CCL23-Eval 任务6系统报告：基于CLS动态加权平均和数据增强的电信网络诈骗案件分类

刘天昀，张兴华，宋梦潇，柳厅文
中国科学院信息工程研究所/ 北京市海淀区树村路19号
liutianyun, zhangxinghua, songmengxiao, liutingwen@iie.ac.cn

摘要

电信网络诈骗领域的案件分类作为文本分类的一项落地应用，其目的是为相关案件进行智能化的分析，有助于公安部门掌握诈骗案件的特点，针对性的预防、制止、侦查。本文以此问题为基础，从模型设计、训练过程、数据增强三个方面进行了研究，通过CLS动态加权平均、Multi-Sample Dropout、对抗训练FGM、回译等方法显著提升了模型对诈骗案件描述的分类性能。

关键词： 文本分类；电信网络诈骗

System Report for CCL23-Eval Task 6: Classification of Telecom Internet Fraud Cases Based on CLS Dynamic Weighted Average and Data Augmentation

Tianyun Liu, Xinghua Zhang, Mengxiao Song, Tingwen Liu
Institute of Information Engineering，CAS / Beijing, China
liutianyun, zhangxinghua, songmengxiao, liutingwen@iie.ac.cn

Abstract

The case classification in the field of telecommunications network fraud, as a practical application of text classification, aims to intelligently analyze relevant cases, help public security departments grasp the characteristics of fraud cases, and provide targeted prevention, suppression, and investigation. Based on this issue, this article conducts research from three aspects: model design, training process, and data enhancement. Through methods such as CLS dynamic weighted average, Multi Sample Dropout, adversarial training FGM, and backtranslation, the classification performance of the model in describing fraud cases has been significantly improved.

Keywords: Text Classification , Telecom Internet Fraud

1 引言

文本分类是自然语言处理领域的基础任务，面向电信网络诈骗领域的案件分类对智能化案件分析具有重要意义。诈骗案件分类是打击电信网路诈骗犯罪过程中的关键一环，根据不同的诈骗方式、手法对案件进行分类，有助于公安部门掌握当前电信网络诈骗案件的分布特点，进

第二十二届中国计算语言学大会论文集，第179页-第183页，哈尔滨，中国，2023年8月3日至5日。
(c) 2023 中国中文信息学会计算语言学专业委员会

而能够对不同类别的诈骗案件作出针对性的预防、监管、制止、侦查等措施。本评测的任务是将给定的案件描述文本进行分类。案件文本包含对案件的整体描述（经过脱敏处理），案件对应的类别共有12类。

本次测评我们采用了哈工大讯飞联合实验室发布的基于全词Mask的中文预训练模型Chinese-Bert-wwm-ext (Cui et al., 2020)作为基底模型。在模型设计方面，我们采用对基底模型中CLS表示动态加权平均和Multi-Sample Dropout的方式；在训练方面，我们通过FGM对抗训练增强模型泛化能力；在数据方面我们通过回译的方法对模型难以分类的案例进行数据增强。最终利用5折交叉融合的方式在宏平均F1值指标上达到了86.0242%。

1.1 赛题分析

电信网络诈骗，是指以非法占有为目的，利用电信网络技术手段，通过远程、非接触等方式，诈骗公私财物的行为[0]，给人民群众造成了巨大经济的损失，严重影响了广大人民群众的安全感、幸福感、获得感。通过对以往案件的分类对公安部门进行反诈宣传、风险防控具有重要的社会意义和应用价值。

面向电信网络诈骗领域的案件分类其本质为文本分类任务，旨在通过设计模型实现对案件的描述精准分类。该任务将诈骗领域案件分为了12个大类，提供82210样本用于模型训练，训练集数据分布情况见Table 1。其中“刷单返利类”数据占了总训练集的27.56%，而“网黑案件”和“冒充军警购物类”训练数据少于千条，由此我们可以考虑通过数据增强的方法来解决数据的长尾分布问题。此外，我们对数据的长度分布进行统计，其中80%的数据长度小于478，由此可以通过设置基础模型的最大处理长度加快训练、降低padding对最终性能的影响。

在此次测评中，使用宏平均F1值作为电信网络诈骗案件分类的评价指标，该指标首先分别计算每个类别的F1值，然后再进行平均得到最终的打分，所以当提升模型分类短板时可以有效提升模型性能。

