大模型（LLM）蒸馏技术解析：应用、实现等等

原创 南七无名式 PyTorch研习社 2025年02月13日 08:00 安徽

蒸馏（Distillation）是一种 LLM 训练技术，通过该技术，较小且更高效的模型（如 GPT-4o mini）被训练来模仿更大、更复杂的模型（如 GPT-4o）的行为和知识。

大型语言模型（LLM）在复杂性和规模上持续增长，部署这些模型带来了显著的挑战。

LLM 蒸馏作为一种强有力的解决方案应运而生，它能够将更大、更复杂的语言模型（“教师”）的知识转移到一个更小、更高效的版本（“学生”）上。

AI 领域中的一个最新例子是从 GPT-4o（教师）蒸馏出 GPT-4o mini（学生）。这个过程可以类比为教师向学生传授智慧，目标是**在不携带大型模型复杂性的情况下，提取出核心知识**。

什么是LLM蒸馏?

LLM 蒸馏是一种旨在**在减少规模和计算需求的同时，复制大型语言模型性能的技术**。

可以将其比作一位经验丰富的教授与新学生之间的知识传授。教授代表教师模型，传授复杂的概念和见解，而学生模型则学习以更简化和高效的方式模仿这些教学内容。
这一过程不仅保留了教师模型的核心能力，同时也优化了学生模型，使其能够更快速、更灵活地应用。

为什么 LLM 蒸馏很重要?

LLM 日益增长的规模和计算需求限制了它们的广泛应用和部署。高性能的硬件和日益增加的能耗通常会限制这些模型的可访问性，尤其是在资源受限的环境中，如移动设备或边缘计算平台。

LLM 蒸馏通过生成更小、更快的模型来解决这些挑战，使它们非常适合在更广泛的设备和平台上进行集成。

这一创新不仅使得先进的 AI 技术更加普及，还支持了对速度和效率要求较高的实时应用。通过使 AI 解决方案更加可访问和可扩展，LLM 蒸馏有助于推动 AI 技术的实际应用。

LLM 蒸馏如何工作：知识传递过程

LLM蒸馏过程涉及多种技术，**确保学生模型在高效运行的同时保留关键信息**。下面我们将探讨使这一知识传递过程有效的关键机制。

教师-学生范式

教师-学生范式是 LLM 蒸馏的核心概念，是驱动知识传递过程的基础。在这一结构中，一个更大、更先进的模型将其知识传授给一个更小、更轻量级的模型。

教师模型通常是一个经过广泛训练且拥有强大计算资源的最先进语言模型，它作为丰富的信息来源。而学生模型则被设计成通过模仿教师的行为并内化其知识来学习。

学生模型的主要任务是复制教师的输出，同时保持更小的规模和更低的计算需求。这个过程涉及学生观察并学习教师对各种输入的预测、调整和响应。

通过这种方式，学生可以达到与教师相当的表现和理解水平，从而适用于资源受限的环境中进行部署。

蒸馏技术

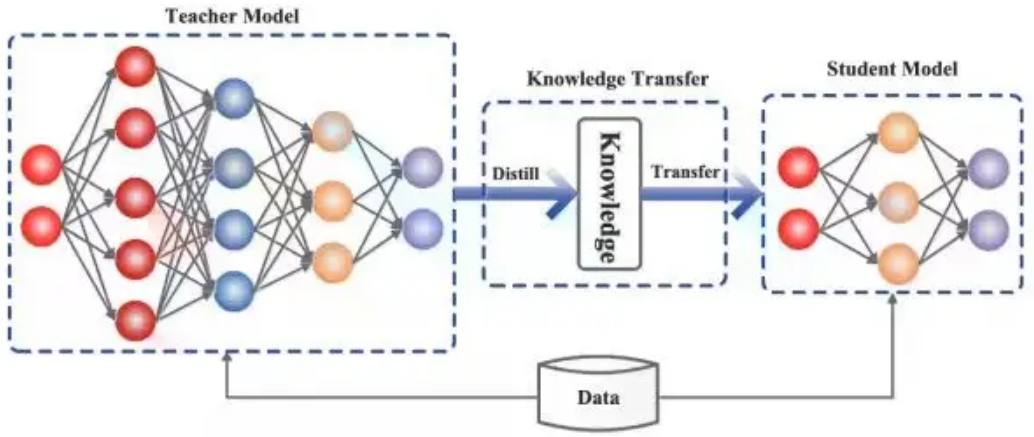
为了实现从教师到学生的知识传递，采用了多种蒸馏技术。这些方法确保学生模型不仅高效学习，还能保留教师模型的核心知识和能力。以下是 LLM 蒸馏中使用的一些最突出技术。

