计算机科学
COMPUTER SCIENCE

基于检索增强分类与解耦表示的NLP对抗鲁棒性提升方法

张鎟, 张道娟, 陈凯, 赵宇飞, 张英杰, 费克雄

引用本文

张鎟, 张道娟, 陈凯, 赵宇飞, 张英杰, 费克雄. 基于检索增强分类与解耦表示的NLP对抗鲁棒性提升方法[J]. 计算机科学, 2025, 52(12): 428-434.

ZHANG Peng, ZHANG Daojuan, CHEN Kai, ZHAO Yufei, ZHANG Yingjie, FEI Kexiong. Enhancing NLP Robustness Against Attacks with Retrieval-augmented Classification and Decoupled Representations [J]. Computer Science, 2025, 52(12): 428-434.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250500005

基于检索增强分类与解耦表示的 NLP 对抗鲁棒性提升方法

张 鹏 张道娟 陈 凯 赵宇飞 张英杰 费克雄
中国电力科学研究院有限公司电力网络安全防护与监测技术实验室 北京 102209
(zhangpeng@epri.sgcc.com.cn)

摘 要 虽然自然语言处理(Natural Language Processing,NLP)模型在各类文本分类任务中表现优异,但面对对抗性攻击时依然存在较大脆弱性。为应对这一问题,提出了一种创新性的检索增强分类方法,有效提升了模型在对抗环境下的鲁棒性。该方法引入了 k-最近邻(K-Nearest-Neighbor,KNN)检索机制,将模型自身的标签预测结果与检索到的相似样本标签分布相结合,使模型在遭受攻击时能做出更为稳健的判断。该方法的一大创新在于将分类与检索所用的表示空间分开设计,从而避免了共享表示带来的性能下降和训练不稳定。通过在多种基准数据集和多样化对抗攻击场景下的实验,证明了所提出的方法显著提升了模型的鲁棒性:在对抗攻击下,可使模型准确率下 30 个百分点到 40 个百分点,即使在强烈攻击下依然能够保持较为稳定的表现。大量实验进一步验证了该方法的有效性,表明检索增强分类和解耦表示对于构建更可靠的系统具有重要意义。

关键词:对抗性防御;检索增强分类;自然语言处理;模型鲁棒性;KNN 检索;表征学习
中图分类号 TP391

Enhancing NLP Robustness Against Attacks with Retrieval-augmented Classification and Decoupled Representations

ZHANG Peng,ZHANG Daojuan,CHEN Kai,ZHAO Yufei,ZHANG Yingjie and FEI Kexiong
State Grid Laboratory of Power Cyber-Security Protection and Monitoring Technology,China Electric Power Research Institute Co.,Ltd.,Beijing 102209,China

Abstract While NLP models have achieved state-of-the-art performance across various classification tasks,their vulnerability to adversarial attacks remains a significant challenge. This paper introduces a novel retrieval-augmented classification approach designed to enhance model robustness against such attacks. By leveraging KNN retrieval mechanism,this method interpolates the predicted label distributions with those of retrieved instances,strengthening the model's decision-making process in adversarial settings. A key innovation of this work is the decoupling of the representation spaces used for classification and retrieval,which mitigates performance degradation and training instability caused by shared representations. The proposed method is evaluated across a range of benchmark datasets under various adversarial attack scenarios,demonstrating substantial improvements in model robustness. Specifically,the accuracy drops typically observed under adversarial conditions are reduced by 30 percentage points to 40 percentage points,with the proposed approach maintaining performance stability even under intense attacks. Comprehensive experiments validate the effectiveness of the proposed method,highlighting the impact of both retrieval-augmented classification and decoupled representations in creating more resilient and reliable systems.

Keywords Adversarial defense,Retrieval-augmented classification,Natural language processing,Model robustness,KNN retrieval,Representation learning

1 引言

近年来,自然语言处理(NLP)模型的飞速发展极大地推动了问答[1-3]、语义解析[4-6]、代码生成[7-9]和文本分类[10-12]等多种应用的进步 。然而,随着模型能力的提升,其对抗攻击[13]的脆弱性也日益突出,成为自然语言处理安全研究中亟待解决的重要问题[14]。对抗攻击通过对输入做出微小但巧妙的修改,往往不易被人察觉,却能极大地干扰模型预测,严重威胁如情感分析[15]、内容审核[16]、金融欺诈检测[17]等高风险场景的安全性和可靠性。因此,提升自然语言处理模型在对抗攻击下的预测准确性和鲁棒性,成为当前机器学习安全领域的研究重点。

尽管已有研究提出了多种防御机制,如对抗训练、输入改写、梯度掩蔽与防御优化等策略[13-14,17],试图提升模型在对抗

到稿日期:2025-05-06 返修日期:2025-09-03
基金项目:国家电网有限公司科技项目:面向电力人工智能模型的攻击防御方法研究(5700-202358708A-3-3-JC)
This work was supported by the Science and Technology Project of State Grid Corporation of China:Research on Attack and Defense Methods for Electric Power Artificial Intelligence Models (5700-202358708A-3-3-JC).
通信作者:张道娟(zhangdaojuan@epri.sgcc.com.cn)