Table 1: 训练集数据分布情况统计

类别名称	样本数量
刷单返利类	22656
冒充电商物流客服类	8804
虚假网络投资理财类	7539
贷款、代办信用卡类	7121
虚假征信类	5432
虚假购物、服务类	4581
冒充公检法及政府机关类	2920
冒充领导、熟人类	2811
网络游戏产品虚假交易类	1389
网络婚恋、交友类（非虚假网络投资理财类）	1064
冒充军警购物类	687
网黑案件	764
总计	82210

2 方法思路

在模型层面，为了使模型的语义表征能力更强，我们将基座模型CLS位置上的各层表示进行动态加权平均，同时采用Multi-Sample Dropout (Inoue, 2019)方法加速训练，增强模型的泛化能力，并采用交叉熵作为损失函数，模型架构如图 1所示。在训练方面，为了增强模型的泛化能力，我们采用对抗训练FGM对输入增加微小扰动，训练模型去区分样例是真实样例还是对抗样本。在数据层面，我们发现模型在“网络婚恋、交友类（非虚假网络投资理财类）”和“虚假

[0]http://www.npc.gov.cn/npc/c30834/202209/faadac81d2e94aa0bd7574efc9862cd0.shtml

购物、服务类”两类中分类效果极差，其中“网络婚恋、交友类（非虚假网络投资理财类）”在上述模型设计下仅能达到59%左右的性能，为此我们设计了回译的方法来增强该类。为了在有限的数据下提高最终的预测能力，我们通过5折交叉验证的方法进行模型融合，使不同子空间可以各取其长，从而提升模型的鲁棒性和泛化能力。

2.1 模型设计

2.1.1 CLS动态加权平均

在以往工作中 (Jawahar et al., 2019)，探究了Bert (Devlin et al., 2018)每一层的编码能力，证明了Bert深层较浅层学习到更为丰富的语言学信息；但是Bert无法在一层非常全面的学习到文本中的语言学信息，特定的层可能只会学到一些特定的语言学信息。同时由于Chinese-Bert-wwm-ext采用了与Bert相同的模型结构，基于此我们将所有Transformer层和编码层中的CLS位置权重进行动态加权平均，以增强向量的语义表征能力从而提升效果。即通过可学习的参数$W = [W_0, W_1, ..., W_n]$，设基座模型的层数m和Embedding层数l，如Chinese-Bert-wwm-ext共12层，Embedding有1层，故$n = m+l = 13$。设Embedding层表示为EL，Transormer层的表示为$TL_m, m = 12$。因该任务为分类任务，故我们选取模型每层输出的第一个token [CLS]的表示进行计算，故案件描述的分类表示为：

$$\mathbf{H} = sum(softmax(W) * [EL[0], TL_0[0], TL_1[0], ..., TL_m[0]]) \tag{1}$$

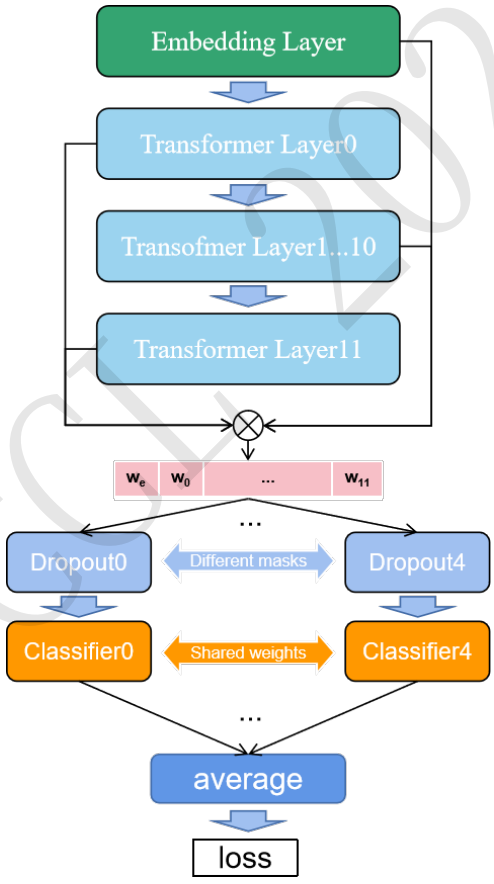

Figure 1: 模型架构图

2.1.2 Multi-Sample Dropout

参考以前的比赛经验，我们采用Multi-Sample Dropout对表示层的输出进行多次Dropout操作，一方面可以大大减少训练的迭代次数，另一方面可以使模型更加鲁棒，增强泛化能力。在获取到案件描述的表示$\mathbf{H}$后经过多次Dropout操作，并最终利用MLP作为分类头，得到最后的

