Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

平面多标签文本分类方法

硕士研究生 马西洋

2023年12月24日

问题回溯

- 相关内容
  - **2022.06.06** 吴杭颐《层次多标签文本分类方法》
  - **2021.08.22** 吴杭颐《多标签学习》
  - **2020.12.13** 张睿智《大规模多标签分类方法》

内容提要

- 预期收获
- 题目内涵解析
- 研究背景与意义
- 研究历史与现状
- 知识基础
- 算法原理
  - **LR-GCN**
  - **LACO**
- 特点总结与工作展望
- 参考文献

预期收获

- 预期收获
  - 掌握多标签文本分类的基本概念
  - 了解多标签文本分类的常用方法
  - 了解领域实际应用和发展方向

内涵解析与研究目标

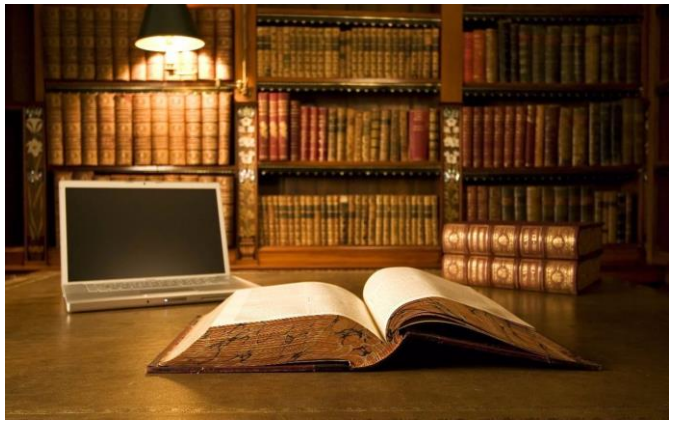
- 内涵解析
  - 多标签分类：指为每个实例标记与之相关的标签集
  - 平面多标签分类：标签间无明显关联的多标签分类

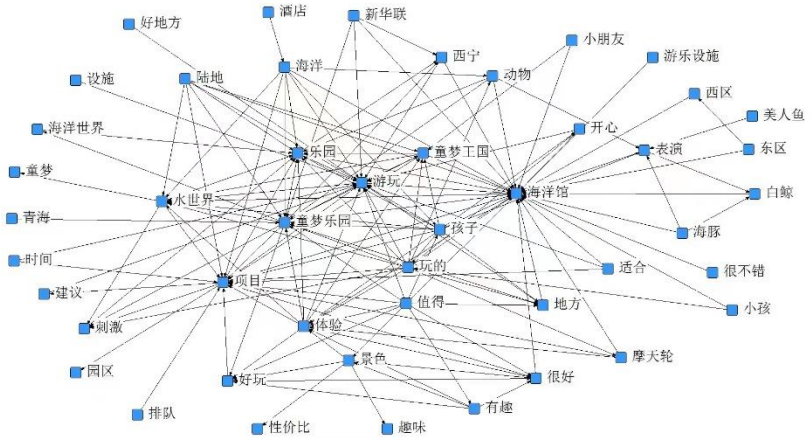
- 研究目标
  - 面向具有多个标签的文本数据
  - 结合图表示学习、多任务学习、深度学习等理论
  - 实现对给定文本的准确分类

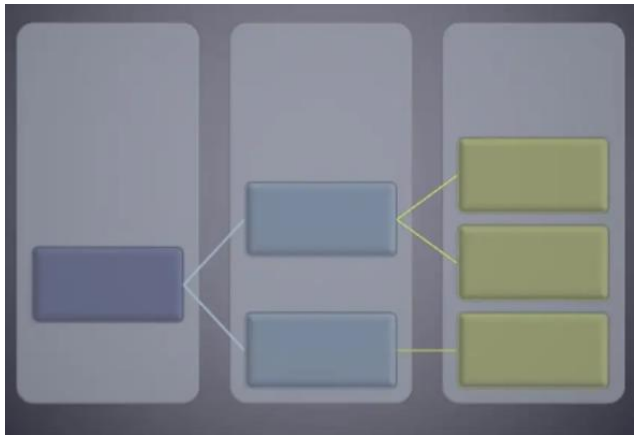
Text	Labels
The mutual information of two random variables is commonly used in learning bayesian nets as well as in other fields ...	math.ST math.IT stat.TH cs.IT cs.AI
Mutual information is widely used, to measure the stochastic dependence of categorical random variables in order to address questions ...	math.ST math.IT stat.TH cs.IT cs.AI cs.LG

研究背景

- 研究背景
  - 现实世界中的实例较为复杂，一般同时具有多种含义
  - 广泛应用于网页标注、新闻分类、文献组织、风格识别等领域
- 研究意义
  - 多标签文本分类是对文本信息进行组织、利用和检索的有效手段，能够提高数据处理效率，具有重要的实际价值
  - 目前样本间关系多样、不同标签的样本数量分布不均匀、标签间关系挖掘不充分等问题使多标签文本分类面临众多挑战，具有重要的理论意义







研究历史

Clare等人首先应用算法适应的思想，提出基于决策树的方法，递归地构建一颗决策树，预测时向下遍历决策树的结点，直到叶子节点

Zhang等人提出多标签K最近邻方法，先识别已标注样本的**K**个最近邻实例，然后从这些实例的标签集中获得统计信息，利用最大后验概率进行推理预测

Read等人提出分类器链方法，先对标签排序，预测时顺序调用分类器，后续分类器需要用到之前分类器预测出的标签集，方法性能受到标签顺序的影响

Google提出了**BERT**预训练模型，通过上下文计算的文本的深度双向表示，只需一个额外的输出层进行微调即可完成多标签文本分类任务

Xu等人针对标签长尾分布的问题提出了特定标签特征增强框架，只增强尾标签的正特征标签对，增强尾部标签的特征

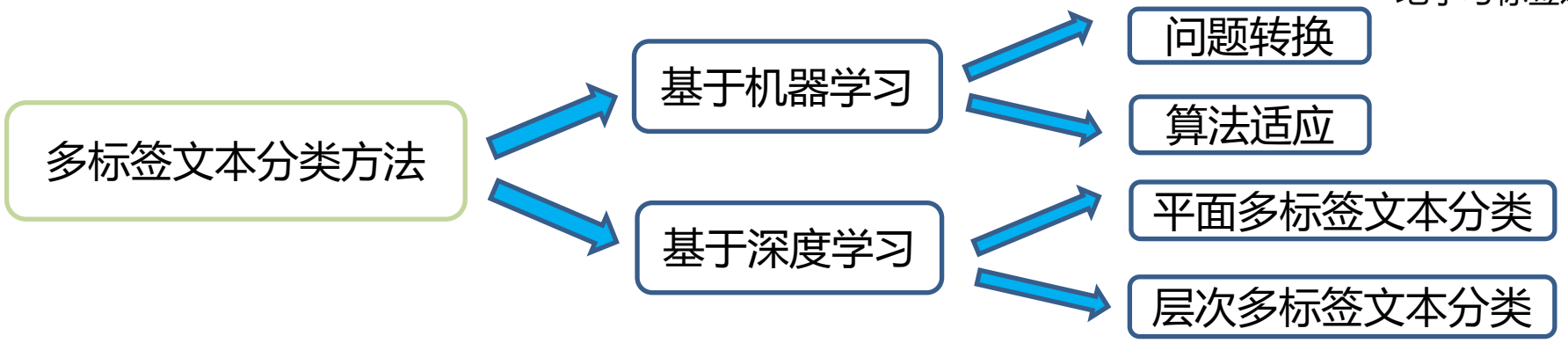
2001 2004 2007 2010 2011 2014 2018 2020 2023

Boutell等人应用问题转换的思想，提出二元关联方法，为每个实例随机选择一个标签而丢弃其他标签；或丢弃多标签的实例，只保留单标签的实例

Tsoumakas等人提出标签幂集分解方法，通过对标签进行组合后使用二元分类器进行多标签文本分类，方法无法泛化到未见标签组合

Kim首次将卷积神经网络引入多标签文本分类领域中，在预先训练好的词向量基础上训练卷积神经网络，通过微调词向量适应多标签文本分类任务

Ankit提出了一种基于图注意力网络的模型，使用特征矩阵和相关性矩阵来捕捉和探索标签之间的关键依赖关系，通过注意力机制为每个标签的邻接节点分配不同的权重，从而让系统隐式地学习标签之间的依赖关系