扰动下的表现，但这些方法通常存在以下局限：1）缺乏对分类决策过程本身的直接加强，导致在面对巧妙构造的对抗输入时，仍容易出现严重预测错误；2）过度依赖特定攻击模式，泛化能力有限，难以应对新型、不可预知的对抗手段；3）防御机制常常带来额外训练开销或推理延迟，影响实际部署。此外，针对分类任务的对抗防御研究相对不足，现有方法更多集中于生成式任务，导致分类场景下模型仍较为脆弱。

针对这些挑战，本文提出了一种提升语言模型对抗攻击鲁棒性的全新方法，该方法受到了最初在机器翻译领域[1,18]提出的基于检索的方法的启发并对其进行了扩展，创新性地将检索增强预测与基于 K-最近邻（KNN）的新型解耦表示技术[19]相结合，在面对对抗输入时显著提升了模型的预测准确率。所提出的技术专为提升分类模型对抗样本的鲁棒性而设计，填补了当前自然语言处理安全研究中的关键空白。本文方法的核心在于：借助与输入相似且已知安全的历史样本信息，让模型的决策过程更稳健，从而有效抵御有针对性的输入扰动。

近年来，基于 KNN 的检索增强方法在语言建模[20]、机器翻译[21]和多标签分类[22]等任务中展现出优异性能。这类方法通常利用带标签数据集构建键值对存储，在推理时检索出 k 个最相近的样本，并结合其标签信息优化预测分布。本文在此基础上进行进一步扩展，提出了一种本质上更具对抗鲁棒性的分类框架。该框架的核心机制是 KNN 解耦新方案，专门解决分类与检索共享表示在对抗攻击下易导致的性能下降和训练不稳定问题。具体做法是引入专用层和新型损失函数，将检索和分类的表示空间分开处理。这样即便某一表示空间受损，系统整体依然能保持健壮。方法流程包括：对预训练语言模型进行微调，集成解耦机制，利用检索表示构建数据存储，并在预测阶段同时用到两种不同的表示。最终输出通过融合模型自身预测与检索到的标签分布，实现了对抗攻击下的高鲁棒性。引入历史安全样本的信息，并采用双重防御机制，有效降低了对抗扰动带来的性能损失，在某些情况下显著缓解了准确率下降的问题。

为了全面验证方法的有效性，本文在 6 个中文和 6 个英文数据集上，针对多种对抗攻击场景开展了大量实验。实验结果表明，所提出的分类方法和解耦模块在面对对抗样本时依然能够保持优异表现。进一步分析还揭示了各个模块对整体鲁棒性的具体贡献。本文方法为构建安全、可靠的自然语言处理系统提供了坚实基础，尤其适用于需要抵御恶意输入的各类应用场景。本文方法将内容理解与安全防护深度融合，为打造更强健的语言模型提供了全新思路，能够有效应对当前和未来人工智能安全领域的各种挑战。本文创新性地结合了检索增强策略与前沿安全机制，为相关研究提供了重要参考价值。其意义不仅仅体现在性能提升上，更为安全 NLP 系统的设计与实际部署奠定了坚实基础，助力行业应对不断变化的安全威胁。