知识蒸馏（KD）

知识蒸馏（KD，Knowledge Distillation）是 LLM 蒸馏中最具代表性的技术之一。在 KD 中，学生模型使用教师模型的输出概率（称为软目标）与真实标签（称为硬目标）一起进行训练。

软目标提供了教师预测的细致视角，它呈现的是可能输出的概率分布，而不是单一的正确答案。这些额外的信息帮助学生模型捕捉教师回答中隐含的微妙模式和复杂知识。

通过使用软目标，学生模型可以更好地理解教师的决策过程，从而实现更准确、更可靠的性能。这种方法不仅保留了教师模型的关键信息，还使学生的训练过程更加平滑和高效。



其他蒸馏技术

除了知识蒸馏（KD）之外，还有一些其他技术可以改善LLM蒸馏过程：

- **数据增强**

  数据增强通过使用教师模型生成额外的训练数据来进行。这种方法通过创建更大、更全面的数据集，使学生能够接触到更广泛的场景和例子，从而提高其泛化能力。

- **中间层蒸馏**

  与仅关注最终输出不同，这种方法将知识从教师模型的中间层转移到学生模型。通过学习这些中间表示，学生可以捕捉到更详细、更结构化的信息，从而提升整体性能。

- **多教师蒸馏**
  学生模型可以通过向多个教师模型学习而受益。通过整合来自不同教师的知识，学生可以获得更全面的理解和更强的鲁棒性，因为它能够融合不同的视角和见解。

LLM 蒸馏的好处

LLM 蒸馏提供了一系列显著的好处，这些好处增强了语言模型的可用性和效率，使它们在各种应用中变得更加实用。

以下是一些关键优势的探讨。

减少模型大小

LLM 蒸馏的主要好处之一是生成明显更小的模型。通过将知识从大型教师模型转移到较小的学生模型，最终的学生模型保留了教师模型的许多能力，同时其大小大大减少。
这种模型大小的减少带来以下好处：

- **更快的推理速度**：较小的模型处理数据的速度更快，响应时间更短。
- **减少存储需求**：较小的模型占用更少的存储空间，尤其在存储容量有限的环境中，便于存储和管理。

提高推理速度

蒸馏模型的较小尺寸直接转化为更高的推理速度。这对于需要实时处理和快速响应的应用至关重要。
以下是该好处的体现：

- **实时应用**：更快的推理速度使蒸馏模型能够应用于实时应用，如聊天机器人、虚拟助手和互动系统，延迟成为关键因素。
- **资源受限设备**：蒸馏模型可以在资源有限的设备上运行，如智能手机、平板电脑和边缘设备，而不影响性能。

降低计算成本

LLM 蒸馏的另一个显著优势是降低计算成本。较小的模型需要更少的计算能力运行，这带来了多个领域的成本节约：

- **云环境**：在云环境中运行较小的模型可以减少对昂贵高性能硬件的需求，降低能耗。
- **本地部署**：较小的模型意味着对于选择本地部署的组织来说，基础设施和维护费用较低。

更广泛的可访问性和部署

蒸馏后的 LLM 更具多样性和可访问性，可以跨平台进行部署。这种扩展的覆盖范围带来了几个方面的影响：

- **移动设备**：蒸馏模型可以在移动设备上部署，使先进的 AI 功能能够以便捷的、用户友好的格式提供。
- **边缘设备**：在边缘设备上的运行使 AI 功能更加接近数据生成的地方，减少了对持续连接的需求，并增强了数据隐私。
- **更广泛的应用**：从医疗保健到金融再到教育，蒸馏模型可以集成到众多应用中，使更多行业和用户能够接触到先进的 AI。

蒸馏 LLM 的应用

LLM 蒸馏的好处不仅限于模型效率和成本节约。蒸馏后的语言模型可以应用于广泛的自然语言处理（NLP）任务和行业特定的用例，使 AI 解决方案能够跨各个领域提供服务。

高效的 NLP 任务

蒸馏后的 LLM 在许多自然语言处理任务中表现优异。其较小的体积和增强的性能使其成为需要实时处理和较低计算能力的任务的理想选择。

- **聊天机器人**：蒸馏后的 LLM 使得开发更小、更快的聊天机器人成为可能，这些机器人能够顺畅地处理客户服务和支持任务，实时理解并回应用户查询，提供无缝的客户体验。
- **文本摘要**：基于蒸馏 LLM 的摘要工具可以将新闻文章、文档或社交媒体信息浓缩成简洁的摘要，帮助用户快速抓住要点，而无需阅读冗长的文本。
- **机器翻译**：蒸馏模型使翻译服务变得更快速、更易于跨设备访问。它们可以部署在手机、平板电脑甚至离线应用程序中，提供实时翻译，减少延迟和计算开销。

