基于大型语言模型微调技术的反诈骗信息识别方法研究

Study on Anti-fraud Information Identification Method Based on Fine-tuning Techniques of Large Language Models

彭成智，谢园园，吕光旭（中讯邮电咨询设计院有限公司，北京 100048）
Peng Chengzhi，Xie Yuanyuan，Lü Guangxu（China Information Technology Designing & Consulting Institute Co.，Ltd.，Beijing 100048，China）

摘　要：
针对反诈骗信息识别，对大型语言模型（LLMs）的微调技术进行了深入的实验研究。选取了3种不同规模的LLMs基础模型，并采用了LoRA和p-tuning v2 2种先进的微调技术，以适应特定的反诈骗信息识别任务。通过多个维度的实验评估，微调策略不仅能够显著提升模型在反诈骗信息识别上的性能，还能够在一定程度上保持模型的通用性。此外，探讨了LLMs在少样本情况下的学习能力，并分析了不同微调策略下的资源消耗情况。

关键词：
大型语言模型；微调技术；反诈骗信息识别；LoRA；p-tuning v2；少样本学习
doi：10.12045/j.issn.1007-3043.2024.08.011
文章编号：1007-3043（2024）08-0053-05
中图分类号：TP391
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
Aiming at the anti-fraud information identification，it conductes in-depth experimental research on fine-tuning techniques of large language models（LLMs）. It selectes three LLMs base models of different scales and employes two advanced fine-tuning technologies，LoRA and p-tuning v2，to adapt to specific anti-fraud information identification tasks. Through experimental evaluations across multiple dimensions，fine-tuning strategies not only significantly enhances the models' performance in anti-fraud information identification，but also maintains the universality of the model to a certain extent. Additionally，it explores the learning capabilities of LLMs under low-sample conditions and analyzes the resource consumption under different fine-tuning strategies.

Keywords：
LLMs；Fine-tuning techniques；Anti-fraud information identification；LoRA；p-tuning v2；Few-shot learning

引用格式：彭成智，谢园园，吕光旭. 基于大型语言模型微调技术的反诈骗信息识别方法研究[J]. 邮电设计技术，2024（8）：53-57.

0 前言

在数字化时代，网络诈骗给信息安全带来了挑战。大型语言模型（Large Language Models，LLMs）在自然语言处理（Natural Language Processing，NLP）领域，特别是基于Transformer的模型，对文本分析和模式识别显示出潜力。然而，直接训练LLMs成本高，微调成为适应特定任务的有效策略。本研究评估了LoRA（Low-Rank Adaptation）和p-tuning v2微调技术，为LLMs在反诈骗任务中的应用提供了实证基础和研究方向。

收稿日期：2024-07-02

1 LLMs背景

在数字化时代，网络诈骗行为的多样性和隐蔽性给个人和社会组织带来了前所未有的挑战。为了有效识别和防范网络诈骗，人工智能技术，尤其是LLMs，

在NLP领域的应用尤为重要。LLMs基于Transformer架构，通过深度学习算法，不仅能够处理和生成自然语言文本，而且在理解语言的复杂模式方面展现出了卓越的能力。这使得LLMs在机器翻译、文本摘要、问答系统等多个领域得到了广泛应用，尤其在反诈骗信息识别领域，LLMs的潜力正逐渐被挖掘。

1.1 基于Transformer架构的大型语言模型

Transformer[1]模型自2017年提出以来，已成为NLP领域的基础架构之一。而基于Transformer，当前涌现了海量的优秀LLMs，包括ChatGPT[2]在内的闭源大模型，以及ChatGLM[3]、Llama[4-5]、Baichuan[6]、Qwen[7]、Mistral[8]在内的开源大模型。LLMs通过堆叠多个Transformer单元，显著增强了模型的处理能力，并扩大了模型规模。LLMs是基于编码器（Encoder）或编码器-解码器（Encoder-Decoder）架构的，这些架构在处理特定类型的反诈骗任务时可能更为有效，但不同架构对LLMs性能的影响尚需进一步研究。

