 PaSIG 的性能差异显著。BertGCN 凭借直推式学习与完整的单词交互媒介,仍保持相对稳定,其 Tail $F1$ 在多数数据集中处于中等偏上水平。

而 PaSIG 因依赖训练集内单词构建子图,在数据极度稀疏时长尾性能严重衰退,凸显了在少样本长尾任务中维持图结构媒介完整性的关键作用。

值得注意的是,部分 PLM 模型(如 DeBERTa)在长尾设置下性能大幅下降,Tail *F1* 多趋近于零,反映了类别不均衡对模型的直接损害。与之相比,融合 LIS-Net 的 DeBERTa 版本在 AGNews 上展现出强大的纠偏能力,将 Tail *F1* 从 59.11% 显著提

升至 76.48%,说明 LIS-Net 不仅能增强图结构鲁棒性,还可帮助 PLM 有效抵御长尾分布的负面影响。

总体而言,LIS-Net 在长尾场景中发挥稳定与提升作用。实验表明,成功的关键在于融合 PLM 的深层语义理解与一个能应对数据稀疏的、结构完整的图学习框架。LIS-Net 通过构建文本异构图,有效避免了依赖有限词汇建图的缺陷,从而在长尾场景下取得优越性能。

表 2 长尾划分下各方法性能比较

单位:%

类别	方法	Ohsumed		20NG		R8		R52		AGNews	
		<i>F1</i>	Tail <i>F1</i>	<i>F1</i>	Tail <i>F1</i>	<i>F1</i>	Tail <i>F1</i>	<i>F1</i>	Tail <i>F1</i>	<i>F1</i>	Tail <i>F1</i>
直推式	TextGCN	16.83	18.34	23.74	29.95	65.00	64.95	35.42	36.05	47.07	49.72
	BertGCN	2.95	4.31	37.56	28.74	75.13	60.53	30.55	21.90	78.09	74.40
归纳式	TextING	7.92	10.05	28.10	31.05	54.92	50.77	14.51	18.18	74.71	71.56
	HyperGAT	3.26	4.65	2.83	0.62	14.16	8.55	14.68	1.75	22.83	26.45
	DADGNN	7.77	5.46	12.06	9.41	56.25	33.26	19.69	5.56	66.32	56.93
	TextSSL	7.35	12.56	17.70	14.47	55.43	45.30	6.86	3.59	66.62	60.29
	TextFCG	5.15	7.83	16.37	29.46	55.80	53.48	12.75	13.47	74.13	72.21
	PaSIG	5.83	6.08	19.49	15.31	39.89	14.20	13.03	8.78	64.02	60.47
PLM	BERT	14.67	13.33	22.33	9.90	70.52	62.63	45.59	36.26	70.11	68.40
	BERT+LIS-Net	18.51	13.71	28.77	15.16	75.68	68.89	44.57	37.39	71.14	70.21
	RoBERTa	30.00	26.90	54.11	27.33	74.73	55.94	61.95	48.94	72.40	67.94
	RoBERTa+LIS-Net	30.90	24.76	55.07	32.57	75.63	58.32	66.38	54.02	76.33	72.49
	DeBERTa	5.75	0.96	14.72	5.66	37.08	8.20	8.62	0.85	68.38	59.11
	DeBERTa+LIS-Net	5.74	1.04	16.99	6.89	36.60	6.75	11.52	1.85	80.76	76.48

3.6 消融研究

为深入分析 LIS-Net 各核心组件的贡献,在长尾划分场景下进行系统性消融研究。以 RoBERTa-base 为基座模型,通过逐一移除关键模块,在 5 个数

据集上进行性能评估,结果如表 3 所示。其中:GNN 表示 GNN 模块;CA 表示跨模态融合模块;LD 表示标签描述初始化;LP 表示标签传播;TE 表示主题更新。

表 3 消融不同组件对模型性能的影响

单位:%

模型	Ohsumed		20NG		R8		R52		AGNews	
	<i>F1</i>	Tail <i>F1</i>	<i>F1</i>	Tail <i>F1</i>	<i>F1</i>	Tail <i>F1</i>	<i>F1</i>	Tail <i>F1</i>	<i>F1</i>	Tail <i>F1</i>
Full	31.8	26.5	55.1	27.9	76.0	59.3	66.1	52.3	77.1	74.2
w/o GNN	30.6	25.1	54.0	27.3	75.2	57.9	64.0	51.2	76.3	72.4
w/o CA	25.5	18.9	53.2	27.4	61.5	21.3	38.5	25.7	42.3	25.5
w/o LD	24.4	23.3	48.5	17.6	54.0	12.8	26.0	13.8	76.8	73.9
w/o LP	31.6	25.3	54.2	27.4	75.8	58.3	65.3	52.2	76.3	72.5
w/o TE	30.9	26.4	54.2	27.9	73.4	51.8	65.7	50.8	74.9	70.2

