大型语言模型的模型压缩综述

朱训予[1,2]，李健[1,2]*，刘勇[3]，可以马[1,2]，王卫平[1,2]

[1]中国科学院信息工程研究所

2中国科学院大学网络安全学院

3中国人民大学高岭人工智能学院

{zhuxunyu, lijian9026, macan, wangweiping}@iie.ac.cn, liuyonggsai@ruc.edu.cn

抽象的

大型语言模型 (LLM) 彻底改变了自然语言处理任务，取得了巨大的成功。然而，他们的强大

尺寸和计算需求显着
实际部署面临的挑战，特别是在
资源受限的环境。随着这些挑战变得越来越相关，

模型压缩已成为缓解这些限制的关键研究领域。这

论文提出了一项全面的调查，探讨了专为法学硕士量身定制的模型
压缩技术的前景。满足高效部署的迫切需求，

我们深入研究各种方法，包括量化、剪枝、知识蒸馏、

和更多。在每一种技术中，我们
突出有助于不断变化的格局的最新进展和创新方法

法学硕士研究。此外，我们探索了基准测试策略和评估指标

对于评估压缩法学硕士的有效性至关重要。通过提供最新的见解

这项调查对于研究人员和实践者来说都是宝贵的资源。随着法学硕士继续

这项调查旨在促进提高效率和现实世界的适用性，建立一个

为该领域未来的发展奠定了基础。

1 简介

大型语言模型 (LLM) [Zhao et al., 2023;黄
和张，2023； Chang et al., 2023]始终表现出
在各项任务中表现出色。尽管如此，它们的卓越能力也伴随着巨大的挑战，这些挑战源于其庞大的规模和计算能力。

要求。例如，GPT-175B 模型 [Brown 等人
al., 2020]，具有令人印象深刻的 1750 亿个参数，需要至少 320GB（使用 1024 的倍数）
以半精度 (FP16) 格式存储。此外，部署该模型进行推理至少需要五个

A100 GPU，每个具有 80GB 内存，可高效
管理运营。为了解决这些问题，一种流行的方法被称为模型压缩 [Deng et al., 2020;
He et al., 2018] 提供了一个解决方案。模型压缩涉及将大型资源密集型模型转换为
紧凑型版本适合在受限的移动设备上存储
设备。此外，它还可能涉及优化模型
以最小的延迟更快地执行或实现平衡
这些目标之间。

除了技术方面之外，法学硕士还引发了
关于环境和道德问题的讨论。这些
模型给发展中国家的工程师和研究人员带来了重大挑战，因为这些国家的资源有限
阻碍访问模型执行所需的基本硬件 [Lin
等人，2023]。此外，法学硕士的大量能源消耗也会导致碳排放，这凸显了

人工智能研究中可持续实践的重要性。 A
这些挑战的有希望的解决方案在于利用模型
压缩技术，该技术展示了在不显着影响排放的情况下减少排放的潜力

性能 [Luccioni 等人，2022]。通过实施模型
压缩，我们可以解决环境问题，增强
可访问性，并促进法学硕士部署的包容性。

在我们的论文中，我们的主要目标是阐明专为法学硕士量身定制的模型压缩技术领域的最新进展。我们的工作需要对方法、指标和基准进行详尽的调查，

我们精心地将其组织成一个创新的分类法。
如图 1 所示，我们提出的分类法提供了
用于理解景观的结构化框架
LLM 的模型压缩方法。这种探索包括对成熟技术的彻底检查，包括但不限于修剪、知识蒸馏、量化和低阶分解。此外，我们的研究揭示了当前的挑战，并让我们一睹该领域未来潜在的研究轨迹。

不断发展的领域。我们提倡内部协作努力
社区为具有生态意识的人铺平道路，
法学硕士的包罗万象、可持续的未来。尤其，
我们的工作是专门针对法学硕士模型压缩领域的首次调查。

*通讯作者。

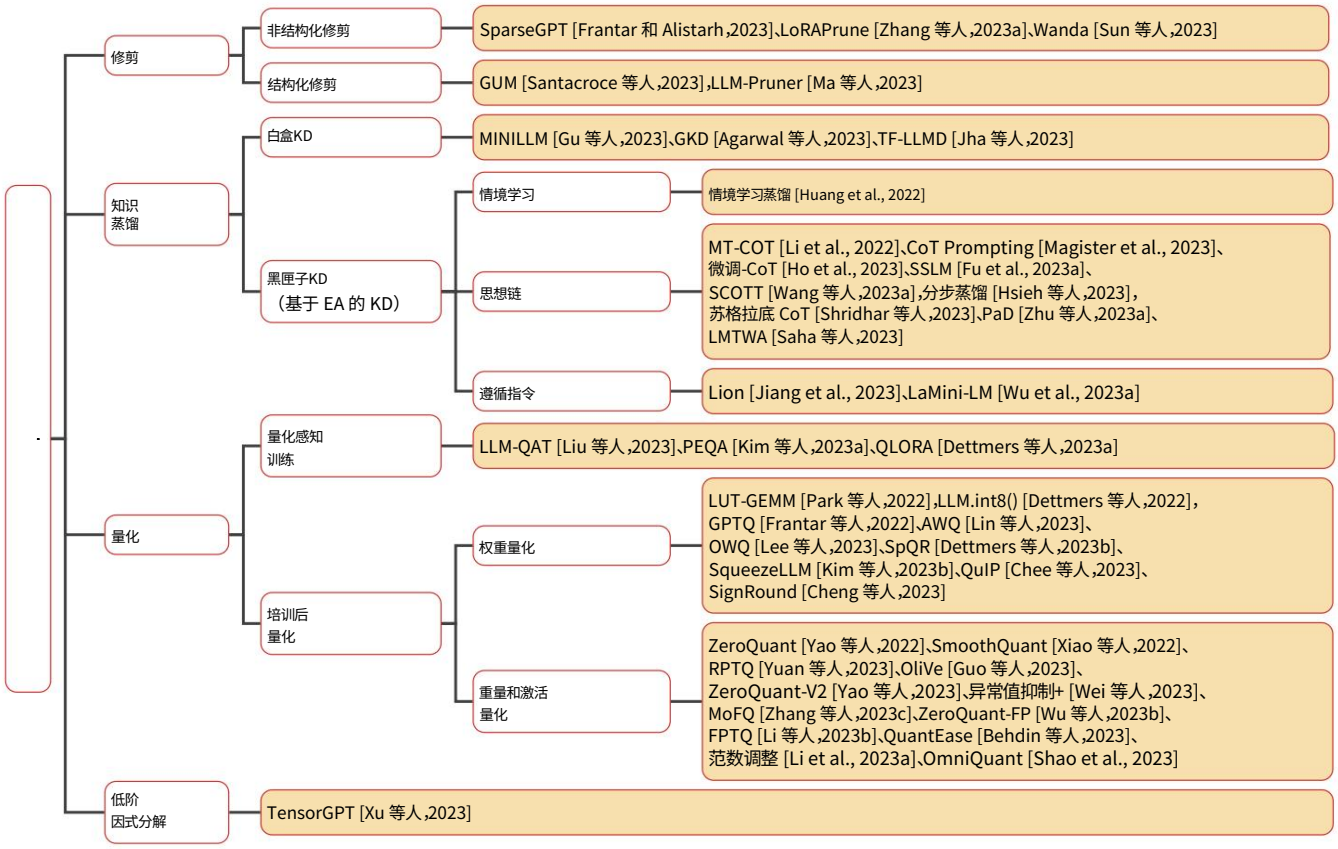

图 1:大型语言模型的模型压缩方法的分类。

2 方法

2.1 修剪

剪枝是一种强大的技术,可以通过删除不必要或冗余的内容来减小模型的大小或复杂性
组件 [LeCun 等人,1989;韩等人,2015;李等人,
2017]。众所周知,有很多冗余参数
对性能影响很小甚至没有影响
模型,因此,模型的性能将最少
直接剪掉这些冗余参数后就下降了。在
同时,剪枝可以使模型存储友好[Ar-dakani et al., 2019]、内存效率高[Han et al., 2015;
Yang et al., 2017],计算效率[Li et al., 2017]。
剪枝可以分为非结构化剪枝[Zhang
等,2018； Gordon 等人,2020 年] 和结构化剪枝[Anwar 等人,2017 年； Fang 等人,2023]。结构化剪枝和非结构化剪枝的主要区别

在于修剪目标和最终的网络结构。结构化剪枝根据特定规则删除连接或层次结构,同时保留

整体网络结构。另一方面,非结构化
修剪会修剪各个参数,从而产生不规则的稀疏结构。最近的研究工作致力于将法学硕士与修剪技术相结合,旨在

解决与法学硕士相关的巨大规模和计算成本。在本节中,我们系统地分类

这些工作基于他们是否采用结构化或非结构化

结构化修剪策略。

非结构化修剪

非结构化修剪通过删除特定参数而不考虑其内部结构来简化法学硕士。

这种方法针对的是个体权重或神经元
LLM,通常通过应用阈值将低于其的参数归零。然而,这种方法忽略了整体
LLM结构,导致不规则的稀疏模型组成。这种不规则性需要专门的压缩技术来有效存储和计算

修剪后的模型。非结构化修剪通常涉及法学硕士的大量再培训以重新获得准确性,这对于法学硕士来说尤其昂贵。这方面的创新方法

域是 SparseGPT [Frantar 和 Alistarh,2023]。它引入了一种不需要重新训练的一次性修剪策略。该方法将剪枝视为广泛的稀疏回归问题,并使用近似稀疏回归来解决它

回归求解器。 SparseGPT 实现了显着的非结构化稀疏性,在最大的 GPT 模型上甚至高达 60%
像 OPT-175B 和 BLOOM-176B 一样,增加最小
困惑。与此相反,Syed 等人。提出迭代
在修剪过程中微调模型的修剪技术
以最少的训练步骤。另一个进步是 Lo-RAPrune [Zhang et al., 2023a],它将参数高效调整 (PEFT) 方法与剪枝相结合,以增强下游任务的性能。它引入了一种独特的pa-

使用值和梯度的参数重要性标准
低秩适应（LoRA)[Hu et al., 2022]。作为回应
仍然需要资源密集型权重更新过程
SparseGPT,Wanda [Sun et al., 2023] 提出了一种新的剪枝
公制。万达根据产品评估每个重量
其大小和相应输入激活的范数,使用小型校准数据集进行近似。这

度量用于线性层内的局部比较
输出,从而能够去除较低优先级的权重
法学硕士。

结构化修剪

结构化修剪通过删除整个 LLM 来简化 LLM
结构组件,例如神经元、通道或层。
这种方法同时针对整组权重,提供
具有降低模型复杂性和内存使用量,同时保持整体 LLM 结构完整的优点。到

探索结构化剪枝方法在法学硕士、GUM 中的应用和功效 [Santacroce et al., 2023]
分析了 NLG 任务上仅解码 LLM 的几种结构化剪枝方法,并发现建立了

