文章编号:1001-9383(2025)02-0094-03

科研动态

大模型知识蒸馏方法研究进展

李 通,羊红光,刘 康,路 凯,刘 龙

(西安理工大学,陕西 西安 710048)

DOI:10.16191/j.cnki.hbkx.2025.02.004

深度神经网络(DNNs)在各类任务中取得了显著成就。然而,高性能深度神经网络模型往往包含大量的参数,在推理阶段存在巨大的计算开销。大模型知识蒸馏技术将大型、复杂模型(教师模型)的知识迁移到较小、高效模型(学生模型)中,显著降低了模型的计算和存储需求。DeepSeek 的成功让蒸馏技术愈加引人瞩目。OpenAI 基于 GPT-4o 蒸馏出的专用小模型 GPT-4o-mini 具有很好的效果。李飞飞团队基于 Qwen2.5 做数据蒸馏出的新模型 s1-32B,竟取得了与 OpenAI o1、DeepSeek R1 等模型相当的数学和编码能力。可见,知识蒸馏技术在大模型发展中承担着重要的角色,不仅提升了大模型的预测性能,推动大模型技术快速迭代,还带来更高效、更经济的应用解决方案,推动大模型技术广泛应用。

知识蒸馏方法的发展历经传统知识蒸馏、大语言模型知识蒸馏两个发展阶段。

1 传统知识蒸馏

知识蒸馏的核心在于如何设计、提取和迁移“知识”。HINTON 等人[1]提出了知识蒸馏的概念,并将教师模型的输出概率分布作为“暗知识”迁移到学生模型中。ROMERO 等人[2]提出了 FitNet 方法,将教师模型中间层的特征定义为“知识”以供学生网络学习,开创了基于中间层深度特征的知识蒸馏方法研究。传统知识蒸馏是基于这两种方法进行各种改进来提升模型性能。算法 AT[3]提出神经网络的注意力机制,将教师模型的注意力信息作为知识传递给学生网络。YIM 等人[4]认为给出网络解决问题的过程比直接给出最终结果更加有用,并将相邻特征图之间的相关性作为知识传递给学生。算法 DML[5]引入互学习方式同时训练多个学生网络,实现了无预训练教师模型条件下的知识蒸馏。算法 DKD[6]将原始 KD 方法中的蒸馏损失解耦为目标类别蒸馏和非目标类别蒸馏损失两类,使得知识传递更加高效。算法 CAT-KD[7]提出了基于类注意力迁移的蒸馏方法。这些方法主要面向卷积神经网络(CNNs),并显著提高了轻量级模型在图像分类任务中的准确性。

2 大语言模型知识蒸馏

近年来,大语言模型(LLMs)展现出了惊人的能力,从智能聊天到复杂的文本生成,从精确的图像识别到高效的数据分析,极大地改变了人们的生活和工作方式。大语言模型的成功也得益于知识蒸馏的应用,特别是 DeepSeek 通过知识蒸馏在模型性能优化、计算密度降低等方面做了令人惊叹的示范。一方面,DeepSeek 利用强大的教师模型生成、优化数据,以实现数据增强、伪标签生成和数据分布优化等功能。另一方面,DeepSeek 通过监督微调将教师模型的知识迁移到学生模型中,并结合数据蒸馏和模型蒸馏,实现了推理性能的显著提升。

LLMs 知识蒸馏的核心是将高参数量的 LLMs 的知识传递给低参数量的 LLMs 或小语言模型(SLM)。根据是否将 LLMs 的涌现能力(EA)作为知识进行迁移,LLMs 知识蒸馏可分为两大类:标准知识蒸馏(KD)和基于涌现能力的知识蒸馏(EA-based KD)。

标准知识蒸馏利用大型语言模型知识指导学生模型学习,看起来与传统的 KD 方法相似,但区别在于教师模型和学生模型均为 LLMs。这类方法的创新主要集中在对输出概率分布、中间层特征信息等方面的处理。算法 MINILLM[8]认为教师模型输出概率分布中远离类别的区间出现过高概率是由最小化 Kullback-Leibler (KL)散度导致的,将其替换为最小化 MINILLM 损失后,阻止了学生模型对教师分布中低概率区域的高估,提高了生成样本的质量。算法 GKD[9]通过优化逆向 KL 散度、采样输出序列分布实现了自回归模型蒸馏,解决了学生模型在部署阶段生成的输出序列与训练阶段的输出序列间的分布不匹配、学生模型无法匹配教师分布等关键问题。