1.2 微调方法

由于LLMs的参数数量庞大，直接训练整个模型需要巨大的计算资源。为了解决这一问题，研究者们提出了多种微调（Fine-tuning）方法，旨在使用少量数据对预训练模型进行调整，以适应特定的反诈骗信息识别任务。微调方法主要分为3类。

a）附加（Addictive）方法。该方法通过添加层或参数进行训练，如适配器（Adapter[9]）、提示调整（Prompt-Tuning[10]）等。这些方法通过在模型中引入额外的组件，增强模型对特定任务的适应性。

b）重参数化（Reparametrization-Based）方法。典型的例子包括LoRA[11]。通过将权重矩阵分解为2个低秩矩阵并执行张量乘法，有效微调预训练模型，对优化反诈骗信息识别任务中的模型尤为关键。而基于LoRA也有许多进一步的优化方案[12-13]。

c）选择性（Selective）方法。选择性方法通过选择模型的特定层进行训练，以减少训练成本。典型方法包括P-Tuning[14]。这种方法适用于资源受限的研究和应用场景，尤其是需要快速部署反诈骗信息识别模型的情况。

本研究重点关注2种主流的微调方法：LoRA和p-tuning v2。LoRA通过冻结预训练模型中的所有参数，并设置2个较小的低秩矩阵为可训练参数，显著减少了可训练参数的数量，提升了模型在反诈骗信息识别任务中的性能。p-tuning v2作为提示调整的一种形式，通过在模型输入中插入可训练的层（如长短期记忆网络LSTM），将伪标记（pseudo tokens）映射为可训练参数，提供了更多的可训练参数选项，为模型在处理复杂的反诈骗信息识别任务时提供了更大的灵活性。

1.2.1 LoRA

LoRA是一种高效的微调技术，它通过在预训练模型的权重矩阵中引入低秩矩阵来实现参数的重参数化。其优势在于能够在保持模型规模不变的同时，为模型提供适应新任务的能力。

1.2.2 p-tuning

p-tuning是一种提示（prompt）基础的微调方法，它通过在模型输入中添加可训练的提示来调整模型的行为。p-tuning的关键优势是其简洁性和灵活性，它允许研究者通过改变提示来快速适应不同的任务，无需对模型的主体结构进行大规模的修改。

2 方法论

本研究采取了一种系统化的方法来探究和评估LLMs在反诈骗信息识别任务中的微调策略。鉴于实验资源和技术条件的限制，本文选择了2种具有代表性的LLMs[15]——ChatGLM和Baichuan，并针对这2种模型应用了p-tuning和LoRA这2种微调策略。下面将详细描述实验目标、模型选择、微调参数配置、实验设计和评估方案。

2.1 任务定义

基于文本的数据集设定反诈骗信息识别任务，这些数据集来源于实际的网络诈骗案例，包括电话通话记录和文本消息。为使数据集适合模型输入，需进行必要的预处理，包括文本清洗、分词和格式化等步骤，整理后数据的格式如图1所示，而具体至下游任务的数据集大小如表1所示。任务的目标是判断给定文本是否包含诈骗信息，并识别出诈骗信息的具体类型。该任务被构建为一个2个阶段的文本分类问题（见图2），其中输入文本T和提示P共同输入至LLMs Φ，以预测文本的异常性N和异常内容类型L。诈骗识别任务可以被表述为式（1）。