本文提出的 KNN 检索方法在数据存储构建与最终标签预测流程中的原理如图 1 所示。图中展示了方法的关键步骤，突出了解耦机制与鲁棒预测框架的集成。

2　本文方法

本章将详细阐述所提出的方法。首先介绍基于 k-最近邻（KNN）的分类流程（见图 1），随后介绍用于提升系统鲁棒性的解耦机制。

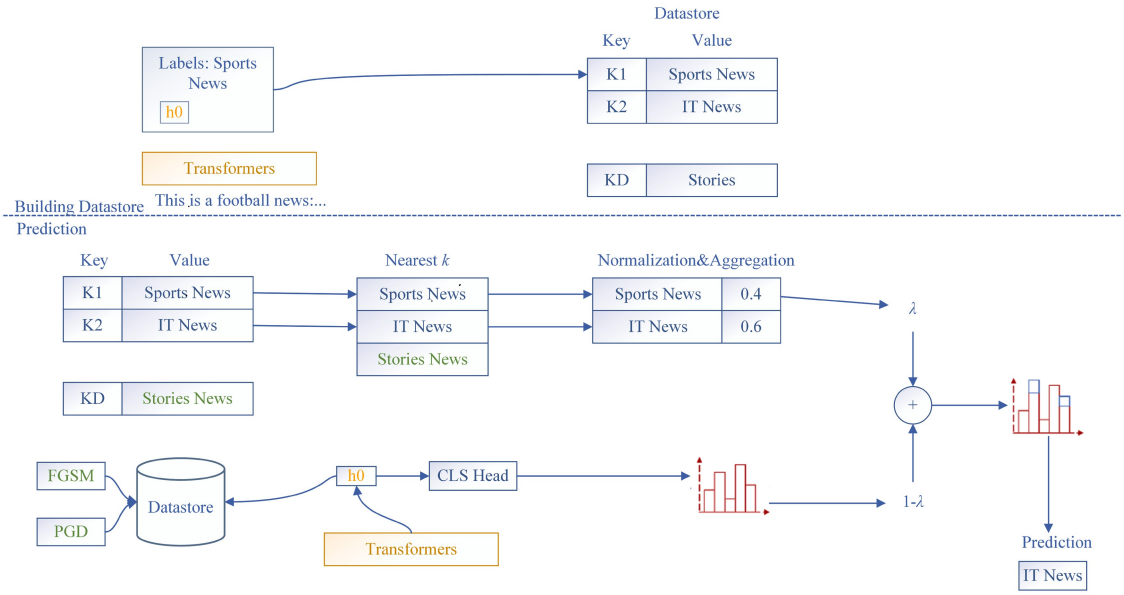

图 1　本文方法构建数据存储和最终标签预测过程的示意图

Fig. 1　Schematic diagram of the proposed method for construction of data storage and final label prediction process

2.1　基于 KNN 的分类方法

本文方案以一个分类模型为基础，该模型能够提取高质量的实例特征表示，并为每个输入给出标签分布预测。基于该模型，构建了一个数据存储库，如图 1 所示，存储库中保存着输入实例与其标签分布的键值对。在实际预测时，模型会从存储库中检索出与当前输入最相似的 k 个实例，然后将这些相似实例的标签分布与模型自身的预测结果结合，得到更加稳健和可靠的最终分类输出。

2.1.1　PLM 微调

为了获得理想的模型表现，首先在训练集 $\mathcal{D}=\{s_i,l_i\}_{i=1}^{N}$

上对一个预训练语言模型(PLM)进行微调。其中,N 表示训练样本数,s_i 是输入句子,l_i 是对应标签。具体流程是:输入句子 s_i 经过 PLM $\mathcal{M}_\theta$ 编码后,得到一组表示 $\{h_0,\cdots,h_L\}=\mathcal{M}_\theta(s_i)$。本文选用 h_0 作为分类头的输入,计算标签的概率分布。

$$\mathcal{P}_{\mathrm{CLS}}(y|s_i)=\mathrm{Softmax}(\sigma(\boldsymbol{W}^o\cdot h_0)) \tag{1}$$

其中,$\boldsymbol{W}^o$ 是权重矩阵,$\sigma(\cdot)$ 为激活函数。最终,通过对真实标签 l_i 和预测分布 $\mathcal{P}_{\mathrm{CLS}}(y|s_i)$ 计算交叉熵损失 $\mathcal{L}_{\mathrm{CE}}$,以更新模型参数 θ。

2.1.2 数据存储库构建

数据存储库的构建过程如图 1 上方所示。该存储库由一系列键值对 $(k_i,v_i)\in(\mathcal{K},\mathcal{V})$ 组成,其中键 k_i 是输入样本 s_i 通过模型 $\mathcal{M}_\theta$ 得到的表示 h_0,值 v_i 则是其真实标签 l_i。具体实现时,会为每个分类任务的数据集独立构建一个数据存储库,所有键值对均来源于其训练集。

2.1.3 预测

在推理阶段,对于测试集中的任意输入 s_i,模型 $\mathcal{M}_\theta$ 首先将其编码为 $\boldsymbol{h}_0$。随后,$\boldsymbol{h}_0$ 作为查询向量 $\boldsymbol{h}_q$,按照平方 L^2 距离 d,在数据存储库中检索出最接近的 k 个邻居 (k_j,l_j)。本文采用了高效的最近邻搜索库 FAISS[1)] 来完成这一过程。

如图 1 下方所示,检索到的 k 个邻居的标签会通过对缩放后的负距离进行 softmax 归一化,并将概率累加到同一标签上,最终得到一个标签概率分布。其计算式为:

$$\mathcal{P}_{\mathrm{KNN}}(y|s_i)\propto\sum_{(k_j,l_j)\in\mathcal{N}}\mathbb{I}_{y=l_i}\exp\left(\frac{-d(k_j,h_q)}{T}\right) \tag{2}$$

其中,$d(\cdot)$ 表示平方 L^2 距离;T 为温度系数(本实验设置为 10),用于调整距离对概率分布的影响。最后,将 KNN 检索得到的分布与模型自身预测的分布按比例进行加权融合,具体如下:

$$\mathcal{P}(y|s_i)=\lambda\,\mathcal{P}_{\mathrm{KNN}}(y|s_i)+(1-\lambda)\mathcal{P}_{\mathrm{CLS}}(y|s_i) \tag{3}$$

其中,λ 是一个超参数,用于控制 KNN 检索分布在最终预测中的权重。

2.2 解耦机制

实验表明,如果分类和检索共用同一套向量表示,模型训练会出现不稳定,且整体性能会下降。为此,本文提出了解耦机制,包含一个解耦层和专门的训练损失。具体做法是,采用独立的表示 r_i,将检索功能与 h_0 分离,确保 h_0 只用于标签预测(见式(1)),而 r_i 仅作为检索用的实例表示。r_i 通过一个简单的 MLP 层从 h_0 映射得到,即 $r_i=MLP(h_0)$。