结构化剪枝方法没有考虑神经元的独特性,留下了过多的冗余。

为了解决这个问题,GUM 引入了概念验证
方法通过基于网络组件的全局移动和修剪来最大化敏感性和独特性

局部独特性得分。 LLM-Pruner [Ma et al., 2023] 需要
一种压缩法学硕士的通用方法,同时保护其多任务解决和语言生成能力。 LLM-Pruner 还解决了以下问题带来的挑战：

用于法学硕士的大量培训数据,可以导致
重要的数据传输和训练后模型大小。
为了克服这些挑战,LLM-Pruner 结合了
依赖性检测算法可查明相互依赖性
模型内的结构。它还实现了一个高效的
考虑一阶的重要性估计方法
信息和近似的 Hessian 信息。这
策略有助于选择最佳修剪组,从而
改进压缩过程。

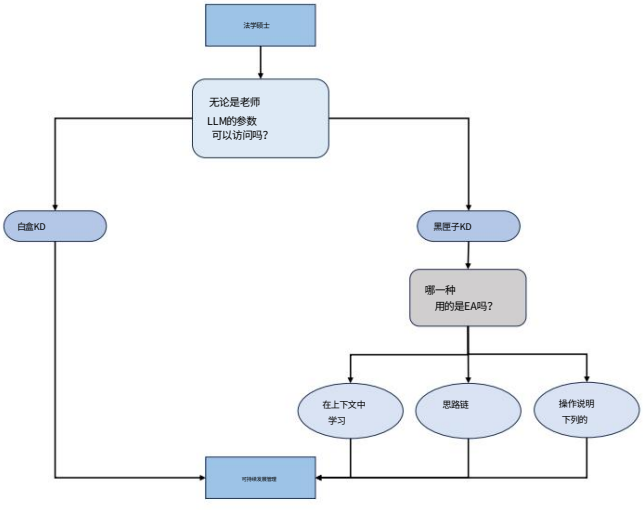

图 2:法学硕士知识蒸馏的简要分类。

2.2 知识蒸馏

知识蒸馏（KD)[Hinton 等人,2015;金和
拉什,2016； Tung 和 Mori,2019] 是一台有价值的机器
旨在提高模型性能的学习技术
和概括。它通过从复杂模型（称为教师模型)转移知识来实现这一目标，

到一个更简单的对应模型,称为学生模型。这
KD背后的核心思想是将教师模型的综合知识转化为更精简、更有效的表示。在
本节中,我们提供

以法学硕士为教师的蒸馏方法概述。我们将这些方法分为两个不同的组：

黑盒 KD,其中只有老师的预测
以及白盒 KD,其中教师的参数可供使用。对于视觉表示，

图 2 提供了法学硕士知识蒸馏的简要分类。

白盒KD

在White-box KD中,不仅仅是老师LLM的预测
可以访问,但可以访问和利用教师的法学硕士学位
参数也是允许的。该方法使学生
LM能够更深入地了解教师LLM的内部结构和知识表示,通常会带来更高水平的
绩效提升。白盒

KD 通常用于帮助较小的学生 LM 学习
并复制更大、更多的知识和能力
强大的法学硕士教师 [Gou et al., 2021;帕克等人,2019；
赵等人,2022;刘等人,2021a]。一个说明性的例子是 MINILLM [Gu et al., 2023],它深入研究了白盒生成法学硕士的蒸馏。它观察到最小化前向 Kullback-Leibler 散度的挑战

(KLD) - 这可能会导致不太可能发生的概率过高
教师分布的区域,在自由运行生成过程中导致不可能的样本。为了解决这个问题,MINILLM
选择最小化反向 KLD。这种方法可以防止
来自高估低概率区域的学生
教师的分布,从而提高生成样本的质量。相比之下,GKD [Agarwal et al., 2023] 探索了自回归模型的蒸馏,其中白盒生成 LLM 是一个子集。该方法识别两个

关键问题:输出序列之间的分布不匹配
训练期间和学生在部署期间生成的那些,以及模型规格不足,其中学生

模型可能缺乏与老师相匹配的表达能力
分配。 GKD 通过在训练期间对学生的输出序列进行采样来处理分布不匹配。它
还通过优化反向 KL 等替代散度来解决模型规格不足的问题。实现任务无关

法学硕士的零样本评估蒸馏,无需访问
最终任务微调数据,TF-LLMD [Jha et al., 2023] 使用
带有来自较大模型的层子集的截断模型
用于初始化,并使用语言建模目标在预训练数据上训练模型。

黑匣子KD

在Black-box KD中,只有老师做出的预测
LLM 是可以访问的。最近,黑匣子KD展现了
在 LLM API 生成的提示响应对上微调小模型方面取得了有希望的结果 [Li et al., 2022;

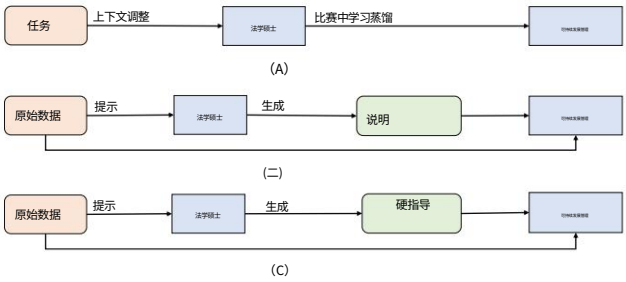

图 3:黑盒 KD（基于 EA 的 KD)概述。 (a) 情境学习蒸馏,(b) 思想链蒸馏,(c)

蒸馏后的说明。

何等人,2023； Hsieh 等人,2023]。同时，
最近的研究 [Wei 等人,2022a;谢弗等人,2023；
赵等人,2023]强调,重点是
由于扩大了模型尺寸,与 BERT（330M 参数)和 GPT-2（1.5B 参数)等较小模型相比,GPT-3（175B 参数)和 PaLM（540B 参数)等 LLM 展示了独特的行为。这些法学硕士在处理复杂的任务时表现出令人惊讶的能力,称为紧急能力。涌现能力包含几个有趣的方面,包括情境学习 (ICL) [Dong et al., 2023; Wang 等人,2023b],思想链 (CoT) [Wei 等人,2022b； Wang 等人,2023c；

Shi et al., 2023],以及指令遵循 (IF) [Ouyang et al., 2023]
等,2022;布鲁克斯等人,2023]。在我们的论文中,我们根据紧急情况对黑盒 KD 方法进行了进一步分类。
能力被利用。因此,我们也将黑盒 KD 称为
基于 EA 的 KD。有关视觉概述,请参阅图 3,其中
提供了基于 EA 的知识蒸馏概念的简明表示。

ICL 采用结构化自然语言提示
包含任务描述和可能的一些任务示例
作为示威活动。通过这些任务示例,法学硕士可以
无需明确的指导即可掌握并执行新任务
梯度更新。黄等人的作品。介绍
ICL 蒸馏,将上下文中的小样本学习和语言建模功能从 LLM 转移到 SLM。

这是通过将上下文学习目标与传统语言建模目标相结合来实现的。到

为了实现这一目标,他们探索了两次少量注射下的 ICL 蒸馏
学习范式:元上下文调优（Meta-ICT)和
多任务上下文调优（多任务-ICT）。在元信息通信技术中，
语言模型经过不同的元训练
使用情境学习目标的任务。这使它能够
通过情境学习适应看不见的任务,从而扩展其解决问题的能力。另一方面，

多任务-ICT 使用 ICL 目标对模型进行微调
目标任务中的一些示例。随后,它
采用上下文学习来对这些进行预测
任务。比较这两种范式,多任务-ICT 展示
优于 Meta-ICT 的性能。然而,它在任务适应过程中确实需要更多的计算资源,使其计算更加密集。

与 ICL 相比,CoT 采用了不同的方法,它结合了中间推理步骤,这可以导致
最终输出进入提示,而不是使用简单的输入输出对。 MT-COT [Li et al., 2022] 旨在利用
法学硕士为加强培训而做出的解释
较小的推理机。它利用多任务学习框架使较小的模型具有强大的推理能力以及生成解释的能力。钴酸甲酯

Prompting [Magister et al., 2023] 探讨了可迁移性
通过知识蒸馏将这种推理能力应用于较小的模型,并发现这是之间的权衡

模型和数据集大小对推理能力的影响。 Fine-tune-CoT [Ho et al., 2023]
更进一步,通过随机抽样从 LLM 生成多个推理解决方案。训练数据的增强有助于学生模型的学习过程。 SSLM [Fu et al., 2023a] 确定了多维能力之间的权衡

语言模型并提出微调指令调整模型。他们从大量的数据中提取 CoT 推理路径
改进分布外泛化的教师模型。
Distilling Step-by-Step [Hsieh et al., 2023] 采用 LLM 原理作为训练较小模型的额外指导
在多任务框架内。苏格拉底 CoT [Shridhar 等
al., 2023] 训练两个精炼模型:问题分解器
和一个子问题求解器。分解器分解一个
将原始问题分解为一系列子问题,而
子问题求解器处理解决这些子问题。为了
基本原理忠实性,SCOTT [Wang et al., 2023a] 采用
对比解码,将每个基本原理与答案联系起来。它鼓励老师提出相关的理由。此外,引导学生参与反事实的活动

基于导致不同答案的基本原理的推理和预测。在 PaD [Zhu et al., 2023a] 中,学生模型
通过程序辅助推理得到加强并得到辅助
通过自动错误检查克服错误的推理步骤。 LMTWA [Saha et al., 2023] 探索了使用
法学硕士作为教师以提高弱势群体的表现
通过自然语言解释代理。具体来说,它
引入了学生-教师框架,调查何时
以及老师应该如何进行解释。这
研究提出了一种个性化的心理理论方法
和预算意识的教学,展示了长期
教师解释的影响,并警告教师故意不一致的潜在负面影响

误导学生。
IF 致力于仅基于阅读任务描述来增强语言模型执行新任务的能力,而不依赖于少数样本。通过使用一系列以指令表示的任务进行微调,语言模型展示了准确地

执行以前未见过的指令中描述的任务。为了
例如,Lion [Jiang et al., 2023] 利用适应性
法学硕士的本质是提高学生模型的表现。它
提示法学硕士识别并生成“硬”指令,然后利用这些指令来增强学生模型的

能力。这种方法利用了法学硕士的多功能性
指导学生模型的学习以解决复杂的问题
指示和任务。 LaMini-LM [Wu et al., 2023a] 解决了资源密集型语言模型的挑战，
需要大量的计算能力和内存，
使许多研究人员和开发人员无法接触到它们

呃。为了解决这个问题,LaMini-LM 开发了一个包含 258 万条指令的广泛集合,包括
现有的和新生成的指令。这些说明
用于微调各种模型,提供
这一问题的有效解决方案。