$$\langle N, L \rangle = \Phi(P,T) \tag{1}$$

2.2 模型选择

在模型选择方面，基于模型的规模、性能和可获得性，本文选择了ChatGLM和Baichuan 2种模型。ChatGLM模型以其较小的规模和高效的性能著称，而

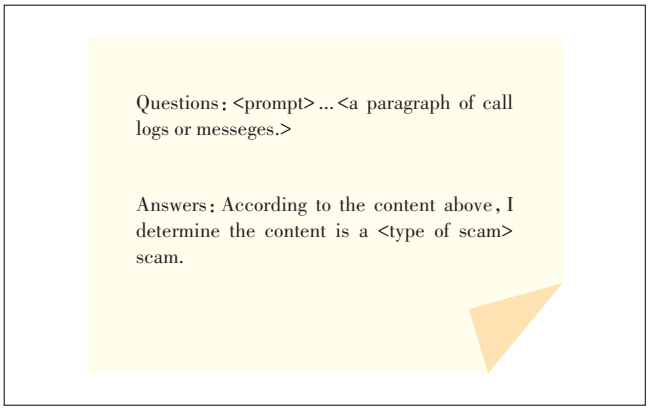

图1　数据集结构

表1　数据集大小

数据集	平均长度/词	数据集大小
通话记录	704	52 434
短信	92	20 796

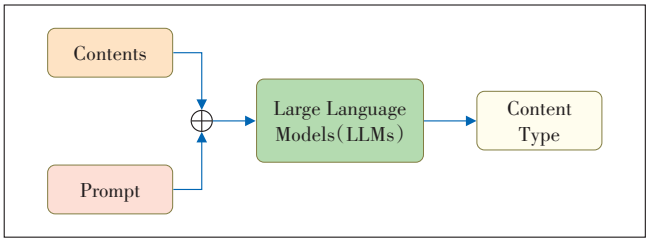

图2　任务工作流

Baichuan模型则以其强大的语言表示能力和广泛的应用场景受到关注。这2种模型在自然语言处理领域有着广泛的应用，并代表了不同的设计哲学和架构特点。

2.3 微调策略

基于当前研究领域内的主流方法进行微调策略的选择，本文采用p-tuning和LoRA 2种策略。p-tuning通过在模型输入中引入可训练的提示来适应任务需求，这种方法简便易行，能够快速适应新任务。而LoRA通过重参数化技术，在保持模型规模不变的情况下进行微调，这种方法通过引入低秩矩阵来调整模型的权重，以实现对特定任务的优化。

2.4 参数配置

对于每种微调策略，本研究设计了一系列的参数配置实验，以探索不同设置对模型性能的影响，其中共用参数如表2所示。这些参数包括但不限于p-tuning中的最大源长度和目标长度，以及LoRA中的低秩矩阵分解参数。通过调整这些参数，试图找到最优的微调配置，以提高模型在反诈骗信息识别任务上的性能。

表2　不同模型及微调策略中相同的参数

共用的参数和模型	值
Learning Rate	1×10^{-4}
Optimizer	Adam
Batch Size	1
Precision	Bfloat16

2.5 实验设计

实验设计考虑了多个因素，包括模型规模、微调策略、参数配置、输入令牌长度等，以确保能够全面评估微调策略的效果。所有实验都在单个Nvidia A100 GPU上执行，以控制实验条件和资源消耗。

2.6 评估方案

评估方案旨在全面评价微调后模型的性能。为了评估模型的通用性，计划使用C-Eval等标准化评估工具。同时，评估时也需考虑模型在不同数据集上的迁移学习能力，以及在面对新任务时的适应性。

3 实验结果

3.1 特定下游任务的微调性能

LLMs的通用结构如图3所示。尽管可用数据有限，但实验结果表明，微调策略在提高模型针对特定任务的性能方面发挥了积极作用，Baichuan以及ChatGLM 2的微调结果如表3所示。对于LoRA方法，将LoRA dropout设置为0.1，并将alpha设置为较小的值，可以获得较好的微调效果。对于p-tuning方法，将最大源长度设置为256，最大目标长度设置为128，在特定任务上的表现更佳。在有限的实验数据中，Baichuan2-13B模型在多种组合中表现最佳，平均准确率提

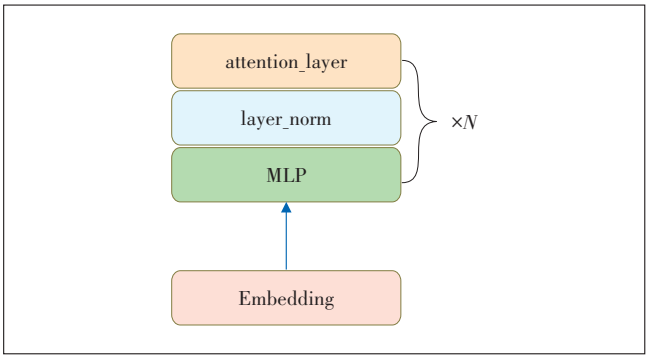

图3　大型语言模型的通用堆叠模块结构

表3 实验结果

模型	微调方法	数据集	准确度/%
ChatGLM 2	p-tuning v2	通话记录	95
ChatGLM 2	p-tuning v2	短信	92
Baichuan 13B	LoRA	通话记录	97
Baichuan 13B	LoRA	短信	93

升了约10%，这一发现为未来的研究提供了有价值的参考。

3.2 微调后的通用性损失

为了分析微调方法对模型通用性的影响，使用C-Eval基准进行评估。由于数据限制，虽然无法进行全面的通用性评估，但初步结果显示，微调后的模型在C-Eval得分上可能低于其对应的预训练模型，这表明微调过程可能会对模型的通用性造成一定影响。值得注意的是，Baichuan2系列模型在保持通用性方面的表现更好，这可能与模型的结构和微调策略的特性有关。

