Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

层次多标签文本分类方法

吴杭颐　硕士研究生

2022年06月05日

内容提要

- 背景简介
- 基本概念
- 算法原理
  - HiMatch
  - HGCLR
- 应用总结
- 参考文献

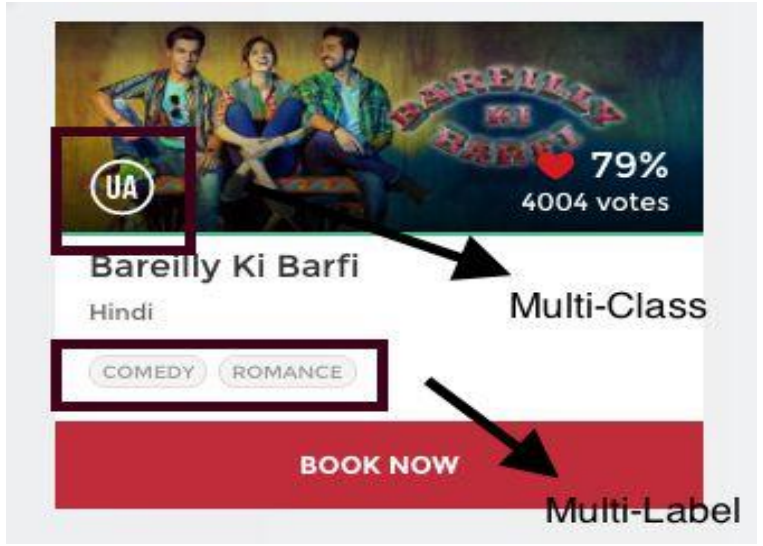
预期收获

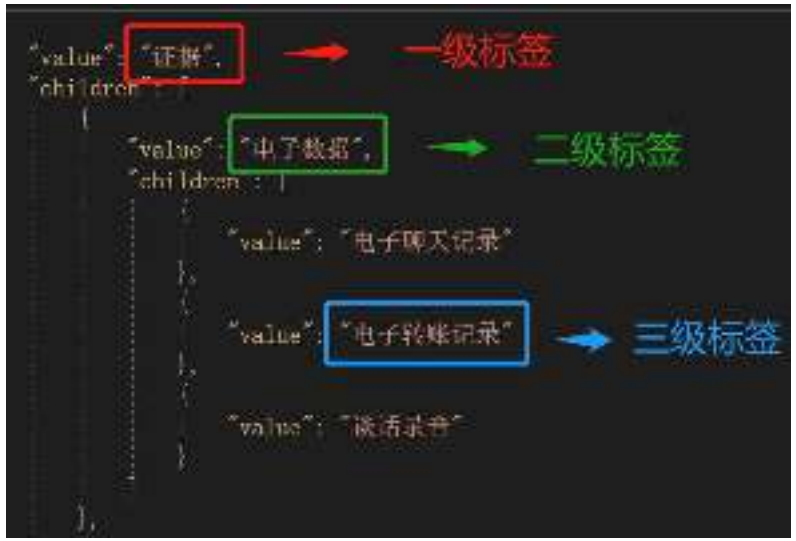
- 预期收获
  - 1. 了解层次多标签基本概念
  - 2. 了解层次多标签分类的常用方法
  - 3. 掌握层次感知方法的算法原理
  - 4. 了解领域实际应用和发展方向

背景简介

背景简介

- 多分类（Multi-Class）
  - 也称多元分类，指在分类任务中有2个以上类别
  - 各标签互斥，每个样本有且仅有一个标签
- 多标签分类（Multi-Label）
  - 给每个样本一系列的目标标签，各标签不互斥
  - 如一部电影可同时被打上动作片和犯罪片标签
- 层次多标签分类（Hierarchical Multi-Label）
  - 多标签分类下的子任务，标签类别被组织成层次结构
  - 如一篇裁判文书可被打上证据/电子证据/电子转账记录标签





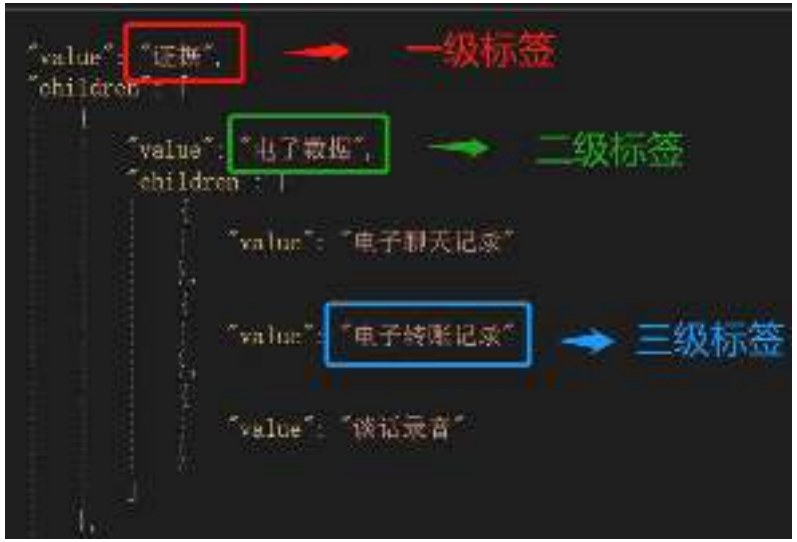
背景简介

- 层次多标签分类方法：

基本概念

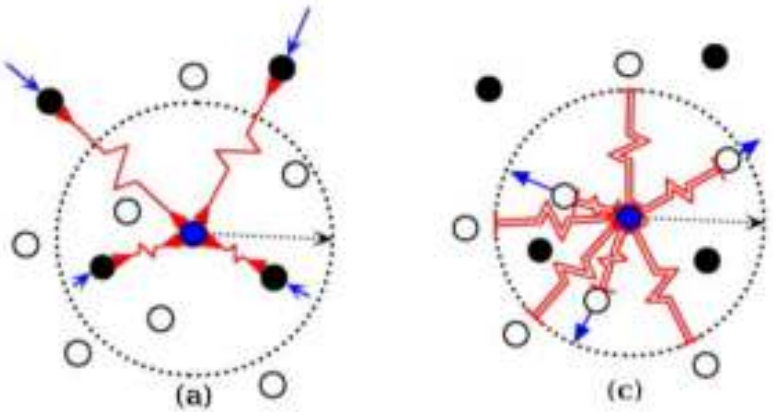
基本概念

- 层次多标签分类中的标签粒度定义
  - 细粒度标签
    - 是用于描述输入样本的最合适的标签
    - 通常是最底层的叶子标签
  - 粗粒度标签
    - 通常是粗粒度标签或细粒度标签的父节点
    - 通常是除叶子标签外的其它层标签



基本概念

- 对比学习
  - 思想：将样本和与它语义相似的例子（正样本）以及与它语义不相似的例子（负样本）进行对比
  - 目标：使得正样本对在表征空间中接近，负样本对在表征空间中远离
  - 难点：正负样本的构造和对比损失的设计



