CCL23-Eval任务6系统报告：基于原型监督对比学习和模型融合的电信网络诈骗案件分类

熊思诗 张劼 赵宇 刘欣璋 宋双永
中国电信数字智能科技分公司
{xiongsishi, zhangj157, zhaoy11, liuxz2, songshy}@chinatelecom.cn

摘要

本文提出了一种基于原型监督对比学习和模型融合的电信网络诈骗案件分类方法。为了增强模型区分易混淆类别的能力，我们采用特征学习与分类器学习并行的双分支神经网络训练框架，并通过领域预训练、模型融合、后置分类等策略优化分类效果。最终，本文方法在CCL2023-FCC评测任务上取得了Macro-F1为0.8601 的成绩。

关键词： 文本分类；原型监督对比学习；模型融合；电信网络诈骗

System Report for CCL23-Eval Task 6: Classification of Telecom Network Fraud Cases Based on Prototypical Supervised Contrastive Learning and Model Fusion

Sishi Xiong Jie Zhang Yu Zhao Xinzhang Liu Shuangyong Song
China Telecom Corporation Ltd. Data&AI Technology Company
{xiongsishi, zhangj157, zhaoy11, liuxz2, songshy}@chinatelecom.cn

Abstract

We propose a method based on prototypical supervised contrastive learning and model fusion for telecom network fraud case classification (FCC) tasks. We introduce a parallel framework of feature learning and classifier learning, which enhances model capability of distinguishing confusing classes. We also take advantage of domain-specific pre-training, multi-model integration and post-classification modules to improve the overall performance. Our method achieves a final score of 86.01% Macro-F1 value on CCL2023-FCC evaluation task.

Keywords: Text Classification , Prototypical Supervised Contrastive Learning , Model Fusion , Telecom Network Fraud

1 引言

电信网络诈骗案件分类任务旨在将给定的电信诈骗案件描述文本自动归类到12个不同类别上。其中，易混淆的案件主要集中在几个特定类别之间，如“冒充电商物流客服类”、“虚假征信类”、“虚假购物、服务类”、“虚假网络投资理财类”等。

在参与CCL2023评测[0]的过程中，我们基于原型监督对比学习思想，提出一种特征学习与分类器学习并行的双分支神经网络训练框架，同时针对易混淆类别设计后置分类、模型融合方案，有效提升了系统评测指标。

[0]评测网址：https://github.com/GJSeason/CCL2023-FCC

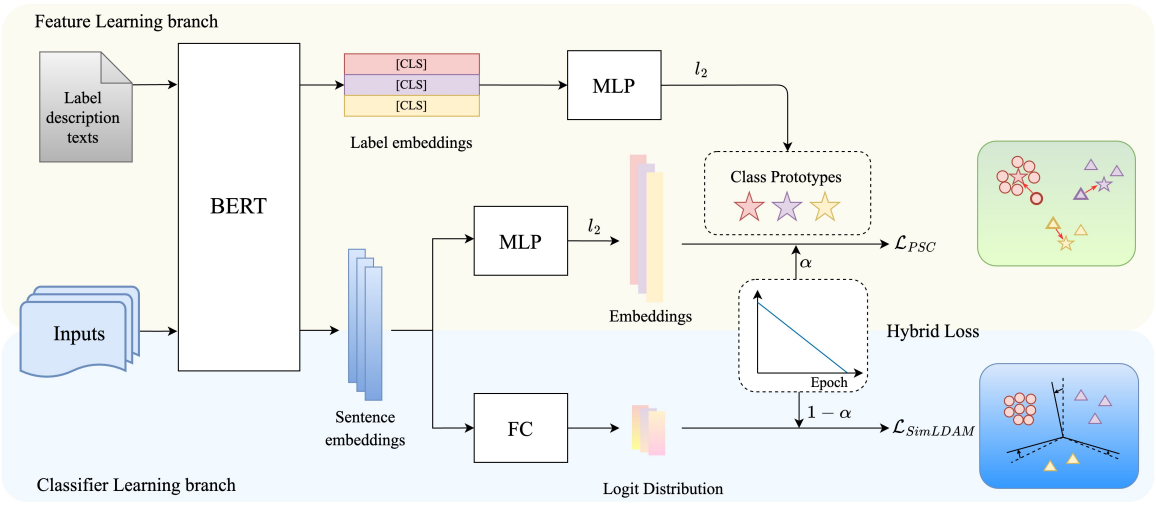

Figure 1: 基于原型监督对比学习的双分支网络模型

2 模型介绍

2.1 原型监督对比学习

原型对比学习(Prototypical Contrastive Learning, PCL)由Li等人 (2021)提出，该方法统一了聚类学习和对比学习思想，将属于同一类的样本聚集在原型附近。Wang等人 (2021)首次将原型对比学习与监督对比学习 (Khosla et al., 2020)结合起来，提出原型监督对比学习(Prototypical Supervised Contrastive Learning, PSCL)，不仅具有PCL的优势，还具备更强的语义辨析能力。