3.3 少样本学习能力

在少样本学习阶段，将数据集随机打乱，并仅使用5%或10%的数据对模型进行微调，以验证LLMs的少样本学习能力。基于现有数据，可观察到LLMs在少样本学习后表现出色，能够从有限的训练集中有效提取关键信息，这表明LLMs具有很好的适应性和灵活性。

3.4 资源成本

在资源成本方面，GPU内存使用情况的结果显示，Baichuan2模型在微调过程中占用的GPU内存较少，但耗时较长。例如，一个具有13亿参数的Baichuan2-13B模型在调用记录数据集进行微调时需要11 h的GPU时间和32 GB的内存。此外，不同微调策略的资源开销也存在差异，对于特定的模型，LoRA的平均开销比p-tuning高约30%，这一差异可以为研究者在选择微调策略时提供一定的指导。

4 分析与讨论

4.1 对比不同微调策略的效果和资源消耗

实验结果显示了LoRA和p-tuning 2种微调策略在提升模型特定任务性能方面的潜力。LoRA通过引入低秩矩阵减少了训练参数的数量，而p-tuning通过调整提示来更好地适应任务需求。LoRA在资源消耗方面平均比p-tuning高约30%，这可能是因为LoRA在训练过程中需要进行更复杂的矩阵运算，但Baichuan2-13B模型在LoRA微调下显示出了较好的性能提升。这一结果表明，在资源允许的情况下，LoRA微调可能为特定任务带来更高的性能收益。

4.2 微调对模型通用性的影响

微调的目的是提升模型在特定任务上的表现，但这可能会以牺牲模型的通用性为代价。C-Eval评估结果表明，微调后的模型在通用任务上的表现普遍有所下降，这可能是因为在微调过程中模型过度学习了特定任务的特征，但Baichuan2系列模型在保持通用性方面表现较好，这可能与模型架构或训练过程中的正则化策略有关。实验结果表明微调时应考虑模型的通用性，保持其对多样化任务的处理能力。

4.3 少样本学习能力的重要性

少样本学习能力是LLMs适应新任务的关键特性，尤其是在数据受限的情况下。实验结果强调了少样本学习在提高模型性能方面的重要性。在实际应用时，需仔细调整训练策略，以实现性能和泛化之间的平衡。

4.4 资源消耗与效率的平衡

在实际应用中，选择微调策略需平衡其性能与资源消耗。LoRA和p-tuning提供了不同的权衡点，研究者和工程师可以根据具体任务的需求和可用资源来选择最合适的策略。例如，对于计算资源较为紧张的应用场景，p-tuning可能是一个更经济的选择。此外，微调策略的选择还应考虑模型部署环境和预期应用频率。

4.5 未来研究方向

尽管本研究提供了微调策略的系统评估方法，但仍有许多问题有待进一步探索。例如，如何设计更高效的微调策略以减少资源消耗，如何平衡模型在特定任务和通用任务上的性能，以及如何利用少样本学习进一步提升模型的适应性等。未来，可在此基础上进行深入的研究探讨，推动LLMs在更广泛领域的应用。此外，考虑到LLMs在社会和经济中的重要作用，对其伦理和可解释性的研究也非常重要。

5 结论与展望

5.1 结论

本研究通过一系列系统化的实验，深入评估了LLMs在不同微调策略下的性能表现。研究结果表明，微调策略能够显著提升LLMs在特定下游任务，即反

诈骗信息识别上的性能。特别是LoRA和p-tuning v2这2种方法，它们通过减少可训练参数的数量，有效地降低了训练成本，同时保持了模型的性能。这一发现对于资源有限的研究和应用场景具有重要意义。然而，微调也带来了模型通用性的损失，这一现象在所有评估的模型中均有体现。这提示在微调时应权衡模型在特定任务上的性能提升与其通用性之间的关系。此外，实验还表明，LLMs展现出了良好的少样本学习能力，能够从有限的训练数据中快速学习和适应新任务。在资源消耗方面，不同模型和微调策略之间存在显著差异，Baichuan2模型在LoRA微调下虽然显示出较好的性能提升，但资源消耗也相对较高。这些研究结果为在资源有限的情况下选择最合适的微调策略提供了指导。

