基于不确定性校准的云边协同推理框架

鲁飞鸿[1)2)] 罗杨一飞[1)2)] 高士淇[1)2)] 邰振赢[1)2)]

周号益[1)3)4)] 孙庆赟[1)2)] 李建欣[1)2)4)]

[1)](北京市大数据科学与脑机智能高精尖中心, 北京 100191)

[2)](北京航空航天大学计算机学院, 北京 100191)

[3)](北京航空航天大学软件学院, 北京 100191)

[4)](中关村实验室, 北京 100191)

摘 要 近年来,随着深度学习的发展,预训练模型由于其出色的泛化性和鲁棒性,广泛应用在各种分类、识别等下游任务中。但随着预训练模型性能的不断增强,其参数规模也呈指数级增长,由此给计算资源受限的边侧设备带来了巨大挑战,使得直接部署大规模预训练模型变得不切实际。为解决这一问题,本文提出了一种基于不确定性校准的云边协同推理框架。该框架在边侧设备上部署轻量化模型,在云侧部署高性能的大参数量模型,同时边侧模型和云侧模型通过证据学习方法可获得推理信心程度评估能力。当遇到低信心程度样本时,边侧模型会自动向云侧模型发起协同推理请求,以获得更准确的预测结果。这种协同机制不仅充分利用了边侧计算的实时性和云计算的高性能优势,还通过智能决策最小化了通信开销。实验结果表明,在不增加大量云侧推理开销的情况下,我们的方法在图像分类任务中的精度平均提升了 13.57%,在文本分类任务中的精度平均提升了 2.92%,这为移动设备或边缘计算等资源受限环境下的智能应用提供了一种高效且可行的解决方案。

关键词 云边协同;不确定性校准;不确定性量化;证据学习;模型轻量化;

中图法分类号 TP391

An Uncertainty-Calibrated Cloud-Edge Collaborative Inference Framework

Feihong Lu[1) 2)] Yangyifei Luo[1) 2)] Shiqi Gao[1) 2)] Zhenying Tai[1) 2)]

Haoyi Zhou[1) 3) 4)] Qingyun Sun[1) 2)] Jianxin Li[1) 2) 4)]

[1)](Beijing Advanced Innovation Center for Big Data and Brain Computing,Beijing,100191)

[2)](School of Computer Science and Engineering,Beihang University,Beijing,100191)

[3)](School of Software,Beihang University,Beijing,100191)

[4)](Zhongguancun Laboratory,Beijing,100191)

Abstract In recent years, with the rapid development of deep learning methods, pre-trained models have been widely used in multiple downstream tasks such as classification, recognition, and decision-making due to their excellent generalization capabilities and cross-domain robustness. By training on massive datasets, these models are able to capture complex patterns and representations, enabling them to perform well in tasks with limited labeled data. At the same time, pre-trained models have demonstrated excellent performance in a variety of fields, such as natural language processing and computer vision, thereby promoting the continuous advancement of

本课题得到国家自然科学基金杰出青年基金(62225202)、国家自然科学基金青年科学基金项目基于关键子图的社交网络恶意群体检测与分析方法研究(62302023)、面向工业大数据的长序列预测方法研究(62202029)资助. 鲁飞鸿，博士研究生，中国计算机学会（CCF）学生会员，主要研究领域为人工智能安全、不确定性量化、多模态认知智能. 罗杨一飞，硕士研究生，中国计算机学会（CCF）学生会员，主要研究领域为多模态认知智能.高士淇，博士研究生，中国计算机学会（CCF）学生会员，主要研究领域为人工智能安全. 邰振赢，硕士，副研究员，主要研究领域为网络大数据分析.周号益，博士，助理教授，中国计算机学会（CCF）专业会员，主要研究领域为序列大数据分析等.孙庆赟（通信作者)，博士，助理教授，中国计算机学会（CCF）专业会员，主要研究领域为网络大数据分析.李建欣，博士，教授，中国计算机学会（CCF）专业会员，主要研究领域为网络大数据分析.

artificial intelligence technology. However, as the performance of the pre-trained model improves, its parameter size also increases exponentially, resulting in a significant increase in the computing resources required for inference. While increasing model size helps improve accuracy, it poses significant challenges for edge devices, which often have limited computing power, storage space, and energy resources. Therefore, deploying large-scale pre-trained models directly on edge devices is often infeasible and may exceed hardware capabilities, resulting in excessive computational overhead. To solve this problem, this paper proposes a cloud-edge collaborative reasoning framework based on uncertainty calibration. This framework uses the uncertainty of the cloud-side and edge-side model outputs as the standard for collaborative reasoning. Specifically, we deploy lightweight models on edge devices and high-performance, large-parameter models on the cloud side. Through uncertainty calibration using evidence-based learning techniques, edge-side and cloud-side models can dynamically assess the uncertainty of their respective inferences. When the side model encounters a sample with high uncertainty, it will issue a collaborative inference request to the cloud side model to obtain more accurate and reliable predictions. In addition, based on the uncertainty of edge model output, this paper further designs a cloud-edge collaborative optimization strategy. When the prediction uncertainty of a certain sample by the side model exceeds the threshold and the output uncertainty of the cloud side model is lower than the threshold, the side model will be optimized based on the prediction results of the cloud side model. This optimization process not only enhances the adaptability of the edge model but also effectively reduces redundant cloud requests, thereby improving overall system efficiency. Experimental results verify the effectiveness of our proposed framework. Without significantly increasing cloud-side inference overhead, our method improves image classification accuracy by an average of 13.57% compared to traditional edge-side models. Similarly, in the text classification task, the recognition accuracy of the side model also increased by an average of 2.92%. In addition, this paper systematically validates the effectiveness and robustness of the proposed uncertainty calibration-based cloud-edge collaborative framework across multiple dimensions, including model calibration abilities, uncertainty quantification, generalization capability, and interpretability. This is achieved through a comprehensive set of experiments and visualization analyses, such as expected calibration error (ECE), kernel density estimation (KDE) histograms, and heatmaps. These results highlight the potential of our approach for application in resource-constrained environments such as mobile devices and edge computing platforms. By enabling efficient cloud collaboration, our framework provides a high-performance, practical solution for the deployment of complex models in real-world scenarios, ensuring efficient operation even in the most resource-constrained environments.

Key words Cloud-Edge Collaboration, Uncertainty Calibration, Uncertainty Quantification, Evidential Learning, Lightweight Models

1 引言

深度学习的快速发展推动了文本预训练模型[1]和视觉预训练模型[2][3]的应用,其在分类、识别等任务中展现了优异的泛化能力和鲁棒性。这些预训练模型可以通过简单的微调快速适应特定任务,从而显著减少从零开始训练模型所需的时间和计算资源,其在训练数据相对匮乏的情况下预训练模型能够有效利用大规模未标记数据,从而大幅提升模型的泛化能力和效率。在云侧环境中由于计算资源充足,可以部署任意参数量的大规模预训练模型。然而,在资源有限且需要高速推理的边侧环境中,难以部署高准确率、强泛化的大参数量预训练模型,只能部署一些小参数量且推理速度快的模型。为了提升边缘设备模型的性能,云边协同推理技术通过优化资源的高效利用与合理分配,正成为当前研究的热点。

云边协同推理技术通过优化云侧和边侧节点之间的协作方式,结合场景中的资源特点,在训练节点或边侧设备上部署适用于协同推理的模型。该技术能够根据不同任务的需求动态分配资源和模型,从而更高效地完成下游任务。模型轻量化[4][5][6][7]是最常用的云边协同推理方法。其通过模型切割[4]、

剪枝[5]和量化[7]等技术,根据任务需求动态激活模型的特定参数或压缩整体模型参数,减少计算资源的占用和推理时间,使得在资源受限的边侧设备上协同复杂模型进行高效推理成为可能。然而,模型轻量化在提升模型效率的同时,可能会降低模型的鲁棒性和精确度,尤其在处理复杂或变化的输入数据时,轻量化模型的性能可能不如原始模型稳定和准确。为了克服模型轻量化导致的模型退化的风险,结合云边场景自适应地选择合适的模型用于离线推理的模型选择方法开始被广泛研究和应用。

在云边协同推理场景下,模型选择方法[8][9]通常将不同参数量的模型分别部署在云侧与边侧,边侧模型用于初步的推理判别,当边侧模型输出的结果的置信程度不满足给定的阈值时,说明边侧输出的结果不可靠,此时需要利用云侧模型的结果作为推理输出。由于不同任务使用的模型架构各不相同,在推理速度、计算占用开销以及计算精度等方面有所差别,因此如何通过置信度选择适宜的模型成为了该研究的难点。以分类任务为例,传统置信度计算方法通过计算模型输出 Softmax 的熵大小得到当前模型的输出结果是否可靠,然而这种方法预测出的置信度和样本分布关联较大,对于一些分布外泛化的样本无法做到准确估计。因此,如何更准确的预测边侧轻量化模型输出结果的置信度,来及时请求云侧模型协助是当前场景下亟待解决的问题。

不确定性估计常用于评估模型输出的可靠程度,随着深度学习的发展,它逐渐应用于智能驾驶[10][11]和环境感知[12]等边侧模型中,以避免模型过度自信的预测并提供更准确的传感参数估计和路径规划,有效保障车辆乘客和其他道路使用者的安全。然而,这些应用仅考虑边侧输出结果的可靠程度,未在边侧与云侧模型异构的场景下进行模型校准对齐,因此无法建立边侧与云侧模型自适应选择的综合评价标准来判定何时使用边缘模型进行推理,何时使用云端模型进行推理。为此,本文结合基于证据学习的不确定性度量方法,利用校准蒸馏方法使边侧和云侧模型具有输出不确定性的能力,同时将边侧和云侧模型输出的不确定性大小及其差异作为指标,来判断哪些样本是边侧模型难以处理的,并及时与云侧模型进行协同。此外,本文还通过云侧模型优化边侧模型对高不确定性样本的输出,实现了边侧模型对于未见场景难样例的主动学习和优化。

在本篇工作中,我们提出了基于不确定性估计的云边协同推理框架,首次将模型输出的不确定性与云边协同推理相结合。首先,边侧模型通过证据学习方法获取输出不确定性的能力,云侧模型通过模型校准蒸馏的方法获取输出不确定性的能力。在推理过程中,边侧模型首先评估当前输入样本的不确定性,并将不确定性较高的样本传送至云侧。云侧模型接收到这些样本后,输出相应的预测结果。此外,对于边侧模型输出不确定性较高的预测类型,本文提出云边协同优化策略,利用云侧模型的预测结果来优化边侧模型。我们在图像分类、文本分类等任务中对该方法进行了验证,实验结果表明,基于不确定性估计的云边协同推理框架相较于传统方法,能够显著提高模型协同推理的速度和精度。本研究的主要贡献可以总结为以下三点：

(1) 本文提出了一种基于不确定性校准的云边协同推理框架,首次将模型输出的不确定性作为模型选择的衡量指标,实现了高效且准确的云边协同推理和优化;

(2) 本文设计了一种模型校准蒸馏方法,通过在微调后的目标模型上添加校准层,将边缘模型和云端模型的输出不确定性进行映射,从而实现二者在不确定性估计上的一致性和准确性。

(3) 本文在视觉和文本分类任务上进行了实验,结果显示,在 CIFAR-10、CIFAR-100 和 ImageNet 数据集上,模型的精度平均提升了 13.57%;在 SST-5 和 TweetEval 数据集上，模型的精度平均提升了 2.92%。实验结果证明了基于不确定性校准的云边协同推理框架不仅确保了云侧模型和边侧模型输出的不确定性的有效性,还加快了云边模型的速度并提升了模型在下游任务的表现。

本文的结构安排如下：第 2 章回顾了云边协同推理,云边协同优化以及不确定性估计等相关前沿技术;第 3 章详细介绍了本文提出的基于不确定性校准的云边协同推理框架;第 4 章给出了实验设计及评价指标的计算方法;第 5 章展示了具体的实验分析结果;第 6 章对全文进行了总结。

2 研究现状

本章首先介绍了云边协同推理方法,重点讨论了模型分割、网络减枝、模型量化及模型选择等策略,旨在解决边侧设备在执行大模型时的硬件限制。接着,本章回顾了采用迁移学习、元学习和因果推断等技术优化模型的云边协同优化方法。其次,本章探讨了不确定性估计方法,包括基于先验分布推测后