研究现状

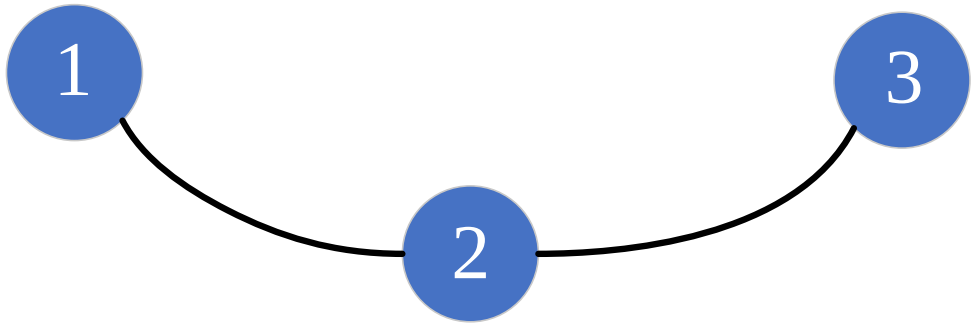
- 基于文本表示学习
  - 引入多种编码器描述文本，如双向LSTM、胶囊网络等
  - 引入注意力机制关注文本中的关键词信息
  - 基于文本表示学习的方法侧重于样本表征的增强，忽视了标签间关联度
- 基于标签关系建模
  - 利用图表示学习捕获标签之间关系
    - 图卷积网络
    - 图注意力网络
    - 超图网络
    - 超图注意力网络
  - 仅利用训练样本中的统计特征无法完整准确地捕获标签特征，难以泛化到测试样本中新出现或少数出现的标签关系

知识基础 图卷积网络

- 图卷积网络用于处理图形数据的神经网络，网络中节点的特征不仅取决于节点自身的特征，还取决于其邻居的特征
- 图$\boldsymbol{G}$可以由特征描述$\boldsymbol{X} \in \mathbb{R}^{n \times d}$和对应的邻接矩阵$\boldsymbol{A} \in \mathbb{R}^{n \times n}$组成，其中$\boldsymbol{n}$表示节点数，$\boldsymbol{d}$表示嵌入维数
- 在图卷积网络中每个节点的权重相同，节点可以通过聚合相邻节点特征进行更新

$$\boldsymbol{X}^{(l+1)} = \boldsymbol{\sigma}(\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{X}^{l}\boldsymbol{\theta})$$

- 在图网络结构中



$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \qquad \boldsymbol{AX} = \begin{bmatrix} x_1 + x_2 \\ x_1 + x_2 + x_3 \\ x_2 + x_3 \end{bmatrix}$$

知识基础 图注意力网络

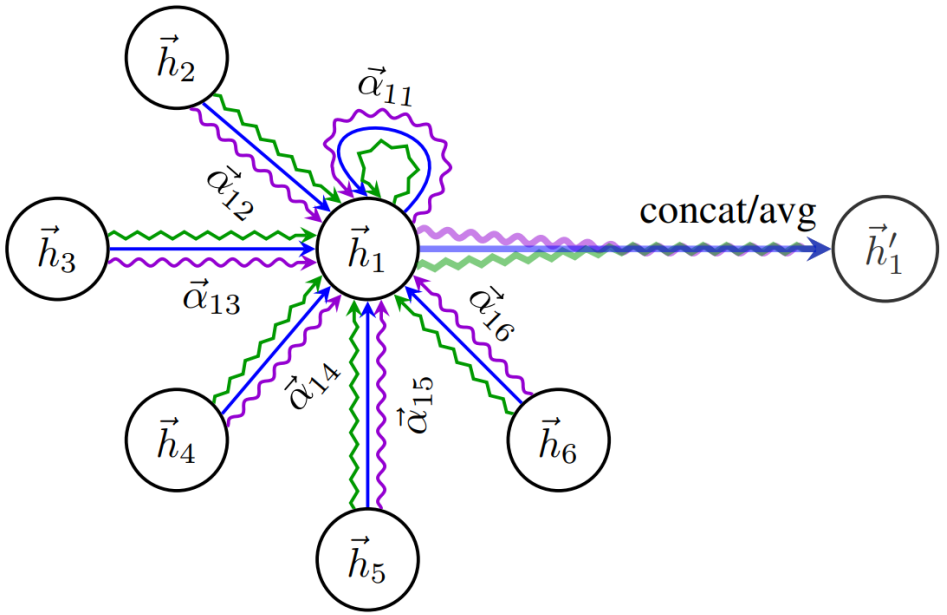
- 在GCN中，节点的邻域以相等或预定义的权重结合在一起，可以通过引入注意力系数使得每个节点在更新自己的特征时，可以根据邻居节点的重要性赋予每个节点不同的权重
- 注意力系数α_{ij}^{l} 用于衡量第j个节点在更新第l个隐层节点i时的重要性

$$\alpha_{ij}^{l} = f(X_i^l\theta, X_j^l\theta)$$

$$e_{ij}^{l} = \sigma\left(X_i^l\theta || X_j^l\theta\right)$$

$$\alpha_{ij}^{l} = \frac{exp(LeakyReLU(e_{ij}^{l}))}{\sum_{k=1}^{n} exp(LeakyReLU(e_{ik}^{l}))}$$

$$X_i^{(l+1)} = \sigma(\frac{1}{\mathrm{K}}\sum_{\mathrm{k}=1}^{\mathrm{K}}\sum_{j=1}^{n} D^{-\frac{1}{2}}\alpha_{ij,k}^{l}D^{-\frac{1}{2}}X_j^l\theta)$$



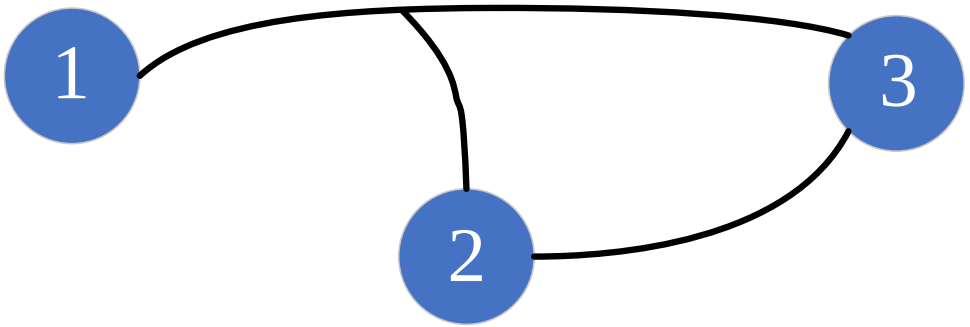
知识基础 超图卷积网络

- 超图是一种广义的图，它的一个边可以连接任意数量的顶点，是节点的任意组合
- 超图$\boldsymbol{G}$可以由特征描述$\boldsymbol{X} \in \mathbb{R}^{n \times d}$、对应的邻接矩阵$\boldsymbol{H} \in \mathbb{R}^{n \times m}$以及权重矩阵$\boldsymbol{W} \in \mathbb{R}^{m \times m}$组成，其中$\boldsymbol{n}$表示节点数，$\boldsymbol{d}$表示嵌入维数，$\boldsymbol{m}$表示超边数
- 超图的更新公式

$$\boldsymbol{X}^{(l+1)} = \boldsymbol{\sigma}(\boldsymbol{D}_v^{-\frac{1}{2}}\boldsymbol{H}\boldsymbol{W}\boldsymbol{D}_e^{-1}\boldsymbol{H}^{\boldsymbol{T}}\boldsymbol{D}_v^{-\frac{1}{2}}\boldsymbol{X}^{\boldsymbol{l}}\boldsymbol{\theta})$$