"涌现能力"是指像 GPT-3(175B)和 PaLM(540B)等大型模型在处理复杂任务时展现出令人惊讶的能力。涌现能力包括三个方面:上下文学习(ICL)、思维链(CoT)和指令遵循(IF)。基于涌现能力的 KD 不仅仅蒸馏 LLMs 的常见知识,还包括蒸馏其涌现能力。ICL 采用结构化自然语言提示,包含任务描述和一些任务示例。通过这些任务示例,LLM 可以掌握并执行新任务,无需显式的梯度更新。HUANG 等人[10]提出了 ICL 蒸馏,通过元上下文调优(Meta-ICT)和多任务上下文调优(Multitask-ICT),将上下文小样本学习和语言建模功能从 LLM 转移到 SLM。在 Meta-ICT 中,语言模型使用上下文学习目标在不同任务中进行元训练,从而使其能够通过上下文学习适应未见过的任务,扩展其问题解决能力。Multitask-ICT 使用 ICL 目标和目标任务中的一些示例进行模型微调,并应用上下文学习进行预测。

CoT 摒弃简单的输入输出对,将中间推理步骤融入到提示中。算法 MT-CoT[11]利用 LLM 生成的解释来强化小模型训练,在多任务学习中赋予小模型更强的推理能力。FU 等人[12]发现大语言模型多维能力之间平衡,从大型教师模型中提取 CoT 推理路径,利用微调指令调优模型,增强了模型的分布外泛化能力。算法 DISCO[13]通过工程化提示生成短语扰动,通过特定任务的教师模型过滤这些扰动,提取到高质量的反事实数据。

指令遵循(IF)在不依赖少量样本下仅基于任务描述来增强大语言模型执行新任务的能力。通过一系列任务指令进行微调,语言模型就能获得未见过任务的识别能力。CHEN 等人[14]利用 LLM 生成"困难"指令来增强学生模型的能力,发挥了 LLM 适应性广泛的作用。

3　大语言模型知识蒸馏技术发展展望

尽管知识蒸馏在提高模型性能方面表现出色,但该技术仍面临一些挑战。首先,当教师模型与学生模型之间的性能差距较大时,蒸馏效果容易受到容量差距问题的影响,导致学生模型难以接近教师模型的性能。其次,多模态数据的复杂性和多样性增加了蒸馏过程的难度,使得在蒸馏过程变得更加复杂。再次,高性能蒸馏方法过分依赖于基于大语言模型的预训练教师,消耗的计算资源巨大。因此,知识传递高效、任务适应广泛、低计算资源消耗的知识蒸馏算法将是大模型知识蒸馏的发展方向。

参考文献:

[1] HINTON G, VINYALS O, DEAN J. Distilling the Knowledge in a Neural Network[EB/OL]. (2025-03-09)[2024-12-12]. http://arxiv.org/abs/1503.02531.

[2] ROMERO A, BALLAS N, KAHOU S E, et al. FitNets: Hints forthin deep nets[C]//3rd International Conference on Learning Representations. ICLR, 2015.

[3] ZAGORUYKO S, KOMODAKIS N. et al. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer[C]//5th International Conference on Learning Representations, ICLR 2017: 1-13.

[4] YIM J, JOO D, BAE J, et al. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA, 2017: 7130-7138.