2.3 量化

在模型压缩领域,量化已经
成为一种广泛接受的技术来缓解
深度学习模型的存储和计算开销 [Liu et al., 2021b;古拉米等人,2022;郭等人,
2020]。虽然传统表示采用浮点数,但量化将它们转换为整数或

其他离散形式。这种转换显着降低了存储需求和计算复杂性。

尽管一些精度损失是固有的,但仔细的量化技术可以实现实质性的模型压缩
精度下降极小。量化可以
可分为三种主要方法： 量化感知训练 (QAT) [Tailor et al., 2021; Kim 等人,2022;
Ding et al., 2022],以及训练后量化（PTQ)[Liu
等人,2021b;内格尔等人,2020;方等人,2020]。这些方法之间的主要区别在于何时应用量化来压缩模型。 QAT 员工

PTQ 在模型训练/微调过程中进行量化,PTQ 在模型完成后对其进行量化
训练。最近的研究工作利用量化来压缩法学硕士,取得了令人印象深刻的成果。这些
努力分为上述两种方法：
量化感知训练和训练后量化。此外,表 1 总结了应用于 LLM 的量化方法的参考。桌子

将这些工作分为 8 位量化和低位量化
量化,基于位数（精度）
LLM的权重。

量化感知训练

在QAT中,量化目标是无缝集成的
进入模型的训练过程。这种方法使得
LLM在训练过程中适应低精度表示,增强其处理精度损失的能力

通过量化。这种调整旨在保持更高的
即使在量化过程之后的性能。例如,LLM-QAT [Liu et al., 2023] 深入研究了获取 LLM 培训数据的挑战。鉴于收集 LLM 培训数据的要求可能很高,LLM-QAT

提出了一个创新的解决方案。它利用预训练模型生成的世代来实现无数据蒸馏。这种方法极大地有助于规避

数据收集挑战。此外,LLM-QAT 还
不仅量化权重和激活,更进一步
还有键值 (KV) 缓存。该策略旨在提高吞吐量并支持更长的序列依赖性。 LLM-QAT 的一个值得注意的成就是它的能力

提取具有量化权重的大型 LLaMA 模型
KV 缓存降至仅 4 位。这一突破性的结果证明了产生精确 4 位的可行性

量化法学硕士。另一方面,PEQA [Kim 等人,
2023a] 和 QLORA [Dettmers et al., 2023a] 均属于

量化感知参数高效微调 (PEFT) 技术的类别 [Liu et al., 2022;丁等人,2023;
Fu 等人,2023b]。这些技术的重点是促进
模型压缩和加速推理。 PEQA 采用双阶段流程。在第一阶段,每个全连接层的参数矩阵被量化为低位整数矩阵和标量向量。在第二

阶段,对每个特定下游任务的标量向量进行微调。 QLORA 引入了创新概念,例如新数据类型、双量化和分页

优化器。这些想法旨在节省内存
不影响性能。 QLORA 使大型
模型在单个 GPU 上进行微调,同时在 Vicuna 基准测试中获得最先进的结果 [Chiang
等人,2023]。

训练后量化

PTQ 涉及量化 LLM 的参数
完成法学硕士的培训阶段。 PTQ 的主要目标是减少存储和计算

法学硕士的复杂性,所有这些都不需要修改法学硕士架构或需要重新培训过程。 PTQ 的主要优势是简单和高效

实现模型压缩。然而,重要的是
请注意,PTQ 可能会带来一定程度的精度损失
由于量化过程。该方法用作
提高 LLM 效率的直接方法
无需进行重大改动或进行大量培训。
在 PTQ 中,某些方法只关注量化
法学硕士的权重可以提高效率并减少计算需求。具体来说,LUT-GEMM [Park 等人,
2022]使用以下方法优化法学硕士内的矩阵乘法
仅权重量化和 BCQ 格式 [Rastegari et
al., 2016],通过以下方式增强延迟减少和性能
提高计算效率。 LLM.int8() [Dettmers 等人
al., 2022] 在 LLM 变换器中采用 8 位量化进行矩阵乘法,有效地将 GPU 内存减半
在推理过程中使用,同时保持性能精度。该方法采用矢量量化

混合精度分解来处理异常值以进行有效的推理。值得注意的是,LLM.int8() 可以进行推理
在具有多达 1750 亿个参数的模型中,性能不会受到影响。 GPTQ [Frantar et al., 2022] 承认上述方法对于低

压缩目标如 8 位权重,但面临挑战
以更高的速度保持准确性。为了解决这些挑战,GPTQ 提出了一种基于近似二阶信息的新型分层量化技术。这

结果是将位宽减少到每个权重 3 或 4 位,其中
与未压缩版本相比,精度损失最小。 Dettmers 和 Zettlemoyer 通过分析推理缩放定律,深入研究了法学硕士中关于零样本性能的模型大小和位精度之间的权衡。他们的

各种 LLM 系列的广泛实验表明,4 位精度几乎普遍适用于

在模型总位数和零样本精度之间实现适当的平衡。 AWQ [Lin et al., 2023] 发现权重
对于法学硕士的表现来说并不同样重要,仅保护 1% 的显着权重就可以大大减少质量

精确	方法
8位量化	LUT-GEMM [Park 等人,2022]、LLM.int8() [Dettmers 等人,2022]、ZeroQuant [Yao 等人,2022]、SmoothQuant [Xiao 等人,2022]
低位量化	LLM-QAT [Liu 等人,2023]、PEQA [Kim 等人,2023a]、QLORA [Dettmers 等人,2023a]、GPTQ [Frantar 等人,2022]、AWQ [Lin 等人,2023]、SpQR [Dettmers 等人,2023b]、RPTQ [Yuan et al., 2023]、OliVe [Guo et al., 2023]、异常值抑制+ [Wei et al., 2023]、OWQ [Lee 等人,2023]、ZeroQuant-FP [Wu 等人,2023b]、ZeroQuant-V2 [Yao 等人,2023]、SqueezeLLM [Kim 等人,2023b]、QuIP [Chee 等人,2023]、FPTQ [Li 等人,2023b]、QuantEase [Behdin 等人,2023]、Norm Tweaking [Li 等人,2023a]、SignRound [Cheng 等人,2023]、OmniQuant [Shao 等人,2023]

表 1:LLM 量化方法总结。我们根据
LLM 权重中的位数（即精度）。

化错误。基于这一见解,AWQ 采用了
通过考虑重要性的激活感知方法
权重通道对应于较大的激活幅度,在处理生命特征中起着关键作用。

该方法结合了每通道缩放技术
确定最小化量化的最佳缩放因子
量化所有权重时的错误。 OWQ [Lee 等人,2023]
对激活异常值如何进行理论分析
可以放大权重量化的误差。从此分析中汲取见解,OWQ 引入了混合精度

量化方案,该方案将更高的精度应用于
权重容易受到激活异常值引起的量化的影响。将精确的 LLM 进一步压缩至每
3-4 位
参数同时保持接近无损,SpQR [Dettmers et
al., 2023b] 识别并隔离异常值权重,存储
它们以更高的精度,并压缩所有其他权重
到 3-4 位。 SqueezeLLM [Kim et al., 2023b] 结合了基于灵敏度的非均匀量
化和密集和稀疏分解,以实现无损压缩

高达 3 位的超低精度。具体来说,基于灵敏度的非均匀量化搜索最佳比特

基于二阶信息的精度分配,以及
密集和稀疏分解存储异常值和敏感值
有效稀疏格式的权重值。动机是
量子化受益于不相干的权重和
Hessian 矩阵,QuIP [Chee et al., 2023] 利用自适应舍入过程来最小化二次代
理
客观高效的预处理和后处理,确保
通过随机乘法实现权重和 Hessian 不相干
正交矩阵实现 LLM 的 2 位量化。到
为了以较低位实现高精度,同时保持成本效益,Norm Tweaking [Li et al.,
2023a] 涉及校正量化激活分布以匹配其浮点数

对应的,这有助于恢复法学硕士的准确性。它包括校准数据生成和通道距离

更新归一化层权重的约束
更好的概括。为了提高仅权重量化的准确性,SignRound [Cheng et al.,

2023]引入了一种轻量级的逐块调整方法
使用有符号梯度下降。

除了上述仅量化权重的工作之外
LLM,PTQ 中的许多工作都试图量化这两个权重
和法学硕士的激活。 ZeroQuant [Yao 等人,2022]
集成了硬件友好的量化方案,层
逐层知识蒸馏和优化量化支持,以降低权重和激活精度

基于 Transformer 的模型到 INT8,精度最低
影响。 SmoothQuant [Xiao et al., 2022] 解决了量化激活的挑战,由于异常
值的存在,量化激活通常更加复杂。观察到不同的代币在其渠道中表现出类似的
变化，

SmoothQuant 引入了每通道缩放变换,可以有效地平滑幅度,渲染

模型更适合量化。通过对各种量化方案进行系统检查,模型

系列和量化位精度,ZeroQuant-V2 [Yao
et al., 2023]发现1)激活量化通常更容易受到权重量化的影响,较小的

模型在激活量化方面通常优于较大的模型,2)当前的方法都不能

使用以下任一方法实现量化的原始模型质量
INT4 权重或 INT4 权重和 INT8 激活。解决
针对这些问题,ZeroQuant-V2 引入了一种称为
低秩补偿（LoRC）,采用低秩
量化误差矩阵上的矩阵分解,以最小的增加来增强模型质量恢复

模型尺寸。认识到法学硕士中量化激活的复杂性,RPTQ [Yuan et al., 2023] 揭
示了这一点
不同地区范围不均匀所带来的挑战
不同的渠道,此外还存在异常值。
为了解决这个问题,RPTQ 战略性地将渠道安排到
聚类进行量化,有效地减轻通道范围的差异。此外,它将通道重新排序集成到层范
数操作和线性层中

权重以最小化相关开销。 OliVe [Guo 等人，
2023]进一步采用离群值-受害者对（OVP)量化并以低硬件本地处理离群值

开销和高性能收益,因为它发现
异常值很重要,而正常值则紧随其后
不是。异常值抑制+ [Wei et al., 2023] 扩展了这一点
通过确认激活中的有害异常值表现出不对称分布来理解,主要是

专注于特定渠道,并推出新颖的
涉及通道方式移位和缩放操作的策略,以纠正异常值的不对称呈现和

减轻有问题的渠道的影响,并定量分析移动和缩放的最佳值，

考虑到输出的不对称性质

层数和由权重引起的量化误差
接下来的几层。 ZeroQuant-FP [Wu et al., 2023b] 探索
浮点 (FP) 量化的适用性,特别关注 FP8 和 FP4 格式。研究揭示

对于法学硕士来说,FP8 激活始终优于其
FP4 与 INT4 相对应,而在权重量化方面,FP4 表现出与 INT4 相当（即使不是更优)的性能。为了解决权重和激活之间的差异所带来的挑战，