验分布的传统方法与基于证据学习的快速推理方法。最后,本章讨论了云边协同方法和不确定性学习方法在跨域泛化场景下的应用。

2.1 云边协同推理

随着预训练模型在各领域的广泛应用,其出色的兼容性和端到端处理能力使得在边侧设备上部署神经网络执行各类任务(如分类、回归和生成等)逐渐成为研究的热点。然而,受限于硬件资源的制约,在边侧设备推理大参数量的预训练模型成为一大难题。传统的解决方法是将数据全部上传至云侧,由云服务器处理后再将结果反馈至边侧设备。此方法虽能提高推理的准确性,但增加了云服务器的计算负担,在网络环境较差的情况下尤其受限。为了解决上述问题,研究者们提出了云边协同推理的概念,旨在通过边侧和云侧的计算资源的协同完成数据推理。这种协同推理模型不仅减轻了云侧的负载,还能在网络连接不稳定时保持较高的操作效率和准确率,其主要分为模型分割、网络减枝、模型量化以及模型选择等方法。

模型分割、网络减枝和模型量化都是常用的模型轻量化方法。模型切割[4][13]的云边协同推理依据模型切割技术,将模型分解为具有先后以来关系的不同切片,分别部署在边侧与云侧。边侧将部分中间结果经过处理后发送至云侧完成后续的计算,最终由云侧返回推理结果。模型减枝是一种通过删除模型中部分计算路径来减少参数空间的技术,主要包括一次性剪枝[5][14][15]和运行时剪枝[6][16]两种形式,其通过有效地识别并删除不必要的计算节点,从而优化模型的结构,并显著加快推理速度。模型量化的云边协同推理利用量化感知训练[17][18]和训练后量化[7]策略有效的压缩云侧模型的参数量,从而使其适应边侧设备的存储和计算限制,支持了模型的轻量化部署。

在云边协同场景下,模型选择通常采用多种参数量模型的协同部署策略,分别部署在云侧和边侧。为了确保最终推理结果的置信度,通常会设置一个阈值,用于评估边端模型输出结果的置信度。如果边端模型的推理结果置信度低于这个阈值,则说明该结果可能不准确或者存在较高的不确定性,此时需要云侧模型进行输出预测。模型选择分为两类,第一是通过预先训练的预测模型[8][9]对输入的数据进行判别,并选择合适的模型进行推理;第二类方法在多任务推理场景下,对不同领域的模型进行智能选择或模型融合[19][20]。Sniper[21]是一种典型的具有时间感知能力的自更新云边协同推理调度系统,该系统能够根据网络条件和设备状态动态调整深度神经网络的分区策略,有效减少推理延迟并提高资源利用率。Hao 等人[22]提出了一种动态令牌级别的边缘-云协同推理方法,通过在边缘设备上部署小型语言模型并在推理过程中与云侧的大型语言模型进行令牌级交互,以接近大型语言模型的输出质量。

上述云边协同方法虽在一定程度上缓解了云端计算负载,但仍面临精度损失与通信开销等关键挑战。现有模型选择方法尽管提升了协同系统的整体性能,却因过度依赖 softmax 概率作为可信度评估机制,导致其判别能力与样本分布高度耦合,难以对分布外数据进行准确的不确定性估计。本文提出的基于不确定性校准的云边协同方法创新性地建立了系统化的不确定性量化框架,有效度量模型输出的可信程度,准确判断边侧模型难以识别的时刻并提升了系统在资源受限场景下的性能和鲁棒性。

2.2 云边协同优化

参考 Yao 等人[23]的定义,云边协同优化方法常表述为一个两阶段优化问题,其在优化的过程中通过使用迁移学习、元学习和因果推断的方法引导云侧模型优化边侧模型。其中迁移学习的方法针对特征空间差异,使用对称变换[24][25][26]和非对称变换[27][28]的方式实现从云侧模型到边侧模型的迁移;元学习[29][30][31][32][33]主要聚焦于模型如何快速处理新任务的能力,通过云上训练的方法在异构边侧环境中快速学习并迭代新模型;因果推断方法[34][35][36][37]以因果效应为基础,由云侧模型强大的泛化能力消除了边侧模型因输入和结果之间的虚假相关性而导致的过拟合现象。

上述方法虽采用迁移学习、元学习和因果推断等技术优化边侧模型,但因其忽视标签可靠性问题而导致优化效果有限。本文提出的基于不确定性校准的云边协同方法创新性地引入自适应信心程度损失函数,通过精确量化云侧与边侧模型的不确定性差异,动态调整优化策略,有效识别边侧模型学习能力不足的样本,提升了边侧模型在资源受限环境下的性能和泛化能力。

2.3 不确定性估计

随着深度学习模型参数规模的持续增长,如何保证模型输出的内容可靠且准确逐渐成为研究热点,为此,越来越多的学者开始关注利用模型的不确定性来约束输出结果的置信度和可靠性。现有的不

确定性输出的方法大致分为两类：一类是通过模型参数的先验分布推测后验分布,并利用后验分布的高阶矩来解读不确定性[38][39][40][41][42]。这类方法包括蒙特卡洛 dropout[41]和贝叶斯网络（BNN）[42]等。然而,由于上述方法需要多次迭代和估算模型整体参数,适用于参数规模较小的模型;对于参数量较大的模型,此类方法优化速度会较慢,且输出效果可能受到影响。另一类方法则通常使用证据学习[43][44][45][46]的方法来优化模型使其输出不确定性。这类方法只需要单次推理就可以使模型输出不确定性,从而加快了不确定性推理的速度。Sensoy 等人[44]使用迪利克雷分布进行建模,将模型的输出映射到高斯分布，实现分类任务的不确定性输出。Amini 等人[46]关注于回归任务的不确定性输出,通过逆伽马分布建模将模型参数映射到高斯分布,从而使该模型具有输出不确定性的能力。

Fathullah[47]提出了自分布蒸馏,其中教师集成和学生模型同时训练,并将集成知识蒸馏到学生中。这种方法有效地提高了不确定性学习和模型可靠性,而无需对每个组件进行单独训练。VBLL[43]引入了一种确定性变分公式,用于训练贝叶斯最后一层神经网络,从而实现了无采样、单次通过的模型和损失,有效地提升了不确定性估计的准确性。Xu 等人[48]结合了扩散技术和隐式先验,通过利用隐式分布来建模贝叶斯最后一层中的权重先验并结合扩散采样器来近似真实的后验预测,提高模型准确性。

上述的方法采用贝叶斯分布,迪利克雷分布或逆伽马分布直接对各类模型的输出进行建模,在大模型场景中存在计算复杂度高和训练稳定性差的问题,不用适用于引导大参数的模型输出准去的不确定性。相比之下,本文提出的校准蒸馏方法通过将边侧模型的不确定性知识蒸馏给云侧模型,避免了对复杂分布的直接建模,从而降低了计算复杂度和资源消耗。

2.4 云边协同和不确定性学习的跨域泛化应用

随着特定领域数据的不断积累,提升模型的跨域泛化能力已成为深度学习领域的重要研究方向。在实际应用中,云边协同与不确定性学习已成为实现跨域泛化的主流技术路径。云边协同方法通过整合云端模型的强大算力与边缘设备的实时感知能力,实现了模型在多源异构环境下的知识共享与自适应,有效提升了模型在不同领域和场景下的泛化能力。Lian 等人[49]设计了一种云边协作的持续适应学习框架,通过从边缘侧收集数据辅助云端模型微调,并采用知识蒸馏机制将云端模型的知识反馈到边缘模型,使其能够适应智能交通场景的多样化和高动态性。He 等人[50]提出的联邦迭代学习算法 FedITA,结合渐进训练与迭代权重更新,增强了不同边侧客户端间的安全交互,有效降低了由极端类别不均衡引起的过拟合风险。此外,不确定性学习则是通过量化模型在未知领域的置信度,帮助模型动态调整决策策略,增强其应对分布转移和任务变化的鲁棒性。He 等人[51]提出了领域推广感知的不确定性反思学习方法,设计了潜在不确定性建模和动量反思学习模块,以应对跨域场景下 3D 分割的准确性。Yin 等人[52]分析不确定性对于实例分割任务的影响,观察到双重注意模块可以减轻语义分割任务中的不确定性,利用 KL 散度估算随机不确定性,有效增强了分割模型在跨域场景下的鲁棒性。Cai 等人[53]提出了一种基于不确定性感知和类别平衡的领域泛化方法用于目标检测中的不确定性估计,解决了智能驾驶场景下的跨域目标检测问题。

尽管上述研究分别在云边协同与不确定性量化两条技术路线取得了一定进展,但前者主要关注于异构算力与多源模型的协作,较少涉及模型输出置信度的精细建模；后者虽能有效度量模型风险,却鲜有工作将其嵌入到云边协同的框架中。为此,本文提出了一种基于不确定性校准的云边协同框架,通过引入置信度估计与模型动态调用机制,实现了对模型输出可靠性的精确量化以及云侧与边侧模型的按需调度,进一步提升了模型的跨域泛化能力。

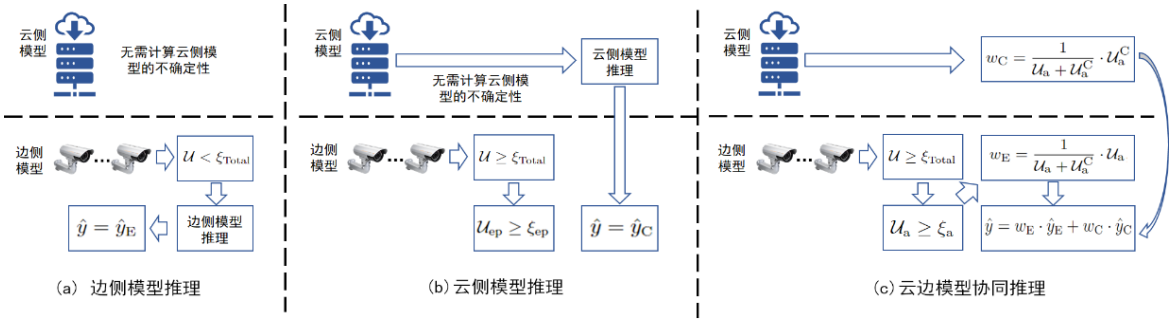


图 1 基于预测不确定性自适应的云边协同推理方法示意图

3 基于不确定性校准的云边协同推理框架

云侧模型部署在计算资源丰富的中心服务器上,而边侧模型则部署在各地区的终端设备上。云侧模型通常为参数较大的模型,而边侧模型是其轻量化版本或小参数量的模型。本文的核心问题是如何利用云侧和边侧模型输出的不确定性,基于模型选择实现自适应的云边协同推理。 基于主流的视觉和文本任务预训练模型,本文设计了基于预测不确定性自适应的云边协同推理方法（见 3.1 节）,如图 1 所示。首先,利用证据学习和模型校准蒸馏方法（见 3.2 节）,使云侧和边侧模型具备输出不确定性的能力。接着,对于边侧输出不确定性较低的样本,采用边侧模型进行处理（见图 1(a)）;对于边侧输出认知不确定性较大的样本,表明边侧模型对当前样本的预测不自信且可能存在较大偏差,采用云侧模型进行处理（见图 1(b)）;对于边侧输出偶然不确定性较大的样本,表明边侧模型的预测受数据采样的影响较大,此时采用云边加权推理方法（见图 1(c)）。随后, 基于边侧模型输出的不确定性,本文进一步设计了云边协同优化策略（见 3.3 节）。当边侧模型对某个输入样本预测的总体不确定性超过阈值而云侧模型输出的总体不确定性小于阈值时,使用云侧模型的输出来优化边侧模型。

3.1 基于预测不确定性自适应的云边协同推理

以分类任务为例,传统的基于模型选择的云边协同推理框架通常依据边侧模型 Softmax 输出层的熵值来判断当前模型的输出是否可靠。但由于 Softmax 计算的概率之和始终为 1,对于每一个输入样本来说不同类别的预测概率是互补的,对于未知领域的样本有很大概率会输出一个错误的高信心程度预测结果。为了解决上述问题,本文通过证据学习的方法引导模型输出不确定性,并以云侧和边侧模型输出的不确定性作为衡量指标来判断当前模型的输出是否可靠,以解决模型预测的置信度在未知领域不稳定的问题。