算法原理

算法原理 HiMatch

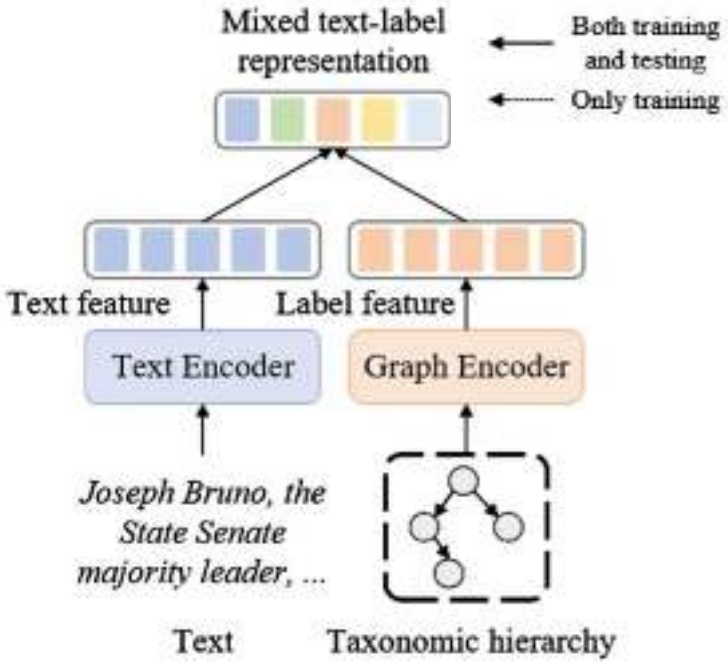
T	预测样本x_i在给定标签层次结构中的多个所属标签
I	文本x_i、先验标签集合L=$\{l_1, l_2, ..., l_N\}$
P	1.对文本进行编码，得到文本表征向量S_t 2.对先验标签体系进行编码，得到标签表征向量S_l 3.将S_t和S_l映射到同一表征空间 4.根据损失函数进行多任务优化
O	文本x_i对应的所有标签

P	对层级性的标签体系进行编码
C	数据集中的标签空间本身带有层次结构
D	如何利用先验层级知识引导和约束网络学习
L	ACL 2021

算法原理 HiMatch

- 核心思想
  - 将文本和标签分别进行表征学习，根据两表征向量定义不同的优化目标，从而提升层次多标签文本分类效果

- 逻辑框架
  - 1. 文本编码
  - 2. 先验标签体系编码
  - 3. 优化目标



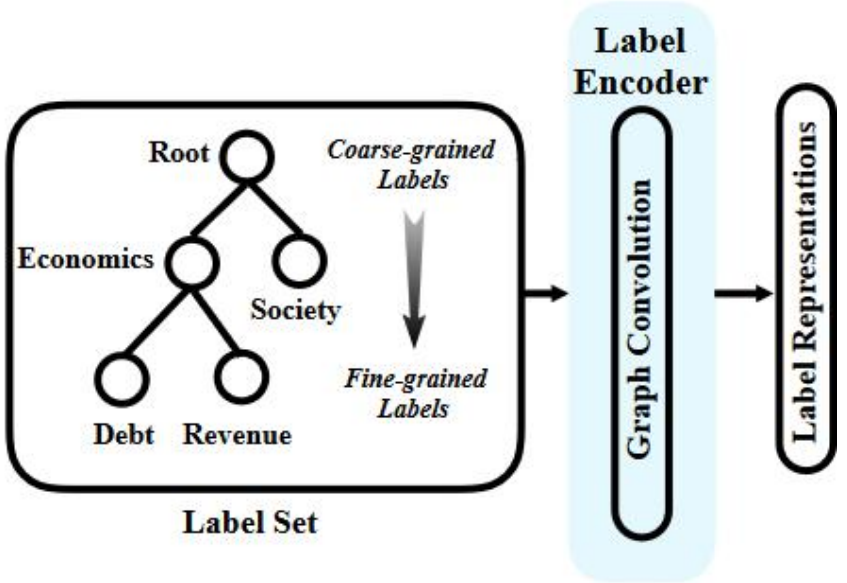
算法原理 HiMatch

- 标签表征
  - 利用先验知识将标签体系定义为图$G = (V_l, \vec{E}, \overleftarrow{E})$，

    V_l是标签表征向量，$\vec{E}$表示父节点到子节点的路径，

    $\overleftarrow{E}$表示子节点到父节点的路径，路径上的值是根据数据集统计而来的先验概率
  - 使用GCN网络进行表征学习

$$S_l = \sigma(\overleftarrow{E} \cdot V_l \cdot W_{g3} + \vec{E} \cdot V_l \cdot W_{g4})$$