从直观理解上来看,r_i 应该能够有效区分不同实例之间的相似性。因此,训练时为 r_i 引入三元组损失,使其更靠近正样本 r_+,远离负样本 r_-。最终的损失函数如下:

$$\begin{aligned}\mathcal{L}&=(1-\beta)\mathcal{L}_{\mathrm{CE}}+\beta\mathcal{L}_{\mathrm{DIS}}\\ \mathcal{L}_{\mathrm{DIS}}&=\max(d(r_i,r_+)-d(r_i,r_-)+\mu,0)\end{aligned} \tag{4}$$

其中,$\mathcal{L}_{\mathrm{CE}}$ 是标签预测用的交叉熵损失;$\mathcal{L}_{\mathrm{DIS}}$ 为三元组损失;$d(\cdot)$ 表示平方 L^2 距离,与 KNN 检索的距离度量方式一致。正负样本的选择基于实例标签。

2.3 对抗攻击方法

本节将介绍两种常用的对抗样本生成技术:快速梯度符号法(FGSM)[23] 和投影梯度下降法(PGD)[24]。

2.3.1 快速梯度符号法(FGSM)

快速梯度符号法(FGSM)是一种简单高效的对抗样本生成方法。其基本思想是利用损失函数为输入的梯度信息生成能够显著增加模型损失的扰动。具体来说,给定参数为 θ 的模型 $f_\theta(x)$ 及损失函数 $L(f_\theta(x),y)$(其中 x 为输入,y 为真实标签),FGSM 会根据梯度方向对输入加以扰动,从而提升损失值。生成的对抗样本 x' 的计算式如下:

$$x'=x+\epsilon\cdot\mathrm{sign}(\nabla_x L(f_\theta(x),y)) \tag{5}$$

其中,ϵ 控制扰动强度,$\nabla_x L(f_\theta(x),y)$ 是损失函数对输入的梯度,sign(·)表示取符号。该方法会在允许的扰动范围内,把输入调整到最能增加模型损失的方向上,从而生成对抗性极强,但与原始输入差异很小的样本。

2.3.2 投影梯度下降法(PGD)

投影梯度下降法(PGD)是在 FGSM 基础上的一种改进方法。与 FGSM 一步生成扰动不同,PGD 通过多次迭代梯度上升,使对抗样本更具攻击性。每次迭代中,PGD 都会沿梯度方向对输入进行微调,然后将结果投影回原输入的 ϵ-邻域内,确保扰动不会超出预设范围。其第 t 步的更新公式为:

$$x_{t+1}=\prod_{x,\epsilon}(x_t+\alpha\cdot\mathrm{sign}(\nabla_x\mathcal{L}(f_\theta(x_t),y))) \tag{6}$$

其中,x_t 是第 t 次迭代的对抗样本;α 为每步的步长;$\prod_{x,\epsilon}(\cdot)$ 为投影操作,保证扰动始终处于 ϵ-球内,即 $\|x_{t+1}-x\|_\infty\leqslant\epsilon$。

PGD 由于能够通过多次迭代对扰动进行逐步优化,因此相比 FGSM,通常能生成更具破坏力的对抗样本。

3 实验与讨论

3.1 训练设置

表 1 中列出了所有数据集的参数设置。其中默认参数值为:β 为 0.5,最近邻数量 k 为 64,温度系数 T 为 10。

表 1 实验参数设置

Table 1 Experimental parameter settings

config	ZH						EN					
	OCNLI	TNEWS	AFQMC	IFLYTEK	WSC	CSL	CoLA	SST-2	MRPC	QQP	MNLI-m/mm	QNLI
Best Lambda	0.2	0.8	0.0	0.1	0.1	0.2	0.3	0.0	0.0	0.5	0.0	0.2
Batch Size	64	64	64	64	64	64	32	32	32	32	32	32
Learning Rate	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}
Init Model	MacBERT	MacBERT	MacBERT	MacBERT	MacBERT	MacBERT	RoBERTa	RoBERTa	RoBERTa	RoBERTa	RoBERTa	RoBERTa
Warmup Steps	100	300	100	100	00	300	100	100	100	100	100	100
Training Epochs	3	5	3	3	3	3	3	3	3	3	3	3
Max Len	150	200	128	512	200	512	128	128	128	128	128	128

1) https://github.com/facebookresearch/faiss

3.2　评估设置

本文在 6 个中文和 6 个英文分类数据集上评估了所提出的方法,这些数据集包括 OCNLI[25],TNEWS,AFQMC,IFLYTEK,WSC,CSL,CoLA,SST-2,MRPC,QQP,MNLI-m/mm[26]和 QNLI。对于中文任务,选择 MacBERT-large[27-28]作为基础预训练语言模型;对于英文任务,则使用 RoBERTa-large[29]。在测试阶段,为每个输入实例检索 64 个最相似的邻居样本来调整预测分布。实验中,超参数 β 设定为 0.5,温度系数 T 设定为 10。