考虑一个边侧模型$\mathcal{M}_{\mathrm{E}}$,该模型接收输入x并生成预测输出$\hat{y}$及其对应的不确定性u。对于包含N个输入输出样本对的数据集$\mathcal{D}=\{(x_i,y_i)\}_{i=1}^{N}$,我们利用该数据集来训练边侧模型$\mathcal{M}_{\mathrm{E}}$。 当提供新的测试输入$x^*$时,本文的目标是预测相应的输出$\hat{y}^*$及其相关的不确定性$u^*$。

由于不确定性的高低和输入样本x^*是否来自域内而有很大关系。来自域内的x^*通常显示较低的不确定性,表明模型对当前预测较为自信。相反,如果x^*来自域外,则会显示较高的不确定性,反映出模型对当前预测的信心较低。设每类样本的证据表示为ε,模型输出的总体不确定性表示为$\mathcal{U}$。参考 Amini[46]等人对不确定性分析的思路,模型输出的不确定性$\mathcal{U}$包含由输入数据原因导致的偶然不确定性$\mathcal{U}_{\mathrm{a}}$和由于模型自身原因导致的认知不确定性$\mathcal{U}_{\mathrm{ep}}$,因此对$\mathcal{U}$进行分解,可得:

$$\mathcal{U}=\mathcal{U}_{\mathrm{a}}+\mathcal{U}_{\mathrm{ep}}\#(1)$$

其中偶然不确定性也称为统计或数据不确定性,代表每次运行相同实验时都会不同的未知数,其通常通过预测的熵(entropy)来衡量:

$$\mathcal{U}_{\mathrm{a}}=-\sum\left(p_i\cdot\log(p_i)\right),\#(2)$$

其中p_i是类别i的预测概率。

认知不确定性描述了预测中估计的不确定性,该不确定性可以通过 Dirichlet 分布的方差来计算:

$$\mathcal{U}_{\mathrm{ep}}=\frac{\sum\left(\alpha_i\cdot(\xi-\alpha_i)\right)}{\xi^2\cdot(\xi+1)},\#(3)$$

其中α_i是 Dirichlet 分布的参数, ξ是所有α_i之和。

由于$\mathcal{U}_{\mathrm{a}}$源自数据采样过程,其仅与输入数据及其分布特性相关,而不涉及模型本身的认知局限,因此不直接影响模型的内在性能表现。相比之下,$\mathcal{U}_{\mathrm{ep}}$则反映了模型对当前输入样本预测结果的质量评估,体现了模型在任务理解和推理能力方面的不足,通常与模型的知识获取和表达能力密切相关。鉴于$\mathcal{U}_{\mathrm{a}}$和$\mathcal{U}_{\mathrm{ep}}$在概念上相互独立且不相交,本文提出通过综合考量边侧模型的总体不确定性$\mathcal{U}$、偶然不确定性$\mathcal{U}_{\mathrm{a}}$和认知不确定性$\mathcal{U}_{\mathrm{ep}}$,构建一个联合判别框架,用于评估当前模型输出结果的可靠性。具体而言,考虑以下三种情况:

A. 边侧总体不确定性低

当边侧模型的总体不确定性$\mathcal{U}$较低时,可以直接使用边侧模型进行推理,无需云侧模型干预:

$$\mathcal{U}<\xi_{\mathrm{Total}}\Rightarrow\hat{y}=\hat{y}_{\mathrm{E}},\#(4)$$

其中,ξ_{Total}是总体不确定性阈值,表示当$\mathcal{U}$低于该阈值时,选择边侧模型的输出 $\hat{y}_{\mathrm{E}}$作为最终结果。

B. 边侧总体不确定性高,认知不确定性也高

当边侧模型的总体不确定性$\mathcal{U}$和认知不确定性

$\mathcal{U}_{\text{ep}}$都较高时,边侧模型输出的结果不可靠,此时将任务交由云侧模型进行推理：

$$\mathcal{U} \geq \xi_{\text{Total}} \text{ and } \mathcal{U}_{\text{ep}} \geq \xi_{\text{ep}} \Rightarrow \hat{y} = \hat{y}_{\text{C}}, \#(5)$$

其中,ξ_{ep}是认知不确定性的阈值,当$\mathcal{U}$和$\mathcal{U}_{\text{ep}}$均高于各自阈值时,选择云侧模型的输出$\hat{y}_{\text{C}}$作为最终结果。

C. 边侧总体不确定性高,偶然不确定性高

当边侧模型的总体不确定性$\mathcal{U}$较高且偶然不确定性$\mathcal{U}_{\text{a}}$也较高时,即$\mathcal{U}_{\text{a}} \geq \xi_{\text{a}}$,说明当前输入数据的噪声比较大,此时通过边侧和云侧模型的加权平均来综合两者的推理结果：

$$\hat{y} = w_{\text{E}} \cdot \hat{y}_{\text{E}} + w_{\text{C}} \cdot \hat{y}_{\text{C}}, \#(6)$$

其中,权重w_{E}和w_{C}依据边侧模型的偶然不确定性$\mathcal{U}_{\text{a}}$和云侧模型的偶然不确定性$\mathcal{U}_{\text{a}}^{\text{C}}$计算：

$$w_{\text{E}} = \frac{1}{\mathcal{U}_{\text{a}} + \mathcal{U}_{\text{a}}^{\text{C}}} \cdot \mathcal{U}_{\text{a}}^{\text{C}}{}_{\text{a}}, \; w_{\text{C}} = \frac{1}{\mathcal{U}_{\text{a}} + \mathcal{U}_{\text{a}}^{\text{C}}} \cdot \mathcal{U}_{\text{a}} \#(7)$$

通过这种方式,可以有效整合边侧和云侧模型的优势,获得更快速且准确的推理结果。

D. 阈值选择策略

上述决策机制中的关键参数是不确定性阈值ξ_{Total}, ξ_{ep}和ξ_{a}的设定。为确保阈值参数的科学合理性,本文采用了数据驱动的方法：从测试集中随机抽取 10%的样本构建校准集,并基于该校准集分别确定认知不确定性和偶然不确定性的阈值,即选取其70%分位数作为相应判别标准。对于总体不确定性阈值ξ_{Total},则通过第 5.3 节中的阈值选择实验进行确定。这种阈值设定策略在模型可靠性和系统效率之间实现了最优权衡,既能准确识别边侧模型预测中的高不确定性样本,又避免了过度依赖计算资源丰富但通信延迟较高的云侧模型,从而有效提升了边云协同推理系统的整体性能。

算法 1. 基于不确定性自适应的云边协同推理算法

输入:

数据: 图像数据X_V(图像分类)或文本数据X_T(文本分类)

输入: 云侧模型$\mathcal{M}_{\text{C}}$,边侧模型$\mathcal{M}_{\text{E}}$

输出：协同推理预测$\hat{y}$

$\mathcal{U} = \mathcal{M}_{\text{E}}(x)$ // 边侧模型总体不确定性

$\mathcal{U}_{\text{ep}} = \sum(\alpha_i \cdot (S - \alpha_i))/S^2 \cdot (S+1)$ //边侧模型认知不确定性

$\mathcal{U}_{\text{a}} = -\sum(\mathcal{M}_{\text{E}}(x_i) \cdot \log(\mathcal{M}_{\text{E}}(x_i)))$ //边侧模型偶然不确定性

SET $\xi_{\text{Total}}, \xi_{\text{ep}}, \xi_{\text{a}}$ // 设置不确定性阈值

IF $\mathcal{U} < \xi_{\text{Total}}$ THEN

RETURN $\hat{y} = \hat{y}_{\text{E}}$ // 边侧模型推理

ELSE IF $\mathcal{U}_{\text{ep}} \geq \xi_{\text{ep}}$ THEN

RETURN $\hat{y} = \hat{y}_{\text{C}}$ // 云侧模型推理

ELSE IF $\mathcal{U}_{\text{a}} \geq \xi_{\text{a}}$ THEN

$\mathcal{U}_{\text{a}}^{\text{C}} = -\sum(\mathcal{M}_{\text{C}}(x_i) \cdot \log(\mathcal{M}_{\text{C}}(x_i)))$ //云侧模型偶然不确定性

$w_{\text{C}} \leftarrow \frac{1}{\mathcal{U}_{\text{a}} + \mathcal{U}_{\text{a}}^{\text{C}}} \cdot \mathcal{U}_{\text{a}}$ // 计算云侧模型权重

$w_{\text{E}} \leftarrow \frac{1}{\mathcal{U}_{\text{a}} + \mathcal{U}_{\text{a}}^{\text{C}}} \cdot \mathcal{U}_{\text{a}}^{\text{C}}$ // 计算边侧模型权重

RETURN $\hat{y} = w_{\text{E}} \cdot \hat{y}_{\text{E}} + w_{\text{C}} \cdot \hat{y}_{\text{C}}$ // 云边协同推理

END IF

计算复杂度分析: 如算法 1 所示，基于不确定性校准的云边协同推理框架的计算复杂度可从以下三个主要环节进行分析：①边侧模型推理,②不确定性量化,以及③在特定条件下云侧模型的推理。在模型推理阶段,边侧模型推理的复杂度为$O(M_E)$,其中M_E为边侧模型的计算量。这是边端侧的主导计算开销,通常通过模型轻量化技术进行优化。不确定性计算阶段 (包括总体不确定性U、认知不确定性U_{ep}和偶然不确定性U_a) 的复杂度为$O(K)$,其中K为类别数,涉及类别级别的操作且计算量较低。云侧模型推理计算的复杂度为 $O(M_C)$。权重计算与融合(w_C 和 w_E)的复杂度为 $O(1)$,仅涉及简单的除法与乘法操作,开销可忽略。综上,总体的边侧计算复杂度为 $O(M_E + K) + O(log(M_C))$。与单纯使用云侧进行推理的计算复杂度$O(M_C)$相比,边侧计算开销显著降低,即$O(M_E + K) + O(log(M_C)) \ll O(M_C)$,从而在保证推理精度的同时提升了计算效率。

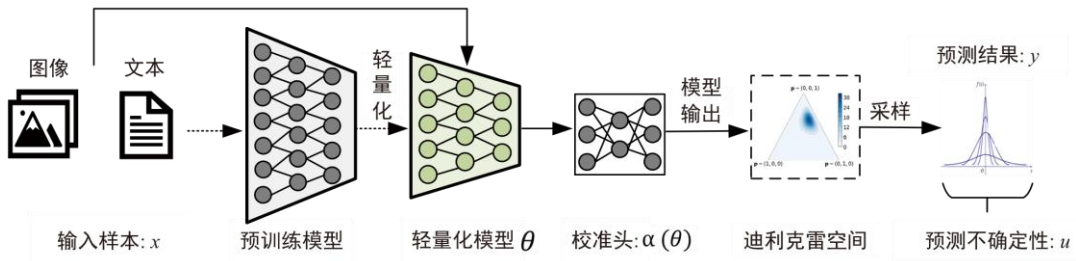


图 2 边侧模型推理流程图

3.2 云边模型校准对齐

随着模型轻量化研究[17][18]的不断深入,边侧模型可以通过轻量化或者蒸馏的方式实现与云侧模型相似的功能,因此云侧模型输出不确定性的能力理论上可以通过模型轻量化的方法迁移到边侧模型中。然而云侧模型由于参数量庞大,直接应用证据学习难以有效训练其不确定性输出能力。对此,本文设计云边模型校准蒸馏策略,利用迁移学习更有效的将输出不确定性的能力从易于优化的小参数量模型蒸馏到难以优化的大参数量预训练模型中。

本文将模型校准蒸馏定义为将模型输出不确定性的能力从一个模型应用扩展到另一个不具有不确定性输出能力模型的过程,目标是使云侧模型能够通过校准蒸馏输出准确、稳定的不确定性估计,同时保持其在原始任务上的性能,边侧模型输出不确定性的过程如图 2 所示, 其中云侧模型和边侧模型输出不确定性的方式近似,但缺少了模型轻量化这一步骤。