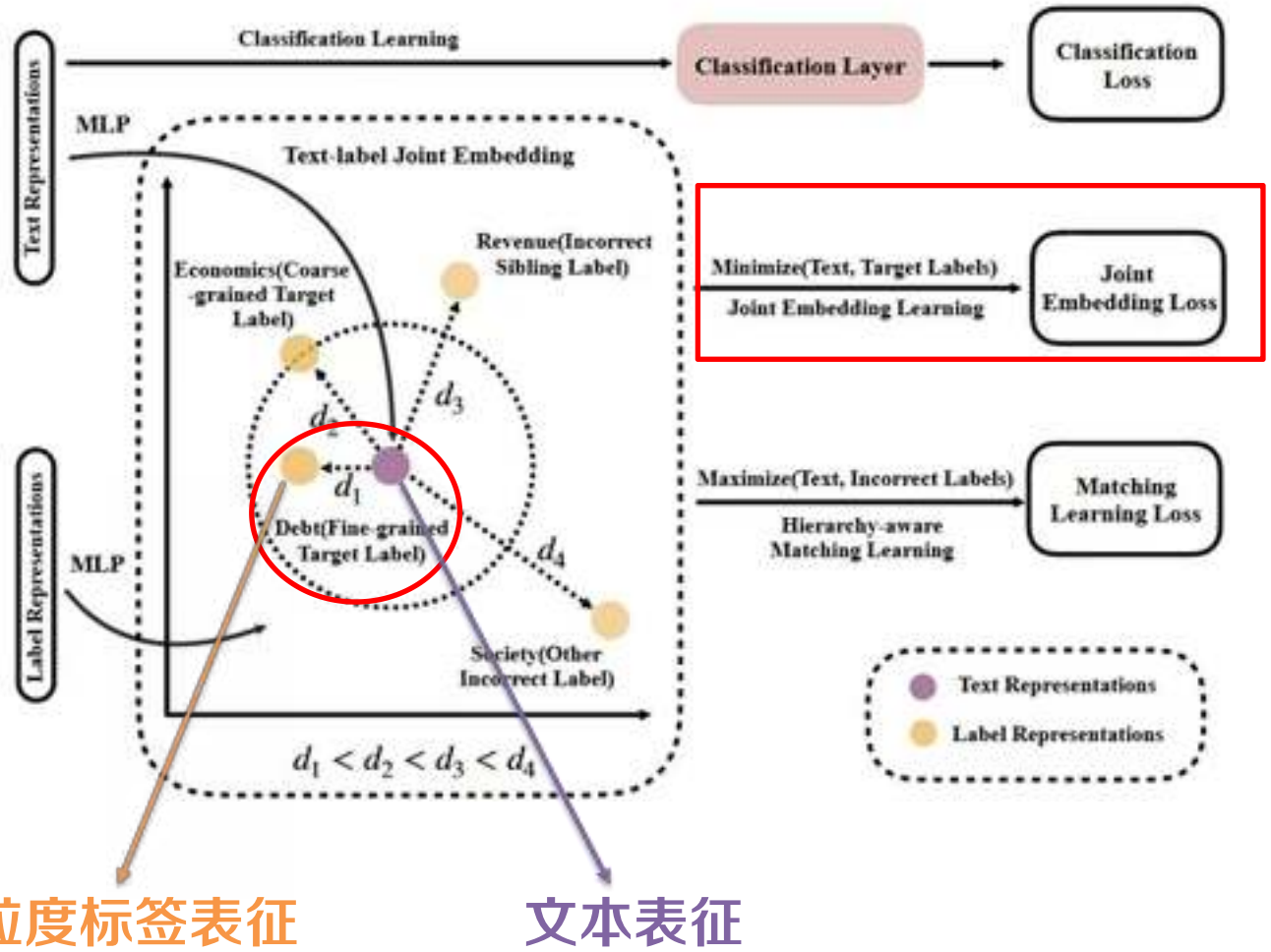


算法原理 HiMatch

- 优化目标：
  - 将文本表征S_t和先验标签表征S_l映射到同一表征空间

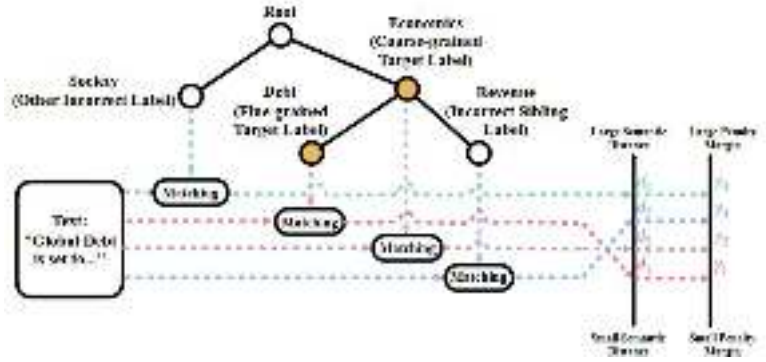
$$\Phi_t = FFN_t(S_t) \quad , \quad \Phi_l = FFN_l(S_l)$$

- 多任务损失：
  - 联合嵌入损失
    - 目标：希望表征的文本语义向量与它对应的真实标签表征的语义向量越近越好
    - $\mathcal{L}_{joint} = \sum_{p \epsilon P(y)} ||\Phi_t - \Phi_l^p||_2^2$
    - $P(y)$是目标标签集