第二十二届中国计算语言学大会论文集，第179页-第183页，哈尔滨，中国，2023年8月3日至5日。

预测结果。其中在该任务中$mh = 4$，故Multi-Sample Dropout的输出结果为：

$$\mathbf{MH} = average(MLP(\sum_{i=0}^{m} hdropout(\mathbf{H}))) \tag{2}$$

2.2 训练过程

对抗训练是一种引入噪声的训练方法，在尽量不改变原样本的分布，对样本增加扰动，使得模型能够忽视这种扰动，从而提升模型的鲁棒性。为此我们采用了FGM (Miyato et al., 2016)在训练过程中进行模型泛化，以此来提升模型对不同数据的泛化能力，提升分类效果。

2.3 数据增强

通过数据增强可以有效增加训练样本，缓解数据不平衡造成的模型偏见问题。通常由同义词替换、随机插入、随机替换与删除以及回译的方式进行增强，相关论文 (Wei and Zou, 2019)已经证明了该方法可以显著提升模型性能。在此次评测过程中，我们尝试了数据增强，并最终选择回译和重复采样的方法对模型难区分类别“网络婚恋、交友类（非虚假网络投资理财类）”和“虚假购物、服务类”进行了训练数据的扩充。回译分别采用德文、日文、韩文三种语言，如德文设置下进行“中文-德文-中文”回译。

3 实验

3.1 实验设定

本次参赛仅使用了官方提供的数据集，并未使用其他额外数据集。在训练过程中我们将学习率设置为3e-5，并采用10%的步数进行linear warmup，共训练5个epoch，将batch size设置为16，最大长度为512。关于数据增强，我们重复采样了“网络婚恋、交友类（非虚假网络投资理财类）”1064条数据和虚假购物、服务类4581条数据，并通过利用百度翻译 [0]API回译“网络婚恋、交友类（非虚假网络投资理财类）”描述产生了1159条（由于仅使用免费服务，故产生数据较少）。

3.2 评测结果

在此次测评中我们采用上述方法进行了模型的设计与训练，最终在宏平均F1指标上达到了0.8602。模型在不同设置下的实验结果如Table 2所示，其中Model Ours为Chinese-Bert-wwm-ext + CLS动态加权平均+ Multi-Sample Dropout：

Table 2: 实验结果表

	Model	Macro Avg F1
method 1	TextCNN (Chen, 2015)	0.8464
method 2	Bert-base	0.8503
method 3	Ours	0.8537
method 4	method 3 + FGM	0.8589
method 5	method 4 + 回译	0.8602

4 总结

在本次电信网络诈骗案件分类任务中，本队伍使用了基于Chinese-Bert-wwm-ext的模型，并通过CLS动态加权平均与Multi-Sample Dropout的方式进行模型改进，并通过回译、对抗训练、五折交叉验证等方式进一步提升了模型性能。评测结果表明，本队伍提出的方法均可以使模型性能得到明显的提升，最终在测试集上的宏平均F1为0.8602%，较baseline方法有一定提升。但本次方法依然存在一些不足。例如，因时间关系数据增强样本不足，未对数据进行预处理（案件描述脱敏后存在大量相同信息，对分类而言无关）。此外，我们未针对任务进行特殊的模型设计，未来可以研究如何利用任务数据特色改进模型结构提升分类性能。

[0]https://fanyi-api.baidu.com/

参考文献

Yahui Chen. 2015. Convolutional neural network for sentence classification. Master's thesis, University of Waterloo.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online, November. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hiroshi Inoue. 2019. Multi-sample dropout for accelerated training and better generalization. *arXiv preprint arXiv:1905.09788*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China, November. Association for Computational Linguistics.