因此，我们基于BERT (Devlin et al., 2019)编码器和PSCL思想，设计了一个特征学习与分类器学习并行的双分支网络模型，如图1所示。特征学习分支由原型监督对比学习损失函数指导训练，公式如下：

$$\mathcal{L}_{PSC} = -log\frac{exp(v \cdot p_y/\tau)}{\sum_{j=1}^{C} exp(v \cdot p_j/\tau)} \tag{1}$$

其中，y为样本所属的真实类别，v表示样本的特征向量，p_y、p_j分别表示类别y和类别j的原型表征向量。从式(1)可以看出，$\mathcal{L}_{PSC}$鼓励样本尽可能向真实类别的原型靠近、远离负类原型，从而使类内特征分布更为紧凑，有助于模型学习到更为均衡且易分离的表征空间。

具体地，我们参考公开资料[1]，为每个标签归纳了一段描述文本(Label Description)作为先验知识。把描述文本输入到双分支共享的、可训练的编码器，经过MLP映射层与L_2归一化后，可以获得各个类别的原型语义表征。标签描述的语义特征代表了同类样本共有的特征和规则，能够帮助模型更好地区分容易混淆的类别。

分类器学习分支由分类损失函数指导训练。我们基于原型结构提出了一种新的分类损失函数SimLDAM。它基于LDAM损失函数 (Cao et al., 2019)改进，公式如下：

$$\mathcal{L}_{SimLDAM} = -log\frac{e^{(z_y-\Delta_y)}}{e^{(z_y-\Delta_y)} + \sum_{j\neq y} e^{z_j+\gamma s(p_y,p_j)}} \tag{2}$$

其中，z_j、z_y分别表示类别j和真实类别y的逻辑分数，Δ_y表示在真实类别y上的附加边距，是一个和类别y的样本数量负相关的常数，用于缓解样本分布的不均衡性。同时，考虑到评测任务的标签粒度较细，SimLDAM在LDAM基础上融合了原型相似度惩罚$\gamma s(p_y, p_j)$，以缓解原型相近的类别之间的混淆程度，其中$s(\cdot)$为余弦相似度函数，γ为超参数。

假设类别y与类别j之间存在高度混淆。PSCL拉近样本与真实类别原型的距离，同时使其与负类原型的距离增加，促使不同类别的样本分离开。若类别y与类别j的原型表征相近，SimLDAM损失函数添加了一项与原型相似度相关的惩罚项，从而能抑制模型在易混淆负类j上的逻辑输出值，进一步降低混淆程度。

[1]公开资料来源：https://baijiahao.baidu.com/s?id=1733668985271273518

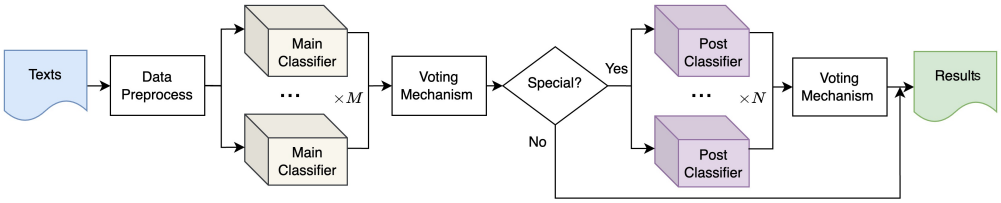

Figure 2: 系统框架

模型的训练过程采用渐进学习策略 (Zhou et al., 2020)，整体损失是对比损失$\mathcal{L}_{PSC}$和分类损失$\mathcal{L}_{SimLDAM}$加权的混合损失，公式如下：

$$\mathcal{L} = \alpha\mathcal{L}_{PSC} + (1-\alpha)\cdot\lambda\mathcal{L}_{SimLDAM} \tag{3}$$