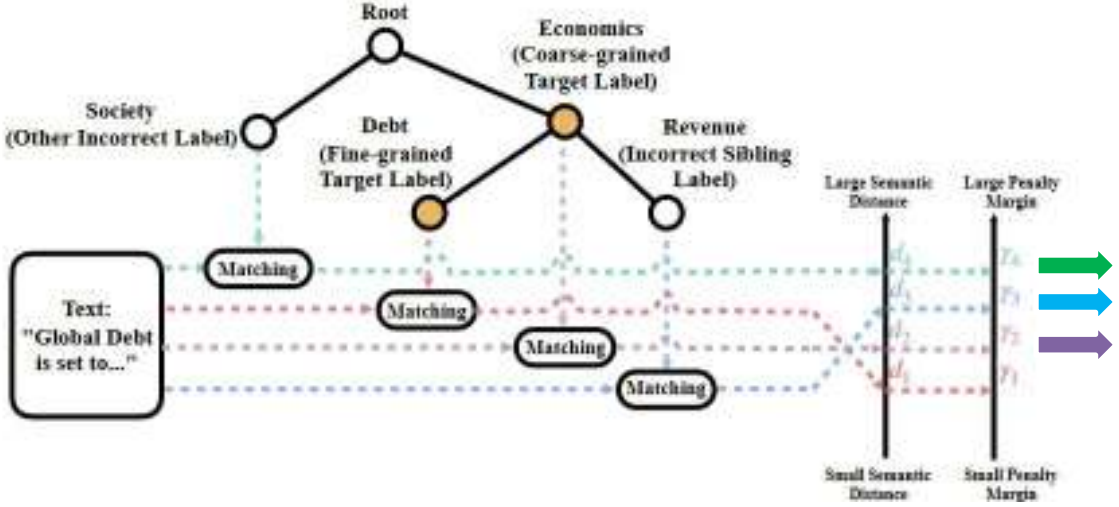
算法原理 HiMatch

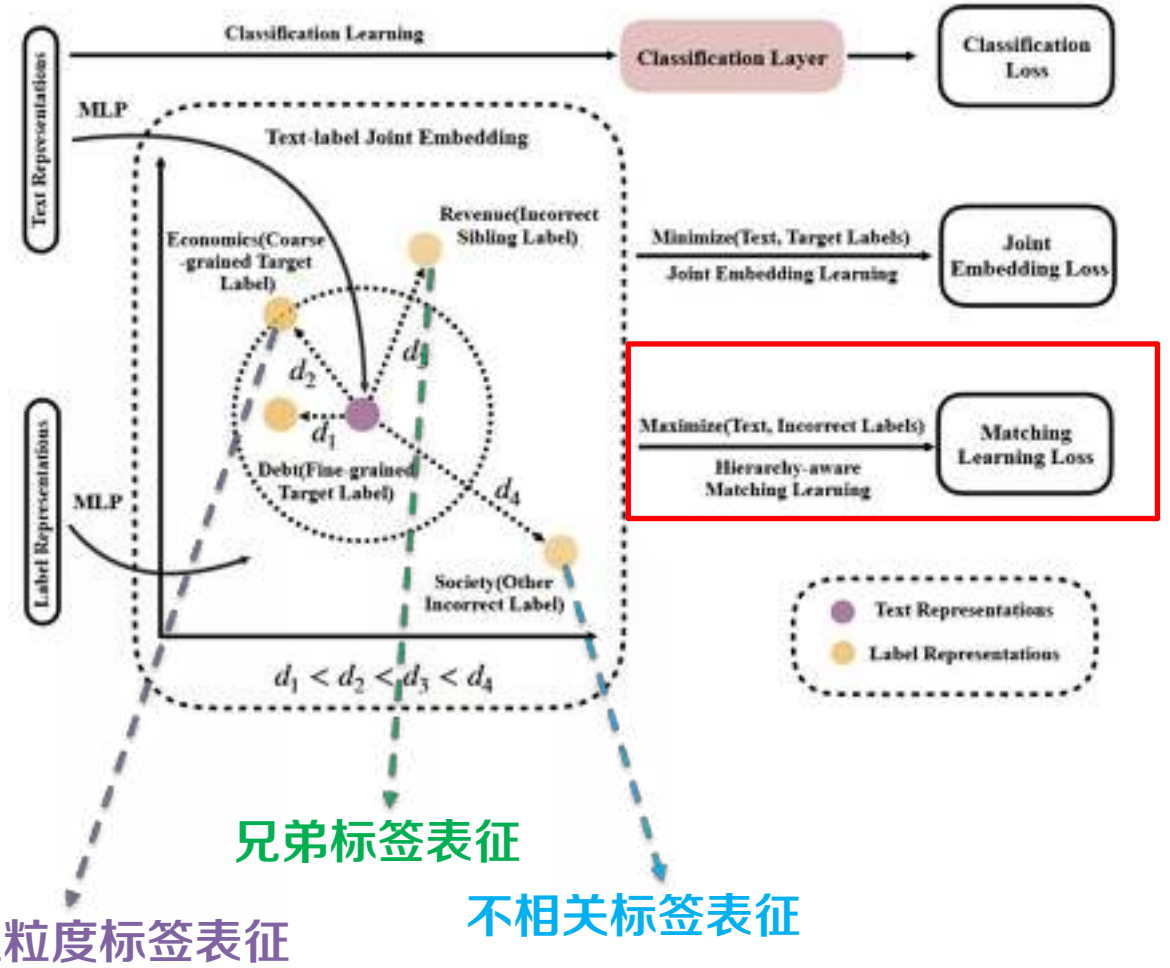
- 多任务损失：
  - 层次感知匹配损失—层次感知抽样
    - 由于HMTC任务的大量标签集，为每个标签计算匹配损失尤为耗时，故为每个细粒度标签y抽样其所有粗粒度标签、一个兄弟标签和一个不相关标签作为y的负标签集$n \epsilon N(y)$
    - 文本语义应该与细粒度标签最匹配，拥有最短的文本-标签距离d_1，文本-粗粒度标签的语义匹配距离d_2稍大，文本-兄弟标签的语义匹配距离d_3更大，文本-不相关标签的语义匹配距离d_4最大
    - 引入层次感知的匹配惩罚边际γ_1、γ_2、γ_3、γ_4建模上述可比关系，若希望语义匹配距离更小，则对应更小的匹配惩罚边际$\gamma_2 = \alpha\gamma$； $\gamma_3 = \beta\gamma$； $\gamma_4 = \gamma$ $(0 < \alpha < \beta < 1)$
    - 由于文本语义和细粒度标签的匹配关系已在联合嵌入学习中被考虑，故忽略γ_1



算法原理 HiMatch

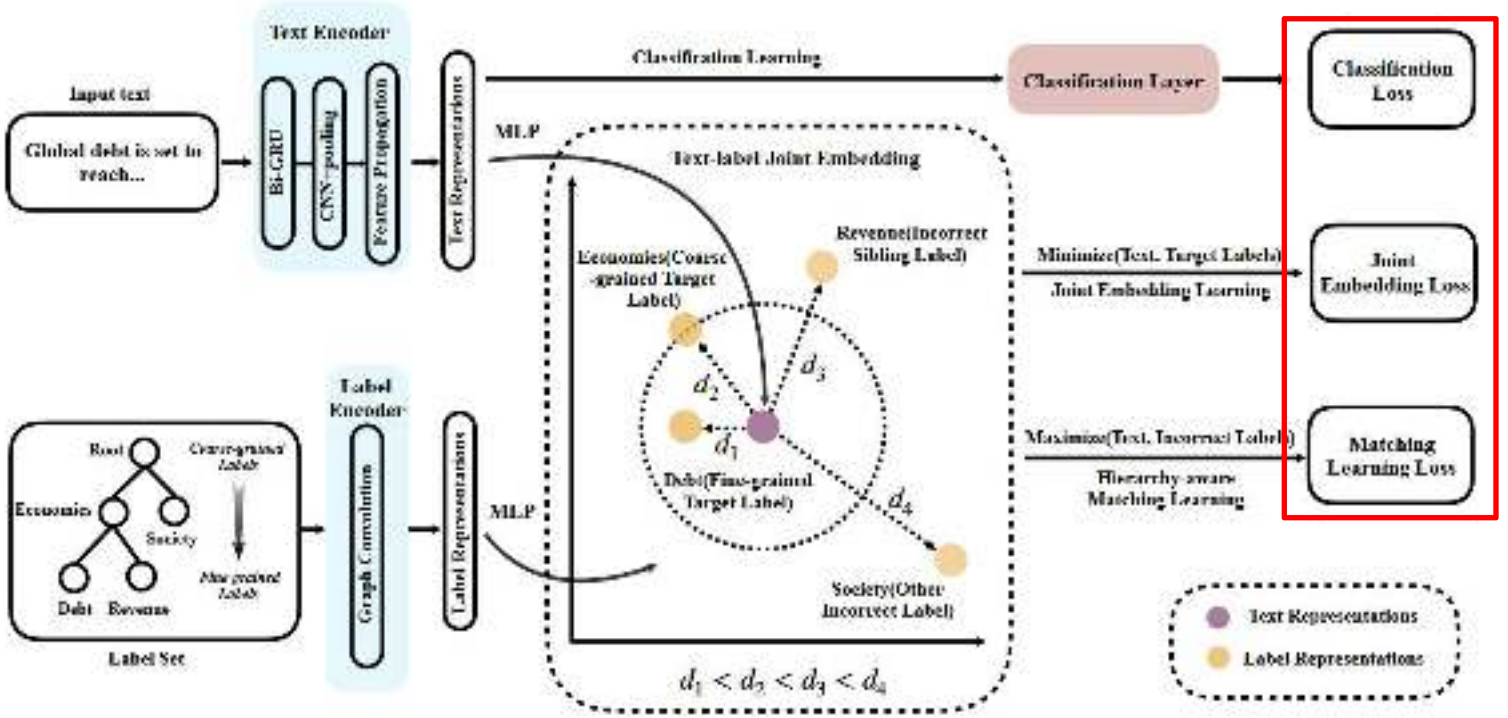
- 多任务损失：
  - 层次感知匹配损失
    - 目标：希望表征的文本向量不仅要与对应的真实标签向量越近越好，还要与非真实标签向量越远越好
    - $\mathcal{L}_{match} = \max(0, D(\Phi_t, \Phi_l^p) - D(\Phi_t, \Phi_l^n) + \gamma)$
    - Φ_l^p表示目标标签语义
    - n是非真实标签，对应为负标签集合