1)w/o GNN。移除 GNN 后,模型退化为仅基于 PLM 与标签描述的检索基线,性能全面下降。例如在 R52 上,*F1* 由 66.1% 降至 64.0%,Tail *F1* 由 52.3% 降至 51.2%。这表明 GNN 通过子图消息聚合,有效整合了主题与标签的上下文知识。

2)w/o CA。将交叉注意力替换为残差连接导致

性能衰退,例如在 AGNews 上 Tail *F1* 由 74.2% 降至 25.5%,表明交叉注意力能够实现动态、有选择的知识融合,其交互能力优于简单的残差连接。

3)w/o LD。移除语义预热模块会损害性能,尤其是尾部类别。例如在 R8 上 Tail *F1* 由 59.3% 降至 12.8%,即使在训练中联合优化,仍与完整模型存

在差距。这凸显了标签描述对构建可靠图结构的基础性作用。

4) w/o LP。移除标签关联传播模块会引起轻微但普遍的下降,例如在AGNews上Tail $F1$ 从74.2%降至72.5%。利用标签间依赖关系有助于优化标签表示,对达到最优性能是必要的。

5) w/o TE。固定潜在主题而不进行动态更新同样会导致性能下滑,例如在R8上Tail $F1$ 由59.3%降至51.8%。这证明了动态更新能使主题节点自适应地捕捉细粒度语义,避免静态聚类偏差。

各组件均对模型性能具有明确贡献,尤其以标签描述初始化和跨模态融合机制的影响最为显著,如图2所示。

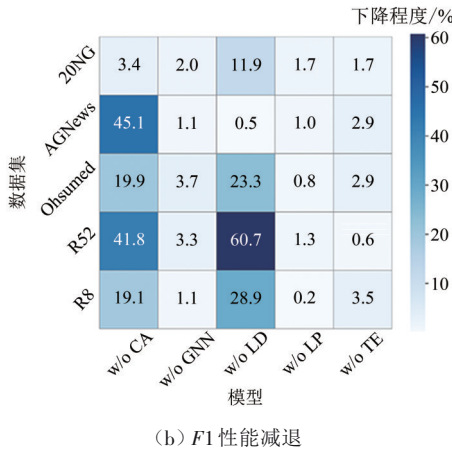
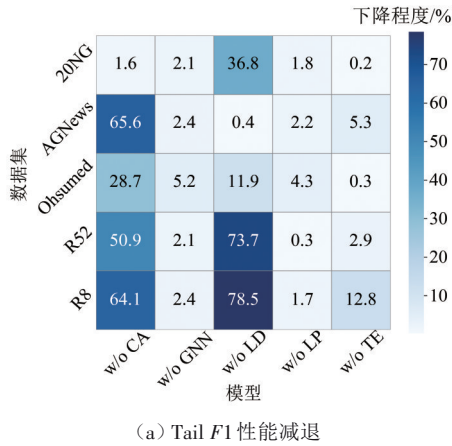


图2 移除组件后的性能下降程度

3.7 敏感性分析

系统分析LIS-Net在潜在主题数、GNN层数及近邻连接数3个核心结构参数下的鲁棒性。所有实验均在长尾场景下基于RoBERTa-base进行。

1) 潜在主题数。将主题数设为类别数的1倍、2倍和3倍。图3显示,模型性能对主题数变化高度鲁棒。例如:在R8与R52数据集上,不同设置下的Tail $F1$ 完全一致;在20NG上仅出现不足0.5个

百分点的轻微波动。潜在主题数的选择基于语义覆盖度与计算效率的综合考量。一方面,主题作为跨类别共享的语义基,当数量与类别数同量级时,已能有效建模类内子簇与类间重叠模式。另一方面,将主题数从1倍增至3倍,模型性能无明显提升,但子图构建与推理开销线性增长。因此,选取1倍类别数作为默认配置。