本文首先让校准代理模型$f^{'}$在$\mathcal{D}$上学习输出预测结果的同时估计其不确定性的能力。接着,在$\mathcal{D}$上训练云侧模型f_{C}。当云侧模型的训练达到收敛后,将其参数冻结,并添加一个校准层ϕ。校准层ϕ的学习过程利用校准代理模型$f^{'}$已学的不确定性评估能力,将其转移到云侧模型f_{C}的预测分布中,而不改变云侧模型f_{C}的原始结构。

本文将$y^{'}$和y_{C}定义为$f^{'}$和f_{C}的实现。 考虑一个校准代理模型的参数为$\theta^{'}$,其中$\theta^{'}$是通证据学习获得的,允许模型输出其当前预测的不确定性$u^{'}$,并且$u^{'}$的值是 0 到 1 之间的实数。对于尚未进行不确定性能力训练且参数为θ_{C}的云侧模型y_{C},定义y_{C}输出的不确定性为u_{C}。模型校准蒸馏学习的主要意图是使 y_{C}能够估计不确定性,从而有效地缩小 u_{C}和$u^{'}$之间的差距。将参数θ_{C}分成两部分,即$\theta_{\mathrm{C}}=[\theta_{\psi};\theta_{\phi}]$。第一部分由模型的原始预训练模型参数组成,称之为$\theta_{\psi}$;第二部分由几个校准层组成,称为$\theta_{\phi}$。在模型的初始训练阶段,使用交叉熵损失函数用真实标签ℓ训练θ_{ψ}。随后,将θ_{ϕ}连接到θ_{ψ}并冻结θ_{ψ},使用 KL 散度约束优化θ_{ϕ}:

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}(u^{'}\|u_{\mathrm{C}})&=\mathrm{D}_{\mathrm{KL}}(y^{'}(x)||y_{\mathrm{C}}(x))\\&=\mathrm{D}_{\mathrm{KL}}(P(u^{'}|x,\theta^{'})||P(u_{\mathrm{C}}|x,\theta_{\mathrm{C}})),\end{aligned}\quad\#(8)$$

模型校准蒸馏的整体损失函数如下:

$$\begin{aligned}\mathcal{L}_t=&\lambda_1\times\mathcal{L}_{\mathrm{CE}}(P(y|x,\theta_{\psi}),\ell)\\&+\lambda_2\times\mathrm{D}_{\mathrm{KL}}(y^{'}(x)||y_{\mathrm{C}}(x)),\end{aligned}\quad\#(9)$$

其中λ_1和λ_2是训练超参数。在θ_{ψ}的初始训练阶段,λ_1设置为 1,λ_2设置为 0。在θ_{ϕ}的后续训练阶段,λ_1设置为 0,λ_2设置为 1。在建立校准迁移学习框架后,我们进一步推导边侧模型不确定性产生的过程。以分类任务为例, 假设各样本类别的证据量表示为ε。参照 Sensoy 等人[44]对不确定性的定义,对于K分类任务存在以下数学关系:

$$u=1-\sum_{k=1}^{K}\frac{\varepsilon_k}{S},\#(10)$$

其中 $S=\sum_{i=1}^{K}(\varepsilon_i+1)$ 表示每个类别中不确定性分布的归一化系数。结合公式 1, 给定一个狄利克雷分布,我们定义K分类网络的概率值的置信度分布为:

$$D(p|\alpha)=\begin{cases}\dfrac{1}{B(\alpha)}\prod_{i=1}^{K}p_i^{\alpha_i-1} & p\in S_K\\ 0 & else\end{cases},\qquad(11)$$

其中α_i表示对样本i进行分类的狄利克雷分布的参数,用于将模型的输出分布映射到高斯分布空间。

通过这些步骤,可以很容易地使用 $\varepsilon_k=(\alpha_k-1)/S$ 和 $S=\sum_{i=1}^{K}\alpha_i$ 从相应的狄利克雷分布的参数中推断出证据。然后,我们将α带入方程 Eq 11 中,即可推到出$f^{'}$的不确定性:

$$u=1-\sum_{k=1}^{K}\frac{(\alpha_k-1)}{\alpha_k}=\frac{K}{S},\qquad(12)$$

然后,在推导出获取不确定性的方程之后,校准代理模型$f^{'}$的优化损失函数定义为:

$$\mathcal{L}_D(\alpha_i)=\sum_{j=1}^{K}y_{ij}\left(\log(S_i)-\log(\alpha_{ij})\right),\#(13)$$

其中$y_{ij}=1,y_{ik}=0$,其表示第i个样本输出属于第j个类别的概率。

本文利用损失函数 Eq。11 优化校准代理模型$y^{'}$,使其输出不确定性$u^{'}$。假设$y^{'}$的狄利克雷分布用 $D(p^{'}|\alpha^{'})$表示,其中 $p^{'}$ 为输出概率,$\alpha^{'}$为分布参数。更进一步,Eq.9 可以推导为:

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}(u^{'}\|u_{\mathrm{C}})&=\mathrm{D}_{\mathrm{KL}}(D(p^{'}|\alpha^{'})||D(p_C|\alpha_C))\\&=\mathrm{D}_{\mathrm{KL}}(P(u^{'}|x,\theta^{'})\|P(u_C|\psi(x),\theta_{\phi})),\end{aligned}\quad\#(14)$$

通过最小化D_{KL},可以确保云侧模型y_{C}的不确定性输出更接近校准代理模型$y^{'}$的不确定性输出。

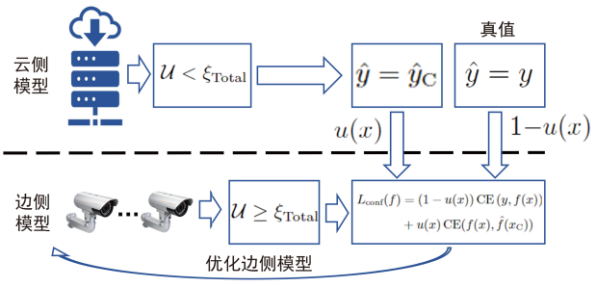


图 3 云边协同优化原理图

3.3 云边协同优化策略

由于边侧模型通常是对云侧模型进行小型化或轻量化处理得到的,因此其鲁棒性和泛化能力通常无法与云侧模型相比。根据公式 4 的定义,认知不确定性指的是模型在面对其未能充分学习或理解的样本时所表现出的较高不确定性。换言之,当模型对某一类样本的学习能力不足时,它会产生较大的预测不确定性。基于这一概念,本文提出通过利用模型输出的认知不确定性（即信心程度）来引导云侧模型优化边侧模型,如图 3 所示。

对于传统的信心程度损失优化方法,我们定义其在云边协同推理优化场景下的损失函数为:

$$L_{conf}(f) = (1-\lambda)\text{CE}(f(x), y) + \lambda\text{CE}(f(x), \hat{f}(x)), \#(15)$$

其中f表示需要优化的边侧模型,$\hat{f}(x)$表示云侧模型对输入样本x预测的硬标签,y是真值标签。λ是一个折中系数,用于平衡从真值标签中学习和从云侧模型中学习的程度。

然而,上述方法仅使用折中系数λ来平衡从真值标签中学习和借助云侧模型进行优化的程度,忽略了关键问题：当前真值标签以及云侧模型的输出结果是否准确可靠。

为了解决人工标注的真值标签的固有局限性以及预训练模型输出的硬标签可能带来的不准确性。参考 Guo 等人的方法[54],本文使用云侧模型与边侧模型输出的不确定性差异作为边侧模型优化的评判标准。当云侧模型与边侧模型输出的不确定性紧密一致且输出的不确定性较低时,表明边侧和云侧对自身输出的信心更高,此时使用真值标签优化边侧模型;当云侧模型输出的不确定性较低而边侧模型输出的不确定性较高时,说明边侧模型对于当前样本的预测结果不自信,需要云侧模型进行优化和更新。基于此,本文引入了基于不确定性的自适应信心程度损失,可根据云边模型输出的不确定性动态调整:

$$L_{\text{conf}}(f) = (1-u(x))\text{CE}(y, f(x)) + u(x)\text{CE}(f(x), \hat{f}(x)), \#(16)$$

$$u(x) = |\mathcal{U}_C - \mathcal{U}|, \#(17)$$

其中$u(x)$表示云侧模型和边侧模型对于输入样本x输出不确定的差异程度,用于计算信心程度权重。$\hat{f}(x)$则是云侧模型中对x的硬标签预测。这个权重帮助确定从云侧模型学习与依赖真值标签进行预测之间的平衡。

通过信心程度调整权重,边侧模型能够灵活判断在何时应当优先考虑基于真值标签的优化,而不是单纯依赖云侧模型的指导,反之亦然。这种适应性是克服两种模型中的不准确和局限性的关键,有效提升边侧模型的整体性能。

4 实验

本章节对本文所使用的数据集、基线模型、实验设置和训练超参数、评价指标等进行详细介绍。

4.1 实验数据

本文评估了基于不确定性校准的云边协同推理框架在视觉和文本分类任务上的表现。在视觉分类任务上对 CIFAR-10, CIFAR-100 和 ImageNet200 数据集进行了实验,在文本分类任务上对 SST-5 和 TweetEval 数据集进行了实验。

CIFAR-10 和 CIFAR-100[55]是最常用的图像分类数据集,分别包含 10 类和 100 类图像,大小为 32×32。每个数据集包含 50,000 张用于训练的图像和 10,000 张用于验证的图像。ImageNet200 [56]是一个用于分类任务的数据集,其中包含 200 个类别的 100,000 张图像, 每个类别有 500 张训练图像、50 张验证图像和 50 张测试图像。 SST-5[57]是常用的情感分类的数据集,具有 5 类标签,本文对数据集进行了预处理,并将这五个类别合并为三个类别。TweetEval[58]由七个异构任务组成,本文仅使用该基准中情感分类部分的数据用于评估模型。

表 1 云侧模型和边侧模型在视觉和文本分类数据集上的对比实验结果

数据集	云侧模型	ACC (云侧)	ECE (云侧)	Size (MB, 云侧)	边侧模型 (蒸馏)	ACC (边侧)	ECE (边侧)	Size (MB, 边侧)
CIFAR10	ViT	0.977	0.310	983	ResNet34	0.911	0.421	243

	ResNet50	0.955	0.161	270	MobilenetV2	0.934	0.274	30
CIFAR100	ViT	0.861	0.231	987	ResNet34	0.643	0.565	244
	ResNet50	0.760	0.182	274	MobilenetV2	0.682	0.390	30
ImageNet	ViT	0.666	0.396	1023	ResNet34	0.482	0.279	245
200	ResNet50	0.492	0.306	289	MobilenetV2	0.466	0.258	30
	BERT-Base	0.739	0.353	417.7	BERT-tiny	0.669	0.311	16.7
SST	RoBERTa	0.774	0.384	475.5	BERT-tiny	0.662	0.216	16.7
	BERT-Large	0.775	0.248	1278.6	BERT-Base	0.671	0.343	417.7
	BERT-Base	0.698	0.305	417.7	BERT-tiny	0.652	0.262	16.7
TweetEval	RoBERTa	0.659	0.271	475.5	BERT-tiny	0.605	0.173	16.7
	BERT-Large	0.706	0.232	1278.6	BERT-Base	0.653	0.255	417.7

4.2 基线模型

本文评估了一系列图像分类与文本分类基线模型,包括用于视觉分类的 MobileNet-v2[59], ResNet34、ResNet50[2]和 ViT[3],以及用于文本分类的 LSTM[60], Bert-tiny、BERT-Base、BERT-Large[1]和 RoBERTa[61]。

MobileNet-v2 基于倒置残差结构,采用轻量级深度卷积来有效减少中间扩展层的特征参数量,从而实现更少的推理参数和更好的识别效果。ResNet34 和 ResNet50 基于深度残差网络结构,缓解了深层网络训练中的梯度消失问题,提高了模型的训练效率和分类精度。