其中，λ为分类损失权重参数，$\alpha = 1 - T/T_{max}$为渐进权重，T、T_{max}分别为当前训练轮次和最大训练轮次。在训练初期，特征学习分支占主导地位，强调不同类别特征空间的差异化。随着训练的进行，分类器学习分支的权重逐步增大，更侧重于调整分类决策边界。

2.2 领域预训练

本次评测中，我们进行了任务持续预训练(TAPT) (Sun et al., 2019; Gururangan et al., 2020)，选取的底座以BERT类模型为主，包括但不限于BERT、RoBERTa (Liu et al., 2019)、ALBERT (Lan et al., 2020)等。我们尝试了不同的预训练模型，并针对训练语料特点设计了特殊的训练策略 (Pan et al., 2020; Dong et al., 2023)，包括：

1. 词表适应：将[unused]替换为在单一分类下高频出现的、类别独有的任务强相关词汇；
2. MLM任务适应：修改mask替换概率，高频词汇被mask概率更高；
3. NSP任务适应：使用整句截断代替字随机截断，与下游任务的长文本处理方式对齐。

2.3 模型融合

如图2所示，我们在模型融合部分采用K-Fold交叉验证的方法，旨在提升系统整体的泛化性。具体地，我们在训练集上随机划分出K个不同的开发集，并使用不同的底座训练得到多个基础分类器。评测时，我们选取效果最好的M个基础分类器，在每个测试样本上进行多数决策投票，获得模型集成后的最终预测结果。最终提交版本的参数为M=7，其中6个模型基于BERT$_{BASE}$预训练，1个模型基于BERT$_{LARGE}$预训练。

2.4 后置分类

在模型迭代过程中，我们统计了模型集成后的投票结果与人工标注的分布差异，得到不同类别之间的混淆矩阵。我们发现，已有方案在“冒充电商物流客服类”、“虚假征信类”两个类别上相互混淆程度最高，误识别的样本最多，如图3中的红框所示。

Figure 3: 混淆矩阵图

基于上述分析，我们在原有方案基础上增加一个后置二分类模块(Post Classifier)以解决这两个类别识别效果欠佳的问题，如图2所示。我们仅使用这两个类别的训练数据训练多个二分类模型，选取最好的5个模型进行多数决策投票。预测阶段，如果主模型(Main Classifier)投票结果属于这两个分类其中之一，则使用后置分类器进行二次补充判别。

第二十二届中国计算语言学大会论文集，第201页-第205页，哈尔滨，中国，2023年8月3日至5日。

3 评测结果

3.1 实验设置

本次评测中，我们采用领域预训练过的BERT作为基座，base版学习率为2×10^{-5}，large版学习率为2×10^{-6}，为使模型训练更稳定，学习率采用了WarmUp (He et al., 2016)与余弦退火衰减策略。优化器使用AdamW (Loshchilov and Hutter, 2019)，最大句子长度设置为512。模型训练过程中还采用了FGM对抗训练 (Miyato et al., 2017)和R-Drop (Liang et al., 2021)方法以增强泛化能力。

3.2 分析与讨论

表1展示了我们采用的不同策略的真实评测指标结果。可以看出，原型监督对比学习方法的总体效果提升最为明显，获得了0.68% 的性能提升。基座方面，Large效果略优于Base。

框架	策略	效果	提升
单模型	BERT_{BASE}	0.8425	
	BERT_{BASE} + 领域预训练	0.8458	+0.33%
	BERT_{BASE} + 领域预训练+ PSCL	0.8526	**+0.68%**
	BERT_{LARGE} + 领域预训练+ PSCL	0.8538	+0.12%
多模型	模型融合	0.8577	+0.39%
	模型融合+ 后置分类器	0.8581	+0.04%
	模型融合+ 后置分类器+ FGM & R-Drop	**0.8601**	+0.20%

Table 1: 不同策略结果对比(Macro-F1)

表2对比了我们采用的不同策略在不同类别上的表现。可以看出，原型监督对比学习方法的正向收益最稳定，在大多数类别上都有性能提升，在部分类别上单模型效果甚至超过了多模型融合版本。同时，后置二分类模块在选定的两个分类上均有正向提升，验证了图3中错误样例分析的正确性。最终版本使用的FGM对抗训练和R-Drop方法也在四个类别上取得了效果提升。