ZeroQuant-FP 要求所有比例因子均为幂
2 并将缩放因子限制在单个计算组内。值得注意的是,ZeroQuant-FP 还集成了低
等级补偿（LoRC)策略进一步增强
其量化方法的有效性。 FPTQ [Li 等人，
2023b]结合了n两种配方的优点W8A8
和 W4A16 为可用的开源 LLM 设计了一种新颖的 W4A8 训练后量化方法,并将细粒度权重量化与分层激活量化策略相结合,该策略针对大多数棘手的层提供了新颖的对数麦克风均衡消除

进一步微调的必要性。 QuantEase [Behdin 等人，
2023],是一个创新的分层量化框架
这涉及到个体的不同量化过程
层。核心挑战是离散结构的
非凸优化问题,通过应用坐标下降（CD)技术得出有效的解决方案。
OmniQuant [Shao et al., 2023] 由两个突破性的组件组成：

裁剪 (LWC) 和可学习的等效变换
（让）。这些组件旨在微调范围
有效地量化参数。 OmniQuant 运营
在可微的框架内,采用逐块误差最小化,并且擅长在各种量化配置中实现令人印象深刻的性能。

2.4 低阶因式分解

低阶分解 [Cheng et al., 2017;波维等人，
2018； Idelbayev 和 Carreira-Perpin~an,2020 年压缩技] 是一个模型
术,旨在逼近给定的
权重矩阵,将其分解为两个或多个较小的
维数显着降低的矩阵。核心理念
低阶因式分解背后涉及寻找因式分解
将一个大的权重矩阵 W 分解为两个矩阵 U 和 V,这样
$W \approx UV$ 其中 U 是 $m \times k$ 矩阵,V 是
$k \times n$ 矩阵,其中 k 远小于 m 和 n。这
U和V的乘积近似于原始权重矩阵,从而大幅减少参数数量和计算开销。在LLM领域

研究中,低阶分解已被广泛采用
有效地微调法学硕士,例如 LORA [Hu et al., 2022]
及其变体 [Valipour 等人,2023;张等人,2023b;
Chavan 等人,2023]。与以上作品不同的是，
我们重点关注这些使用低阶分解的工作
压缩法学硕士。 TensorGPT [Xu et al., 2023] 存储大量
以低阶张量格式嵌入,减少空间
法学硕士的复杂性并使其可在边缘设备上使用。具体来说,TensorGPT 有效地压缩了
使用张量训练分解（TTD)在 LLM 中嵌入层。通过将每个令牌嵌入视为一个矩阵

Product State (MPS),嵌入层可压缩
高达 38.40 倍,同时仍保持或
与之前相比,甚至还提高了模型的性能
原来的法学硕士。

3 指标和基准

3.1 指标

法学硕士的推理效率可以使用各种指标来衡量,这些指标捕获性能的不同方面。

这些指标通常与准确性一起呈现
和零样本能力来综合评估LLM。

参数数量

参数数量 [Ma et al., 2023;达斯古普塔等人，
2023]在法学硕士中指的是可学习权重的总数
或法学硕士在培训期间需要优化的变量。
在法学硕士中,参数代表连接中的权重
神经元或注意力层之间。一般来说,法学硕士的参数越多,它的表现力就越强,但它也
两者都需要更多的计算资源和内存
训练和推理。

型号尺寸

模型尺寸 [Shridhar et al., 2023;李等人,2022； Magist-ter et al., 2023] 通常指磁盘空间或内存
存储整个 LLM 所需的空间,包括权重，
偏见和其他必要的组成部分。模型尺寸为
与参数的数量密切相关,因为更多的参数通常会导致更大的模型尺寸。然而,其他因素,例如用于表示参数的数据类型和

模型架构,也会影响整体尺寸。

压缩率

压缩比 [Frantar 和 Alistarh,2023;陶等人，
2023]表示未压缩的LLM的原始大小与压缩的LLM的大小之间的比率。 A

更高的压缩比表明更有效的压缩,因为 LLM 的大小已显着减小,而

保留其功能和性能。

推理时间

推理时间（即延迟)[Kurtic et al., 2023;弗兰塔尔等
al., 2022]测量法学硕士处理所花费的时间
并在推理或预测期间生成输入数据的响应。推理时间对于现实世界尤其重要

法学硕士需要响应用户查询的应用程序
或实时处理大量数据。

浮点运算 (FLOP)

失败 [Dettmers 和 Zettlemoyer,2022;袁等人,2023;
Wei et al., 2023] 测量涉及浮点数（通常是 32 位或

16 位）,LLM 在处理输入数据时执行。
FLOP 提供了一种有用的方法来估计计算量
法学硕士的要求并比较不同法学硕士或压缩技术的效率。

3.2 基准和数据集

这些基准测试和数据集的主要目标是衡量压缩的有效性、效率和准确性
法学硕士与未压缩的法学硕士相比。
这些基准和数据集通常包含不同的
涵盖一系列自然语言处理挑战的任务和数据集。

通用基准和数据集

大多数研究评估压缩法学硕士
完善的 NLP 基准和数据集。例如,GLUE [Wang et al., 2019b] 和 SuperGLUE [Wang
et al., 2019a] 旨在评估性能
各种自然语言理解（NLU)任务的语言模型。 LAMBADA [Paperno et al., 2016] 是
旨在评估上下文相关的理解
语言模型。 LAMA [Petroni 等人,2019] 和Strate-gyQA [Geva 等人,2021] 都旨在评估
语言模型的推理能力。小队 [拉杰普尔卡
et al., 2016] 是为机器阅读理解而设计的
（MRC)任务。

大长凳

BIG-Bench (BBH) [Srivastava et al., 2022] 是一个基准
专为 LM 设计的套件,涵盖 200 多个 NLP 任务,例如，
文本理解任务、推理任务、数学
推理任务。 BBH 的目的是评估 LM 在这些各种复杂任务中的表现。这

压缩法学硕士使用 BBH 来衡量现实世界任务的一般能力。这种方法提供了模型性能和效率的多维视角。 BBH 促进富有洞察力的评估和方法评估。

看不见的指令数据集

未见过的指令数据集的目的是用来评估
法学硕士在面对任意任务时的表现。那里
是两个著名的数据集,即 Vicuna-Instructions [Chi-ang et al., 2023] 和 User-Oriented-Instructions [Wang et al., 2023]
2023d]。 Vicuna-Instructions 数据集由 GPT-4 生成,包含 80 个旨在挑战的复杂问题
基线模型。它涵盖九个不同的类别,包括通用的、基于知识的、角色扮演的、

常识、费米、反事实、编码、数学、
和写作任务。面向用户的指令数据集是
精心策划的合集,包含 252 条说明。
该数据集的灵感来自 71 个面向用户的应用程序,包括 Grammarly、StackOverflow、Overleaf,而不是
而不是以广泛研究的 NLP 任务为中心。
这些数据集旨在衡量压缩法学硕士在面对看不见的指令时的表现,以审查他们处理和执行任意任务的能力。

4 挑战和未来方向

更高级的方法

法学硕士模型压缩技术研究
仍处于早期阶段。这些压缩的法学硕士,如
先前的研究已证明 [Frantar 和 Alistarh,2023；
刘等人,2023； Ho et al., 2023],与未压缩的相比,继续表现出显着的性能差距

同行。通过深入研究专为法学硕士量身定制的更先进的模型压缩方法,我们有潜力
提高这些未压缩的 LLM 的性能。

性能与尺寸的权衡

先前的研究 [Magister 等人,2023； Dettmers 和 Zettle-moyer，2022] 强调了大型之间的微妙平衡
语言模型 (LLM) 性能和模型大小。分析这种权衡可以在以下范围内实现最佳性能
硬件限制。然而,当前的工作缺乏对这种权衡的理论和实证见解。未来法学硕士

压缩研究应进行综合分析,以指导先进技术。了解性能和规模之间的关系可以帮助研究人员

开发定制的压缩方法,有效地在设计空间中导航以获得高效的解决方案。

动态LLM压缩

尽管当前的压缩方法取得了进步，
他们仍然依靠手动设计来确定压缩
法学硕士的规模和结构。这通常涉及基于输入数据或任务要求的试错方法。

这个过程在场景中变得特别具有挑战性
就像知识蒸馏一样,需要多次尝试
在计算限制内找到合适的学生模型。这种手动操作存在实际障碍。

神经网络集成中出现了一个有前途的解决方案
架构搜索 (NAS) 技术 [Elsken 等人,2019；
佐夫和勒,2016;朱等人,2021； Zhu et al., 2023b] 进入
压缩法学硕士领域。 NAS 拥有潜力
减少对人类设计架构的依赖，
潜在地彻底改变 LLM 压缩以改进
效率和有效性。

可解释性

早期研究 [Stanton 等人,2021;徐等人,2021]
引起了人们对可解释性的严重担忧
应用于预训练语言的压缩技术
模型 (PLM)。值得注意的是,这些同样的挑战还延伸到
LLM压缩方法也是如此。例如,有
没有解释为什么 CoT 蒸馏可以使
SLM拥有CoT能力,并且在推理任务中取得了良好的表现。因此,整合可解释的

压缩方法成为至关重要的必要条件
LLM压缩申请的进展。而且，
采用可解释的压缩不仅解决了
可解释性问题,同时也简化了评估
压缩模型的过程。这反过来又增强了
整个模型的可靠性和可预测性
生产阶段。

5 结论

在这次彻底的调查中,我们探索了模型压缩
大型语言模型（LLM)技术。我们的覆盖范围涵盖压缩方法、评估指标和

基准数据集。通过深入研究 LLM 压缩,我们已经

强调了其挑战和机遇。随着法学硕士压缩的进步，人们明确呼吁研究专门针对法学硕士的先进方法，以释放他们的潜力。

跨应用程序的潜力。这项调查旨在提供有价值的参考，提供对当前形势的见解

并促进对这一关键主题的持续探索。

致谢

李健的工作得到了国家自然科学基金（No.62106257）、中国博士后科学基金（No.2023T160680）和中国科学院信息工程研究所卓越人才计划的部分支持。支持刘勇的工作

部分国家自然科学基金项目（No.
62076234），北京市杰出青年科学家计划
（编号：BJJWZYJH012019100020098）联通创新
生态合作计划、CCF-华为杨树
格罗夫基金。

参考

[Agarwal 等人，2023] Rishabh Agarwal、Nino Vieillard、Pi-otr Stanczyk、Sabela Ramos、Matthieu Geist 和 Olivier
巴赫姆。 GKD：广义知识蒸馏
自回归序列模型。 CoRR，abs/2306.13649，
2023 年。

[Anwar 等人，2017] Sajid Anwar、Kyuyeon Hwang 和
宋元勇.深度卷积的结构化剪枝
神经网络。 ACM J. Emerg。技术。计算。系统，
13(3):32:1–32:18, 2017。

[Ardakani 等人，2019] Arash Ardakani，季正云，
肖恩·C·史密森、布雷特·H·迈耶和沃伦·J·格罗斯。
学习循环二元/三元权重。在第七届国际学习表征会议上，