ViT 利用 Transformer 架构将图像分割为小块,并利用自注意力机制有效捕捉图像中各部分的全局依赖关系,广泛应用于视觉分类任务。

LSTM 是一种能学习长期依赖信息的循环神经网络,在长序列文本分类任务上有较好的效果。

BERT-base 是主流文本预训练模型,通过深层双向语言表示显著提升自然语言处理任务的性能;BERT-tiny 是 BERT 的轻量化版本,适合资源受限环境及快速推理任务。

BERT-large 是 BERT 的增强版本,具备更多隐藏层和参数,适用于复杂的语言建模任务;

RoBERTa 通过优化预训练过程和调整超参数进一步提高预训练 BERT 模型的性能。

4.3 实验设置和训练超参数

本研究的实验在配备了 2 块 Nvidia A100 GPU（每块 80GB 显存）和 2 块 Nvidia V100 GPU（每块 40GB 显存）的计算机上进行,其他硬件配置包括 2×Intel Xeon Gold 6148 CPU、512GB DDR4 RAM 和 2×240GB M.2 SSD。在实验计算资源分配方面,云侧模型充分利用全部可用计算资源; 而对于边侧环境,我们采用精确受控的资源分配策略,在单块 V100 GPU 上实施严格的计算资源限制（仅分配 25%的计算单元）,同时对内存占用与功耗进行精确约束,以真实模拟边缘设备的资源受限特性。

在图像分类任务中,批大小设置为 128,优化器采用 Adam。学习率在[1e-6, 1e-5, 2e-5]之间选择,模型校准蒸馏的训练周期为 50,云边协同优化的训练周期为 20。

对于文本分类任务,批大小设置为 300,优化器采用 AdamW。学习率在[1e-6, 1e-5, 2e-5]之间选择,模型校准蒸馏的训练周期为 100,云边协同优化的训练周期为 50。

4.4 评价指标

在评估模型预测的结果好坏时,本文采用准确率(ACC)作为评价指标。此外,为了评估模型输出的不确定性好坏,本文使用预期校准误差（ECE）作为衡量标准。

定义 1 准确率(ACC)是衡量模型分类性能的一个常用评价指标,其定义为模型正确预测的样本数占总样本数的比例。具体计算公式表示如下：

$$\mathrm{ACC}=\frac{1}{N}\sum_{i=1}^{N}1(\hat{y}_i=y_i),\#(18)$$

其中,$1(\cdot)$ 是指示函数,当括号内的条件为真时值为 1,否则为 0。

定义 2 预期校准误差(ECE)定义为多个箱体之间的准确度和不确定性之间的预期差异。首先需定义每个箱体 B_m 的准确度和不确定性,如下所示：

$$\mathrm{ACC}(B_m)=\frac{1}{|B_m|}\sum_{i\in B_m}\mathbf{1}(\hat{y}_i=y_i)$$
$$\mathrm{Conf}(B_m)=\frac{1}{|B_m|}\sum_{i\in B_m}(1-\hat{u}_i),\quad \#(19)$$
$$B_m=\{0,\Upsilon\}$$

其中,y_i 和 $\hat{y}_i$ 分别表示样本 i 的真实标签和预测标签,而 $\hat{u}_i$ 是样本 i 的预测的不确定性, Υ表示不确定性的阈值,当模型输出的不确定性小于当前阈值时该样本参与计算,否则不参与计算。由此得出预期校准误差如下：

$$\mathrm{ECE}=\sum_{m=1}^{M}\frac{|B_m|}{n}|\mathrm{ACC}(B_m)-\mathrm{Conf}(B_m)|,\#(20)$$

其中n表示样本的总数,m表示不确定性的区间数量,B_m表示落入当前区间m内的样本数量。

5 实验结果分析

本文将全面评估基于不确定性估计的云边协同推理方法在视觉分类和图像分类任务中的有效性与通用性,并深入探讨其在跨域泛化能力方面的表现。

5.1 云边协同推理实验

本文首先对大参数量的云侧模型（ResNet50、ViT、BERT-base、BERT-Large 和 RoBERTa）与轻量化的边侧模型（MobileNet-v2、ResNet34 和 BERT-tiny）进行了对比分析,实验结果如表 1 所示。结果显示,云侧模型在五个数据集上均表现出较强的识别能力。然而,轻量化的边侧模型识别性能相较于云侧模型下降明显,这表明虽然轻量化模型能够减少参数量并提升推理效率,但不可避免地会带来性能下降的风险。

本文还在表 1 中发现边侧模型输出不确定性的能力并不会随着模型轻量化的过程退化。这表明,在不增加额外训练成本的情况下,轻量化边侧模型仍能保持与云侧模型相似的输出不确定性能力,展现出模型校准蒸馏方法的强大鲁棒性,同时验证了将云侧模型与边侧模型的输出不确定性作为推理依据的可行性。

表 2 视觉模型的云边协同推理实验结果

数据集	云侧	边侧（协同）$\xi_{\mathrm{Total}}=0.7$	
		ResNet34	MobilenetV2
CIFAR-10	ResNet50	0.913 (+0.002)	0.936 (+0.002)
	ViT	0.915 (+0.004)	0.934 (+0.000)
CIFAR-100	ResNet50	0.756 (+0.113)	0.741 (+0.059)
	ViT	0.860 (+0.217)	0.682 (+0.000)
ImageNet 200	ResNet50	0.499 (+0.017)	0.490 (+0.024)
	ViT	0.642 (+0.160)	0.658 (+0.192)

表 3 文本模型的云边协同推理实验结果

数据集	云侧	边侧（协同）$\xi_{\mathrm{Total}}=0.7$
		BERT-tiny/ BERT-Base
SST-5	BERT-Base	0.689 (+0.020)
	RoBERTa	0.677 (+0.015)
	BERT-Large	0.737 (+0.075)
TweetEval	BERT-Base	0.667 (+0.015)
	RoBERTa	0.621 (+0.016)
	BERT-Large	0.665 (+0.012)

接着,本文对这些图像与文本的云侧模型和边侧模型进行了协同推理实验,如表 2 和表 3 所示。总体而言,本文提出的基于不确定性校准的云边协同推理框架在所有测试的模型和数据集上均提升了边侧模型的识别准确率。在视觉分类任务中,云边协同推理方法表现尤为显著,例如边侧模型 ResNet34 在 CIFAR-100 数据集上的识别准确率从 0.643 提升至 0.860。在文本分类任务中,类似的提升也得到了验证,边侧模型 BERT-tiny 在 SST-5 数据集上的识别准确率从 0.669 提升至 0.689。说明基于不确定性校准的云边协同推理框架能够有效识别边侧模型易于处理和难以处理的样本,优化云边协同推理的整体性能。

与此同时,为了进一步验证基于不确定性校准的云边协同推理方法的有效性,本文将提出的方法与现有模型轻量化方法剪枝、量化和蒸馏方法进行了对比,其中在视觉任务上使用 ViT 模型、文本任务上使用 BERT-Large 模型, 如表 4 所示。 结果表明,基于不确定性校准的云边协同推理框架在各项评价指标上均显著优于模型轻量化方法,在 SST-5 和 CIFAR-100 数据集上相较于最优的模型轻量化方法提升了 2.1%, 这一结果证实了相较于传统的剪枝、蒸馏、量化等模型轻量化技术,融合不确定性估计的云边协同推理框架能够更有效地保持模型的泛化能力和鲁棒性,尤其在面对复杂文本场景时表现出明显优势。

表 4 不同模型轻量化方法在视觉和文本分类任务上的效果

数据集	云侧模型	剪枝	量化	蒸馏	协同
SST-5	0.775	0.564	0.524	0.671	0.689
CIFAR100	0.861	0.622	0.745	0.643	0.756

此外,本研究对边侧偶然不确定性 $\mathcal{U}_a$ 和云侧偶然不确定性 $\mathcal{U}_a^c$ 的权重分配策略进行了对比实验。本文选择 SST-5 作为测试数据集,以参数量较大且泛化能力较强的 BERT-Base 作为云侧模型,参数量较小的 BERT-Tiny 作为边侧模型,采样并分析了不同偶然不确定性权重比例（2:8, 5:5 和 8:2）对基于不确定性校准的云边协同推理方法性能的影响,如表 5 所示。实验结果表明,采用公式 7 所定义的自适应权重分配策略,模型在各种权重配置下均能保持稳定且接近最优的性能表现,证明了本文所提出的权重计算方法具有显著的鲁棒性和有效性,能够在不同的不确定性分布情境下自适应地整合边侧和云侧模型的预测结果。

表 5 不同模型轻量化方法在视觉和文本分类任务上的效果

数据集	云侧/边侧权重比例		
	2:8	5:5	8:2
SST-5	0.672	0.669	0.671

表 6 云侧模型 ViT 和 BERT-Base 的校准蒸馏实验结果

		ViT					ViT(校准蒸馏)				
数据集	阈值	1	0.8	0.6	0.4	0.2	1	0.8	0.6	0.4	0.2
CIFAR-10	ACC	0.795	-	-	-	-	0.977	0.977	0.977	0.977	-
	ECE	0.695	-	-	-	-	0.31	0.31	0.31	0.312	-
	Num	10,000	0	0	0	0	10,000	10,000	10,000	8,112	0
CIFAR-100	ACC	0.662	-	-	-	-	0.862	0.893	0.905	0.915	0.83
	ECE	0.631	-	-	-	-	0.233	0.219	0.217	0.219	0.14
	Num	10,000	0	0	0	0	10,000	9,537	9,326	9,055	6
ImageNet200	ACC	0.591	0.659	-	-	-	0.6	0.719	0.928	-	-
	ECE	0.614	0.642	-	-	-	0.396	0.449	0.659	-	-
	Num	10,000	192	0	0	0	10,000	7,805	561	0	0
		BERT-Base					BERT-Base(校准蒸馏)				
SST-5	ACC	0.733	0.733	0.74	0.948	-	0.741	0.744	0.797	0.912	-
	ECE	0.184	0.184	0.185	0.342	-	0.224	0.225	0.253	0.263	-
	Num	2,210	2,208	2,178	466	0	2,210	2,196	1,796	753	0
TweetEval	ACC	0.663	0.665	0.672	0.776	-	0.602	0.614	0.681	0.798	-
	ECE	0.091	0.089	0.088	0.145	-	0.15	0.122	0.12	0.147	-
	Num	12,284	12,233	11,782	5,136	0	12,284	10,789	8,018	2,545	0

- 当一个区间中没有样本时 (Num 为 0),准确率 (ACC) 和预期校准误差 (ECE) 均为"-".

5.2 模型校准蒸馏实验

为了评估模型校准蒸馏方法在云侧模型上的应用效果,本文对视觉预训练模型 ViT 和文本预训练模型 BERT-Base 进行了实验,将利用公式 13 训练模型的输出结果和利用模型校准蒸馏方法训练模型的在不同阈值下的输出结果进行了对比,其中阈值表示当不确定性小于当前阈值时,样本参与计算,否则不参与计算。如表 6 所示,当预训练模型 ViT、BERT-Base 使用公式 13 进行训练时,虽然模型能够具备输出不确定性的能力,但是其识别效果会下降。随后,通过使用校准迁移学习的方法,不仅成功保证了模型的识别准确率不下降,还赋予了模型输出不确定性的能力,即不确定性越低的区间模型预测的结果越好。我们还发现越是参数量较大的模型,其校准蒸馏的效果越明显,诸如 ViT 通过校准蒸馏后的效果要优于直接使用公式 13 的方法,验证了本文设计的校准蒸馏方法应用于参数量较大的云侧模型是可行且有效的。

为了进一步验证模型校准蒸馏方法的有效性,本文将使用 Dirichlet 分布的模型校准蒸馏方法与贝叶斯估计和 Softmax 熵估计方法进行了对比分析,其中文本分类任务使用 BERT-Large 作为云侧模型、BERT-Base 作为边侧模型, 图像分类任务使用 ViT 作为云侧模型、ResNet34 作为边侧模型, 如表 7 所示。结果显示,使用 Dirichlet 分布的模型校准蒸馏方法的准确率与贝叶斯估计方法相近,但推理时间要远远小于贝叶斯估计方法。这表明校准蒸馏方法在计算效率与资源消耗平衡方面具有显著的优势,能