其他任务

蒸馏 LLM 不仅对常见的 NLP 任务有价值，还在一些需要快速处理和准确结果的专业领域中表现出色。

- **情感分析**：分析文本的情感（如评论或社交媒体帖子）变得更加快捷，企业能够迅速了解公众意见和客户反馈。
- **问答**：蒸馏模型能够支持准确、及时地回答用户问题，增强虚拟助手和教育工具等应用中的用户体验。
- **文本生成**：无论是内容创作、讲故事还是自动化报告生成，蒸馏后的LLM能够简化生成连贯和上下文相关的文本的过程。

行业用例

蒸馏后的 LLM 不仅限于一般的 NLP 任务，它们还能在许多行业中产生深远的影响，改善流程、提升用户体验，并推动创新。

- **医疗健康**：在医疗行业中，蒸馏后的 LLM 可以更高效地处理患者记录和诊断数据，帮助医生和医疗专业人员做出更快、更准确的诊断。它们可以部署在医疗设备中，支持实时数据分析和决策。
- **金融**：金融领域从蒸馏模型中受益，通过升级的欺诈检测系统和客户互动模型，蒸馏 LLM 能够快速解读交易模式和客户查询，帮助防止欺诈活动，并提供个性化的财务建议和支持。
- **教育**：在教育领域，蒸馏后的 LLM 促进了自适应学习系统和个性化辅导平台的创建。这些系统可以分析学生表现并提供量身定制的教育内容，提升学习成果，使教育变得更加可达和有效。

LLM 蒸馏的实现

实施 LLM 蒸馏需要一系列步骤，并使用专门的框架和库来支持该过程。以下是实施蒸馏过程所需的工具和步骤。

框架和库

为简化蒸馏过程，有几种框架和库可供使用，每个框架和库提供独特的功能，支持 LLM 蒸馏。

- **Hugging Face transformers**：Hugging Face transformers 库是实现 LLM 蒸馏的流行工具，包含一个 Distiller 类，用于简化将知识从教师模型转移到学生模型的过程。
- **其他库**：除了 Hugging Face Transformers，还有许多其他库支持 LLM 蒸馏：
  - TensorFlow 模型优化：提供模型修剪、量化和蒸馏工具，是创建模型的多功能选择。
  - PyTorch distiller：专为深度学习模型的压缩而设计，支持蒸馏技术，提供一系列工具来管理蒸馏过程并提高模型效率。
  - DeepSpeed：由微软开发的DeepSpeed是一个深度学习优化库，包含模型蒸馏的功能，支持大型模型的训练和部署。

实施步骤

实施 LLM 蒸馏需要谨慎规划和执行。以下是蒸馏过程中的关键步骤。

1. **数据准备**：准备适合训练学生模型的数据集，确保数据集具有代表性，能够帮助学生模型良好地进行泛化。
2. **教师模型选择**：选择合适的教师模型，它应当是一个在目标任务上表现良好的预训练模型，教师模型的质量直接影响学生模型的表现。
3. **蒸馏过程**：包括初始化学生模型，配置训练环境，使用教师模型生成软标签（概率分布），并结合硬标签（真实标签）一起训练学生模型。

评估指标

评估蒸馏模型的性能是确保其达到预期标准的关键。常见的评估指标包括：

- **准确度**：衡量学生模型相对于真实标签的预测正确率。
- **推理速度**：评估学生模型处理输入并生成输出的时间。
- **模型大小**：评估模型大小的减少及其在存储和计算效率上的优势。
- **资源利用率**：监控学生模型在推理过程中所需的计算资源，确保满足部署环境的限制。

LLM 蒸馏：挑战与最佳实践

虽然 LLM 蒸馏带来了许多好处，但它也存在一些挑战，需要加以解决，以确保成功实施。

知识损失

LLM 蒸馏的主要难题之一是潜在的知识损失。在蒸馏过程中，教师模型的一些细节信息和特征可能无法完全被学生模型捕捉，从而导致性能下降。这个问题在需要深入理解或专业知识的任务中尤为突出。