ICLR 2019，美国路易斯安那州新奥尔良，2019 年 5 月 6-9 日。
OpenReview.net，2019。

[Behdin 等人，2023] Kayhan Behdin、Ayan Acharya、Aman
古普塔、萨蒂亚·科尔蒂·塞尔瓦拉吉和拉胡尔·马宗德。
Quantease：基于优化的语言量化
模型 - 一种高效且直观的算法。钻RR，
绝对/2309.01885，2023。

[Brooks 等人，2023] Tim Brooks、Aleksander Holynski 和
阿列克谢·埃夫罗斯. Instructpix2pix：学习遵循图像编辑说明。 IEEE/CVF
会议记录
计算机视觉和模式识别会议，
第 18392–18402 页，2023 年。

[Brown 等人，2020] Tom B. Brown、Benjamin Mann、Nick
莱德、梅兰妮·苏比亚、贾里德·卡普兰、普拉芙拉·达里瓦尔、阿文
德·尼拉坎坦、普拉纳夫·希亚姆、吉里什·萨斯特里、
阿曼达·阿斯克尔、桑迪尼·阿加瓦尔、阿里尔·赫伯特-沃斯、
格雷琴·克鲁格、汤姆·赫尼汉、Rewon Child、阿迪亚
Ramesh、Daniel M. Ziegler、Jeffrey Wu、Clemens Winter、
Christopher Hesse、Mark Chen、Eric Sigler、Ma-teusz Litwin、
Scott Gray、Benjamin Chess、Jack Clark、
克里斯托弗·伯纳、山姆·麦坎迪什、亚历克·雷德福、伊利亚
苏茨克韦尔和达里奥·阿莫代。语言模型是小样本学习者。雨果·拉罗谢尔 (Hugo Larochelle)、马克·奥雷里奥·兰扎托 (Marc Aurelio Ran-zato)、拉亚·哈塞尔 (Raia Hadsell)、玛丽亚·弗洛琳娜·巴尔坎 (Maria-Florina Balcan) 和宣天 (Hsuan-Tien)
林，编辑，《神经信息处理进展》
Systems 33：神经信息年会
处理系统 2020，NeurIPS 2020，12 月 6-12 日，
2020 年，虚拟，2020 年。

[Chang et al., 2023] 常玉鹏、王旭、金东
王元、朱凯杰、陈浩、杨林一、易小奥源、王存祥、王一东、叶伟、岳

张，易昌，Philip S. Yu，杨强，谢兴。
大语言模型评估调查。钻RR，
绝对/2307.03109，2023。

[Chavan 等人，2023] Arnav Chavan、Zhuang Liu、Deepak K.
古普塔、Eric P. Xing 和沈志强。一对一：通用 lora，用于参数高效的微调。钻RR，
绝对/2306.07967，2023。

[Chee 等人，2023] Jerry Chee、Yaohui Cai、Volodymyr
库列绍夫和克里斯托弗·德萨。 Quip：有保证的大型语言模型的 2 位量化。
钻RR，
ABS/2307.13304，2023。

[Cheng et al., 2017] Yu Cheng、Duo Wang、Pan Zhou 和
张涛.深度神经网络模型压缩和加速的调查。 CoRR，abs/1710.09282，

2017年。

[Cheng et al., 2023] 程文华、张伟伟、海浩
沉，蔡一阳，何鑫，吕考考。优化
通过带符号梯度下降进行权重舍入，以实现 LLMS 的量化。 CoRR，abs/
2309.05516，2023。

[Chiang 等人，2023] Wei-Lin Jiang、Zzhuohan Li、Zi Lin、
盛颖、吴张浩、张浩、郑连民、
庄思源、庄永浩、Joseph E. Gonzalez、
离子·斯托伊卡 (Ion Stoica) 和埃里克·P·邢 (Eric P. Xing)。 Vicuna：一款
开源聊天机器人，以 90%* chatgpt 质量给 gpt-4 留下深刻印象，3 月
2023 年。

[Dasgupta 等人，2023] Sayantan Dasgupta、Trevor Cohn，
和蒂莫西·鲍德温。具有成本效益的大型蒸馏
语言模型。安娜·罗杰斯、乔丹·L·博伊德-格雷伯，
和 Naoaki Okazaki，编辑，计算语言学协会的调查结果：ACL 2023，多伦多，

加拿大，2023 年 7 月 9 日至 14 日，第 7346-7354 页。协会
计算语言学，2023。

[Deng et al., 2020] 邓雷、李国琪、韩松、路平
石、谢元.神经网络的模型压缩和硬件加速：综合调查。

过程。 IEEE，108(4)：485–532，2020。

[Dettmers 和 Zettlemoyer，2022] Tim Dettmers 和 Luke
泽特尔莫耶。 4 位精度的情况：k 位推理
缩放定律。 CoRR，abs/2212.09720，2022。

[Dettmers 等人，2022] Tim Dettmers、Mike Lewis、Younes
贝尔卡达和卢克·泽特尔莫耶。 Llm.int8()：大规模 Transformer 的 8 位矩阵乘法。钻RR，
绝对/2208.07339，2022。

[Dettmers 等人，2023a] Tim Dettmers、Artidoro Pagnoni，
阿里·霍尔兹曼和卢克·泽特莫耶。 Qlora：高效
量化 llms 的微调。 CoRR，abs/2305.14314，
2023 年。

[Dettmers 等人,2023b] Tim Dettmers、Ruslan Svirschevski、
Vage Egiazarian、丹尼斯·库兹内代列夫、埃利亚斯·弗兰塔、萨利赫
阿什克布斯、亚历山大·博尔祖诺夫、托斯顿·赫夫勒和
丹·阿利斯塔. Spqr:近无损 LLM 权重压缩的稀疏量化表示。钻RR，

绝对/2306.03078,2023。

[Ding 等人,2022] Shaojin Ding,Phoenix Meadowlark，
何彦章、Lukasz Lew、Shivani Agrawal 和 Oleg
雷巴科夫。具有本机量化感知功能的 4 位一致性
语音识别训练。在韩石高和
John HL Hansen,编辑,Interspeech 2022,国际语音通信协会第
23 届年会,韩国仁川,2022 年 9 月 18-22 日，

第 1711-1715 页。伊斯卡,2022。

[Ding et al., 2023] 丁宁,秦雨佳,杨光，
魏富超,杨宗瀚,苏玉生,胡胜定，
陈玉林、陈志敏、陈伟泽、静怡、伟林
赵、王晓志、刘志远、郑海涛、剑飞
陈,刘阳,唐杰,李娟子,孙茂松。
大规模预训练的参数高效微调
语言模型。纳特。苹果。情报,5（3）:220–235,2023。

[Dong et al., 2023] Qingxiu Dong,Lei Li,Damai Dai，
郑策、吴志勇、常宝宝、孙旭、晶晶
徐,李雷,隋志方。情境学习调查。 CoRR,abs/2301.00234,2023。

[Elsken 等人,2019] Thomas Elsken、Jan Hendrik Metzen，
和弗兰克·哈特。神经架构搜索:一项调查。
J.马赫.学习。研究,2019 年 20:55:1–55:21。

[Fang 等人,2020] Jun Fang、Ali Shafiee、Hamzah Abdel-Aziz、
David Thorsley、Georgios Georgiadis 和 Joseph
哈松.训练后分段线性量化
深度神经网络。安德里亚·维达尔迪、霍斯特·比绍夫、
Thomas Brox 和 Jan-Michael Frahm，《计算机》编辑
愿景 - ECCV 2020 - 第 16 届欧洲会议,英国格拉斯哥,2020 年 8 月
23-28 日,计算机科学讲义第 II 部分,第 12347 卷,页数

69–86。施普林格,2020。

[Fang et al., 2023] 方功凡,马欣银,宋明丽，
迈克尔·毕米,王新超。深度图:朝向
任何结构性修剪。 CoRR,abs/2301.12900,2023。

[Frantar 和 Alistarh,2023] Elias Frantar 和 Dan Alistarh。
Sparsegpt:海量语言模型可以准确
一次性修剪。 CoRR,abs/2301.00774,2023。

[Frantar 等人,2022] Elias Frantar、Saleh Ashkboos、Torsten
赫夫勒和丹·阿利斯塔。 GPTQ:生成预训练变换器的精确训练后量
化。 CoRR,abs/2210.17323,2022。

[Fu et al., 2023a] Yao Fu、Hao Peng、Litu Ou、Ashish Sab-harwal 和
Tushar Khot。将较小的语言模型专门用于多步骤推理。钻RR，

ABS/2301.12726,2023。

[Fu et al., 2023b] Zihao Fu、Haoran Yang、Anthony Man-Cho
So、Wai Lam、Lidong Bing 和 Nigel Collier。在
参数高效微调的有效性。在

Brian Williams、Yiling Chen 和 Jennifer Neville,编辑,第三十七届
AAAI 人工智能会议,AAAI 2023,第三十五届创新会议

人工智能的应用,IAAI 2023,第十三届人工智能教育进展研讨会

情报,EAAI 2023,美国华盛顿特区,二月
2023 年 7-14 日,第 12799-12807 页。 AAAI 出版社,2023 年。

[Geva 等人,2021] Mor Geva、Daniel Khashabi、Elad Segal、
图沙尔·科特、丹·罗斯和乔纳森·贝兰特。亚里士多德使用笔记本电脑
吗?问答基准
隐性推理策略。跨。副教授。计算。语言学,9:346–361,2021。

[Gholami 等人,2022] Amir Gholami、Sehoon Kim、Zhen
董、姚哲伟、Michael W Mahoney 和 Kurt
科伊策。高效量化方法综述
神经网络推理。在低功耗计算机视觉中，
第 291-326 页。查普曼和霍尔/CRC,2022 年。

[Gordon 等人,2020] Mitchell A. Gordon、Kevin Duh 和
尼古拉斯·安德鲁斯.压缩 BERT:研究权值剪枝对迁移学习的影响。位
于斯潘达纳
盖拉、约翰内斯·韦尔布尔、马雷克·雷、法比奥·彼得罗尼、帕特里克
SH Lewis、Emma Strubell、Min Joon Seo 和 Han-naneh
Hajishirzi,编辑,第五届研讨会论文集
关于 NLP 的表示学习,RepL4NLP@ACL
2020 年,在线,2020 年 7 月 9 日,第 143-155 页。协会
计算语言学,2020。

[Gou 等人,2021]Jianping Gou、Baosheng Yu、Stephen J.
马来亚银行和大成涛。知识蒸馏:A
民意调查。国际。 J. 计算机。访问,129（6）:1789–1819,2021。

[Gu et al., 2023] Yuxian Gu、Li Dong、Furu Wei 和 Minlie
黄。大语言模型的知识蒸馏。
CoRR,abs/2306.08543,2023。

[Guo 等人,2020] RuiqiGuo、Philip Sun、Erik Lindgren、
耿权、David Simcha、Felix Chern 和 Sanjiv Ku-mar。利用各向异性
加速大规模推理
矢量量化。第 37 届国际机器学习会议论文集,ICML 2020,13-

2020 年 7 月 18 日,虚拟活动,会议记录第 119 卷
机器学习研究,第 3887-3896 页。 PMLR,
2020.