够在保证不确定性估计质量的同时,大幅降低推理延迟,更适合基于模型选择的云边协同推理任务。

表 7 不同不确定性量化方法对云边协同推理的影响

数据集	评价指标	Softmax 熵估计	贝叶斯估计	校准蒸馏
SST-5	准确率	0.623	0.679	0.689
	速度 ms/s	27.900	93.350	28.320
CIFAR 100	准确率	0.704	0.782	0.756
	速度 ms/s	0.528	3.432	0.638

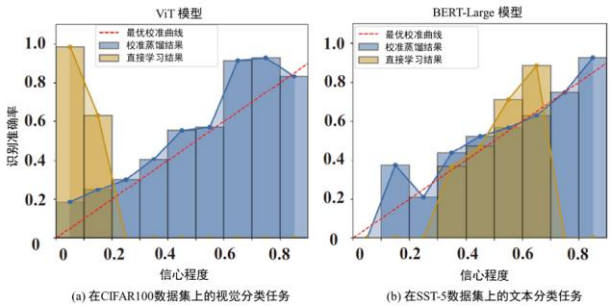


图 4 模型校准实验结果

信心程度指模型对其自身预测结果的确信程度,通常用“1-预测的不确定性”来表示,数值越高代表模型对该预测结果越有把握。为了验证本文提出的基于不确定性校准的云边协同推理框架中模型预测的不确定性和识别准确率之间的关联,本文在 CIFAR-100 和 SST-5 数据集上分别对云侧模型 ViT 和 BERT-Large 进行了模型校准实验来验证,如图 4 所示。其中黄色柱状分布表示模型直接学习得到的不确定度分布,蓝色柱状分布表示经过校准蒸馏后的结果,红色虚线为理想校准基线,分布曲线越贴近该虚线说明校准越充分。从实验结果可以看出,尽管云侧模型参数规模庞大,其直接输出的不确定性与真实识别准确率仍存在明显偏差,说明仅靠不确定性损失函数训练难以获得良好的不确定度估计能力。采用校准蒸馏后,仅需微调校准层即可显著提升不确定性和识别准确率一致性,既保留了模型的核心表征能力,又提高了预测可靠性。

为了分析训练轮次对模型校准蒸馏方法收敛性和最终性能的影响,本文在 SST-5 和 CIFAR-100 数据集上分别对 BERT-Base、BERT-Large、ResNet34 以及 ViT 模型在 [1, 5, 20, 50, 100, 200] 轮次范围内进行测试,如图 5 所示,模型精度在初始阶段（1-5 轮）呈现显著的上升趋势。随后在 20-100 轮区间内,精度增长逐渐放缓逐渐达到最高点,表明模型已基本达到收敛状态。当训练轮次超过 100 轮后,模型性能趋于稳定,额外的训练轮次（100-200 轮）会导致模型在测试集上的精度下降,此时说明模型已经过拟合了。 基于上述结果,本文在将图像分类任务中模型校准蒸馏训练周期设置为 50,将文本分类任务中的模型校准蒸馏训练周期设置为 100。

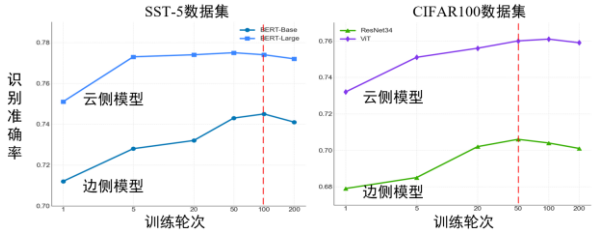


图 5 不同校准蒸馏训练轮次下云边模型的识别结果

5.3 不同不确定性阈值下的云边协同推理实验

本文进一步探讨了在不同总体不确定性阈值下,边侧模型与云侧模型协同推理后准确率的差异以及推理速度的差异。鉴于每个样本输出的不确定性为 0 到 1 之间的数值,我们将不确定性区间均匀划分为 10 个等份,即总体不确定性阈值ξ_{Total}分别设置为[0.1,0.2, ...,0.9]。并采用“1 – 总体不确定性”的“信心程度”作为横坐标展示不同阈值下云边协同推理框架的准确率(实线)与推理速度(推理完成一个任务所需要的时间,单位"秒",虚线)的变化情况(图 5)以及吞吐量的变化情况(图 6)。当边侧模型输出的信心程度大于或等于设定阈值时,直接使用边侧模型进行推理;若输出信心程度低于阈值,则使用云侧模型进行推理。

不确定性阈值与模型推理时间和准确率的对比：如图 6 所示,可以观察到模型推理时间与信心程度阈值呈正相关关系。随着信心程度阈值的增加（从 0 到 0.3）,云侧模型参与计算的比例也随之增大,从而导致推理时间明显增长。在较低信心程度阈值下,云边协同推理框架能够顺畅地进行任务切换,推理时间与准确率之间保持较为理想的平衡。然而,当阈值超过某一临界值时,推理时间开始显著上升,同时推理速度下降,准确率趋于稳定。这表明,尽管提高阈值有助于提升准确率,但过度依赖云侧模型时,若云侧与边侧模型之间存在较高的通信延迟,推理性能将受到较大影响。

不确定性阈值与吞吐量的对比：如图 7 所示,可以观察到,随着信心程度阈值的增加（即不确定性阈值的降低）,模型的吞吐量逐渐下降。在信心程度阈值位于 0 到 0.1 之间时,边侧模型进行推理的比例较大,云侧模型的参与度较低,因此云侧模型的计算负载较轻,吞吐量基本保持稳定。然而,随着阈值的不断增加,云侧模型需要分配更多计算资源来协助

边侧模型的推理任务,导致云侧模型的计算负担加重,吞吐量出现急剧下降。当信心程度阈值增大至某一临界值后,边侧模型的输出不确定性逐渐增大,云侧模型的参与比例趋于稳定,推理任务的负载也趋于平稳,并最终与仅使用云侧模型进行推理时的吞吐量相当。 为了在推理时间、识别准确率以及模型吞吐量之间取得更好的平衡,本文选择 $\xi_{\text{Total}} = 0.7$作为判断边侧模型输出有效性的标准。该阈值不仅确保了较高的准确率,同时避免了过度依赖云侧模型,从而优化了云边协同推理的效率,显著提升了推理速度,并有效降低了由于通信延迟带来的推理时间瓶颈问题。在此阈值下,推理速度达到了完全由云侧模型进行推理时的 5 倍。

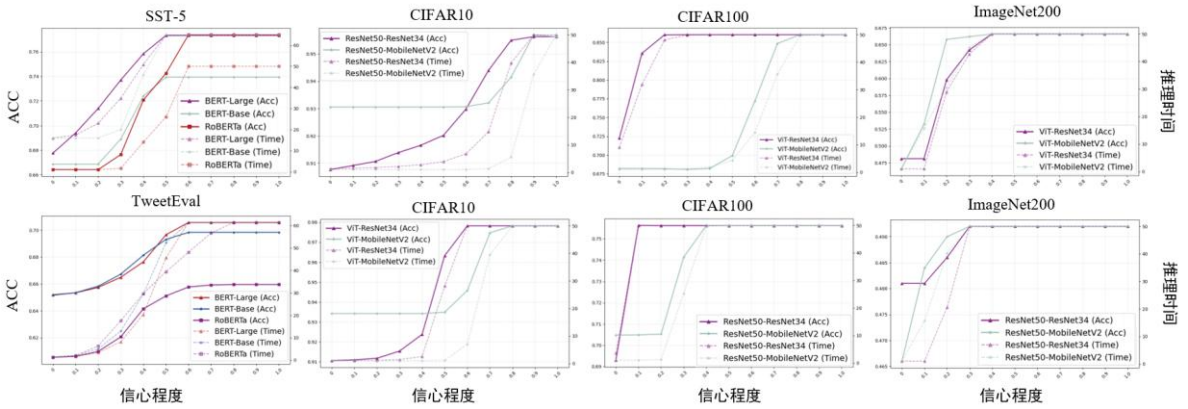


图 6 不同不确定性阈值下的云边协同推理准确率与推理时间实验结果

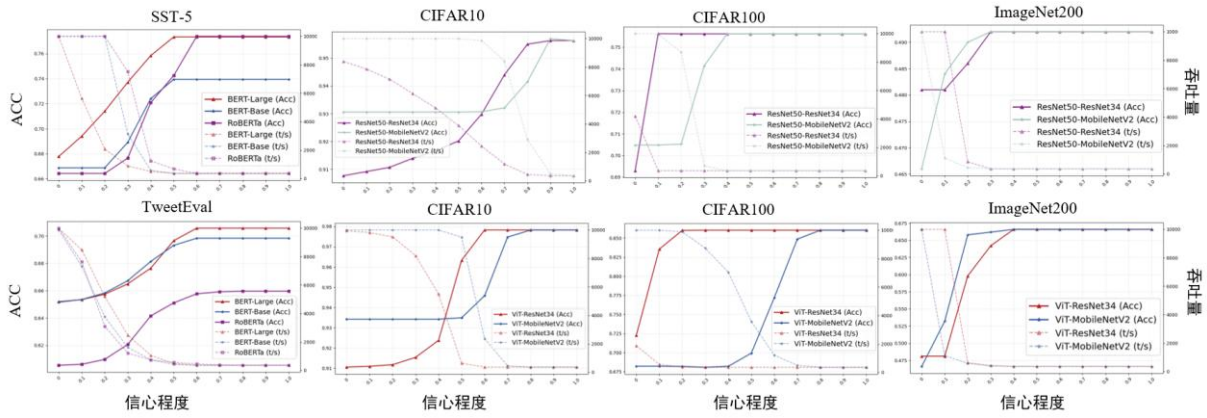
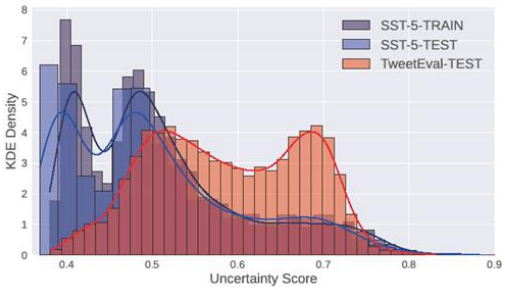


图 7 不同不确定性阈值下的云边协同推理准确率与推理吞吐量实验结果

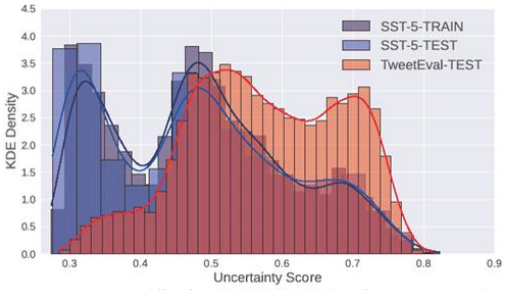
5.4 不确定性的分布外泛化环境实验

为了验证云侧和边侧模型输出的不确定性是否可以应用于分布外泛化的场景,本文在 SST-5 数据集上训练了 BERT 和 RoBERTa 模型,并在 TweetEval 数据集上进行了测试,如图 8 所示。对于域内(ID)数据,例如从 SST-5 训练集到 SST-5 测试集,模型的输出不确定性表现出一致性,倾向于产生具有较低不确定性的预测。然而,对于分布外泛化(OOD)数据,例如从 SST-5 训练集到 TweetEval 测试集,模型的输出不确定性表现出明显的转变,倾向于产生具有较高不确定性的预测。这表明经过校准蒸馏后,模型可以准确区分域内和分布外泛化样本,将较高的不确定性分配给不可靠的 OOD 样本,将较低的不确定性分配给可靠的 ID 样本。

此外,通过分析 BERT 模型和 RoBERTa 模型输出的不确定性趋势,我们发现两个模型在面对分布外数据时的输出不确定性具有一致性。具体来说,在 SST-5 到 TweetEval 的迁移任务中,尽管两个模型的绝对不确定性数值存在差异,但它们的趋势和变化模式相似,即在处理OOD数据时,两者都表现出显著的高不确定性。这表明,校准蒸馏方法能够有效地保证代理模型(如边侧模型)与目标模型(如云侧模型)在分布外数据上的一致性。