- 在图网络结构中，$\boldsymbol{H}$表示那些节点与超边相连，$\boldsymbol{W}$表示超边的权重

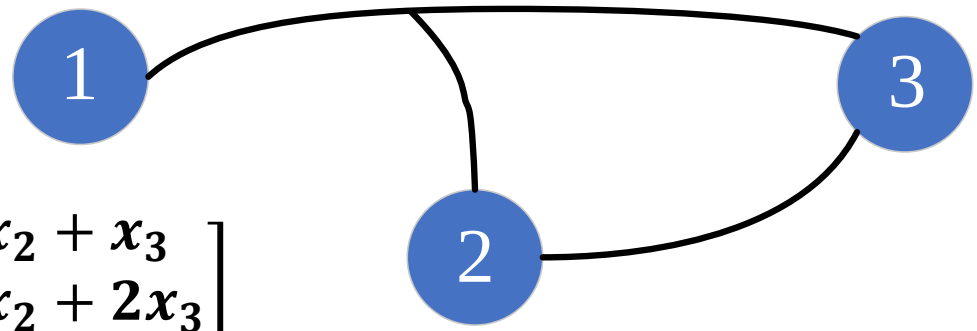
$$\boldsymbol{H} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \quad \boldsymbol{W} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



知识基础 超图卷积网络

- 根据节点特征构建超边特征（将边连接的节点求和）

$$H^T X^l = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_1 + x_2 + x_3 \\ x_2 + x_3 \end{bmatrix}$$



- 汇聚超边特征后完成节点特征更新

$$HWH^T X^l = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_1 + x_2 + x_3 \\ x_2 + x_3 \end{bmatrix} = \begin{bmatrix} x_1 + x_2 + x_3 \\ x_1 + 2x_2 + 2x_3 \\ x_1 + 2x_2 + 2x_3 \end{bmatrix}$$

- 归一化结果

$$D_v^{-\frac{1}{2}} HW D_e^{-1} H^T D_v^{-\frac{1}{2}} X^l = \begin{bmatrix} 0.33x_1 + 0.23x_2 + 0.23x_3 \\ 0.23x_1 + 0.41x_2 + 0.41x_3 \\ 0.23x_1 + 0.41 + 0.41x_3 \end{bmatrix}$$

- **GCN与HGNN虽然计算出来的值有一定的差别，但是他们本质上的思想是差不多的，GCN可以说是一种特殊的HGNN**

LR-GCN

Label-representative graph convolutional network for multi-label text classification

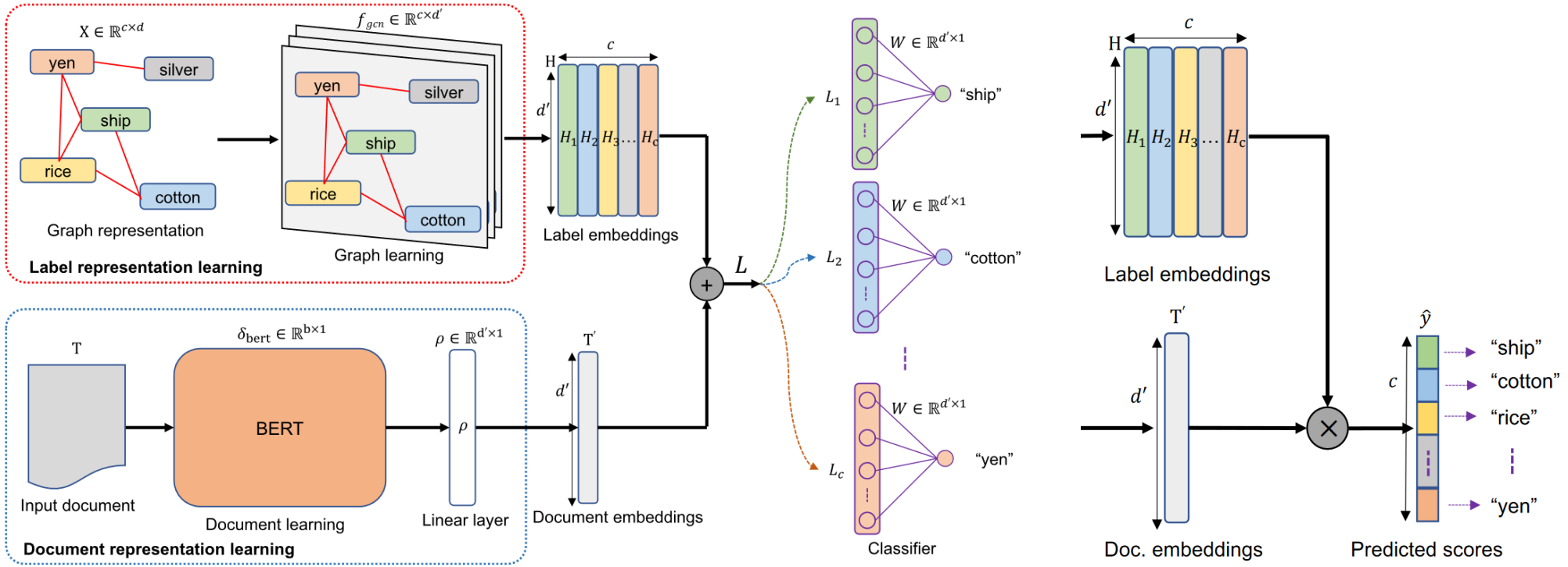
LR-GCN TIPO

T	目标	预测样本x_i所属的多个标签
I	输入	多标签文本分类数据集*2（AAPD、RCV1数据集）
P	处理	1.根据标签共现关系构建邻接矩阵 2.通过图卷积聚合标签节点特征 3.将标签特征与文本特征融合后进行分类
O	输出	与样本相关的标签*n

P	问题	现有多标签文本分类方法未充分考虑标签之间的相关性
C	条件	标签之间存在共现关系
D	难点	如何有效地建立标签间的相关性模型，以提高分类性能
L	水平	**Applied Intelligence（2023 SCI 二区）**

LR-GCN 算法原理图

- 算法原理图
  - 标签表示学习：基于标签共现关系构建邻接矩阵，通过图卷积聚合标签嵌入
  - 文本表示学习：利用**RoBERTa**提取文本嵌入
  - 预测分类：将每个标签的嵌入与文本嵌入拼接计算得到每一类的概率



创新分析 标签表示学习

- 邻接矩阵构建
  - 基于信息学中点互信息的概念构建邻接矩阵

$$PMI(X;Y) = \log \frac{p(X,Y)}{p(X)p(Y)}$$

  - 通过标签出现和共现次数近似计算标签出现概率

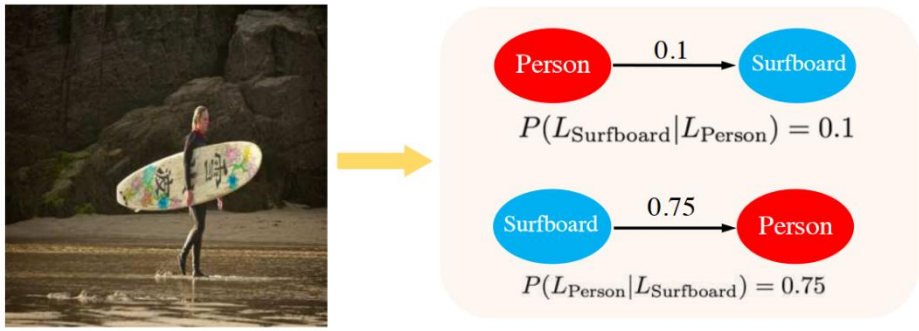
$$A_{ij} = \frac{p(i,j)}{p(i)p(j)} = \frac{\#L_{(i,j)}\#D}{\#L_{(i)}\#L_{(j)}}$$