以下是我们可以实施的一些策略，以减轻知识损失：

- **中间层蒸馏**：从教师模型的中间层转移知识，有助于学生模型捕捉更详细和结构化的信息。
- **数据增强**：使用教师模型生成的增强数据，可以为学生模型提供更广泛的训练示例，帮助其学习过程。
- **迭代蒸馏**：通过多轮蒸馏不断优化学生模型，使其逐步捕捉教师模型的更多知识。

超参数调优

精心调整超参数对于蒸馏过程的成功至关重要。关键超参数，如温度和学习率，对学生模型从教师模型中学习的能力有重要影响：

- **温度**：该参数控制教师模型生成的概率分布的平滑度。较高的温度会生成较软的概率分布，这有助于学生模型更全面地从教师模型的预测中学习。

- **学习率**：调整学习率对于平衡训练过程的速度和稳定性至关重要。适当的学习率确保学生模型在不发生过拟合或欠拟合的情况下收敛到最优解。

评估效果

评估蒸馏模型的效果是保证其达到期望性能标准的不可或缺的一步，特别是与其前辈和替代模型的对比。这需要比较学生模型与教师模型以及其他基准的性能，以了解蒸馏过程在多大程度上保持或提升了模型功能。
评估蒸馏模型效果时，重点关注以下几个指标：

- **准确性**：衡量学生模型的准确性与教师模型和其他基准的对比，了解是否存在精度损失或保持。
- **推理速度**：比较学生模型与教师模型的推理速度，突出处理时间的改善。
- **模型大小**：评估学生模型与教师模型以及其他基准的模型大小差异，评估蒸馏带来的效率提升。
- **资源利用率**：分析学生模型与教师模型的资源使用情况，确保学生模型在不妥协性能的情况下提供更经济的替代方案。

最佳实践

遵循最佳实践可以提高 LLM 蒸馏的效果。这些指导原则强调实验、持续评估和战略实施。

- **实验**：定期尝试不同的蒸馏技术和超参数设置，识别最适合特定用例的配置。
- **持续评估**：通过相关基准和数据集持续评估学生模型的性能。迭代测试和优化是实现最佳结果的关键。
- **平衡训练**：通过结合来自教师模型的软标签与硬标签，确保实施平衡的训练过程。这有助于学生模型在保持准确性的同时，捕捉细微的知识。
- **定期更新**：关注 LLM 蒸馏研究的最新进展，并将新技术和发现纳入蒸馏过程。

研究与未来方向

LLM 蒸馏领域正在迅速发展。本节探讨了 LLM 蒸馏中的最新趋势、当前的研究挑战以及新兴技术。

最新研究与进展

最近，LLM 蒸馏的研究集中在开发新颖的技术和架构，以提高蒸馏过程的效率和效果。一些显著的进展包括：

- **渐进蒸馏**：这种方法涉及逐步蒸馏知识，在各个阶段从教师模型中逐步蒸馏学生模型。该技术在提高最终学生模型的性能和稳定性方面表现出了潜力。
- **任务无关蒸馏**：研究人员正在探索无关任务的蒸馏方法，允许学生模型跨不同任务进行泛化，而无需特定任务的微调。这种方法可以大大减少新应用所需的训练时间和计算资源。
- **跨模态蒸馏**：另一个新兴领域是跨不同模态（如文本、图像和音频）的蒸馏。跨模态蒸馏旨在创建能够处理多种类型输入数据的通用学生模型，扩展了蒸馏模型的应用范围。

未解问题与未来方向

尽管已有显著进展，但 LLM 蒸馏领域仍然存在若干挑战和未解的研究问题：

- **提高泛化能力**：一个关键挑战是提高蒸馏模型的泛化能力。确保学生模型能够在各种任务和数据集上表现良好，仍然是一个持续的研究领域。
- **跨领域知识转移**：不同领域之间有效的知识转移是另一个重要领域。开发可以应用于新领域的蒸馏方法，而不会显著损失性能，是一个重要目标。
- **可扩展性**：扩展蒸馏技术以高效处理日益增大的模型和数据集仍然是一个持续的挑战。研究正专注于优化蒸馏过程，以使其更具可扩展性。

新兴技术

新兴技术和创新不断发展，以应对这些挑战并推动该领域前进。一些有前景的方法包括：

- **零样本和少样本学习适应**：将零样本和少样本学习能力集成到蒸馏模型中是一个新兴的研究领域。这些技术使得模型能够在几乎没有任务特定训练数据的情况下执行任务，从而增强了其通用性和实用性。
- **自蒸馏**：在自蒸馏中，学生模型使用其预测作为软标签进行训练。这种方法通过迭代使用已学习的知识，能够提高模型的性能和鲁棒性。
- **对抗性蒸馏**：将对抗性训练与蒸馏技术相结合是一种创新方法。对抗性蒸馏不仅训练学生模型模仿教师模型，还使其能够抵抗对抗性攻击，从而提高安全性和可靠性。

DeepSeek-R1 的蒸馏模型

DeepSeek 使用了 Qwen 和 Llama 架构从 R1 模型中创建了一系列蒸馏模型。

3.2. Distilled Model Evaluation

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	**1820**
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	**72.6**	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633

Table 5 | Comparison of DeepSeek-R1 distilled models and other comparable models on reasoning-related benchmarks.