	基线	+预训练	+PSC	+融合	+后置	最佳版本
冒充公检法及政府机关类	0.9044	0.8982↓	0.9162↑	**0.9216**↑	0.9216	0.9195↓
冒充军警购物类	0.7531	0.7733↑	0.7803↑	**0.7870**↑	0.7870	0.7870
冒充电商物流客服类	0.7924	0.8001↑	0.7850↓	0.7975↑	0.7981↑	**0.8060**↑
冒充领导、熟人类	0.8892	0.8894	0.8958↑	**0.9045**↑	0.9045	0.9045
刷单返利类	0.9571	0.9578	0.9559↓	**0.9616**↑	0.9616	0.9609
网络婚恋、交友类	0.6142	0.6424↑	**0.6547**↑	0.6500↓	0.6500	0.6545↑
网络游戏产品虚假交易类	0.9167	0.9095↓	**0.9204**↑	0.9130↓	0.9130	0.9170↑
网黑案件	0.9597	0.9675↑	**0.9754**↑	0.9714↓	0.9714	0.9754↑
虚假征信类	0.8180	0.7993↓	0.8175↑	0.8188↑	**0.8231**↑	0.8218↓
虚假网络投资理财类	0.8827	0.8704↓	0.8768↑	0.8890↑	0.8890	**0.8898**↑
虚假购物、服务类	0.6972	0.6945↓	0.7056↑	0.7265↑	0.7265	**0.7322**↑
贷款、代办信用卡类	0.9250	0.9461↑	0.9482↑	0.9518↑	0.9518	**0.9525**↑

Table 2: 不同类别结果对比(Macro-F1)

4 总结

在CCL2023-FCC评测中，我们使用原型监督对比学习、领域预训练、模型融合、后置分类、对抗训练和R-Drop等策略，有效提升了分类效果。其中，原型监督对比学习方法能够有效区分易混淆类别，带来的正向收益最为稳定。另外，我们还尝试过不同的长文本处理、投票机制、数据增强等策略，但是没有取得更好的效果。未来，我们会在这些方向上做进一步的探索和优化。

参考文献

Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. BBN: Bilateral-Branch Network with Cumulative Learning for Long-Tailed Visual Recognition. 2020. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9716-9725.* Seattle, WA, USA.

Chi Sun, Xipeng Qiu, Yige Xu and Xuanjing Huang. How to Fine-Tune BERT for Text Classification?. 2019. *Chinese Computational Linguistics, pages 194-206.*

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. 2019. *7th International Conference on Learning Representations.* New Orleans, LA, USA.

Jacob Devlin, Mingwei Chang, Kenton Lee and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4171-4186.* Minneapolis, Minnesota.

Junnan Li, Pan Zhou, Caiming Xiong and Steven Hoi. Prototypical Contrastive Learning of Unsupervised Representations. 2021. *9th International Conference on Learning Representations.*

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Aréchiga and Tengyu Ma. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. 2019. *Proceedings of the 33rd International Conference on Neural Information Processing Systems, pages 1567–1578.* Red Hook, NY, USA.

Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. Deep Residual Learning for Image Recognition. 2016. *Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770-778.* Las Vegas, NV, USA.

Peng Wang, Kai Han, Xiushen Wei, Lei Zhang and Lei Wang. Contrastive Learning Based Hybrid Networks for Long-Tailed Image Classification. 2021. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 943-952.* Nashville, TN, USA.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu and Dilip Krishnan. Supervised contrastive learning. 2020. *Advances in neural information processing systems, pages 18661-18673.*

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. 2020. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8342-8360.*

Takeru Miyato, Andrew M. Dai and Ian J. Goodfellow. Adversarial Training Methods for Semi-Supervised Text Classification. 2017. *5th International Conference on Learning Representations.* Toulon, France.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang and Tieyan Liu. R-Drop: Regularized Dropout for Neural Networks. 2021. *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, pages 10890-10905.*

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai and Xuanjing Huang. Pre-trained models for natural language processing: A survey. 2020. *Science China Technological Sciences, 63(10): 1872-1897.*

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019. *CoRR,abs/1907.11692.*

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. 2020.

Zican Dong, Tianyi Tang, Lunyi Li and Wayne Xin Zhao. A Survey on Long Text Modeling with Transformers. 2023. *CoRR,abs/2302.14502.*