算法原理 HiMatch

- 多任务损失：
  - 分类损失
    - 目标：使用交叉熵损失，用于衡量模型分类的预测值与真实值之间的不一致程度
  - 总优化目标
    - $\mathcal{L} = \mathcal{L}_{cls}(y, \hat{y}) + \lambda_1 \mathcal{L}_{joint} + \lambda_2 \mathcal{L}_{match}$



实验分析

- 数据集
  - WOS
    - web of science中的论文数据集，包含出版论文的摘要和与之相关的主题，采用2级标签结构、标签总数达到141的标签体系
  - RCV1-V2
    - 新闻语料库数据集，包含大量路透社新闻故事，采用4级标签结构、标签总数达到103的标签体系

- 评价指标
  - Macro-F1
    - $F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$
    - $Macro-F1 = \frac{1}{N}\sum_{i=1}^{N} F1_{Ci}$
  - Micro-F1
    - $Precision_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i}$
    - $Recall_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i}$
    - $Micro-F1 = 2 \cdot \frac{Precision_{micro} \cdot Recall_{micro}}{Precision_{micro} + Recall_{micro}}$

实验分析 实验结果

- 对比实验

Models	Micro	Macro
Baselines		
TextRCNN (Zhou et al., 2020)	81.57	59.25
TextRCNN-LA (Zhou et al., 2020)	81.88	59.85
SGM (Zhou et al., 2020)	77.30	47.49
Hierarchy-Aware Models		
HE-AGCRCNN (Peng et al., 2019)	77.80	51.30
HMCN (Mao et al., 2019)	80.80	54.60
Htrans (Banerjee et al., 2019)	80.51	58.49
HiLAP-RL (Mao et al., 2019)	83.30	60.10
HiAGM (Zhou et al., 2020)	83.96	63.35
HiMatch	**84.73**	**64.11**
Pretrained Language Models		
BERT (Devlin et al., 2018)	86.26	67.35
BERT+HiMatch	**86.33**	**68.66**

RCV1-V2实验结果

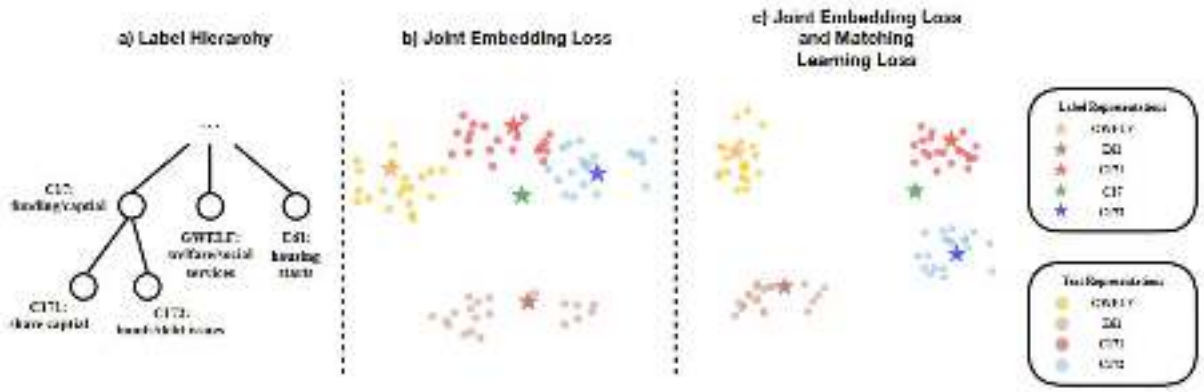
Models	Micro	Macro
Baselines		
TextRNN (Zhou et al., 2020)	77.94	69.65
TextCNN (Zhou et al., 2020)	82.00	76.18
TextRCNN (Zhou et al., 2020)	83.55	76.99
Hierarchy-Aware Models		
HiAGM (Zhou et al., 2020)	85.82	80.28
HiMatch	**86.20**	**80.53**
Pretrained Language Models		
BERT (Devlin et al., 2018)	86.26	80.58
BERT+HiMatch	**86.70**	**81.06**

WOS实验结果

- 实验结果分析
  - 以前的方法都忽略文本和标签的语义关系
  - HiMatch通过以层次感知的方法捕获文本-标签之间的匹配关系，达到了最好的结果

- 消融实验

Ablation Models	Micro	Macro
TextRCNN	81.57	59.25
HiMatch	**84.73**	**64.11**
- w/o Joint Embedding Loss	84.49	62.57
- w/o Matching Learning Loss	84.46	63.58
- w/o Hierarchy-aware Sampling	84.67	63.45



- 实验结果分析
  - 只引入Joint Embedding Loss，文本表征接近于相应的标签表征。但忽略了标签间的匹配关系，不同标签的文本表征可能会重叠
  - 再引入Matching Learning Loss，不同标签的文本表征更分离
  - HiMatch能捕获文本与不同粒度标签之间的语义关系

实验分析 实验结果

- 超参数实验

No.	γ	α	β	Micro	Macro
HiMatch					
①	0.2	0.01	0.5	**84.73**	**64.11**
Fine-tuning γ					
②	0.02	0.01	0.5	84.51	63.26
③	2	0.01	0.5	84.69	63.55
Fine-tuning α, β					
④	0.2	0.5	0.01	84.52	63.35
⑤	0.2	1	1	84.37	63.45
⑥	0.2	0.01	0.01	84.49	63.20
⑦	0.2	0.5	0.5	84.47	64.02

匹配惩罚边际$\gamma_2 = \alpha\gamma$； $\gamma_3 = \beta\gamma$； $\gamma_4 = \gamma$ $(0 < \alpha < \beta < 1)$

- 实验结果分析
  - 实验②③微调γ，得到最佳γ为0.2
  - 实验④⑤⑥⑦微调α、β
    - ④违反了层次结构，性能降低
    - ⑤忽略了层次结构，性能降低
    - ⑥⑦忽略了粗粒度标签和兄弟标签的关系，性能降低
  - 为粗粒度标签设置一个小的匹配惩罚边际，为不相关标签设置一个大的匹配惩罚边际是必要的