基于 Qwen 的蒸馏模型

DeepSeek 的基于 Qwen 的蒸馏模型注重效率和可扩展性，在性能和计算需求之间提供平衡。

DeepSeek-R1-Distill-Qwen-1.5B

这是最小的蒸馏模型，在 MATH-500 上的得分为 83.9%。MATH-500 测试解决高中文化数学问题的能力，包括逻辑推理和多步解答。该结果表明，尽管模型体积小，但它在处理基本数学任务时表现良好。

然而，它在 LiveCodeBench 上的表现显著下降（16.9%），LiveCodeBench 是评估编程能力的基准，突出显示了该模型在编程任务中的局限性。

DeepSeek-R1-Distill-Qwen-7B

Qwen-7B 在 MATH-500 上表现出色，得分 92.8%，展示了其强大的数学推理能力。它在 GPQA Diamond（49.1%）上的表现也不错，该基准评估事实问答能力，表明它在数学推理和事实推理之间取得了良好的平衡。
然而，它在 LiveCodeBench（37.6%）和 CodeForces（1189评分）上的表现表明，它在复杂的编程任务上相对较弱。

DeepSeek-R1-Distill-Qwen-14B

该模型在 MATH-500 上的表现良好（93.9%），反映了它处理复杂数学问题的能力。它在 GPQA Diamond 上的得分为 59.1%，也表明了其事实推理能力。
然而，在 LiveCodeBench（53.1%）和 CodeForces（1481评分）上的表现显示它在编程和编程特定推理任务上还有提升空间。

DeepSeek-R1-Distill-Qwen-32B

作为最大的 Qwen 基础模型，它在 AIME 2024 上的得分最高（72.6%），该基准评估高级多步骤数学推理。它在 MATH-500（94.3%）和 GPQA Diamond（62.1%）上的表现也非常出色，展示了其在数学推理和事实推理方面的优势。
在 LiveCodeBench（57.2%）和 CodeForces（1691评分）上的结果表明它具有广泛的适用性，但在编程任务上仍然没有像专注于编程的模型那样优化。

基于 Llama 的蒸馏模型

DeepSeek 的基于 Llama 的蒸馏模型优先考虑高性能和先进的推理能力，特别是在需要数学和事实精度的任务中表现优异。

DeepSeek-R1-Distill-Llama-8B

Llama-8B 在 MATH-500 上的得分为 89.1%，在 GPQA Diamond 上表现合理（49.0%），表明它能够处理数学和事实推理。然而，在 LiveCodeBench（39.6%）和 CodeForces（1205分）等编程基准测试中的得分较低，突显了它在编程任务中的局限性，相比于 Qwen 基础模型更为明显。

DeepSeek-R1-Distill-Llama-70B

最大的蒸馏模型，Llama-70B，在 MATH-500上 表现最佳（94.5%），所有蒸馏模型中表现最强，并且在 AIME 2024 上取得了 86.7% 的好成绩，使其成为先进数学推理的优选模型。

它在 LiveCodeBench（57.5%）和 CodeForces（1633评分）上的表现也不错，表明它在编程任务上比大多数其他模型更具竞争力。在这方面，它与 OpenAI 的 o1-mini 或 GPT-4o 相当。

总结

LLM 蒸馏在使大型语言模型变得更加实用和高效方面起着至关重要的作用。通过将复杂教师模型的核心知识转移到较小的学生模型中，蒸馏能够在减少模型大小和计算需求的同时，保持模型的性能。

这一过程使得 AI 应用更加快速、便捷，能够跨多个行业提供服务，从实时的 NLP 任务到医疗和金融等领域的专业应用。实施 LLM 蒸馏需要谨慎的规划和适当的工具，但其带来的好处，如降低成本和更广泛的部署，具有巨大的潜力。

随着研究的不断进展，LLM 蒸馏将在促进 AI 普及和应用方面发挥越来越重要的作用，使强大的模型在多种环境中变得更加可访问和实用。