

DeepSeek发布R1模型，好好恶补文章中提到的“蒸馏技术”

原创 智能体AI 智能体AI 2025年01月31日 07:02 湖南

DeepSeek-R1 发布，性能对标 OpenAI o1 正式版

深度求索 DeepSeek 2025年01月20日 20:10 北京

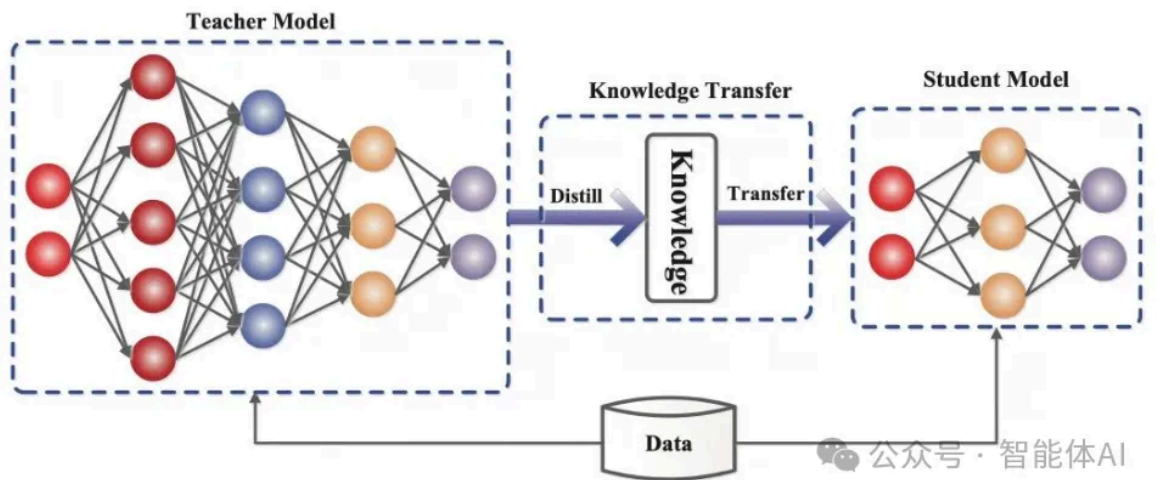
今天，我们正式发布 DeepSeek-R1，并同步开源模型权重。

- DeepSeek-R1 遵循 MIT License^Q，允许用户通过蒸馏技术借助 R1 训练其他模型。
- DeepSeek-R1 上线 API，对用户开放思维链输出，通过设置 `model='deepseek-reasoner'` 即可调用。
- DeepSeek 官网与 App 即日起同步更新上线。

公众号 · 智能体AI

上周，DeepSeek发布了其全新R1模型，一时间引发了AI圈的巨大轰动。这款国产模型在各类测试中表现亮眼，多个指标甚至直逼甚至超越OpenAI的o1系列，成为了行业的新标杆。消息一出，AI爱好者们纷纷涌向各大平台讨论R1模型的卓越表现，研究人员也开始深入研究其技术报告，试图解开其背后强大能力的秘密。

然而，在大家为R1模型的出色表现赞叹时，我却被困在了官方介绍的第二行——“蒸馏技术”上。这是个什么概念？作为一名AI从业者，我决定先好好补补课，理清这个技术的原理和实际应用。



一、什么是“蒸馏技术”？

知识蒸馏（Knowledge Distillation，简称KD）是由AI领域的三位大佬Geoffrey Hinton、Oriol Vinyals和Jeff Dean在2015年提出的技术，旨在通过将复杂教师模型的知识迁移到较简单的学生模型中，使学生模型在保持高性能的同时，能够实现更小的模型规模和更快的推理速度。

比喻：学生向老师学习



简单来说，蒸馏技术就像是学校里的学习过程：老师拥有丰富的知识和经验，学生通过学习老师的知识逐渐成长。AI中的“教师模型”相当于知识渊博的老师，它通过大量的数据训练，掌握了复杂的模式和特征；而“学生模型”则像是一个刚刚起步的学生，虽然知识面不如老师丰富，但通过学习老师的“思考方式”和“经验”，可以在没有过多计算资源的情况下取得不错的成绩。

具体应用：从图像识别到移动端应用

例如，在图像识别任务中，假设你有一张照片，要判断其中是“猫”还是“狗”。大模型（教师模型）能够准确判断，输出的结果可能是“猫”有80%的可能性，“狗”有10%的可能性，“其他动物”有10%的可能性。而小模型（学生模型）如果直接训练，可能很难达到这么高的精度，但通过蒸馏技术，它可以学习到大模型的判断方法，最终在手机等计算能力有限的设备上，也能够高效地进行图像识别。