优劣分析

- 横向对比
  - 通过多任务优化提升F1值
  - 单独建模文本和先验标签，在模型测试阶段仍需编码先验标签

- 纵向对比
  - 提出了层次感知的语义匹配网络，充分考虑了文本与粗/细粒度标签关系
  - 首次将HMTC任务视为语义匹配任务

算法原理

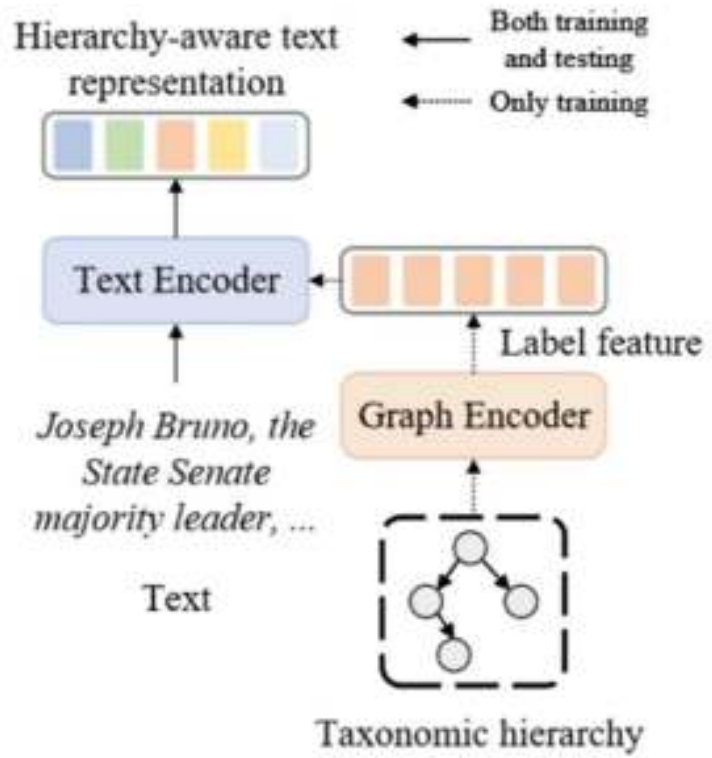
算法原理 HGCLR

T	预测样本x_i在给定标签层次结构中的多个所属标签
I	文本x_i、先验标签集合L=$\{l_1, l_2, ..., l_N\}$
P	1.对文本和先验标签体系进行编码 2.构造正样本 3.对比学习模块 4.分类和损失计算
O	文本x_i对应的所有标签

P	在层次标签的指导下构造正样本
C	数据集中的标签空间本身带有层次结构
D	如何利用先验层级知识引导和约束网络学习
L	ACL 2022

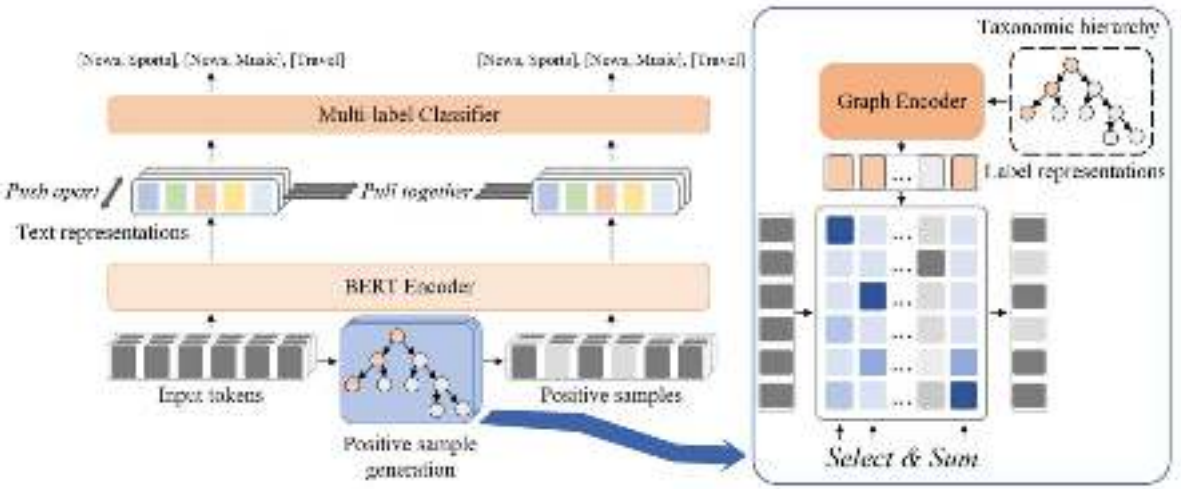
算法原理 HGCLR

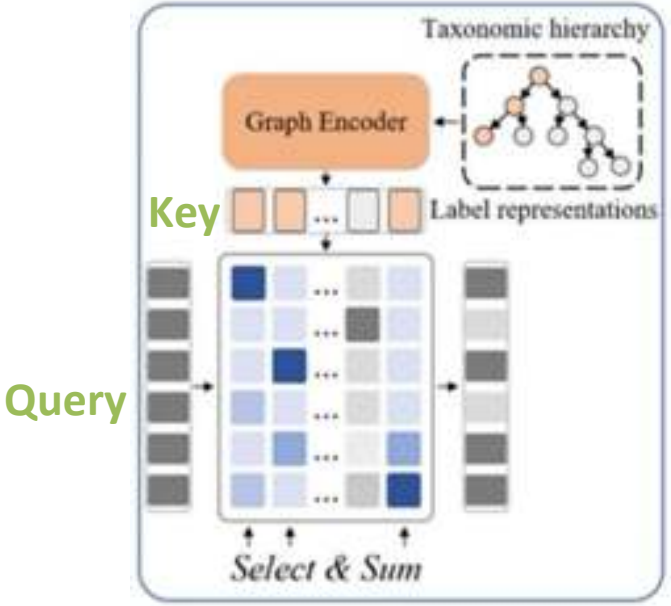
- 核心思想
  - 将层次嵌入到文本编码器中，在训练时HGCLR在标签层次结构的指导下，为输入文本构建正样本。通过将输入文本和它的正样本放在一起，文本编码器学习到独立地生成支持层次结构的文本表示
- 逻辑框架
  - 1. 文本编码
  - 2.先验标签体系编码
  - 3.正样本生成
  - 4.对比学习模块
  - 5.分类和目标函数



算法原理 HGCLR

- 文本编码
  - $\{e_1, e_2, ..., e_n\} = BERT\ (x)$
  - e_i为每个样本词嵌入
- 先验标签编码
  - 利用图编码，得到l_j作为每个标签嵌入
- 正样本生成
  - 目标：在保留样本部分词的同时保留样本标签
    - $q_i = e_i W_Q$， $k_j = l_j W_K$， $A_{ij} = \frac{q_i k_j^T}{\sqrt{d_h}}$
    - $P_i = \sum_{j \in y} P_{ij}$
    - $\hat{x} = \{x_i\ if\ P_i > \gamma\ else\ 0\}$
    - $\widehat{H} = BERT(\hat{x})$
  - Query：样本词嵌入；Key：标签嵌入