[Guo et al., 2023] 郭从、唐家明、胡伟明、
冷静文、张晨、杨帆、刘云馨、敏一
郭,朱宇豪。 Olive:加速大语言
通过硬件友好的异常值-受害者对量化来建立模型。载于 Yan Solihin
和 Mark A. Heinrich,编辑,第 50 届年度国际研讨会论文集

计算机体系结构,ISCA 2023,奥兰多,佛罗里达州,美国，
2023 年 6 月 17 日至 21 日,第 3:1–3:15 页。美国CM,2023。

[Han 等人,2015] Song Han、Jeff Pool、John Tran 和
威廉·J·达利.学习权重和连接
高效的神经网络。在《科琳娜·科尔特斯》中,尼尔·D.
劳伦斯、丹尼尔·D·李、杉山雅史和罗曼
Garnett,编辑,神经信息处理系统进展 28:神经信息年会

2015 年处理系统,2015 年 12 月 7-12 日,加拿大魁北克省蒙特利尔，
第 1135-1143 页,2015 年。

[He et al., 2018] 何一辉,林吉,刘志坚,王瀚瑞,
李丽佳,韩松。 AMC:用于移动设备上模型压缩和加速的 automl。维
托里奥·法拉利 (Vitto-rio Ferrari)、马夏尔·赫伯特 (Martial Hebert)、
克里斯蒂安·斯明奇塞斯库 (Cristian Sminchisescu) 和
Yair Weiss,计算机视觉编辑 - ECCV 2018 - 第 15 届
欧洲会议,德国慕尼黑,9 月 8 日 -
2018 年 12 月 14 日,《计算机科学讲座笔记》第 VII 部分,第 11211
卷,第 815-832 页。施普林格,
2018.

[Hinton 等人,2015] Geoffrey E. Hinton、Oriol Vinyals 和
杰弗里·迪恩.在神经网络中提取知识。 CoRR,abs/1503.02531,
2015。

[Ho 等人,2023] Namgyu Ho、Laura Schmid 和 Se-Young
云。大型语言模型是推理老师。
安娜·罗杰斯、乔丹·L·博伊德-格雷伯和直明
冈崎,编辑,第 61 届年会记录
计算语言学协会（卷
1:长论文）,ACL 2023,加拿大多伦多,7 月 9 日至 14 日,
2023 年,第 14852–14882 页。计算协会
语言学,2023。

[Hsieh 等人,2023] Cheng-Yu Hsieh、Chun-Liang Li、Chih-Kuan
Yeh、Hootan Nakhost、Yasuhisa Fujii、Alex Ratner、
兰杰·克里希纳、李振宇和托马斯·菲斯特。一步一步蒸馏!超越更大的语
言模型
训练数据较少,模型尺寸较小。在安娜
Rogers、Jordan L. Boyd-Graber 和 Naoaki Okazaki,编辑,计算语
言学协会的调查结果:ACL 2023,加拿大多伦多,2023 年 7 月 9-14 日,

第 8003–8017 页。计算语言学协会,2023。

[Hu 等人,2022] Edward J. Hu、Yelong Shen、Phillip Wallis,
Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang、
还有陈伟柱。 Lora:大的低阶适应
语言模型。第十届国际会议
关于学习表示,ICLR 2022,虚拟活动,
2022 年 4 月 25 日至 29 日。OpenReview.net,2022 年。

[Huang 和 Chang,2023] 黄杰和 Kevin Chen-Chuan
张。在大型语言模型中进行推理:A
民意调查。安娜·罗杰斯 (Anna Rogers)、乔丹·博伊德·格雷伯 (Jordan L. Boyd-Graber) 和
Naoaki Okazaki,编辑,协会的调查结果
计算语言学:ACL 2023,加拿大多伦多,
2023 年 7 月 9 日至 14 日,第 1049-1065 页。计算语言学协会,2023。

[Huang et al., 2022] Yukun Huang,Yanda Chen,Zhou Yu,
和凯瑟琳·R·麦基翁。上下文学习蒸馏:迁移预训练的小样本学习能力

语言模型。 CoRR,abs/2212.10670,2022。

[伊德尔巴耶夫和卡雷拉-佩尔皮安,2020年] 叶尔兰·伊德尔巴耶夫
和米格尔·A·卡雷拉-佩尔皮恩 (Miguel A. Carreira-Perpi nān)。神经
网络的低秩压缩:学习每一层的秩。 2020年
IEEE/CVF 计算机视觉与模式会议
表彰,CVPR 2020,美国华盛顿州西雅图,6 月 13 日至 19 日,
2020 年,第 8046–8056 页。计算机视觉基础 /
IEEE,2020。

[Jha 等人,2023] Ananya Harsh Jha、Dirk Groeneveld,
艾玛·斯特鲁贝尔和伊兹·贝尔塔吉。大语言模型
蒸馏不需要老师。 CoRR,abs/2305.14864,
2023 年。

[Jiang et al., 2023] Yuxin Jiang、Chunkit Chan、Mingyang
陈、王伟. Lion:闭源大语言模型的对抗性蒸馏。钻RR,

绝对/2305.12870,2023。

[Kim 和 Rush,2016] Yoon Kim 和 Alexander M. Rush。
序列级知识蒸馏。泽维尔在简苏
卡雷拉斯 (Carreras) 和凯文·杜 (Kevin Duh),《2016 年会议记录》编辑
自然语言经验方法会议
处理,EMNLP 2016,美国德克萨斯州奥斯汀,11 月
1-4,2016 年,第 1317-1327 页。计算语言学协会,2016。

[Kim 等人,2022] Minsoo Kim、Sihwa Lee、Sukjin Hong、
张斗成、崔政旭。理解和
改进知识蒸馏以实现量化感知
大型变压器编码器的训练。在约夫·戈德堡中,
Zornitsa Kozareva 和 Yue Zhang,编辑,会议记录
2022 年自然经验方法会议
语言处理,EMNLP 2022,阿布扎比,美国
阿拉伯联合酋长国,2022 年 12 月 7 日至 11 日,第 6713-6725 页。
计算语言学协会,2022。

[Kim 等人,2023a] Jeonghoon Kim、Jung Hyun Lee、Sung-dong
Kim、Joonsuk Park、Kang Min Yoo、Se Jung Kwon、
还有李东洙。通过低于 4 位整数量化对压缩大型语言模型进行内存高
效微调。 CoRR,abs/2305.14152,2023。

[Kim 等人,2023b] Sehoon Kim、Coleman Hooper、Amir
古拉米、董震、李秀玉、沉盛、Michael W.
马奥尼和库尔特·科策。 Squeezellm:密集和稀疏量化。 CoRR,abs/
2306.07629,2023。

[Kurtic 等人,2023] Eldar Kurtic、Elias Frantar 和 Dan Al-istarh。
Ziplm:硬件感知的语言模型结构化修剪。 CoRR,abs/2302.04089,
2023。

[LeCun 等人,1989] Yann LeCun、John S. Denker 和
萨拉·A·索拉.最佳脑损伤。 David S. Touret-zky,《神经信息处理进
展》编辑
Systems 2,[NIPS 会议,美国科罗拉多州丹佛市,
1989 年 11 月 27-30 日],第 598-605 页。摩根·考夫曼,1989。

[Lee 等人,2023] Changhun Lee、Jungyu Jin、Taesu Kim,
金亨俊和朴恩赫。 OWQ:课程
从权重量化的激活异常值中学习
大语言模型。 CoRR,abs/2306.02272,2023。

[Li et al., 2017] 李浩,Asim Kadav,Igor Durdanovic,
哈南·萨梅特和汉斯·彼得·格拉夫。修剪过滤器
高效的卷积网络。在第五届国际会议上
学习代表,ICLR 2017,法国土伦,
2017 年 4 月 24-26 日,会议记录。开放评论网,2017 年。

[Li et al., 2022] Shiyang Li、Jianshu Chen、Yelong Shen、
陈志宇、张新禄、李泽坤、王宏、静
钱,彭宝林,毛毅,陈文虎,严西峰。
大语言模型的解释使小推理变得更好。 CoRR,abs/2210.06726,
2022。

[Li 等人,2023a]Liang Li、Qingyuan Li、Bo Zhang 和
褚香香。规范调整:高性能
大型语言模型的低位量化。钻RR,
绝对/2309.02784,2023。

[Li et al., 2023b] 李清源、张一凡、李亮、彭
姚、张波、褚香香、孙业瑞、杜丽、
谢雨辰. FPTQ:大型语言模型的细粒度训练后量化。 CoRR,abs/
2308.15987,
2023 年。

[Lin et al., 2023] Ji Lin、Jiaming Tang、Haotian Tang、Shang
杨,党星宇,宋瀚。 AWQ:LLM 压缩和加速的激活感知权重量化。
CoRR,abs/2306.00978,2023。

[Liu et al., 2021a] Yuanxin Liu、Fandong Jia、Zheng Lin、
王卫平、周杰.边际效用递减:
探索BERT知识的最低限度
蒸馏。承庆宗、夏飞、李文杰、
和 Roberto Navigli,编辑,第 59 届计算语言学协会年会和第 11 届国
际计算语言学联合会议论文集

自然语言处理,ACL/IJCNLP 2021,(第一卷:长论文),虚拟活动,
2021 年 8 月 1-6 日,
第 2928-2941 页。计算语言学协会,2021。

[Liu et al., 2021b] 刘振华、王云鹤、韩凯、魏
张,马思维,高文。视觉变压器的训练后量化。在马克奥雷利奥·兰扎托,

Alina Beygelzimer、Yann N. Dauphin、Percy Liang 和
Jennifer Wortman Vaughan,《神经进展》编辑
信息处理系统 34:年会
神经信息处理系统 2021,NeurIPS
2021 年,2021 年 12 月 6-14 日,虚拟,第 28092-28103 页,
2021 年。

[Liu et al., 2022] 刘浩坤、Derek Tam、Mohammed
Muqeeth、Jay Mohta、Tenghao Huang、Mohit Bansal 和
科林·拉斐尔.少样本参数高效微调是
比情境学习更好、更便宜。在 NeurIPS 中,
2022 年。

[Liu et al., 2023] Zechun Liu、Barlas Oguz、Changsheng
赵、Ernie Chang、Pierre Stock、Yashar Mehdad、
史阳阳、Raghuraman Krishnamoorthi 和 Vikas
钱德拉。 LLM-QAT:大型语言模型的无数据量化感知训练。 CoRR,
abs/2305.17888,
2023 年。

[Luccioni 等人,2022] Alexandra Sasha Luccioni,Sylvain
维吉耶和安妮-劳尔·利戈扎特。估算 Bloom 的碳足迹,这是一个 176b
参数语言模型。
CoRR,abs/2211.02001,2022。