[5] ZHANG Ying, XIANG Tao, HOSPEDALES T M, et al. Deep mutual learning[C]//Proceedings of the IEEE Computer

Society Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 4320-4328.
[6] ZHAO Borui, CUI Quan, SONG Renjie, et al. Decoupled Knowledge Distillation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, 2022: 11953-11962.
[7] GUO Ziyao, YAN Haoman, LI Hui, et al. Class attention transfer based knowledge distillation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2023: 11868-11877.
[8] GU Yuxian, DONG Li, WEI Furu, et al. MiniLLM: Knowledge distillation of large language models[EB/OL]. (2024-04-10)[2024-12-12]. https://arxiv.org/abs/2306.08543.
[9] AGARWAL R, VIEILLARD N, STANCZYK P, et al. Gkd: generalized knowledge distillation for auto-regressive sequence models[EB/OL]. (2024-01-17)[2024-12-12]. https://arxiv.org/abs/2306.13649v1.
[10] HUANG Yukun, CHEN Yanda, YU Zhou, et al. In-context learning distillation: Transferring few-shot learning ability of pre-trained language models[EB/OL]. (2022-12-20)[2024-12-12]. https://arxiv.org/abs/2212.10670.
[11] LI Shiyang, CHEN Jianshu, SHEN Yelong, et al. Explanations from large language models make small reasoners better [EB/OL]. (2022-10-13)[2024-12-12]. https://arxiv.org/abs/2210.06726.
[12] FU Yao, PENG Hao, OU Litu, et al. Specializing smaller language models towards multi-step reasoning[C]//Proceedings of the 40th International Conference on Machine Learning. Honolulu, Hawaii, USA: PMLR 202, 2023: 10421-10430.
[13] CHEN Zeming, GAO Qiyue, BOSSELUT A, et al. DISCO: Distilling Counterfactuals with Large Language Models [C]// 61st Annual Meeting of the Association for Computational Linguistics, ACL, 2023: 5514-5528.
[14] CHEN Hongzhan, QUAN Xiaojun, CHEN Hehong, et al. Knowledge distillation for Closed-Source language models [EB/OL]. (2024-01-13)[2024-12-12]. https://arxiv.org/html/2401.07013v1.

(上接第 84 页)

参考文献：

[1] 何莽，彭菲. 基于流动性与健康关系的康养旅游学体系建构[J]. 旅游学刊，2022，37(3)：13-15.
[2] 杨红英，杨舒然. 融合与跨界:康养旅游产业赋能模式研究[J]. 思想战线，2020，46(6)：158-168.
[3] 马东跃，马伊莎. 品牌创新感知对游客康养旅游行为的影响[J]. 商业经济研究，2022(10)：82-85.
[4] 黄天柱，张恒瑞. 农业农村部《关于拓展农业多种功能促进乡村产业高质量发展的指导意见》解读[J]. 农村实用技术，2022(1)：15-16.
[5] RIDDERSTAAT J. Measuring hidden demand and price behavior from US outbound health tourism spending[J]. Tourism Economics，2023，29(3)：103-126.
[6] 朱冬芳,钟林生,虞虎. 康养旅游研究的国内外对比与展望 [J]. 世界地理研究，2023，32 (11)：167-180.
[7] 王瑞,单莉莉. 乡村康养旅游的发展模式分析 [J]. 中国农业资源与区划，2024，45 (2)：27.
[8] 张广海，董跃蕾. 中国康养旅游政策演化态势及效果评估[J]. 资源开发与市场，2022，38(12)：1491-1496.
[9] 文平. 基于恢复性环境视角的乡村康养旅游发展研究 [J]. 农业经济，2022(5)：98-100.
[10] 李莉，陈雪钧. 康养旅游产业创新发展的影响因素研究[J]. 企业经济，2020，39(7)：116-122.
[11] 张贝尔,黄晓霞. 康养旅游产业适宜性评价指标体系构建及提升策略 [J]. 经济纵横，2020(3)：78-86.
[12] 黄玖琴. 政府作用、游客体验与乡村旅游发展绩效:以贵州省为例[J]. 社会科学家，2021(3)：64-69.
[13] 李莉，陈雪钧. 康养旅游产业创新发展的动力因素研究:基于共享经济视角[J]. 技术经济与管理研究，2021(4)：36-40.
[14] 向程，李环. 基于 AHP-Fuzzy 综合模型的乡村旅游资源评价研究:以四川省通江县为例[J]. 焦作大学学报，2020，34(1)：63-69.
[15] 向程，唐仲霞，李环. 乡村旅游核心利益相关者协调发展评价研究:以青海省海东市互助土族自治县小庄村为例[J]. 西部经济管理论坛，2020，31(6)：42-52，73.