5.2 展望

尽管本研究在LLMs微调方法的评估上取得了进展，但未来研究仍面临诸多挑战。首先，微调策略仍有优化空间，可以探索更高效的算法来进一步减少训练成本，同时最大限度地保持或提升模型性能。其次，如何平衡模型在特定任务和通用任务上的性能，是一个值得深入研究的问题。此外，少样本学习的潜力尚待充分挖掘。研究者可以探索新的训练技术和策略，以提高模型在面对新任务时的适应性和灵活性。未来，随着计算能力的提升和算法的改进，LLMs的微调方法有望实现更广泛的应用。同时，多语言和跨文化场景下的微调策略研究，也是一个重要的研究方向，有助于提升LLMs在全球范围内的应用价值和普及度。最后，考虑到LLMs在社会和经济中的重要作用，对其伦理和可解释性的研究也将越来越重要。随着LLMs在决策支持、内容生成和个性化服务等领域的应用日益增多，确保模型的决策过程透明、公正和可解释，将是未来重要的研究课题，而伦理问题，如隐私保护、偏见和公平性，也需要得到充分考虑和解决。本研究不仅为LLMs在反诈骗信息识别领域的应用提供了实证基础，也为未来LLMs微调策略的研究和发展指明了方向。

参考文献：

[1] LIU X，Ji K X，FU Y C，et al. P-tuning v2：prompt tuning can be comparable to fine-tuning universally across scales and tasks[EB/OL]. [2024-01-24]. https://arxiv.org/abs/2110.07602.

[2] WEI C W，WANG Y C，WANG B，et al. An overview on language models：recent developments and outlook[EB/OL]. [2024-01-24]. https://arxiv.org/abs/2303.05759.

[3] LU X D，LIU Z Y，LIUSIE A，et al. Blending is all you need：cheaper，better alternative to trillion-parameters LLM[EB/OL]. [2024-01-24]. https://arxiv.org/abs/2401.02994.

[4] HOULSBY N，GIurGIU A，JASTRZEBSKI S，et al. Parameter-efficient transfer learning for NLP[C]//Proceedings of the 36th International Conference on Machine Learning. New York：PMLR，2019：2790-2799.

[5] LESTER B，AL-RFOU R，CONSTANT N. The power of scale for parameter-efficient prompt tuning[EB/OL]. [2024-01-24]. https://arxiv.org/abs/2104.08691.

[6] HU E J，SHEN Y L，WALLIS P，et al. LoRA：low-rank adaptation of large language models[EB/OL]. [2024-01-24]. https://arxiv.org/abs/2106.09685.

[7] MUELLER A，LINZEN T. How to plant trees in language models：data and architectural effects on the emergence of syntactic inductive biases[EB/OL]. [2024-01-24]. https://arxiv.org/abs/2305.19905.

[8] ARORA K，SHUSTER K，SUKHBAATAR S，et al. DIRECTOR：generator-classifiers for supervised language modeling[EB/OL]. [2024-01-24]. https://arxiv.org/abs/2206.07694.

[9] WHITE J C，COTTERELL R. Examining the inductive bias of neural language models with artificial languages[EB/OL]. [2024-01-24]. https://arxiv.org/abs/2106.01044.

[10] ELAZAR Y，KASSNER N，RAVFOGEL S，et al. Measuring causal effects of data statistics on language model's 'factual' predictions[EB/OL]. [2024-01-24]. https://arxiv.org/abs/2207.14251.

[11] WEI J，WANG X Z，SCHUURMANS D，et al. Chain-of-thought prompting elicits reasoning in large language models[EB/OL]. [2024-01-24]. https://arxiv.org/abs/2201.11903.

[12] LU K，GROVER A，ABBEEL P，et al. Pretrained transformers as universal computation engines[EB/OL]. [2024-01-24]. https://arxiv.org/abs/2103.05247.

[13] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[EB/OL]. [2024-01-24]. https://arxiv.org/abs/1706.03762.

[14] LEBRUN B，SORDONI A，O'DONNELL T J. Evaluating distributional distortion in neural language modeling[EB/OL]. [2024-01-24]. https://arxiv.org/abs/2203.12788.

[15] FORD N，DUCKWORTH D，NOROUZI M，et al. The importance of generation order in language modeling[EB/OL]. [2024-01-24]. https://arxiv.org/abs/1808.07910.

作者简介：
彭成智，毕业于北京电子科技大学，工程师，学士，主要从事AI语音业务安全、互联网反诈、数据安全、通信安全相关的研究工作；谢园园，毕业于郑州大学，学士，主要从事AI语音业务安全和呼叫中心AI类语音产品的研究工作；吕光旭，毕业于北京交通大学，高级工程师，硕士，主要从事核心网、音视频业务、消息业务的研究工作。