[Ma et al., 2023] Xinyin Ma、Gongfan Fang 和 Xinchao
王. Llm-pruner:关于大型语言模型的结构剪枝。 CoRR,abs/
2305.11627,2023。

[Magister 等人,2023] Lucie Charlotte Magister,乔纳森
Mallinson、Jakub Adamek、Eric Malmi 和 Aliaksei Severyn。教授
小语言模型进行推理。在安娜
Rogers、Jordan L. Boyd-Graber 和 Naoaki Okazaki,编辑,《美国
学会第 61 届年会记录》
计算语言学协会(第 2 卷:简短
论文),ACL 2023,加拿大多伦多,2023 年 7 月 9-14 日,
第 1773-1781 页。计算语言学协会,2023。

[Nagel 等人,2020] Markus Nagel、Rana Ali Amjad、Mart
范巴伦、克里斯托斯·路易佐斯和蒂门·布兰克沃特。
上或下?训练后量化的自适应舍入。第37届国际会议论文集

关于机器学习,ICML 2020,2020 年 7 月 13-18 日,虚拟活动,机器学
习研究论文集第 119 卷,第 7197-7206 页。 PMLR,2020。

[欧阳等人,2022] 欧阳龙,吴杰弗里,江旭,
迪奥戈·阿尔梅达、卡罗尔·L·温赖特、帕梅拉·米什金、
张冲、桑迪尼·阿加瓦尔、卡塔琳娜·斯拉玛、Alex
雷、约翰·舒尔曼、雅各布·希尔顿、弗雷泽·凯尔顿、卢克
米勒、麦迪·西蒙斯、阿曼达·阿斯克尔、彼得·韦林德、
保罗·F·克里斯蒂安诺、简·雷克和瑞安·洛。训练
遵循人类反馈指令的语言模型。在 NeurIPS,2022 年。

[Paperno 等人,2016] Denis Paperno、German Kruszewski、
Angeliki Lazaridou、Quan Ngoc Pham、Raffaella Bernardi、
桑德罗·佩泽尔、马可·巴罗尼、杰玛·博莱达和
拉克尔·费尔南德斯。 LAMBADA 数据集:需要广泛语篇上下文的单词
预测。诉讼中
计算语言学协会第 54 届年会,ACL 2016,2016 年 8 月 7-12 日,柏
林,
德国,第 1 卷:长论文。协会
计算机语言学,2016。

[Park 等人,2019] Wonpyo Park、Dongju Kim、Yan Lu 和
敏苏曹.关系知识蒸馏。在IEEE
计算机视觉和模式识别会议,
CVPR 2019,美国加利福尼亚州长滩,2019 年 6 月 16-20 日,
第 3967-3976 页。计算机视觉基金会/IEEE,
2019.

[Park et al., 2022] Gunho Park、Baeseong Park、Se Jung
权、金秉旭、李英珠和东洙
李。 nuqmm:用于高效推理大规模生成语言模型的量化 matmul。
钻RR,
绝对/2206.09557,2022。

[Petroni 等人,2019] Fabio Petroni、Tim Rocktaschel、Sebastian
Riedel、Patrick SH Lewis、Anton Bakhtin、Yux-iang Wu 和
Alexander H. Miller。语言模型为
知识库?干健太郎、蒋静、吴文森、
和晓军,编辑,2019 年自然语言处理经验方法会议和第九届自然语言
处理国际联合会议论文集,EMNLP-IJCNLP 2019,香港

中国香港,2019 年 11 月 3-7 日,第 2463-2473 页。计算语言学协会,
2019。

[Povey 等人,2018] Daniel Povey、程高峰、一鸣
王,李科,徐海南,Mahsa Yarmhammadi,和
桑吉夫·库丹普尔。半正交低秩矩阵
深度神经网络的因式分解。 B. Yegna-narayana,编辑,Interspeech
2018,国际语音交流协会第 19 届年会,印度海得拉巴,2018 年 9 月
2-6 日,第 3743 页

3747.ISCA,2018。

[Rajpurkar 等人,2016] Pranav Rajpurkar、Jian Zhang、Konstantin Lopyrev 和 Percy Liang。小队:100、000+
机器理解文本的问题。在简苏,
Xavier Carreras 和 Kevin Duh,编辑,Proceedings of
2016年自然经验方法会议
语言处理,EMNLP 2016,奥斯汀,德克萨斯州,美国,
2016 年 11 月 1-4 日,第 2383-2392 页。该协会
计算语言学,2016。

[Rastegari 等人,2016] Mohammad Rastegari、Vicente Or-donez、Joseph Redmon 和 Ali Farhadi。 Xnor-net:使用二元卷积神经网络的 Ima-genet 分类。巴斯蒂安·莱贝、吉里·麦塔斯、尼库·塞贝和马克斯
Welling,计算机视觉编辑 - ECCV 2016 - 第 14 届
欧洲会议,荷兰阿姆斯特丹,2016 年 10 月 11 日至 14 日,会议记录,第四部分,第 9908 卷
计算机科学讲义,第 525-542 页。
施普林格,2016。

[Saha 等人,2023] Swarnadeep Saha、Peter Hase 和 Mo-hit Bansal。
语言模型可以教较弱的智能体吗?
教师的解释通过心理理论提高学生的水平。
CoRR,abs/2306.09299,2023。

[Santacroce 等人,2023] Michael Santacroce,Zixin Wen,
沉叶龙,李远志.在生成语言模型的结构化修剪中什么重要?钻RR,

绝对/2302.03773,2023。

[Schaeffer 等人,2023] Rylan Schaeffer、Brando Miranda,
和桑米·科耶乔。大型语言模型的新兴能力是海市蜃楼吗? CoRR,abs/2304.15004,2023。

[Shao et al., 2023] 邵文琪、陈孟照、朝阳
张,徐鹏,赵丽瑞,李志干,张凯鹏,
彭高、于乔、平罗。 Omniquant:大型语言模型的全方位校准量化。
CoRR,abs/2308.13137,2023。

[Shi et al., 2023] Freda Shi、Mirac Suzgun、Markus Fre-itag、Xuezhi Wang、Suraj Srivats、Soroush Vosoughi、
郑亨元、泰伊、塞巴斯蒂安·鲁德、周丹尼、
迪潘詹·达斯和贾森·魏。语言模型是多语言的思维链推理机。在第十一届国际学习表征会议上,ICLR

2023 年,卢旺达基加利,2023 年 5 月 1-5 日。OpenReview.net,
2023 年。

[Shridhar 等人,2023] Kumar Shridhar、Alessandro Stolfo,
和姆林玛雅·萨尚。提炼推理能力
成更小的语言模型。在安娜·罗杰斯、乔丹·L.
Boyd-Graber 和 Naoaki Okazaki,编辑,《发现》
计算语言学协会:ACL 2023,
加拿大多伦多,2023 年 7 月 9 日至 14 日,第 7059-7073 页。计算语言学协会,2023。

[Srivastava 等人,2022] Aarohi Srivastava、Abhinav Ras-togi、Abhishek Rao、Abu Awal Md Shoeb、Abubakar Abid、
亚当·费什、亚当·R·布朗、亚当·桑托罗、阿迪亚
古普塔、阿德里亚·加里加-阿隆索、阿格涅斯卡·克鲁斯卡、艾托·卢科维奇、阿克沙特·阿加瓦尔、Alethea Power、亚历克斯·雷、
亚历克斯·沃斯塔特 (Alex Warstadt)、亚历山大·科库雷克 (Alexander W. Kocurek)、阿里·萨法亚 (Ali Safaya)、阿里
塔扎夫、爱丽丝·项、艾丽西娅·帕里什、艾伦·涅、阿曼·侯赛因、阿曼达·阿斯克尔、阿曼达·杜苏扎、阿米特·拉哈内、
Anantharaman S. Iyer、Anders Andreassen、Andrea San-tilli、Andreas Stuhlmuller、Andrew M. Dai、Andrew La、Andrew K. Lampinen、Andy Zou、Angela Jiang、Angelica Chen、Anh Vuong、Animesh Gupta、Anna Gottardi、Anton-nio Norelli、Anu Venkatesh、Arash Gholamidavoodi、Arfa
塔巴苏姆、阿鲁尔·梅内塞斯、阿伦·基鲁巴拉扬、阿舍尔·穆尔洛坎多夫、阿什什·萨巴瓦尔、奥斯汀·赫里克、阿维亚·埃弗拉特、
艾库特·埃尔德姆 (Aykut Erdem)、艾拉·卡拉卡斯 (Ayla Karakas) 等人。超越模仿游戏:量化和推断能力
的语言模型。 CoRR,abs/2206.04615,2022。

[Stanton 等人,2021] Samuel Stanton、Pavel Izmailov,
波琳娜·基里琴科、亚历山大·A·阿勒米和安德鲁·戈登·威尔逊。知识蒸馏真的吗
工作?在马克·奥雷利奥·兰扎托、阿丽娜·贝格尔齐默中,
扬·N·多芬 (Yann N. Dauphin)、珀西·梁 (Percy Liang) 和詹妮弗·沃特曼 (Jennifer Wortman)
沃恩,编辑,《神经信息进展》
处理系统 34:神经年会
信息处理系统 2021、NeurIPS 2021、
2021 年 12 月 6 日至 14 日,虚拟,2021 年第 6906-6919 页。

[Sun et al., 2023] 孙明杰、刘壮、安娜·贝尔和
J.济科·科尔特。简单有效的修剪方法
对于大型语言模型。 CoRR,abs/2306.11695,2023。

[Syed 等人,2023] Aaquib Syed、Phillip Huang Guo 和 Vi-jaykaarti Sundarapandiyan。修剪和调整:改进
针对大规模语言模型的高效修剪技术。克里斯塔尔·莫恩 (Krystal Maughan)、罗珊娜·刘 (Rosanne Liu) 和托马斯·F.
Burns,编辑,ICLR 2023 的第一个小型论文轨道,
Tiny Papers @ ICLR 2023,卢旺达基加利,2023 年 5 月 5 日。
OpenReview.net,2023 年。

[Tailor 等人,2021] Shyam Anil Tailor、Javier Fernandez-Marques 和 Nicholas Donald Lane。 Degree-quant:图神经网络的量化感知训练。
第九届学习表征国际会议,ICLR 2021,虚拟活动,奥地利,2021 年 5 月 3-7 日。
OpenReview.net,2021 年。

[Tao et al., 2023] 陶超凡、侯鲁、白浩丽、魏建生、蒋欣、刘群、罗平和黄艺。

用于高效生成预训练语言模型的结构化剪枝。安娜·罗杰斯、乔丹·L·博伊德-格雷伯,
和 Naoaki Okazaki,编辑,计算语言学协会的调查结果:ACL 2023,多伦多,
加拿大,2023 年 7 月 9 日至 14 日,第 10880-10895 页。计算语言学协会,2023。