3.3　准确率的定义与说明

在评估不同数据集上的分类模型性能时,准确率是一个核心指标。作为评估分类模型性能的基础度量,准确率的计算式如下:

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

其中:TP(真正例)指实际为正类且被模型正确预测为正类的样本数;TN(真负例)指实际为负类且被模型正确预测为负类的样本数;FP(假正例)指实际为负类但被模型错误地预测为正类的样本数;FN(假负例)指实际为正类但被模型错误地预测为负类的样本数。该指标直观地反映了模型区分不同类别的能力。在本文的实验中,较高的准确率表明模型能够更有效地做出正确预测,这一点在各种对抗攻击场景下尤为重要。各模型在 12 个分类数据集上的准确率结果如表 2 所列。

表 2　在 12 个分类数据集上的准确率结果

Table 2　Accuracy results on 12 classification datasets

(%)

#id	Methods	中文					
		OCNLI	TNEWS	AFQMC	IFLYTEK	WSC	CSL
1	MacBERT	78.71	58.76	75.49	61.45	87.17	83.97
2	MacBERT+FGSM	47.38	27.83	44.16	32.02	56.02	52.52
3	MacBERT+PGD	42.47	25.23	41.27	28.06	53.07	49.39
4	MacBERT+FGSM+Ours	74.66	55.50	71.85	58.58	83.68	79.87
5	MacBERT+PGD+Ours	74.04	54.77	71.27	57.47	83.20	79.29
#id	Methods	英文					
		CoLA	SST-2	MRPC	QQP	MNLI-m/mm	QNLI
6	RoBERTa	67.49	96.44	90.69	92.18	90.15	94.76
7	RoBERTa+FGSM	36.31	64.67	58.94	61.55	58.94	62.91
8	RoBERTa+PGD	33.70	61.87	55.88	58.19	55.93	60.12
9	RoBERTa+FGSM+Ours	63.92	93.14	86.24	88.13	86.30	90.50
10	RoBERTa+PGD+Ours	63.20	92.63	86.73	88.68	86.92	90.92

3.4　主要结果

根据表 3 中的实验结果,可以得出以下结论。

1)以下 id 组合证明了本文提出的防御方法是有效的:(1,2),(1,3),(6,7),(6,8)。

2)通过对比(2,4),(3,5),(7,9),(8,10)id 组合的效果可以看出,本文的防御方法能够在一定程度上减轻对抗攻击的影响。其中 id5 和 id10 代表了本文方法的最佳表现。

在表 3 所列出的消融实验结果中,以 FGSM 攻击下 OCNLI 数据集的结果为例,可以观察到一个明显的性能递增关系:Ours>KNN_TL(三元组损失,Triplet Loss)>KNN_TL_D(解耦机制,Decoupling)>KNN。这一结果表明,本文方法中的各个组件对于有效抵抗对抗攻击都是必不可少的。

表 3　算法的消融分析

Table 3　Ablation analysis of algorithms

(%)

Methods	OCNLI	TNEWS	AFQMC	SST-2	MRPC	QNLI
Base Model	78.71	58.76	75.49	96.44	90.69	94.76
Base+FGSM+Ours	**74.66**	**55.50**	**71.85**	**93.12**	**85.69**	**85.51**
Base+PGD+Ours	**74.04**	**54.77**	**71.27**	**91.02**	**85.98**	**86.37**
Base+FGSM+KNN	56.99	37.71	53.95	75.65	68.78	73.65
Base+PGD+KNN	52.25	35.25	51.39	72.85	66.66	70.72
Base+FGSM+KNN_TL	62.01	42.70	59.30	80.49	74.51	78.53
Base+PGD+KNN_TL	57.64	40.33	56.63	77.88	71.79	76.06
Base+FGSM+KNN_TL_D	72.00	52.76	69.32	90.78	85.06	84.77
Base+PGD+KNN_TL_D	71.80	52.69	69.00	90.24	84.21	84.04

3.5　讨论

3.5.1　检索表示方式的影响

本文研究了不同检索向量对模型性能的影响,比较了 3 种表示策略:CLS 标记向量($\boldsymbol{h}_0$)、所有标记向量的平均值(MEAN)以及所有标记向量的最大池化(MAX)。分析范围涵盖了标准分类性能和在对抗攻击(FGSM 和 PGD)下的鲁棒性。如表 4 所列,CLS 向量在大多数数据集上表现最佳,尤其是在对抗环境中。在正常条件下,虽然 MEAN 方法在某些数据集(如 SST-2 和 MNLI-m/mm)上偶尔能获得稍好的结果,但 CLS 向量在所有任务上表现得更为稳定。更重要的是,在面对对抗攻击时,基于 CLS 的方法展现出更强的鲁棒性,与其他表示方法相比,在 FGSM 和 PGD 攻击下能够保持更高的准确率。以 OCNLI 数据集为例,在 FGSM 攻击下,CLS 方法能够保持 67.01%的准确率,而 MEAN 和 MAX 方