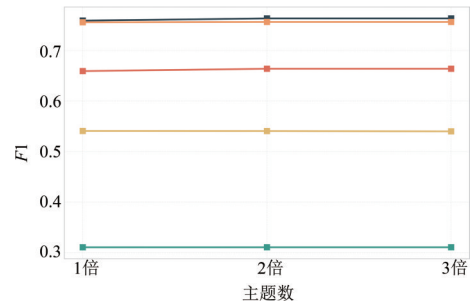
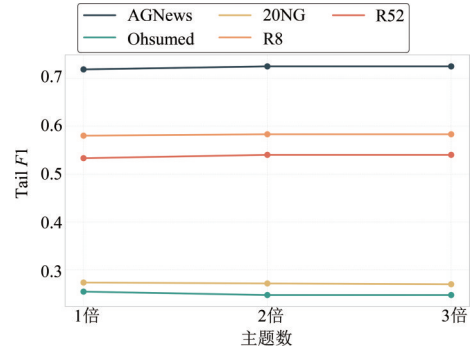
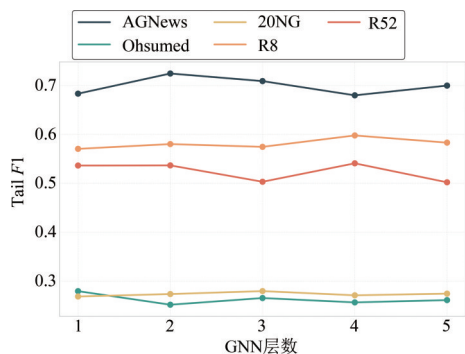


图3 主题数的影响

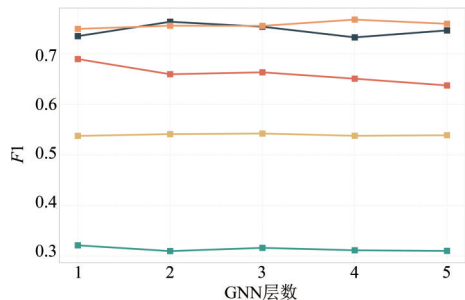
2) GNN层数。图4显示,部分数据集的性能随GNN层数呈现“倒U型”变化,最优层数因不同数据集而异。例如,AGNews在2层时取得最优Tail $F1$,而R8则在4层时表现最佳。总体来看,1层性能普遍较差,2层在多数数据集上表现稳健,可作为通用推荐配置。

3) 近邻连接数。文本—标签连接数 K_{label} 从1增至2时,性能在多数数据集上有显著提升,继续增加则趋于稳定。标签—标签连接数在1~5范围内变化时,性能波动极小,说明标签间先验依赖对模型影响有限。文本—主题连接数在初步测试中显示出极低的敏感性,因此在后续分析中不做讨论。

综合分析表明,潜在主题数与标签—标签连接数均为高鲁棒性参数;GNN层数与文本—标签连接数存在明确最优区间,推荐默认配置分别为2层和 $K_{\text{label}} = 2$ 。实验结果验证了模型中注意力与动态聚合机制的有效性,能够在参数宽范围内保持稳定性能,降低了实际应用中的调参负担。



(a) GNN 层数对 Tail F1 的影响



(b) GNN 层数对 F1 的影响

图 4 GNN 层数的影响

3.8 计算效率分析

为验证 LIS-Net 在计算开销和可扩展性方面的性能,对多种代表性文本图模型在不同数据集下的总体训练时间与显存占用范围(训练阶段的峰值显存消耗)进行对比评估,其中时间和显存均以[最小值,最大值]的形式报告,如表 4 所示。

表 4 各方法计算效率对比

类别	方法	训练时间/s	显存占用/MB
直推式	TextGCN	[14.9,46.3]	[705.4 763]
	BertGCN	[361.7,1 522.3]	[11 235,12 715]
归纳式	TextING	[42.8,150.8]	[569,1 267]
	HyperGAT	[40.5,152.1]	[507,605]
	DADGNN	[24.1,62.4]	[553,881]
	TextSSL	[187.1,1 367.2]	[485,741]
	TextFCG	[306.1,772.2]	[567,993]
	PaSIG	[58.8,66.0]	[553,707]
PLM	BERT	[36.4,295.9]	[2 109,6 769]
	LIS-Net	[113.4,529.8]	[1 708,1 943]

从表 4 可以看出,不同范式的模型在计算开销上呈现出明显差异。传统的直推式全图模型 TextGCN 在小规模数据集上的时间较短,但随着数据规模增大,其显存占用可超过 4 GB,且难以支持在线增量样本的快速推理。基于 PLM 的直推式方法 BertGCN 由于同时维护全局文档一词图和大规模 Transformer 参数,其显存需求在所有模型中最高,超过 11 GB,训练时间也显著长于其他方法。在归

纳式图模型中,TextSSL 与 TextFCG 由于引入复杂的图结构学习或密集的消息传递,其时间开销在大规模数据集上明显增加;相对而言,TextING、HyperGAT、DADGNN 和 PaSIG 在时间和显存上更为轻量。

LIS-Net 在计算效率上处于一个折中且可接受的区间。与单纯微调 PLM 相比,LIS-Net 在引入潜在主题和实例级子图构建后,时间与显存开销有所增加,这是由于图构建和 GNN 推理引入了额外计算;但得益于归纳式的局部子图设计,LIS-Net 无需维护全局文档一词图,其显存占用远低于 BertGCN,且随着数据规模增大,时间与内存的增长更为平缓。与归纳式方法横向比较,LIS-Net 的时间开销高于 DADGNN 和 PaSIG,但低于 TextSSL 和 TextFCG,同时在显存占用方面始终保持在 2 GB 左右的量级,在单卡 GPU 上即可稳定运行。