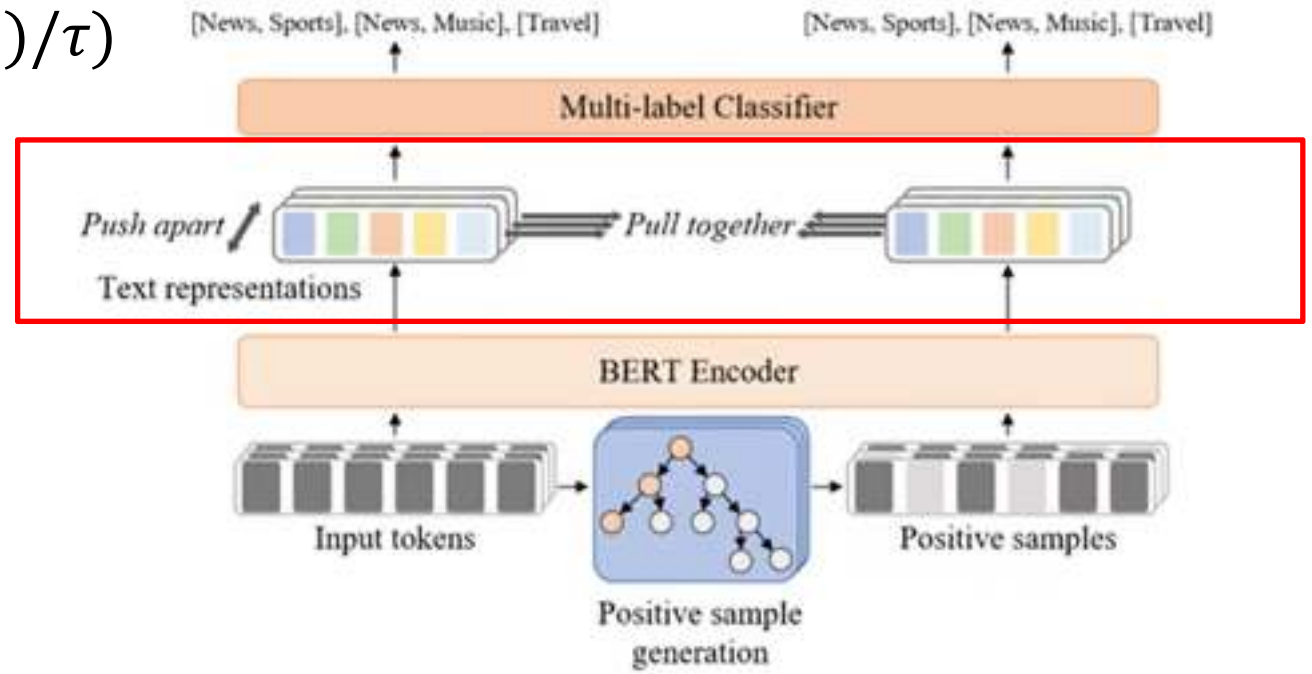
算法原理 HGCLR

- 对比学习模块：

$$\mathcal{L}_m^{con} = -\log \frac{\exp(sim(z_m, \mu(z_m))/\tau)}{\sum_{i=1, i\neq m}^{2N} \exp(sim(z_m, z_i)/\tau)}$$

$$\mu(z_m) = \begin{cases} c_i, & if\ z_m = \hat{c}_i \\ \hat{c}_i, & if\ z_m = c_i \end{cases}$$

$$\mathcal{L}^{con} = \frac{1}{2N} \sum_{m=1}^{2N} \mathcal{L}_m^{con}$$



算法原理 HGCLR

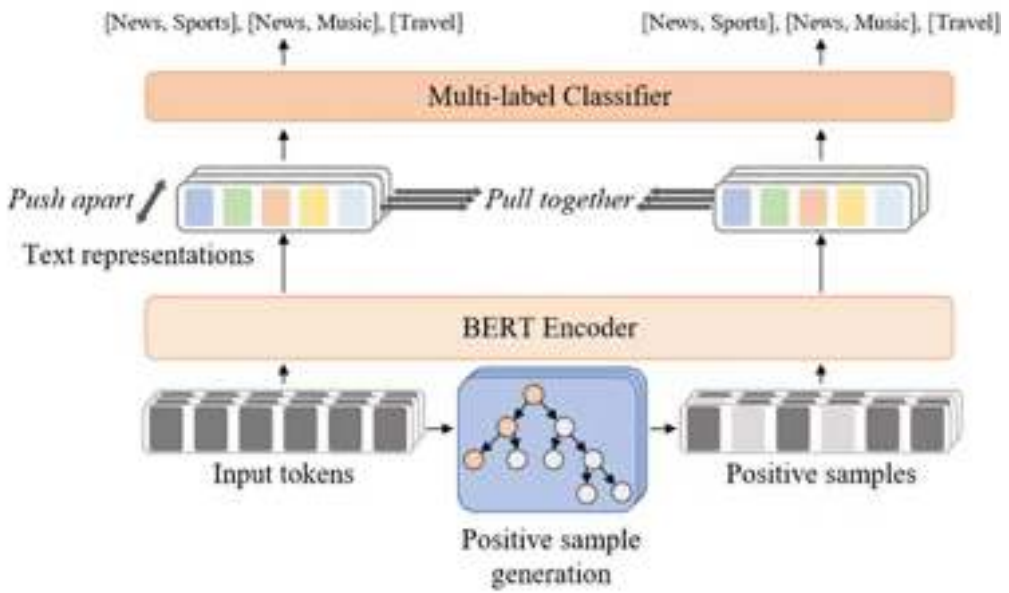
- 分类和目标函数：

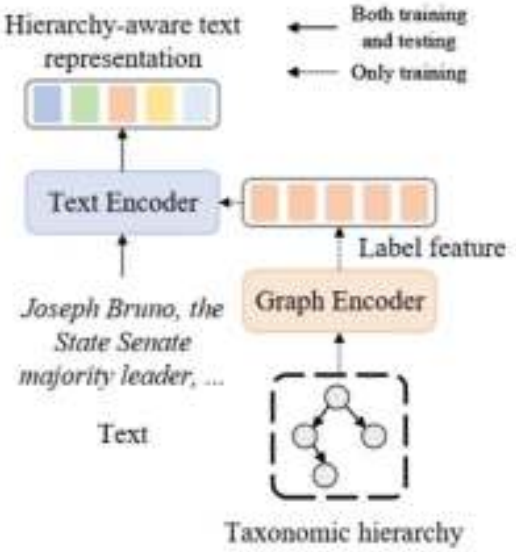
$$L^C$$

$$=-\sum_{i=1}^{N}\sum_{j=1}^{k}[y_{ij}\log(p_{ij})+(1-y_{ij})\log(1-p_{ij})$$

$$\mathcal{L}=\mathcal{L}^C+\widehat{\mathcal{L}^C}+\lambda\mathcal{L}^{con}$$

- 在模型测试阶段，仅用文本编码器进行分类，模型退化为 的Bert模型





实验分析

- 数据集：
  - WOS
    - web of science中的论文数据集，包含出版论文的摘要和与之相关的主题，采用2级标签结构、标签总数达到141的标签体系
  - RCV1-V2
    - 新闻语料库数据集，包含大量路透社新闻故事，采用4级标签结构、标签总数达到103的标签体系
  - NYT
    - 纽约时报数据集，包含大量纽约时报新闻故事，采用8级标签结构、标签总数达到166的标签体系
- 评价指标
  - Macro-F1
  - Micro-F1