  - 以条件概率的形式对标签相关性进行建模，即$p(i|j)$，表示当标签j出现时标签i出现的概率

$$A_{ij} = p(i|j) = \frac{p(i,j)}{p(j)} = \frac{\#L_{(i,j)}}{\#L_{(j)}}$$

“labels”： [[corn, grain, wheat],
[acq, corn, grain],
[corn, grain, oilseed, soybeam]]

$$A_{oilseed,soybeam} = \frac{1*3}{1*1} = 3.0$$



创新分析 标签表示学习

- 标签初始嵌入构建
  - 标签嵌入会与文档表示结合，初始化输入标签的特征节点是一项重要任务
  - 任何单词嵌入方法，如**Word2vec**、**Glove**和**FastText**等能够捕捉到一些句法和语义信息，但仍需对其进行改进
  - 利用外部信息丰富标签嵌入
    - 使用维基百科检索与标签最相关的句子，保留前两句
    - 通过**Sentence-BERT**生成标签嵌入

Input label: “wheat”

Output: “wheat is a grass widely cultivate for its seed, a cereal grain which is a worldwide staple food. The many species of wheat together make up the genus Triticum; the most widely grown in common wheat(T. aestivum)”

实验设计

- 数据资源
  - 数据集如下图所示

数据集	训练样本数	测试样本数	标签数	标签平均样本数	样本平均标签数
AAPD	54840	1000	54	2444.0	2.41
RCV1	23149	781265	103	729.67	3.18

- 对比方法
  - 基于文本的方法
    - XML-CNN、HTTN、DocBERT、VLAWE
  - 利用标签相关性的方法
    - LSAN、AttentionXML、HA-seq2seq、HCSM等
- 评价指标
  - 前k位精确度（P@k）
  - 前k位归一化折损累计增益（nDCG@k）

$$P@k = \frac{1}{k} \sum_{l \in rank_k(\hat{y})} y_l$$

$$DCG@k = \sum_{l \in rank_k(\hat{y})} \frac{y_l}{log(l+1)}$$

$$nDCG@k = \frac{DCG@k}{\sum_{l=1}^{min(k, ||y||_0)} \frac{1}{log(l+1)}}$$

实验设计 对比实验

- 实验结果
  - 与其他方法相比，**LR-GCN**方法在除***P@5***之外得指标均优于其他算法
  - 通过使用提出的标签嵌入方法，可以提高模型的性能
  - 图嵌入方法依赖标签的相关性和语义信息，而AAPD数据集缺乏标签信息的描述

Datasets	Models	P@1(%)	P@3(%)	P@5(%)	nDCG@3(%)	nDCG@5(%)
AAPD	XML-CNN*	74.38	53.84	37.79	71.12	75.93
	HTTN×	83.84	59.92	40.79	79.27	82.67
	DXML*	80.54	56.30	39.16	77.23	80.99
	SGM*	75.67	56.75	35.65	72.36	75.35
	AttentionXML*	83.02	58.72	40.56	78.01	82.31
	EXAM*	83.26	59.77	40.66	79.10	82.79
	LSAN*	85.28	61.12	**41.84**	80.84	84.78
	SLEEC†	81.96	57.48	38.99	77.65	81.59
	LAHA†	84.48	60.72	41.19	80.11	83.70
	LR-GCN (Ours)	**86.50**	**62.43**	41.66	**82.52**	**85.48**

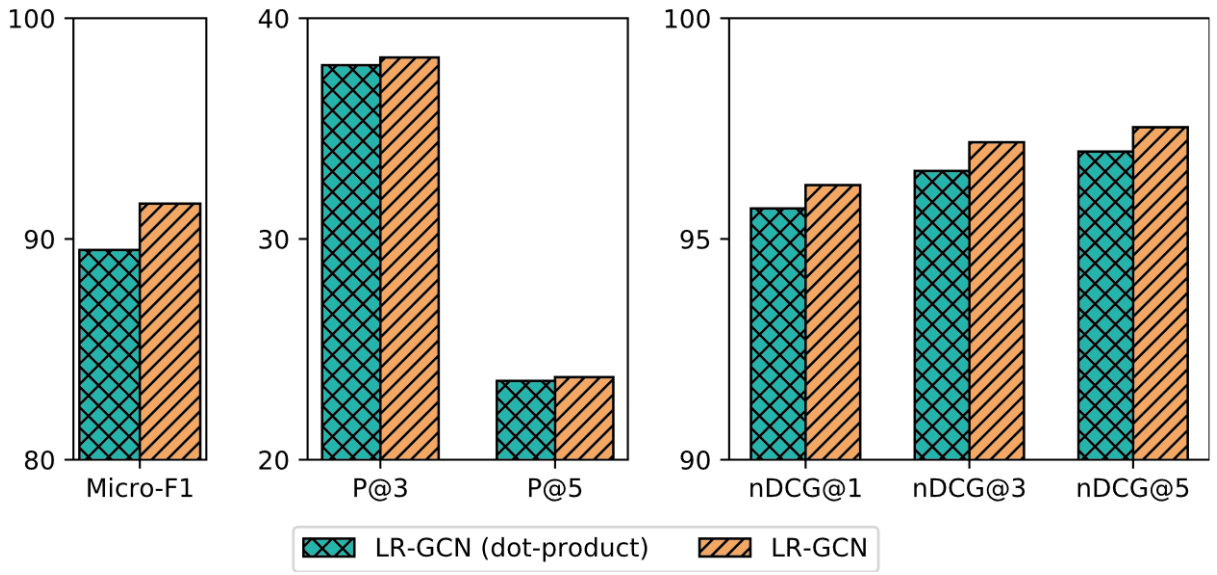
实验设计 对比实验

- 实验结果
  - 与其他方法相比，**LR-GCN方法在所有评价指标下**均优于其他算法
  - RCV1数据集中标签语义信息较容易获取

Datasets	Models	P@1(%)	P@3(%)	P@5(%)	nDCG@3(%)	nDCG@5(%)
RCV1	Bow-CNN$^{\perp}$	96.40	81.17	56.74	92.04	92.89
	Kim-CNN$^{\perp}$	93.54	76.15	52.94	87.26	88.20
	XML-CNN$^{\perp}$	96.86	81.11	56.07	92.22	92.63
	HTTN$^{\times}$	95.86	78.92	55.27	89.61	90.86
	DXML*	94.04	78.65	54.38	89.83	90.21
	SGM*	95.37	81.36	53.06	91.76	90.69
	AttentionXML*	96.41	80.91	56.38	91.88	92.70
	EXAM*	93.67	75.80	52.73	86.85	87.71
	LSAN*	96.81	81.89	56.92	92.83	93.43
	SLEEC$^{\perp}$	95.35	79.51	55.06	90.45	90.97
	FastXML$^{\perp}$	94.62	78.40	54.82	89.21	90.27
	HR-DGCNN$^{\div}$	95.29	50.32	55.38	90.02	90.28
	HG-Transformer$^{\div}$	95.80	80.98	55.96	90.03	91.96
	LR-GCN (Ours)	**97.13**	**84.29**	**58.45**	**94.98**	**95.38**