(a) BERT模型从SST5数据集到TweetEval数据集的泛化性实验



(b) RoBERTa模型从SST5数据集到TweetEval数据集的泛化性实验

图 8 BERT 模型在的 SST5 与 TweetEval 数据集上输出的 KDE 密度直方图(a), RoBERTa 模型在 SST5 与 TweetEval 数据集上输出的 KDE 密度直方图(b)

针对 OOD 场景下的不确定性表现,Ovadia 等人[62]的研究发现面对 OOD 数据时,认知不确定性的增长幅度显著超过偶然不确定性。大规模云侧模型虽然在分布内数据上展现出较低的整体不确定性 (如图 8 蓝色分布所示),但当遇到 OOD 数据时,其认知不确定性呈现出更为明显的相对增长[62]。相比之下,边侧模型尽管整体不确定性水平较高,但其偶然不确定性在分布内外场景中表现出相对稳定的特性[41]。这种差异性表现可归因于模型结构特点：边侧模型由于参数规模有限、网络拓扑结构相对简化,对数据中的随机噪声敏感度较低,从而使偶然不确定性保持更高的稳定性；而云侧模型凭借更大的参数空间能够捕捉更为复杂的数据模式,但这也导致其对数据分布的细微变化表现出更强的敏感性。

5.5 不确定性可视化分析实验

为了将云边模型的不确定性进行直观展示,本文采用了一种新颖的不确定性映射方法。该方法的核心思想是将证据学习中的全局不确定性分数映射到图像的每个局部区域以得到不确定性分布图。该方法可以将全局的不确定性信息与局部的特征信息相结合。通过计算每个通道的空间平均值并与全局不确定性相乘得到权重因子,并将其应用到原始特征图上,从而将全局不确定性"分配"到每个局部区域。具体来说,我们使用以下公式：

$$U_{map} = \text{Normalize}\left(\sum_{c}\left(\frac{1}{H\cdot W}\sum_{h}\sum_{w}F_{c,h,w}\right)\times\left(\frac{K}{\sum_{k}\left(\text{ReLU}\left(M(I)\right)_{k}+1\right)}\right)\cdot F_{c,h,w}\right), (21)$$

其中,U_{map} 指最终的局部不确定性映射,$F_{c,h,w}$ 表示特征图,c为通道,h 和 w 为空间维度,H, W 为特征图的高度和宽度,K 是全局不确定性相关的常数,M(I)表示模型对输入图像 I 的预测。如图 9 所示,图中的热图展示了模型输出的高不确定性区域,颜色越亮表示该区域的不确定性越大。

可以观察到云侧模型和边侧模型在关注高不确定性区域上显示出较高的一致性。这种一致性表明两种模型在识别潜在风险或决策关键点时具有相似的敏感度。尽管如此,云侧模型在整体上展示了较低的不确定性水平,这表明它在处理信息和做出预测决策时更为稳定和可靠。因此,在边侧模型在某些特征上的理解尚不充分,自信程度不足时,云侧模型能够准确定位不确定性区域,协同边侧模型进行推理。基于上述观察和分析,本文提出的协同推理最终显著提升了识别结果的准确性。

5.6 仿真实验分析

为了验证基于不确定性校准的云边协同推理方法在实际环境中的性能,本文构建了仿真实验平台, 模拟了典型边缘设备的硬件特性（4 核 1.5GHz ARM Cortex-A72 处理器, 4GB RAM）和 NVIDIA Jetson Nano（128 核 Maxwell GPU,4GB LPDDR4 内存, 1.43 TFLOPS）。通信特性方面,本文模拟了边缘设备的网络带宽限制（10-20Mbps）和端到端通信延迟（10-20ms）。实验中, 本文特意引入了网络波动（±30%带宽波动）等真实部署环境中的挑战因素,以全面评估方法在复杂场景下的鲁棒性,如表 8 所示。结果表明, 与单一使用云侧模型和边侧模型相比, 本文提出的基于不确定性校准的云边协同推理框架在准确率、计算时间和吞吐量上具有显著的优势, 充分证明了该方法在资源受限的边缘计算环境中的优越性能和实用价值。

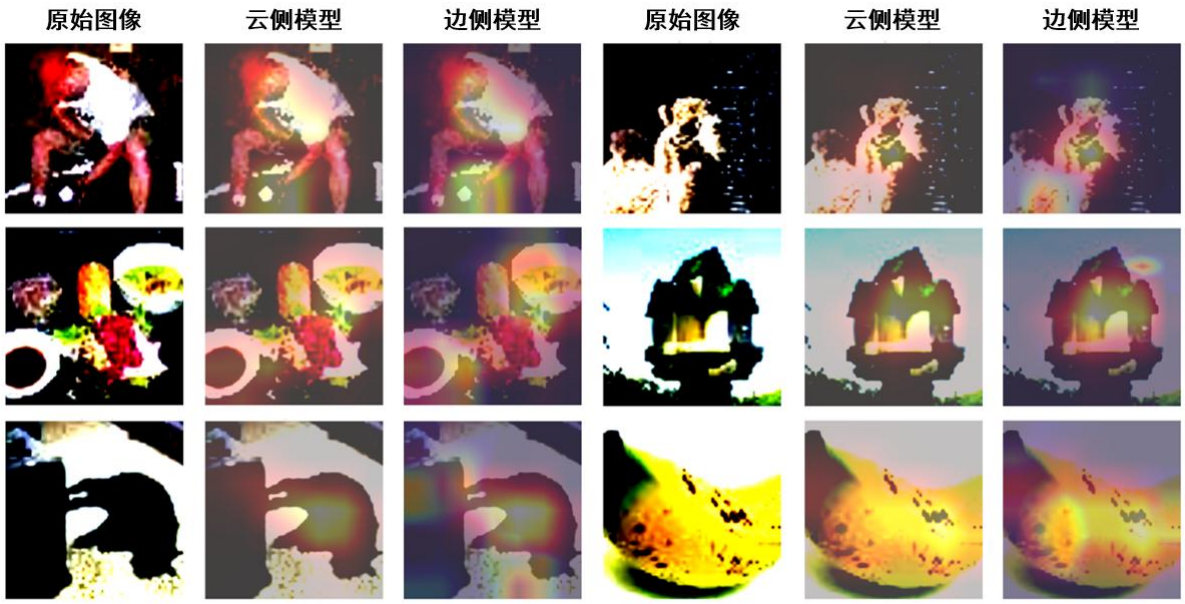


图 9 在云侧模型 ResNet50 和边侧模型 MobilenetV2 上的不确定性可视化实验结果

表 8 云边协同仿真实验

数据集	模型类别	准确率	计算时间	吞吐量
SST-5	云侧 Bert-Base	0.739	1596.1s	1.385/s
	边侧 Bert-Tiny	0.669	883.6s	2.500/s
	本文方法	0.689	1012.3s	2.183/s
CIFAR 100	云侧 ResNet50	0.760	41.2s	15.384/s
	边侧 MobileNetV2	0.682	25.4s	24.390/s
	本文方法	0.741	31.1s	20.408/s

6 总结与展望

本文提出了一种基于不确定性校准的云边协同推理框架,使得在推理过程中能够更加精准地选择使用云侧模型、边侧模型或云边模型协同加权输出。此外,设计了模型校准蒸馏方法,通过迁移学习有效解决了云侧和边侧模型在处理复杂输入时,输出不确定性能力退化的问题。针对云边侧模型输出不确定性的差异,本文还设计了云边协同优化策略,有效提升了边侧模型对难样本的识别能力。实验结果表明,在图像分类和文本分类任务中,该框架显著提高了边侧模型的推理准确性,并通过合理的云边协同优化了资源利用率。

本文首次将模型输出的不确定性与模型选择策略相结合,克服了传统模型选择依赖 Softmax 输出层在面对未知领域输入时,即使模型推理信心不足也会输出概率较高类别,从而导致错误预测的问题。该推理框架主要应用于文本和图像分类任务,未来可以针对回归任务、生成任务进行更详细的设计和扩展,使云边协同推理框架在复杂多变的应用场景中更加灵活智能地进行资源调度和模型选择,为无人驾驶等场景下的分类任务提供更高效的支持。

7 符号定义

为便于读者理解本文中使用的各种符号及其含义,特整理如表 9 所示的符号定义表。该表汇总了文中出现的主要数学符号、专业术语缩写及其对应的定义,以供读者参考,这将有助于更加清晰地理解文中的核心概念和理论推导过程。

表 9 符号定义表

符号	定义	符号	定义
$\mathcal{M}_{\mathrm{E}}$	边侧模型	$\mathcal{M}_{\mathrm{C}}$	云侧模型
ε	证据表示	$\mathcal{D}$	边侧模型的训练数据集
$\mathcal{U}_{\mathrm{a}}$	边侧模型偶然不确定性	$\mathcal{U}$	边侧模型输出的不确定性
$\mathcal{U}_{\mathrm{a}}^{\mathrm{C}}$	云侧模型偶然不确定性	$\mathcal{U}_{\mathrm{e}}$	边侧模型认知不确定性

ξ_{Total}	总体不确定性阈值	α_i	Dirichlet 分布的参数
$\hat{y}_{\text{C}}$	云侧模型推理结果	ξ	所有α_i之和
ξ_{ep}	认知不确定性的阈值	ξ_{a}	偶然不确定性阈值
$\hat{y}_{\text{E}}$	边侧模型推理结果	$\hat{y}_C$	云侧模型推理结果
w_{C}	云侧模型的权重	w_{E}	边侧模型的权重
$f^{'}$	校准代理模型	$\theta^{'}$	校准代理模型的参数
$y^{'}$	校准代理模型的实现	ϕ	校准层
f_{C}	云侧模型	$\mathcal{U}^{'}$	校准代理模型输出的不确定性
$\mathcal{U}_{\text{C}}$	云侧模型输出的不确定性	θ_{ψ}	云侧模型输出的不确定性
y_{C}	云侧模型的实现	θ_{C}	云侧模型的参数
ℓ	真实标签	θ_{ϕ}	云侧模型的校准层参数
λ_1, λ_2	训练超参数,用于平衡不同损失项的权重	λ	折中系数,平衡从真值标签中学习和从云侧模型中学习的程度.
$L_{conf}(f)$	信心程度损失函数	D_{KL}	KL 散度
f	表示需要优化的边侧模型	$D(.)$	狄利克雷分布
$u(x)$	云侧模型和边侧模型对输入样本x输出不确定性的差异程度	L_{D}	模型校准蒸馏的整体损失函数
p_i	类别i的预测概率	y	真值标签
$\hat{f}(x)$	云侧模型中对x的硬标签预测	L_{CE}	交叉熵损失函数

参 考 文 献

[1] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies. Minneapolis, USA, 2019,1: 4171-4186.

[2] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, NV, USA, 2016: 770-778.

[3] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale// Proceedings of the 9th International Conference on Learning Representations,Virtual , Austria, 2021

[4] Kang Y, Hauswald J, Gao C, et al. Neurosurgeon: Collaborative intelligence between the cloud and mobile edge// Proceedings of the Twenty-Second International Conference on Architectural Support for Programming Languages and Operating Systems, {ASPLOS}, 2017. Xi'an, China, 2017: 615-629.

[5] Hassibi B, Stork D G, Wolff G J. Optimal brain surgeon and general network pruning//Proceedings of International Conference on Neural Networks (ICNN'88), San Francisco, USA,1993: 293-299.

[6] Wu Z, Nagarajan T, Kumar A, et al. Blockdrop: Dynamic inference paths in residual networks//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 8817-8826.

[7] Choukroun Y, Kravchik E, Yang F, et al. Low-bit quantization of neural networks for efficient inference//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul, Republic of Korea, 2019: 3009-3018.

[8] Taylor B, Marco V S, Wolff W, et al. Adaptive deep learning model selection on embedded systems//Proceedings of the 19th {ACM} {SIGPLAN/SIGBED} International Conference on Languages, Compilers, and Tools for Embedded Systems, {LCTES} 2018. Philadelphia, USA, 2018: 31-43.

[9] Stamoulis D, Chin T W R, Prakash A K, et al. Designing adaptive neural networks for energy-constrained image classification// Proceedings of the International Conference on Computer-Aided Design, San Diego, USA, 2018: 23.