法的准确率则分别下降至61.39%和56.05%。MAX池化策略在所有测试场景中表现最差，对对抗扰动尤为敏感，在各数据集上均出现了最严重的性能下降。这些结果表明，对于本文提出的基于KNN的分类方法而言，CLS向量提供了最可靠且最鲁棒的检索表示，这一点在需要抵抗对抗攻击的场景中尤为重要。

表4 在正常条件和对抗性攻击(FGSM和PGD)下使用不同检索向量的准确度结果

Table 4 Accuracy results under normal conditions and adversarial attacks (FGSM and PGD) with different retrieval vectors

(%)

Datasets	中文								
	CLS			MEAN			MAX		
	Base	FGSM	PGD	Base	FGSM	PGD	Base	FGSM	PGD
OCNLI	**78.98**	**67.01**	**64.53**	77.96	61.39	58.52	77.01	56.05	53.45
TNEWS	**59.37**	**48.25**	**45.44**	59.17	43.16	40.50	57.89	36.72	34.02
AFQMC	76.18	**64.93**	**62.01**	**76.34**	59.92	57.24	76.04	55.04	52.14
IFLYTEK	**62.25**	**51.30**	**48.68**	61.87	45.61	42.91	61.31	40.41	37.80
WSC	**91.12**	**79.53**	**76.76**	**91.12**	74.43	71.89	90.28	69.01	66.35
CSL	84.30	**72.87**	**70.41**	**84.57**	67.80	65.17	83.96	62.30	59.60
Datasets	英文								
	CLS			MEAN			MAX		
	Base	FGSM	PGD	Base	FGSM	PGD	Base	FGSM	PGD
CoLA	**70.49**	**58.85**	**56.26**	69.24	52.98	50.32	68.79	47.65	45.05
SST-2	95.56	70.51	67.83	**95.99**	**78.88**	**76.13**	94.98	73.11	70.30
MRPC	**90.93**	**78.89**	**76.27**	90.69	74.11	71.38	89.84	68.41	65.71
QQP	92.08	**78.78**	**76.22**	**92.15**	75.40	72.72	91.39	69.92	67.27
MNLI-m/mm	90.50	**78.47**	**75.92**	**90.64**	73.92	71.25	89.74	68.64	65.93
QNLI	**95.15**	**80.62**	**78.07**	94.85	77.87	75.02	94.17	72.64	69.99

3.5.2 正负样本选择策略(超参数影响分析)

本小节详细介绍了如何选择正负样本来学习实例表示。具体来说，将同一批次中具有相同标签的实例视为正样本，将标签不同的实例视为负样本。当无法找到正样本时，直接使用实例本身作为其正样本。当无法找到负样本时，则随机从数据集中采样一个其他实例作为负样本。

本文对检索邻居数量(k)如何影响模型在FGSM和PGD攻击下的鲁棒性进行了系统分析。如图2所示，本文方法在所有k值下均优于基线模型，且性能随k值的增加而提升。在FGSM攻击环境下，当k从2增加到64时，本文模型的准确率从49.0%提升至峰值63.9%，明显高于基线模型的45.0%。同样，在PGD攻击下，本文模型的准确率从44.2%提升至59.4%，而基线模型仅维持在40.0%。这一趋势表明，增加参考邻居数量通常能增强模型抵御对抗攻击的能力。

然而，当k值超过64后，性能提升出现边际递减，甚至略有下降，这表明过多的检索邻居可能会在决策过程中引入噪声。这种现象在两种攻击类型中表现一致，说明$k=64$是鲁棒性和计算效率间的最佳平衡点。FGSM和PGD攻击下观察到的相似趋势进一步证明，k值的影响在不同对抗场景中相对稳定，因此是模型调优的可靠参数。

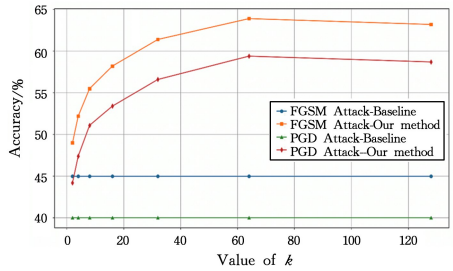

图2 FGSM和PGD攻击下不同k值的性能比较

Fig. 2 Performance comparison with different k values under FGSM and PGD attacks

本文对不同场景下最优插值系数λ进行了深入分析。如表5所列，所有数据集呈现出一致的规律：当模型面临对抗攻击时，最优λ值会上升，且PGD攻击所需的λ值通常大于FGSM攻击。以OCNLI数据集为例，最优λ值从正常条件下的0.3，上升至FGSM攻击下的0.4，在PGD攻击下进一步增至0.5。这一现象表明，面对越强的对抗攻击，模型越需要依赖检索到的外部信息来维持性能，这凸显了检索增强机制在提升模型对抗鲁棒性方面的关键作用。