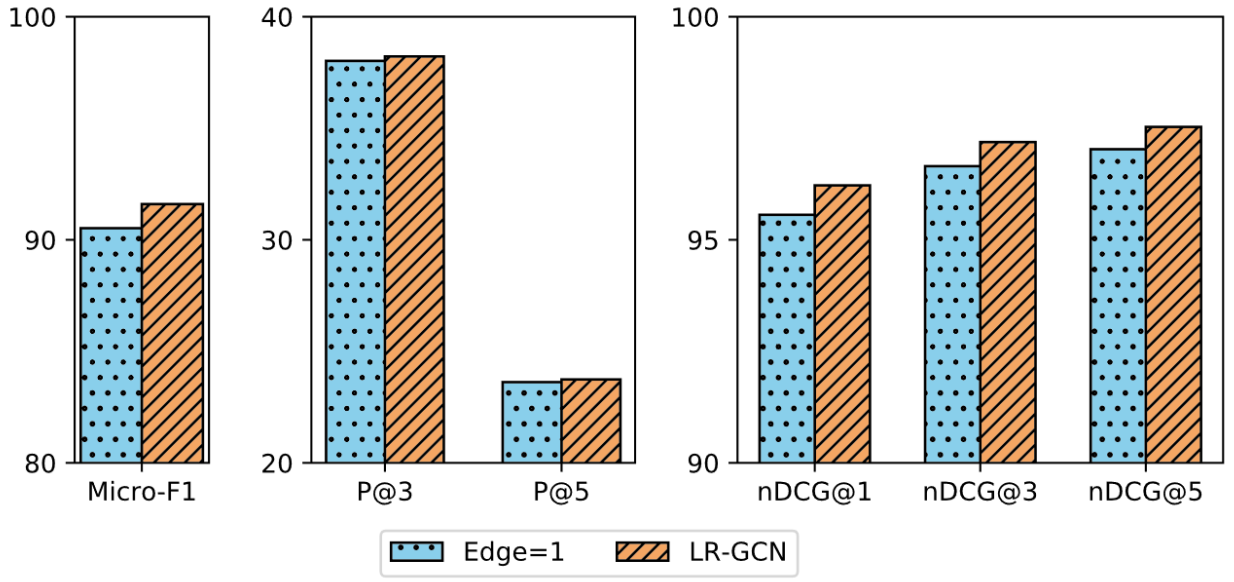
实验设计 消融实验

- 实验目的
  - 使用两种不同的分类方法进行标签分类，验证全连接神经网络作为分类器的有效性
- 实验结果
  - 基于全连接层的模型优于基于点积的模型（**Micro-F1提高2.1%**）



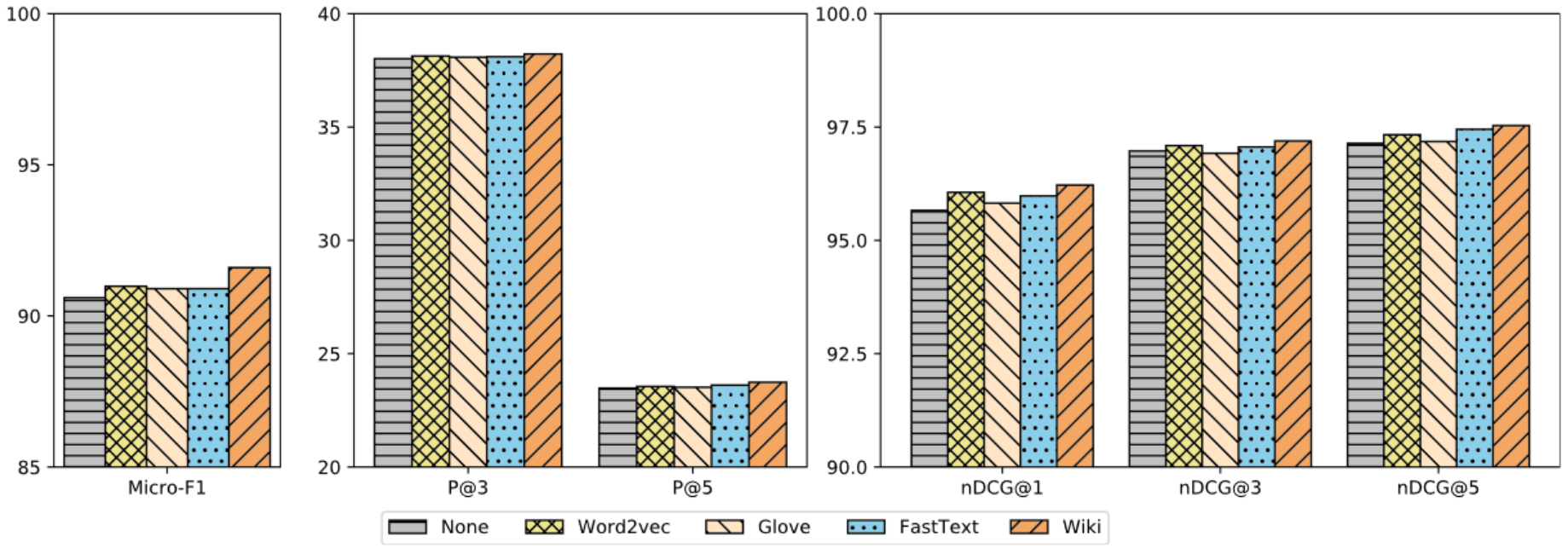
实验设计 消融实验

- 实验目的
  - 使用两种不同方法构建邻接矩阵，验证基于点互信息构建邻接矩阵的有效性
- 实验结果
  - 基于点互相关的邻接矩阵效果更好
  - 将相关性二值化不能够更加准确的描述相关性大小



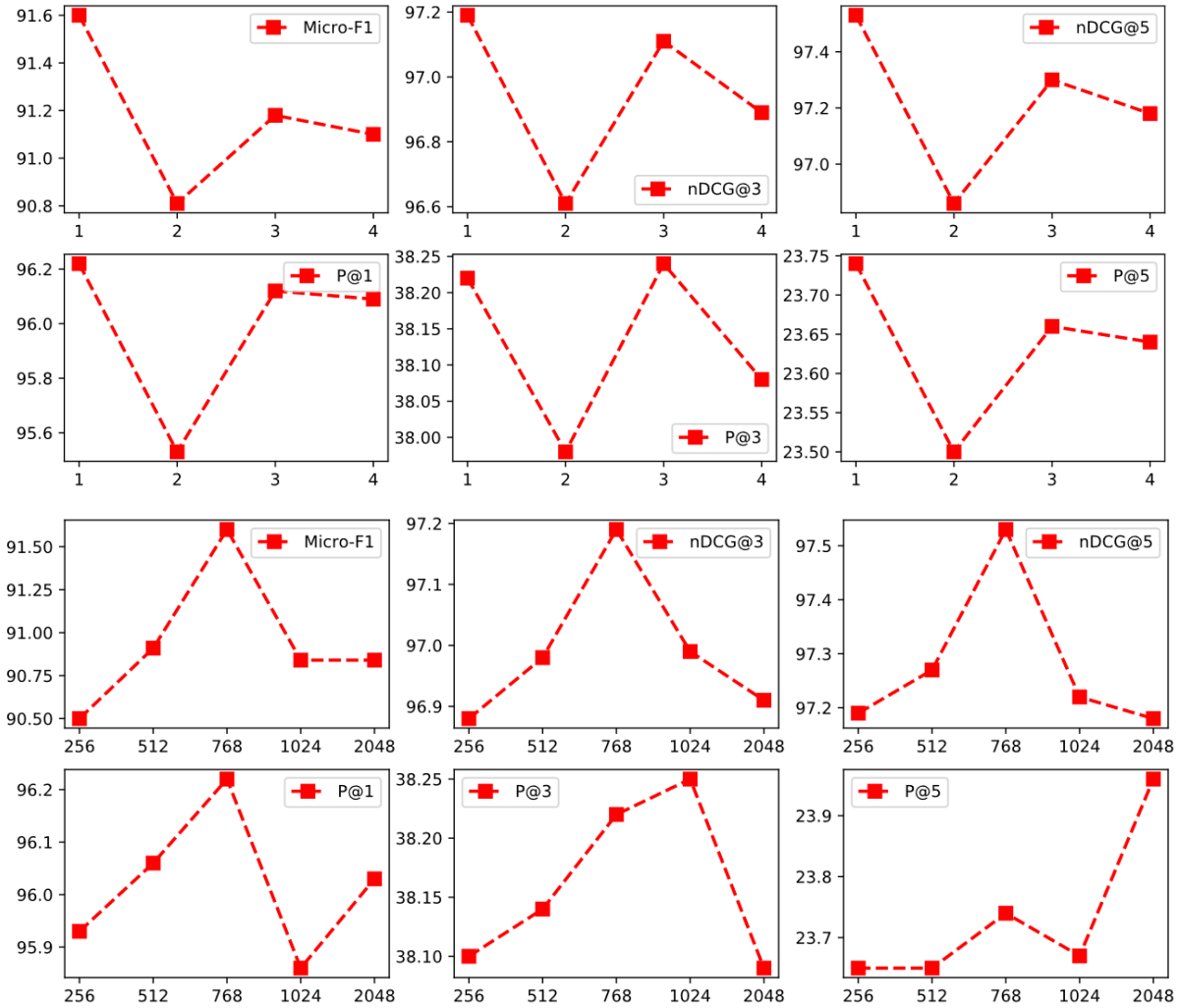
实验设计 消融实验

- 实验目的
  - 验证不同节点嵌入方法的性能
- 实验结果
  - 通过外部信息丰富节点嵌入取得了最好的效果



实验设计 消融实验

- 实验目的
  - 验证图卷积层数和标签嵌入隐藏层数对结果的影响
- 实验结果
  - 只用一个GCN层就达到最佳效果
  - 层数过多会使不同标签嵌入之间的差异减少，导致过度平滑的问题
  - 隐藏层维数为768维时效果最好
  - 过低的嵌入维度会导致没有足够的容量处理复杂任务，而过高嵌入维度可能会导致过拟合