[Tung 和 Mori,2019] Frederick Tung 和 Greg Mori。
保持相似性的知识蒸馏。 2019年
IEEE/CVF 计算机视觉国际会议,
ICCV 2019,韩国首尔(南部),10 月 27 日至 11 月
2,2019 年,第 1365-1374 页。 IEEE,2019。

[Valipour 等人,2023] Mojtaba Valipour、Mehdi Reza-gholizadeh、Ivan Kobyzev 和 Ali Ghodsi。迪洛拉:
使用参数有效地调整预训练模型
动态无搜索低秩自适应。在安德烈亚斯
Vlachos 和 Isabelle Augenstein,编辑,《会议记录》
欧洲分会第十七届会议
计算语言学协会,EACL 2023,
克罗地亚杜布罗夫尼克,2023 年 5 月 2-6 日,第 3266-3279 页。
计算语言学协会,2023。

[Wang 等人,2019a] Alex Wang、Yada Pruksachatkun、Nikita Nangia、Amanpreet Singh、Julian Michael、Felix Hill、Omer Levy 和 Samuel R. Bowman。 Superglue:通用语言理解系统的更具粘性的基准。 Hanna M. Wallach、Hugo Larochelle、Alina Beygelzimer、Florence d Alche-Buc、Emily B. Fox 和 Roman Garnett 编辑,《神经信息处理系统进展 32:2019 年神经信息处理系统年会》、NeurIPS 2019， 2019 年 12 月 8 日至 14 日,加拿大不列颠哥伦比亚省温哥华,第 3261-3275 页,2019 年。

[Wang 等人,2019b] Alex Wang、Amanpreet Singh、Julian Michael、Felix Hill、Omer Levy 和 Samuel R. Bowman。
GLUE:用于自然语言理解的多任务基准测试和分析平台。第七届学习表征国际会议,ICLR 2019,美国洛杉矶新奥尔良,2019 年 5 月 6-9 日。OpenReview.net，2019。

[Wang et al., 2023a] 王培峰、王正阳、李正、高一凡、尹冰和任向。斯科特:自洽的思想链蒸馏。载于 Anna Rogers、Jordan L. Boyd-Graber 和 Naoaki Okazaki,编辑,计算语言学协会第 61 届年会论文集（第 1 卷:长论文）,ACL 2023,加拿大多伦多,2023 年 7 月 9-14 日,第 5546-5558 页。计算语言学协会,2023。

[Wang et al., 2023b] Xinyi Wang、Wanrong Zhu 和 William Yang Wang。大型语言模型隐含地是主题模型:解释并找到上下文学习的良好演示。 CoRR,abs/2301.11916,2023。

[Wang 等人,2023c]Xuezhi Wang、Jason Wei、Dale Schuur-mans、Quoc V. Le、Ed H. Chi、Sharan Narang、Aakanksha Chowdhery 和 Denny Zhou。自我一致性改善了语言模型中的思维链推理。第十一届学习代表国际会议,ICLR 2023,卢旺达基加利,2023 年 5 月 1-5 日。

OpenReview.net,2023 年。

[Wang 等人,2023d] Yizhong Wang、Yeganeh Kordi、Swa-roop Mishra、Alisa Liu、Noah A. Smith、Daniel Khashabi 和 Hannaneh Hajishirzi。自指导:将语言模型与自生成的指令保持一致。载于 Anna Rogers、Jordan L. Boyd-Graber 和 Naoaki Okazaki,编辑,第 61 届计算语言学协会年会论文集（第 1 卷:长论文）,ACL 2023,加拿大多伦多,7 月 9-14 日,2023 年,第 13484–13508 页。计算语言学协会,2023 年。

[Wei 等人,2022a] Jason Wei、Yi Tay、Rishi Bommasani、Colin Raffel、Barret Zoph、Sebastian Borgeaud、Dani Yo-gatama、Maarten Bosma、Denny Zhou、Donald Metzler、Ed H. Chi、Tatsunori Hashimoto、Oriol Vinyals 、珀西·梁、杰夫·迪恩和威廉·费杜斯。大型语言模型的新兴能力。跨。马赫。学习。研究，2022,2022。

[Wei 等人,2022b] Jason Wei、Xuezhi Wang、Dale Schuur-mans、Maarten Bosma、Brian Ichter、Fei Xia、Ed H. Chi、Quoc V. Le 和 Denny Zhou。思路链提示
ing 在大型语言模型中引发推理。在 NeurIPS,2022 年。

[Wei et al., 2023] Xiuying Wei、Yunchen Zhang、Yuhang Li、Xiangguo Chang、Ruihao Kong、JinyangGuo 和 Xian-long Liu。异常值抑制+:通过等效且最优的移位和缩放对大型语言模型进行精确量化。 CoRR,abs/2304.09145,2023。

[Wu 等人,2023a] Minghao Wu、Abdul Waheed、Chiyu Zhang、Muhammad Abdul-Mageed 和 Alham Fikri Aji。
Lamini-lm:从大规模指令中提取出来的多样化模型。 CoRR,abs/2304.14402，2023。

[Wu et al., 2023b] 吴晓霞、姚哲伟和何宇雄。 Zeroquant-fp:使用浮点格式进行 llms 训练后 W4A8 量化的飞跃。 CoRR,abs/2307.09782,2023。

[Xiao 等人,2022]Guangxuan Shaw、Ji Lin、Mickael Seznec、¡Julien Demouth 和 Song Han。 Smoothquant:大型语言模型的准确高效的训练后量化。 CoRR,abs/2211.10438,2022。

[Xu et al., 2021] Canwen Xu、Wangchunshu Zhou、Tao Ge、Ke Xu、Julian J. McAuley 和 Furu Wei。超越保留的准确性:评估 BERT 压缩的忠诚度和稳健性。 Marie-Francine Moens、Xuanjing Huang、Lucia Specia 和 Scott Wen-tau Yih,编辑,2021 年自然语言处理经验方法会议论文集,EMNLP 2021,虚拟活动/多米尼加共和国蓬塔卡纳,7- 2021 年 11 月 11 日,第 10653–10659 页。计算语言学协会,2021。

[Xu et al., 2023] Mingxue Xu、Yao Lei Xu 和 Danilo P.
曼迪奇。 Tensorgpt:基于张量序列分解的 llms 中嵌入层的高效压缩。

CoRR,abs/2307.00526,2023。

[Yang 等人,2017] Tien-Ju Yang、Yu-Hsin Chen 和 Vivi-enne Sze。使用能量感知修剪设计节能卷积神经网络。 2017 年 IEEE 计算机视觉和模式识别会议,CVPR 2017,美国夏威夷州檀香山,2017 年 7 月 21-26 日,第 6071-6079 页。 IEEE 计算机协会,2017 年。

[Yao et al., 2022] Zhewei Yao、Reza Yazdani Aminabadi、Minjia Zhang、Xiaoxia Wu、Conglong Li 和 Yuxiong He。
Zeroquant:大规模 Transformer 的高效且经济实惠的训练后量化。在 NeurIPS，2022 年。

[Yao et al., 2023] Zhewei Yao、Cheng Li、Xiaoxia Wu、Stephen Youn 和 Yuxiong He。 Zeroquant-v2:探索 llms 中的训练后量化,从综合研究到低阶补偿。 CoRR,abs/2303.08302,2023。

[Yuan et al., 2023] 袁志航、牛林、刘家伟、刘文宇、王兴刚、尚玉章、孙光宇、吴强、吴嘉祥和吴秉哲。 RPTQ:大型语言模型基于重排序的训练后量化。 CoRR,abs/2304.01089,2023。

[Zhang et al., 2018] 张天云、叶少凯、张凯奇、唐健、文武杰、Makan Fardad 和 Yanzhi

王.使用乘子交替方向方法的系统 DNN 权重修剪框架。维托里奥·法拉利、马夏尔·赫伯特、克里斯蒂安·斯明奇塞斯库和

Yair Weiss,计算机视觉编辑 - ECCV 2018 - 第 15 届
欧洲会议,德国慕尼黑,9 月 8 日 -
2018 年 12 月 14 日,《计算机科学讲座笔记》第 VIII 部分,第 11212
卷,第 191-207 页。施普林格,
2018.

[Zhang et al., 2023a] Mingyang 张、Hao Chen、Chun-hua
Shen、Zhen Yang、Linlin Ou、Xinyi Yu 和 Bohan
壮。剪枝满足低秩参数高效微调。 CoRR,abs/2305.18403,2023。

[Zhang 等人,2023b] 张庆如,陈敏硕,
亚历山大·布哈林、何鹏程、于成、伟柱
陈,拓昭。适应性预算分配
参数高效的微调。在第十一届国际学习表征会议上,ICLR

2023 年,卢旺达基加利,2023 年 5 月 1-5 日。OpenReview.net,
2023 年。

[Zhang et al., 2023c] Yijia Zhang,Lingran Zhao,Shijie Cao,
王文强、曹挺、杨范、毛杨、张尚航和徐宁一。整数还是浮点数?

大语言低比特量化的新展望
楷模。 CoRR,abs/2305.12356,2023。

[Zhao et al., 2022] 赵博瑞,崔全,宋仁杰,
邱一宇,梁家俊。解耦的知识蒸馏。在 IEEE/CVF 计算机视觉会议上

和模式识别,CVPR 2022,路易斯安那州新奥尔良,
美国,2022 年 6 月 18-24 日,第 11943-11952 页。 IEEE,2022。

[Zhao 等人,2023] Wayne Xin Zhao、Kun Zhou、Junyi Li,
唐天一、王晓蕾、侯玉鹏、民迎千、
张北辰、张俊杰、董子灿、杜一凡、
陈阳、陈玉硕、陈志鹏、蒋金浩、
任瑞阳、李一凡、唐新宇、刘子康、佩玉
刘,聂建云,温继荣。一项大型调查
语言模型。 CoRR,abs/2303.18223,2023。

[Zhu et al., 2021] Xuyu Zhu,Jian Li,Yong Liu,Jun Liao,
还有王卫平。操作级渐进式可微架构搜索。在 James Bailey、Pauli Mi-
ettinen、Yun Sing Koh、Da Cheng Tao 和 Xindong Wu 的著作中,

IEEE 国际数据挖掘会议编辑,
ICDM 2021,新西兰奥克兰,12 月 7-10 日
2021 年,第 1559–1564 页。 IEEE,2021。

[Zhu et al., 2023a] 朱学凯,齐碧清,张凯彦,
龙兴伟,周博文. Pad:程序辅助
蒸馏专门用于推理的大型模型。钻RR,
绝对/2305.13888,2023。

[Zhu et al., 2023b] Xuyu Zhu、Jian Li、Yong Liu 和 Weip-ing Wang。
通过改进可微架构搜索
自蒸馏。 CoRR,abs/2302.05629,2023。

[Zoph 和 Le,2016] Barret Zoph 和 Quoc V. Le。神经
使用强化学习进行架构搜索。钻RR,
绝对/1611.01578,2016。