表5 不同情况下不同数据集的最优λ值

Table 5 Optimal λ values for different datasets under different conditions

Situation	OCNLI	TNEWS	AFQMC	IFLYTEK	WSC	CSL	CoLA	SST-2	MRPC	QQP	MNLI-m/mm	QNLI
Opt. (Normal)	0.3	0.7	0.1	0.2	0.2	0.3	0.4	0.1	0.1	0.6	0.1	0.3
Opt. (FGSM)	0.4	0.8	0.2	0.3	0.3	0.4	0.5	0.2	0.2	0.7	0.2	0.4
Opt. (PGD)	0.5	0.9	0.3	0.4	0.4	0.5	0.6	0.3	0.3	0.8	0.3	0.5

4 相关工作

4.1 NLP领域的对抗攻击

近年来，神经网络在NLP任务中面对对抗攻击的脆弱性引起了广泛关注[13]。研究者提出了多种攻击方法，包括基于梯度的FGSM[23,30]和PGD[24]，以及基于词汇替换的攻击策略[31]。这些攻击手段能在保持语义相似性的同时，显著降低模型性能。因此，本文提出了一个基于检索增强的鲁棒分类

框架，旨在提升模型抵御此类攻击的能力。

4.2　检索增强方法

检索增强技术在多种 NLP 任务中展现出显著成效[32]。在语言建模领域，KNN-LM[20]证明了结合神经预测与最近邻检索的有效性。类似的成功也出现在机器翻译[21]、问答系统[32]、文本摘要[33]、问答任务[34]及其他应用[35-36]中。与这些主要关注增强输入表示的方法不同，本文方法创新性地将检索技术应用于提升分类模型对抗攻击的鲁棒性，为检索增强方法开辟了新的研究视角。

4.3　预训练模型与鲁棒性增强

预训练语言模型（如 BERT[37]，GPT[38]）的兴起为 NLP 任务带来了性能突破，但其在对抗攻击下的鲁棒性仍面临挑战。近期研究表明，预训练模型的深度语义表征虽能捕捉复杂语言模式，但对细微对抗扰动的鲁棒性仍弱于人类认知[39]。

部分研究通过设计抗干扰模块改进预训练模型结构。例如，Yoo 等[40]聚焦于对抗样本检测，提出基于鲁棒密度估计的检测框架，通过联合学习分类任务与对抗样本检测任务，实现模型对对抗输入的早期识别与防御。而本文方法侧重于通过解耦分类表示与检索表示提升鲁棒性。

结束语　本文提出了一种创新的基于 KNN 的检索增强分类方法，该方法利用 KNN 模型从训练数据中检索相关信息。通过将检索到的 k 个实例用于插值预测标签分布，提升了模型的决策性能。为进一步提高基于 KNN 方法的鲁棒性和稳定性，本文引入了一种简单而有效的解耦机制，确保模型在对抗环境下仍能保持良好表现。实验结果表明，本文方法在多种分类任务中显著提升了性能。未来，计划将这一方法扩展到更复杂的 NLP 应用，如问答系统和命名实体识别，以评估其在更广泛领域中的适用性和有效性。

参 考 文 献

[1] LU S Y，LIU M Z，YIN L R，et al. The multi-modal fusion in visual question answering：a review of attention mechanisms[J]. PeerJ Comuter Science，2023，9：e1400.

[2] OMAR R，MANGUKIYA O，KALNIS P，et al. Chatgpt versus traditional question answering for knowledge graphs：Current status and future directions towards knowledge graph chatbots [J]. arXiv：2302.06466，2023.

[3] ZHUANG Y C，YU Y，WANG K，et al. Toolqa：A dataset for llm question answering with external tools[J]. Advances in Neural Information Processing Systems，2023，36：50117-50143.

[4] LI B Z，DONATELLI L，KOLLER A，et al. Slog：A structural generalization benchmark for semantic parsing[J]. arXiv：2310.15040，2023.

[5] ZHUO T Y，LI Z，HUANG Y J，et al. On robustness of prompt-based semantic parsing with large pre-trained language model：An empirical study on codex[J]. arXiv：2301.12868，2023.

[6] CHEN Y R，ZHANG S Y，QI G L，et al. Parameterizing context：Unleashing the power of parameter-efficient fine-tuning and in-context tuning for continual table semantic parsing[C]// Advances in Neural Information Processing Systems. 2024.

[7] HUI B Y，YANG J，CUI Z Y，et al. Qwen2.5-coder technical report[J]. arXiv：2409.12186，2024.

[8] LIU S K，CHAI L Z，YANG J，et al. Mdeval：Massively multilingual code debugging[J]. arXiv：2411.02310，2024.

[9] CHAI L Z，LIU S K，YANG J，et al. Mceval：Massively multilingual code evaluation[J]. arXiv：2406.07436，2024.

[10] GARRIDO-MERCHAN E C，GOZALO-BRIZUELA R，GONZALEZ-CARVAJAL S. Comparing bert against traditional machine learning models in text classification[J]. Journal of Computational and Cognitive Engineering，2023，2(4)：352-356.

[11] BEKAMIRI H，HAIN D S，JUROWETZKI R. Patentsberta：A deep nlp based hybrid model for patent distance and classification using augmented sbert[J]. Technological Forecasting and Social Change，2024，206：123536.