Enhancing Label Correlation Feedback in Multi-Label Text Classification via Multi-Task Learning

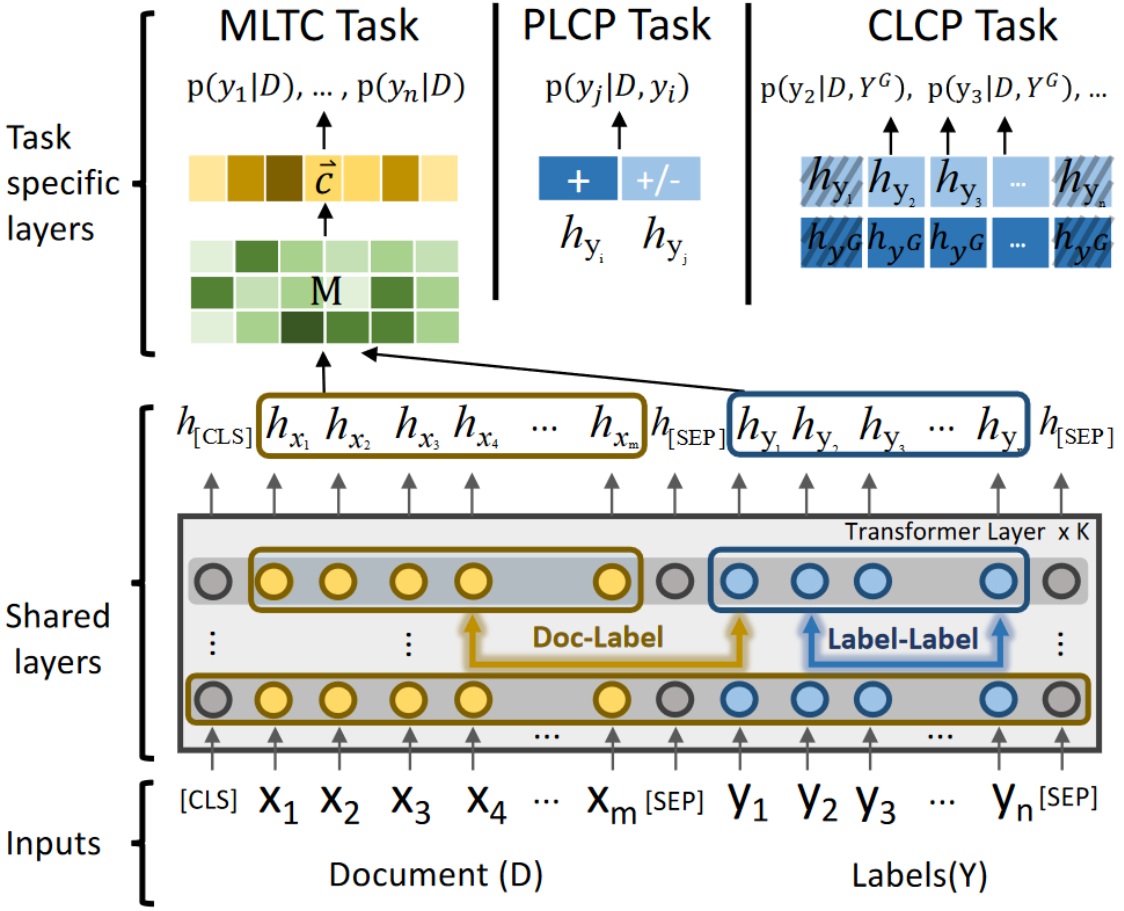
LACO TIPO

T	目标	预测样本所属的多个标签
I	输入	多标签文本分类数据集*2（AAPD、RCV1数据集）
P	处理	1. 利用联合嵌入机制来同时获得文本和标签表示 2.利用文本-标签交叉注意力机制来生成更具区分度的文档表示 3.通过了成对标签共现预测和条件标签共现预测两个任务来增强标签相关性学习
O	输出	与样本相关的标签*n

P	问题	Seq2Seq模型存在过度依赖标签顺序和误差传播等问题
C	条件	标签之间存在相关性
D	难点	设计任务增强对标签相关性的利用
L	水平	ACL（2021 CCFA）

LACO 算法原理图

- 算法原理图
  - 联合嵌入：通过特殊标记[SEP]分割标签与文本，一起输入到BERT预训练模型编码
  - 交叉注意力：通过将文本嵌入与标签嵌入做点积计算注意力
  - 成对标签共现：通过预测标签对是否共现训练模型理解二阶标签关系
  - 条件标签共现：通过让模型在知道一个或多个正确标签的情况下预测其余标签是否与其相关训练模型学习高阶标签关系



创新分析 交叉注意力

- 交叉注意力计算
  - 将标签嵌入与文本嵌入卷积得到相关性矩阵

$$G = H_D H_Y^T$$

  - 对相关性矩阵做卷积，捕获局部矩阵块中的相关性

$$M = g(G_{p-r;p+r} W_1 + b_1)$$

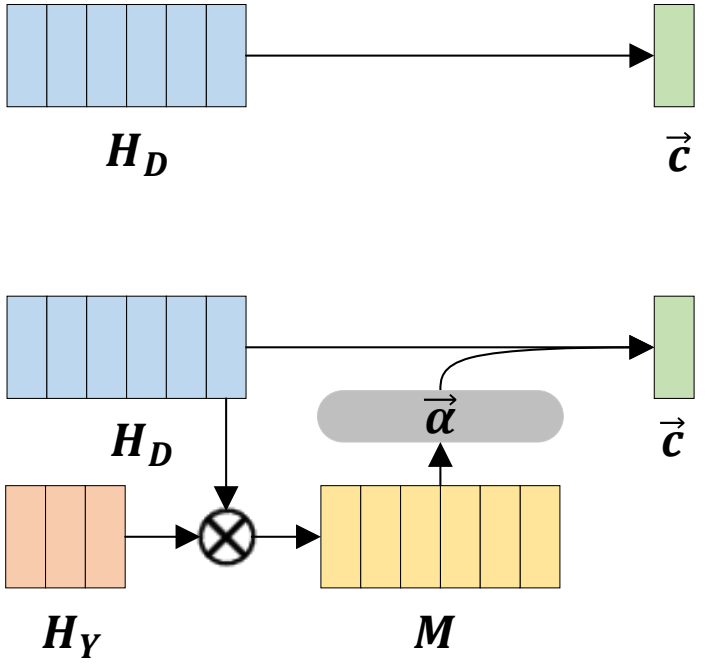
  - 通过softmax和双曲正切将矩阵块压缩为向量

$$\vec{\alpha} = \Omega(M)$$

  - 由注意力向量加权文本嵌入，得到更具区分度的文本表示

$$\vec{c} = \vec{\alpha} \cdot H_D$$

  - 即通过使用高阶文本-标签相关性加权文本嵌入，并充分考虑文本和标签之间的非线性关系



创新分析 标签共现任务

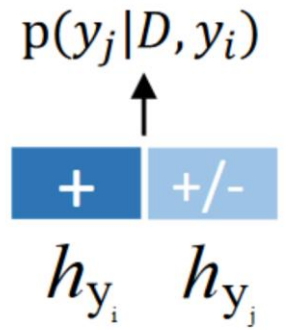
- 成对标签共现任务
  - 假设文档D包含正确的标签集Y+和不正确的标签集Y-
  - 从Y+中选择标签y_i，从Y+和Y-中选择标签y_j
  - 使用额外的二元分类器预测两个标签的状态是否相关
  - 损失函数如下