- 对比实验结果

Model	WOS		NYT		RCV1-V2	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
Hierarchy-Aware Models						
TextRCNN (Zhou et al., 2020)	83.55	76.99	70.83	56.18	81.57	59.25
HiAGM (Zhou et al., 2020)	85.82	80.28	74.97	60.83	83.96	63.35
HTCInfoMax (Deng et al., 2021)	85.58	80.05	-	-	83.51	62.71
HiMatch (Chen et al., 2021)	86.20	80.53	-	-	84.73	64.11
Pretrained Language Models						
BERT (Our implement)	85.63	79.07	78.24	65.62	85.65	67.02
BERT (Chen et al., 2021)	86.26	80.58	-	-	86.26	67.35
BERT+HiAGM (Our implement)	86.04	80.19	78.64	66.76	85.58	67.93
BERT+HTCInfoMax (Our implement)	86.30	79.97	78.75	67.31	85.53	67.09
BERT+HiMatch (Chen et al., 2021)	86.70	81.06	-	-	86.33	**68.66**
HGCLR	**87.11**	**81.20**	**78.86**	**67.96**	**86.49**	68.31

- 消融实验结果

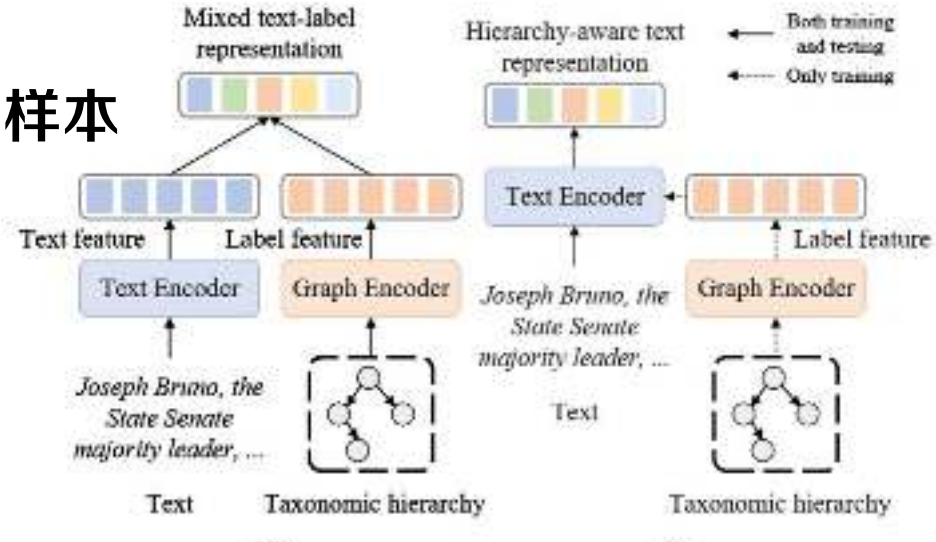
Ablation Models	Micro-F1	Macro-F1
BERT	85.75	79.36
HGCLR	**87.46**	**81.52**
-*r.p.* GCN	87.06	80.63
-*r.p.* GAT	87.18	81.45
-*r.m.* graph encoder	86.67	80.11
-*r.m.* contrastive loss	86.72	80.97

Generation Strategy	Micro-F1	Macro-F1
Hierarchy-guided	**87.46**	**81.52**
Dropout	86.94	79.91
Random masking	87.19	81.16
Adversarial attack	86.67	80.24

优劣分析

- 横向对比
  - 在训练阶段将先验标签注入到文本编码器中，使得文本编码器学习到独立地生成支持层次结构的文本表示，在测试阶段无需编码先验标签

- 纵向对比
  - 首次将对比学习引入HMTC任务
  - 考虑了任务独有性，在标签层次的指导下构造正样本

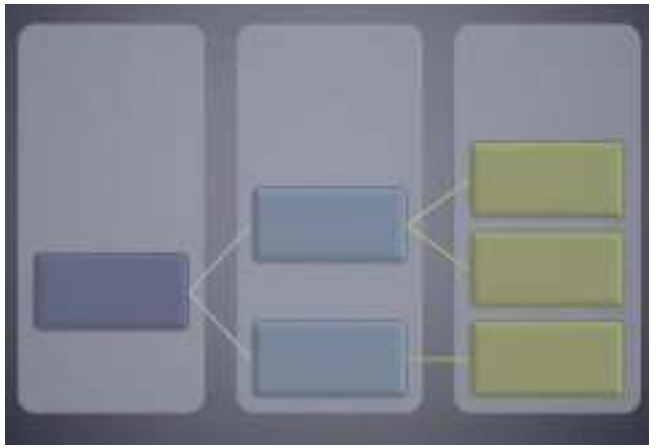


HiMatch HGCLR

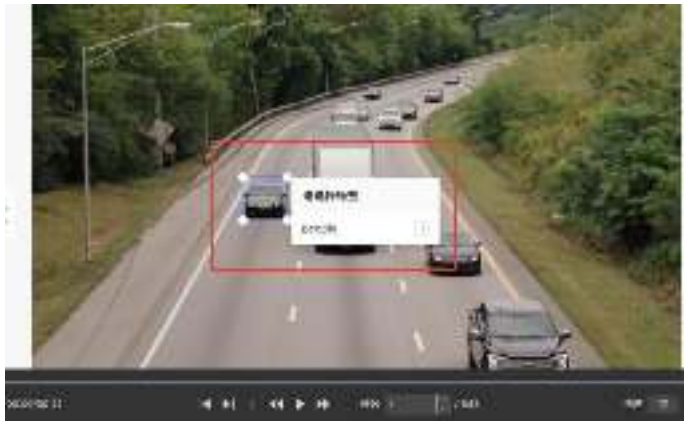
应用总结

应用总结

- 应用领域
  - 文献组织
  - 图像识别
  - 视频注释
  - 基因功能预测



- 未来的发展
  - 时间和空间效率的提升
  - 极端层次多标签的处理



参考文献

[1] Haibin Chen, Qianli Ma, Zhenxi Lin and Jiangyue Yan. Annual Hierarchy-aware Label Semantics Matching Network for Hierarchical Text Classification. Meeting of the Association for Computational Linguistics, 2021.
[2] Zihan Wang, Peiyi Wang, et al. Incorporating Hierarchy into Text Encoder: a Contrastive Learning Approach for Hierarchical Text Classification. Meeting of the Association for Computational Linguistics, 2022.
[3] Jie Zhou, Chunping Ma, Dingkun Long, , et al. Hierarchy-aware global model for hierarchical text classification. Meeting of the Association for Computational Linguistics, 2020.

上善若水。水善利万物而不争，处众人之所恶，故几於道。居善地，心善渊与善仁，言善信，正善治，事善能，动善时。夫唯不争，故无尤。