[10] Ahmed S T, Hefenbrock M, Tahoori M B. Tiny Deep Ensemble: Uncertainty Estimation in Edge AI Accelerators via Ensembling Normalization Layers with Shared Weights//Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design. NJ, USA, 2024, 71: 1-9.

[11] Alharbi M, Karimi H A. Context-aware sensor uncertainty estimation for autonomous vehicles. Vehicles, 2021, 3(4): 721-735.

[12] Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision?//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 5580-5590.

[13] Laskaridis S, Venieris S I, Almeida M, et al. SPINN: synergistic progressive inference of neural networks over device and

cloud//Proceedings of the 26th annual international conference on mobile computing and networking. London, UK, 2020, 37: 1-15.

[14] Vysogorets A, Kempe J. Connectivity matters: Neural network pruning through the lens of effective sparsity. Journal of Machine Learning Research, 2023, 24(99): 1-23.

[15] Han S, Pool J, Tran J, et al. Learning both weights and connections for efficient neural network// Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems, Montreal, Quebec, Canada, 2015, 1135--1143.

[16] Bolukbasi T, Wang J, Dekel O, et al. Adaptive neural networks for efficient inference//Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 2017: 527-536.

[17] Gysel P, Motamedi M, Ghiasi S. Hardware-oriented approximation of convolutional neural networks. arXiv preprint arXiv:1604.03168, 2016.

[18] Rastegari M, Ordonez V, Redmon J, et al. Xnor-net: Imagenet classification using binary convolutional neural networks// Proceedings of the European conference on computer vision, Amsterdam, The Netherlands, 2016: 525-542.

[19] Tang Y, Wang Y, Li H, et al. Mv-net: Toward real-time deep learning on mobile gpgpu systems. ACM Journal on Emerging Technologies in Computing Systems (JETC), 2019, 15(4): 1-25.

[20] Song M, Zhong K, Zhang J, et al. In-situ ai: Towards autonomous and incremental deep learning for iot systems// Proceedings of the 2018 IEEE International Symposium on High Performance Computer Architecture (HPCA). Vienna, Austria, 2018: 92-103.

[21] Zheng Y, Chen Y, Qian B, et al. A review on edge large language models: Design, execution, and applications. ACM Computing Surveys, 2025, 57(8): 1-35.

[22] Hao Z, Jiang H, Jiang S, et al. Hybrid slm and llm for edge-cloud collaborative inference//Proceedings of the Workshop on Edge and Mobile Foundation Models. Tokyo, Japan, 2024: 36-41.

[23] Yao J, Zhang S, Yao Y, et al. Edge-cloud polarization and collaboration: A comprehensive survey for ai. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(7): 6866-6886.

[24] Yeh Y R, Huang C H, Wang Y C F. Heterogeneous domain adaptation and classification by exploiting the correlation subspace. IEEE Transactions on Image Processing, 2014, 23(5): 2009-2018.

[25] Samat A, Persello C, Gamba P, et al. Supervised and semi-supervised multi-view canonical correlation analysis ensemble for heterogeneous domain adaptation in remote sensing image classification. Remote sensing, 2017, 9(4): 337.

[26] Wang X, Ma Y, Cheng Y, et al. Heterogeneous domain adaptation network based on autoencoder. Journal of Parallel and Distributed Computing, 2018, 117: 281-291.

[27] Feuz K D, Cook D J. Transfer learning across feature-rich heterogeneous feature spaces via feature-space remapping (FSR). ACM transactions on intelligent systems and technology (TIST), 2015, 6(1): 1-27.

[28] Xiao M, Guo Y. Feature space independent semi-supervised domain adaptation via kernel matching. IEEE transactions on pattern analysis and machine intelligence, 2014, 37(1): 54-66.

[29] Rosenfeld B, Rajendran B, Simeone O. Fast on-device adaptation for spiking neural networks via online-within-online meta-learning// Proceedings of the 2021 IEEE Data Science and Learning Workshop (DSLW). Toronto, Canada, 2021, 2021: 1-6.

[30] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks// Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 2017: 1126-1135.

[31] Pan H, Wang C, Qiu M, et al. Meta-KD: A Meta Knowledge Distillation Framework for Language Model Compression across Domains//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Virtual, 2021,1: 3026-3036.

[32] Fifty C, Duan D, Junkins R G, et al. Context-Aware Meta-Learning// Proceedings of the Twelfth International Conference on Learning Representations, Vienna, Austria, 2024:1-22.

[33] Vettoruzzo A, Bouguelia M R, Vanschoren J, et al. Advances and challenges in meta-learning: A technical review. IEEE transactions on pattern analysis and machine intelligence, 2024, 46(7): 4763-4779.

[34] Teshima T, Sato I, Sugiyama M. Few-shot domain adaptation by causal mechanism transfer//Proceedings of the 37th International Conference on Machine Learning, Virtual, 2020: 9458-9469.

[35] Chen Y, Bühlmann P. Domain adaptation under structural causal models. Journal of Machine Learning Research, 2021, 22(261): 1-80.

[36] Yue Z, Sun Q, Hua X S, et al. Transporting causal mechanisms for unsupervised domain adaptation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 8579-8588.

[37] Yuan J, Ma X, Xiong R, et al. Instrumental variable-driven domain generalization with unobserved confounders. ACM Transactions on Knowledge Discovery from Data, 2023, 17(8): 1-21.

[38] Blundell C, Cornebise J, Kavukcuoglu K, et al. Weight uncertainty in neural network// Proceedings of the 32nd International Conference on Machine Learning. Lille, France, 2015: 1613-1622.

[39] Kingma D P, Salimans T, Welling M. Variational dropout and the local reparameterization trick. Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems. Montreal, Canada, 2015:2575—2583.

[40] Li Y, Gal Y. Dropout inference in bayesian neural networks with alpha-divergences// Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 2017: 2052-2061.

[41] Gal Y, Ghahramani Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning// Proceedings of the 33nd International Conference on Machine Learning, New York City, USA, 2016: 1050-1059.

[42] MacKay D J C. A practical Bayesian framework for backpropagation networks. Neural computation, 1992, 4(3): 448-472.

[43] Harrison J, Willes J, Snoek J. Variational Bayesian Last Layers// Proceedings of the Twelfth International Conference on Learning Representations, Vienna, Austria, 2024:1-31.

[44] Sensoy M, Kaplan L, Kandemir M. Evidential deep learning to quantify classification uncertainty//Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems. Montreal, Canada, 2018: 3183—3193.

[45] Yu Y, Deng D, Liu F, et al. Adaptive Negative Evidential Deep Learning for Open-set Semi-supervised Learning. arXiv preprint arXiv:2303.12091, 2023.

[46] Amini A, Schwarting W, Soleimany A, et al. Deep evidential regression. Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems. Virtual, 2020: 14927-14937.

[47] Fathullah Y, Gales M J F. Self-distribution distillation: efficient uncertainty estimation//Uncertainty in Artificial Intelligence, Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, Eindhoven, The Netherlands, 2022:663-673.

[48] Xu J, Lin Z, Li S, et al. Flexible Bayesian Last Layer Models Using Implicit Priors and Diffusion Posterior Sampling. arXiv preprint arXiv:2408.03746, 2024.

[49] Lian Z, Lv M, Xu X, et al. Cloud-Edge Collaborative Continual Adaptation for ITS Object Detection// Proceedings of the International Conference on Spatial Data and Intelligence. Nanjing, China, 2024: 15-27.

[50] He Y, Shen W. FedITA: A cloud–edge collaboration framework for domain generalization-based federated fault diagnosis of machine-level industrial motors. Advanced Engineering Informatics, 2024, 62: 102853.

[51] He P, Jiao L, Li L, et al. Domain Generalization-Aware Uncertainty Introspective Learning for 3D Point Clouds Segmentation//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia, 2024: 651-660.

[52] Yin H, Wang P, Liu B, et al. An uncertainty-aware domain adaptive semantic segmentation framework. Autonomous Intelligent Systems, 2024, 4(1): 15.

[53] Cai M, Kezierbieke J, Zhong X, et al. Uncertainty-Aware and Class-Balanced Domain Adaptation for Object Detection in Driving Scenes. IEEE Transactions on Intelligent Transportation Systems, 2020,25(11):15977-15990

[54] Guo J, Chen H, Wang C, et al. Vision superalignment: Weak-to-strong generalization for vision foundation models. arXiv preprint arXiv:2402.03749, 2024.

[55] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. Technical Report, University of Toronto, Toronto, Canada, 2009

[56] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database// Proceedings of the 2009 IEEE conference on computer vision and pattern recognition. Miami, USA, 2009: 248-255.

[57] Socher R, Perelygin A, Wu J, et al. Recursive deep models for semantic compositionality over a sentiment treebank//Proceedings of the 2013 conference on empirical methods in natural language processing. Grand Hyatt Seattle, Seattle, USA, 2013: 1631-1642.

[58] Barbieri F, Camacho-Collados J, Neves L, et al. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. arXiv preprint arXiv:2010.12421, 2020.

[59] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City, USA, 2018: 4510-4520.

[60] Hochreiter S, urgen Schmidhuber J, Elvezia C. LONG SHORT-TERM MEMORY. Neural Computation, 1997, 9(8): 1735-1780.

[61] Liu Z, Lin W, Shi Y, et al. A robustly optimized BERT pre-training approach with post-training// Proceedings of the China national conference on Chinese computational linguistics. Hohhot, China, 2021: 471-484.

[62] Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., ... & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift//Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2019: 13969—13980.

[63] Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles// Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems. Long Beach, USA, 2017:6402--6413

LU Fei-Hong, Ph. D. candidate. His research interest is AI safety, Uncertainty Quantification and Multimodal cognitive intelligence.

LUO Yang-Yi-Fei, Master candidate. His research interest is Multimodal cognitive intelligence.

Gao Shi-Qi, Ph. D. candidate. His research interest is AI safety.

TAI Zhen-Ying, M.S.,Associate Researcher. His research interests include social network and big data.

ZHOU Hao-Yi, Ph. D., Assistant Professor. His research interests include long sequence forcasting etc.

Sun Qing-Yun, Ph. D., Assistant Professor. Her research interests include social network and big data.

Li Jian-Xin, Ph. D., Professor. His research interests include social network and big data.

Background

This paper focuses on the field of cloud-edge collaborative reasoning, with a particular emphasis on model uncertainty prediction and model selection in edge computing environments.

With the widespread use of pre-trained models in various downstream tasks, deploying efficient and accurate models on resource-constrained edge devices has become a critical research topic. Several solutions have already been proposed, such as model lightweighting techniques (including pruning, quantization, and model compression) and model selection methods based on Softmax output credibility. However, these approaches often face challenges related to robustness and accuracy, especially when dealing with complex or unevenly distributed samples.

This paper proposes a cloud-edge collaborative reasoning framework based on uncertainty calibration, aiming to overcome the limitation of traditional model selection relying on the Softmax output layer, especially when facing unknown domain input, even if the model confidence is insufficient, it may still output wrong predictions. By distilling and calibrating the cloud-side model to give it uncertainty prediction capabilities, the framework combines the uncertainty of cloud-side and edge-side model outputs to optimize the model selection and reasoning process, thereby achieving efficient cloud-edge collaborative reasoning.

In visual and text classification tasks, the proposed reasoning framework outperforms existing methods, particularly when handling samples with high cognitive uncertainty. In these cases, the benefits of cloud-edge collaboration are especially significant.

While the cloud-edge collaborative reasoning framework, based on uncertainty adaptation, is primarily designed for text and image classification tasks, it can be further expanded for regression and generation tasks in the future. This would allow the framework to become more flexible and intelligent in resource scheduling and model selection, enabling it to handle dynamic and complex application scenarios and providing more powerful and efficient support for practical applications.

This paper offers both a theoretical foundation and technical support for the next generation of intelligent edge computing systems by optimizing the configuration and collaboration of cloud-edge resources.

This work was funded by the National Natural Science Foundation of China Outstanding Young Scholars Fund "Intelligent Computing of Network Behavior Big Data" (62225202) and the National Natural Science Foundation of China Youth Science Fund Project "Research on Detection and Analysis Methods of Malicious Groups in Social Networks Based on Key Subgraphs" (62302023) and "Research on Long Sequence Prediction Methods for Industrial Big Data" (62202029).