从可扩展性角度看,LIS-Net 通过“按需检索+实例级子图”的机制,将图构建与消息传递的复杂度从与语料规模线性相关,转化为与批次大小和近邻数相关,使得模型可以在不重新构建全局图的情况下,对新增样本进行快速归纳式推理。这一特性使得 LIS-Net 在长尾场景下不仅能够提升尾部类别的识别性能,同时在大规模数据集和在线推理环境中具有更好的工程适用性。

4 结束语

针对少样本长尾文本分类中样本稀缺与类别分布不均衡的双重挑战,提出一种基于潜在语义聚类与图交互的端到端网络(LIS-Net)。该方法引入可学习的潜在主题节点作为语义中继,通过归纳式的实例子图构建动态连接文本、主题与标签,并利用交叉注意力等深度融合机制实现图结构化知识与 PLM 语义表示的协同增强。在多个基准数据集上的实验表明,LIS-Net 在长尾场景下显著提升了对尾部类别的识别能力,其性能增益主要源于语义空间的充分预热、图结构信息的有效利用以及各组件间的深度协同。系统的消融与敏感性分析进一步验证了模型设计的合理性与鲁棒性。本研究为少样本不平衡学习提供了一种融合深度语义理解与结构化关系推理的有效途径。

尽管 LIS-Net 在少样本长尾分类中取得显著提升,但仍存在局限。首先,联合优化 PLM 与 GNN 带来额外计算开销;其次,无监督聚类初始化可能影响收敛效率;最后,当前验证主要限于英文标准

景。未来工作将聚焦模型轻量化、引入外部知识增强尾部泛化,并拓展至跨语言、跨模态等复杂场景。

参考文献:

- [1] 陈静,周刚,但文皓.知识增强的异质图注意力虚假信息检测方法[J].信息工程大学学报,2025,26(3):323-329.
- [2] 卢银鹏,郭凯威,卢记仓,等.基于上下文学习和语义检索增强的零样本立场检测方法[J].信息工程大学学报,2025,26(2):196-202.
- [3] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Stroudsburg, USA: ACL, 2019:4171-4186.
- [4] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach [DB/OL]. (2019-07-26) [2026-01-14]. <https://arxiv.org/abs/1907.11692>.
- [5] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks [DB/OL]. (2017-07-18) [2026-01-14]. <https://arxiv.org/abs/1703.03400>.
- [6] LIU X, JI K X, FU Y C, et al. P-tuning: prompt tuning can be comparable to fine-tuning across scales and tasks [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: ACL, 2022:61-68.
- [7] SNELL J, SWERSKY K, ZEMEL R. Prototypical networks for few-shot learning[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, USA: Curran Associates Inc., 2017:4080-4090.
- [8] YAO L, MAO C S, LUO Y. Graph convolutional networks for text classification[C]//Proceedings of the 33th AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2019:7370-7377.
- [9] LIN Y X, MENG Y X, SUN X F, et al. BertGCN: transductive text classification by combining GNN and BERT [C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Stroudsburg, USA: ACL, 2021:1456-1462.
- [10] ZHANG Y F, YU X L, CUI Z Y, et al. Every document owns its structure: inductive text classification via graph neural networks [C] //Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: ACL, 2020:334-339.
- [11] DING K Z, WANG J L, LI J D, et al. Be more with less: hypergraph attention networks for inductive text classification[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: ACL, 2020:4927-4936.
- [12] LIU Y H, GUAN R C, GIUNCHIGLIA F, et al. Deep attention diffusion graph neural networks for text classification[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: ACL, 2021:8142-8152.
- [13] WANG S Y, ZHOU G, LU J C, et al. Pre-trained semantic interaction based inductive graph neural networks for text classification [C]//Proceedings of the 31st International Conference on Computational Linguistics. Stroudsburg, USA: ACL, 2025:812-827.
- [14] PIAO Y H, LEE S, LEE D, et al. Sparse structure learning via graph neural networks for inductive document classification[C]//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2022:11165-11173.
- [15] WANG Y Z, WANG C X, ZHAN J Y, et al. Text FCG: fusing contextual information via graph learning for text classification[J]. Expert Systems with Applications, 2023, 219:No.119658.
- [16] HE P C, LIU X D, GAO J F, et al. DeBERTa: decoding-enhanced BERT with disentangled attention [DB/OL]. (2021-10-06) [2026-01-14]. <https://arxiv.org/abs/2006.03654>.

(编辑:李志豪)