$$\mathcal{L}_{plcp} = -[q_{ij}\ln(p_{ij}) + (1-q_{ij})\ln(1-p_{ij})]$$

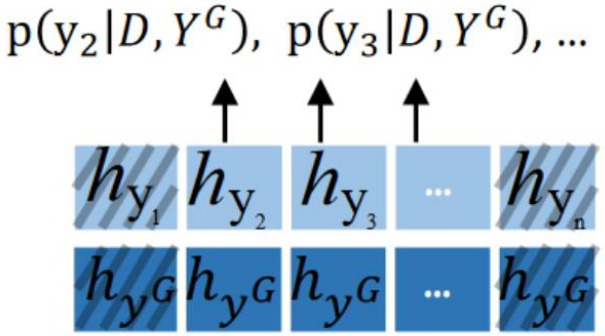
- 条件标签共现任务
  - 从Y+中选择s个标签组成Y^G
  - 使用额外的sigmoid分类器预测所有其余标签是否与之相关
  - 损失函数如下

$$\mathcal{L}_{clcp} = -\sum_{i=1}^{n-s}[q_i\, ln(p_i) + (1-q_i) ln(1-p_i)]$$

PLCP Task



CLCP Task



实验设计

- 数据资源
  - 数据集如下图所示

数据集	训练样本数	测试样本数	标签数	标签平均样本数	样本平均标签数
AAPD	54840	1000	54	2444.0	2.41
RCV1-V2	23149	781265	103	729.67	3.18

- 对比方法
  - 不考虑标签相关性的方法
    - **BR、CNN、LEAM、LSAN、BERT**
  - 考虑标签相关性的方法
    - **CC、SGM、Seq2Seq、OCD等**
- 评价指标
  - 汉明损失（**HL**）
  - 微/宏-**F1**（**Mi/Ma-F1**）、微/宏-精确率（**Mi/Ma-P**）、微/宏-召回率（**Mi/Ma-R**）

实验设计 对比实验

- 实验结果
  - 基于LACO的模型在主要评估指标上都优于所有基线模型
  - 由于clcp任务考虑了高阶相关性在*Ma-F1*表现更好

Algorithm	AAPD dataset						RCV1-V2 dataset					
	HL↓	Mi- P	Mi- R	Mi- F1↑	Ma- P	Ma- R	Ma- F1↑	HL↓	Mi- P	Mi- R	Mi- F1↑	Ma- P	Ma- R	Ma- F1↑
BR†(Boutella et al., 2004)	0.0316	64.4	64.8	64.6	-	-	-	0.0086	90.4	81.6	85.8	-	-	-
CNN†(Kim, 2014)	0.0256	**84.9**	54.5	66.4	-	-	-	0.0089	92.2	79.8	85.5	-	-	-
LEAM(Wang et al., 2018)	0.0261	76.5	59.6	67.0	52.4	40.3	45.6	0.0090	87.1	84.1	85.6	69.5	65.8	67.6
LSAN(Xiao et al., 2019)	0.0242	77.7	64.6	70.6	67.6	47.2	53.5	0.0075	91.3	84.1	87.5	74.9	65.0	68.4
BERT(Devlin et al., 2019)	0.0224	78.6	68.7	73.4	68.7	52.1	57.2	0.0073	**92.7**	83.2	87.7	77.3	61.9	66.7
CC†(Read et al., 2011)	0.0306	65.7	65.1	65.4	-	-	-	0.0087	88.7	82.8	85.7	-	-	-
SGM†♣(Yang et al., 2018)	0.0251	74.6	65.9	69.9	-	-	-	0.0081	88.7	85.0	86.9	-	-	-
Seq2Set†♣(Yang et al., 2019)	0.0247	73.9	67.4	70.5	-	-	-	0.0073	90.0	85.8	87.9	-	-	-
OCD†♣(Tsai and Lee, 2020)	-	-	-	72.0	-	-	58.5	-	-	-	-	-	-	-
ML-R†(Wang et al., 2020)	0.0248	72.6	**71.8**	72.2	-	-	-	0.0079	89.0	85.2	87.1	-	-	-
Seq2Seq$_T^{\clubsuit}$(Nam et al., 2017)	0.0275	69.8	68.2	69.0	56.2	53.7	54.0	0.0074	88.5	**87.4**	87.9	69.8	65.5	66.1
SeqTag$_{Bert}$	0.0238	74.3	71.5	72.9	61.5	**57.5**	58.5	0.0073	90.6	84.9	87.7	73.7	66.7	68.7
LACO	0.0213	80.2	69.6	74.5	70.4	54.0	59.1	0.0072	90.8	85.6	88.1	75.9	66.6	69.2
LACO+*plcp*	**0.0212**	79.5	70.8	**74.9**	68.4	55.8	59.9	**0.0070**	90.8	86.2	88.4	76.1	66.5	69.2
LACO+*clcp*	0.0215	78.9	70.8	74.7	**71.9**	56.6	**61.2**	**0.0070**	90.6	86.4	**88.5**	**77.6**	**71.5**	**73.1**

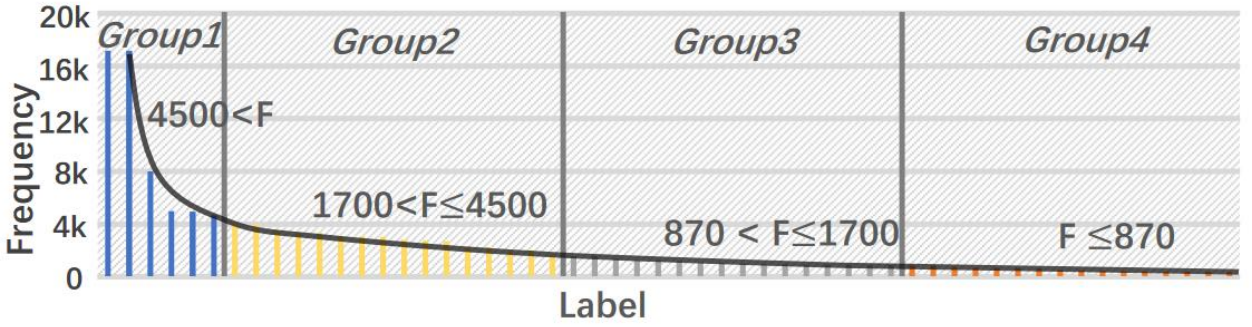
实验设计 消融实验

- 实验目的
  - 验证联合嵌入和交叉注意力的有效性
- 实验结果
  - 联合嵌入和交叉注意力对于获得更具区分度的文本嵌入都很重要
  - 去除联合嵌入和交叉注意力机制后，**AAPD**数据集的性能下降幅度大于**RCV1-V2** 数据集，由于**AAPD**数据集中标签相关性更强

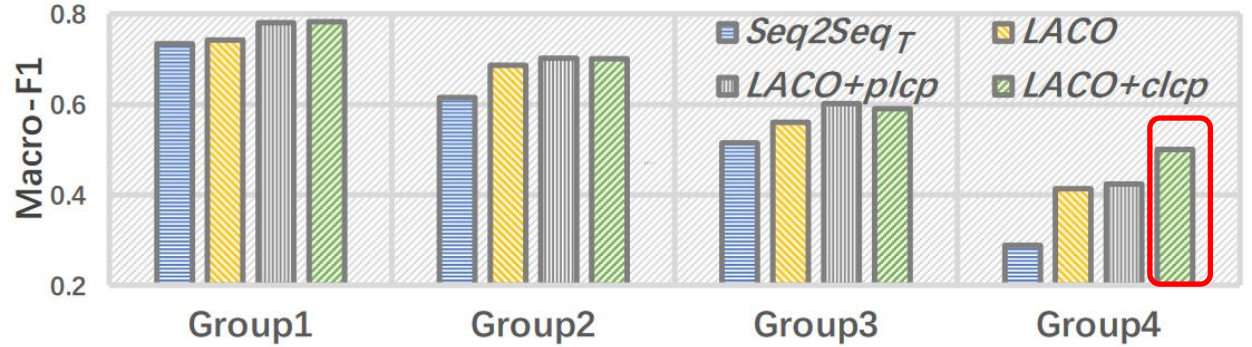
Model	AAPD			RCV1-V2		
	HL	**Mi-F**	**Ma-F**	**HL**	**Mi-F**	**Ma-F**
LACO	0.0213	74.5	59.1	0.0072	88.1	69.2
w/o JE	0.0237	72.6	57.7	0.0077	87.5	68.4
w/o CA	0.0220	73.5	58.4	0.0073	87.8	68.5
w/o JE & CA	0.0224	73.4	57.2	0.0073	87.7	66.7