[12] OLUSEGUN R，OLADUNNI T，AUDU H，et al. Text mining and emotion classification on monkeypox twitter dataset：A deep learning-natural language processing（nlp） approach[J]. IEEE Access，2023，11：49882-49894.

[13] SHAYEGANI E，AL MAMUN M A，FU Y，et al. Survey of vulnerabilities in large language models revealed by adversarial attacks[J]. arXiv：2310.10844，2023.

[14] LIU S B，LIU G R，ZHU B R，et al. Balancing innovation and privacy：Data security strategies in natural language processing applications[C]// 2024 5th International Conference on Machine Learning and Computer Application（ICMLCA）. IEEE，2024：609-613.

[15] TAN K L，LEE C P，LIM K M. A survey of sentiment analysis：Approaches，datasets，and future research[J]. Applied Sciences，2023，13(7)：4550.

[16] KOZYREVA A，HERZOG S M，LEWANDOWSKY S，et al. Resolving content moderation dilemmas between free speech and harmful misinformation[C]// Proceedings of the National Academy of Sciences. 2023.

[17] MOTIE S，RAAHEMI B. Financial fraud detection using graph neural networks：A systematic review[J]. Expert Systems with Applications，2024，240：122156.

[18] GAO Y，CAO Z W，MIAO Z J，et al. Efficient k-nearest-neighbor machine translation with dynamic retrieval[J]. arXiv：2406.06073，2024.

[19] GUO G D，WANG H，BELL D，et al. Knn model-based approach in classification[C]// On The Move to Meaningful Internet Systems 2003：CoopIS，DOA，and ODBAS. Berlin：Springer，2003：986-996.

[20] KHANDELWAL U，LEVY O，JURAFSKY D，et al. Generalization through Memorization：Nearest Neighbor Language Models [C]// International Conference on Learning Representations（ICLR）. 2020.

[21] KHANDELWAL U，FAN A，JURAFSKY D，et al. Nearest neighbor machine translation[C]// International Conference on Learning Representations（ICLR）. 2021.

[22] SU X A，WANG R，DAI X Y. Contrastive learning-enhanced nearest neighbor mechanism for multi-label text classification [C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. ACL，2022：672-679.

[23] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]// Proceedings of the International Conference on Learning Representations(ICLR). 2015.
[24] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]// Proceedings of the International Conference on Learning Representations(ICLR). 2018.
[25] HU H, RICHARDSON K, XU L, et al. OCNLI: Original Chinese Natural Language Inference[C]// Findings of the Association for Computational Linguistics: EMNLP 2020. ACL, 2020: 3512-3526.
[26] WILLIAMS A, NANGIA N, BOWMAN S. A broad-coverage challenge corpus for sentence understanding through inference [C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. ACL, 2018: 1112-1122.
[27] CUI Y M, CHE W X, LIU T, et al. Revisiting pre-trained models for Chinese natural language processing[C]// Findings of the Association for Computational Linguistics: EMNLP 2020. ACL, 2020: 657-668.
[28] CUI Y M, CHE W X, LIU T, et al. Pre-training with whole word masking for chinese bert[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.
[29] LIU Y H, OTT M, GOYAL N, et al. Roberta: A robustly optimized BERT pretraining approach[J]. arXiv: 1907. 11692, 2019.
[30] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 9185-9193.
[31] YE M C, CHEN J, MIAO C L, et al. Leapattack: Hard-label adversarial attack on text via gradient-based optimization[C]// Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022: 2307-2315.
[32] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks[J]. Advances in Neural Information Processing Systems, 2020, 33: 9459-9474.
[33] LIU S J, WU J, BAO J Y, et al. Towards a robust retrieval-based summarization system[J]. arXiv: 2403. 19889, 2024.
[34] SIRIWARDHANA S, WEERASEKERA R, WEN E T, et al. Improving the domain adaptation of retrieval augmented generation models for open domain question answering[J]. Transactions of the Association for Computational Linguistics, 2023, 11: 1-17.
[35] ZHU Y H, REN C Y, XIE S Y, et al. Realm: Rag-driven enhancement of multimodal electronic health records analysis via large language models[J]. arXiv: 2402. 07016, 2024.
[36] WU S Y, XIONG Y, CUI Y F, et al. Retrieval-augmented generation for natural language processing: A survey[J]. arXiv: 2407. 13193, 2024.
[37] 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. 2019: 4171-4186.
[38] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information ProcessingSystems, 2020, 33: 1877-1901.
[39] RIBEIRO M T, WU T, GUESTRIN C, et al. Beyond accuracy: Behavioral testing of NLP models withCheckList[J]. arXiv: 2005. 04118, 2020.
[40] YOO K Y, KIM J, JANG J, et al. Detection of word adversarial examples in text classification: Benchmark and baseline via robust density estimation[J]. arXiv: 2203. 01677, 2022.

ZHANG Peng, born in 1981, master, senior engineer. His main research interests include AI security, intelligent attack and defense, and threat detection.

ZHANG Daojuan, born in 1989, Ph. D, senior engineer. Her main research interests include AI security, intelligent attack and defense, and threat detection.

（责任编辑：柯颖）