实验设计 低频标签性能实验

- 实验目的
  - 验证模型预测低频标签的性能
- 实验结果
  - 将所有标签按频率分为四组，即大头组、高频组、中频组和低频组
  - 这是一种典型的大头长尾分布，所有方法的性能都随着标签出现频率的增加而降低
  - 通过条件标签共现预测任务可以提高低频标签的性能



(a) The label distribution of AAPD



(b)Macro-F1 for the four groups on AAPD

实验设计 标签相关性分析

- 实验目的
  - 验证模型对标签相关性的利用
- 实验结果
  - 利用标签y_a和y_b之间的条件概率$p(y_b|y_a)$来定量表示它们之间的依赖关系
  - 计算$p(y_b|y_a)$的KL散度，衡量模型预测分布（P^p）与训练/测试数据集上的真实分布（P^g）之间的差异

Model	AAPD		RCV1-V2	
	train	**test**	**train**	**test**
Seq2Seq$_T$	**1.27**	1.30	0.08	0.94
SeqTag$_{Bert}$	1.40	1.28	0.09	0.95
LACO	1.40	1.27	0.09	0.94
LACO$_{+plcp}$	1.35	1.28	0.08	**0.76**
LACO$_{+clcp}$	1.32	**1.10**	0.08	0.91

$$KL(P^g||P^p) = \sum P^g(y_b|y_a)\log\frac{P^g(y_b|y_a)}{P^p(y_b|y_a)}$$

特点总结与未来展望

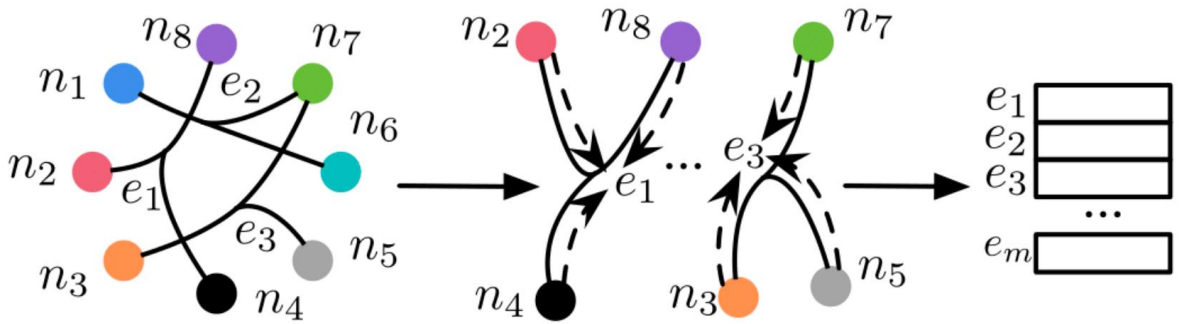
特点总结

- **LR-GCN**
  - 利用图卷积捕获标签的相关性和语义信息，根据标签共现概率建模邻接矩阵
  - 仅通过统计训练样本的静态标签共现次数建模标签间关系，无法完整准确地建模标签特征
- **LACO**
  - 利用文档和标签之间的语义和相关性，获得具有区分性的文本表示
  - 联合嵌入会占用额外的输入长度，且忽略了分类层次结构的重要性
  - 两个辅助任务之间相似度较高，共同使用无法增加 性能
- 两种方法分别从标签关系建模和文本表示学习两个方面，但是未能同时兼顾两个方面，导致效果提升有限

未来展望

- 未来发展
  - 通过注意力机制动态学习标签之间的相关性，建立邻接矩阵
  - 通过超图增强标签之间的相关性，降低图的复杂度
  - 构建提示学习模板让模型隐式学习标签之间相关性
  - 引入额外的标签相关性解码器避免占用输入长度
- 超图注意力网络
  - 以标签特征作为节点，以标签之间的共现关系作为超边构建超图
  - 超边特征聚合

$$\boldsymbol{e_k = Aggregate(\{n_i | i \in m\})}$$



未来展望

- 超图注意力网络
  - 为避免静态构建邻接矩阵无法准确建模标签特征的问题，引入注意力系数
    - 计算注意力系数

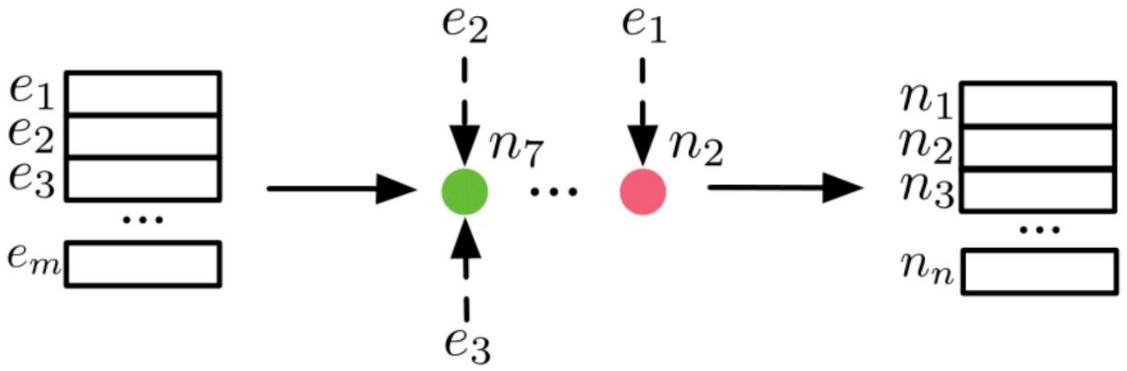
$$\boldsymbol{\alpha_{ij}} = \boldsymbol{\sigma}(\boldsymbol{n_i}||\boldsymbol{e_j})$$

    - 归一化注意力系数：使用 **softmax** 函数对每个节点对应的所有超边的注意力系数进行归一化。

$$\boldsymbol{\alpha_{ij}} = \frac{\boldsymbol{exp}(\boldsymbol{LeakyReLU}(\boldsymbol{\alpha_{ij}}))}{\sum_{\boldsymbol{k=1}}^{\boldsymbol{n}} \boldsymbol{exp}(\boldsymbol{LeakyReLU}(\boldsymbol{\alpha_{ik}}))}$$

    - 特征更新：使用注意力系数对每个节点的超边特征进行加权求和，得到更新后的节点特征

$$\boldsymbol{n_i'} = \sum \boldsymbol{\alpha_{ij}} \boldsymbol{e_j}$$



参考文献

[1] Vu H T, Nguyen M T, Nguyen V C, et al. Label-representative graph convolutional network for multi-label text classification[J]. Applied Intelligence, 2023, 53(12): 14759-14774.

[2] Zhang X, Zhang Q W, Yan Z, et al. Enhancing Label Correlation Feedback in Multi-Label Text Classification via Multi-Task Learning: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021[C]. Online: Association for Computational Linguistics, 2021: 1190-1200.

[3] Pal, Ankit, Muru Selvakumar, Malaikannan Sankarasubbu. MAG-NET: Multi-Label Text Classification using Attention-based Graph Neural Network: Proceedings of the 12th International Conference on Agents and Artificial Intelligence[C]. Florida: ScitePress , 2020, 2: 494–505, 2020.

道德经

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！