二、蒸馏技术的基本原理

1. 教师模型与学生模型的准备

蒸馏的第一步就是准备好“教师模型”和“学生模型”。这就像是组织一场教学活动，需要有经验丰富的老师，也需要有充满潜力的学生。

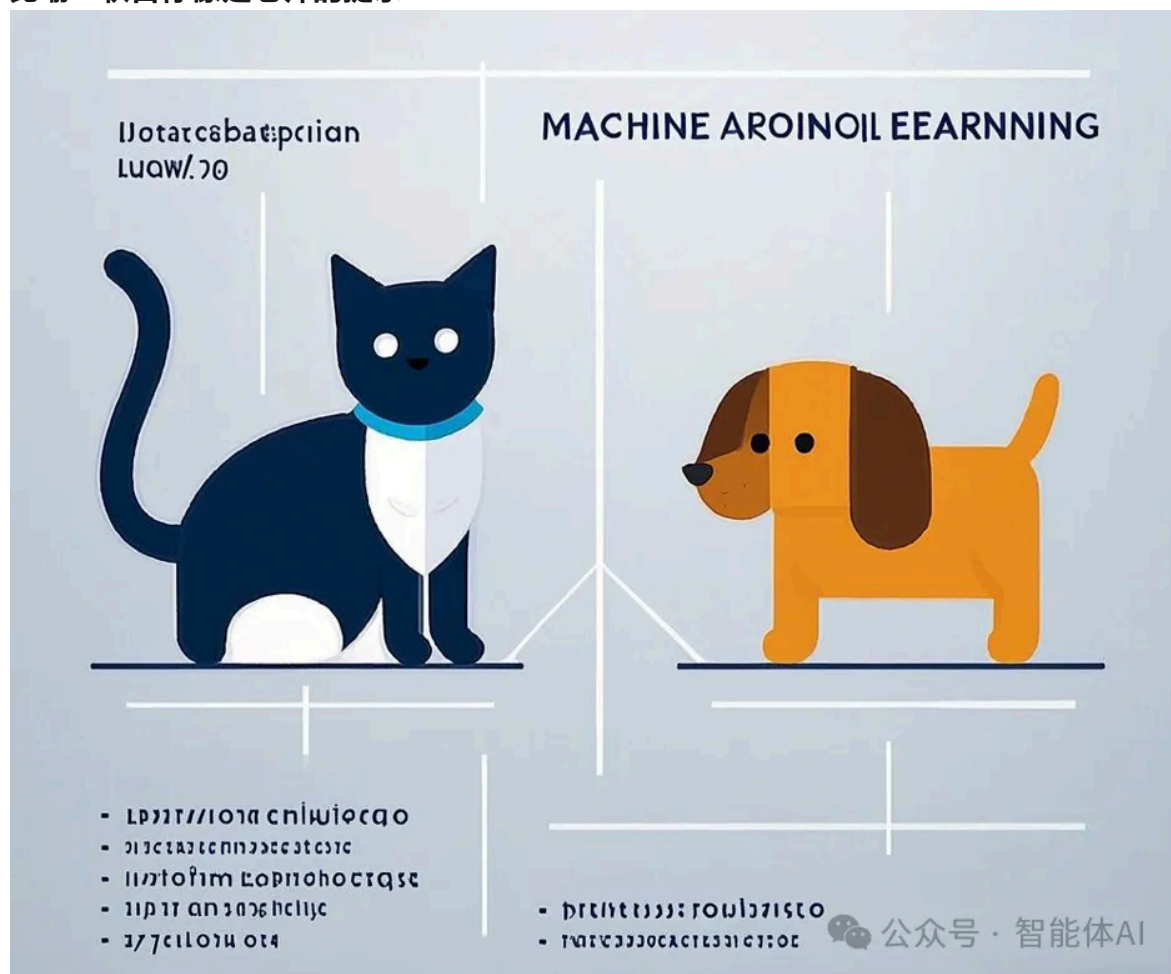
教师模型：教师模型是经过大规模训练的，通常是复杂且深度的网络模型，能够从大量的数据中提取和学习各种特征。例如，ResNet-101这类深度神经网络，通常被作为图像识别任务中的教师模型，它在大规模数据集（如ImageNet）上进行训练，能够准确识别图像中的细微差别。

学生模型：学生模型则是结构相对简单、参数较少的模型。假设教师模型有一百层，而学生模型可能只有十层；教师模型有几千个神经元，学生模型可能只有几百个。虽然学生模型没有教师模型那么强大，但通过蒸馏，它可以逐步学习到教师模型的知识 and 推理能力，最终能够在实际应用中发挥出色的性能。

2. 知识传递的过程

当教师模型和学生模型都准备好后，接下来就进入知识传递的关键阶段。在这一阶段，学生模型不仅要学习原始数据的标签（硬目标），还要学习教师模型的输出（软目标）。软目标与硬目标不同，它包含了更丰富的概率信息，帮助学生模型理解更多的细节。

比喻：软目标像是老师的提示



假设你在做一道题，正确答案是“猫”，这是硬目标。教师模型则给出了更多的信息：它认为这张图片是“猫”的概率是80%，是“狗”的概率是10%，剩下的可能性是其他动物。这个概率分布就是软目标，它帮助学生模型理解：即使我们知道这张图是猫，但也不能完全排除其他可能性。通过这种方式，学生模型不仅仅学习到“猫”的标签，而是学习到整个推理过程。

3. 损失函数与优化

为了确保学生模型能够尽可能接近教师模型的输出，我们需要使用一个损失函数，这个函数可以衡量学生模型和教师模型之间的差异。损失函数通常包含两部分：

- **KL散度 (Kullback-Leibler Divergence)：** 这个指标用来计算两个概率分布的差异。它衡量的是学生模型的输出和教师模型的输出之间的“距离”，目的是让学生模型尽可能模仿教师模型的输出。
- **交叉熵损失：** 用来衡量学生模型预测的标签和真实标签之间的差距。它通常用于分类任务中，表示预测值与真实值之间的误差。

通过不断调整学生模型的参数，使得损失函数最小化，学生模型会逐渐学习到教师模型的“智慧”，提升性能。

三、蒸馏技术的作用

1. 模型部署与计算资源优化

在实际应用中，很多设备（如智能手机、物联网设备等）具有有限的计算能力和内存。如果直接在这些设备上运行大模型，不仅速度慢，甚至可能因为内存不足导致无法运行。通过蒸馏，小模型能够在保证性能的基础上，减少计算资源的消耗，从而顺利运行在这些资源受限的设备上。

比喻：小模型如“迷你版”大模型

想象一下，你家里有一个“迷你版”的智能家居助手，它虽然体积小、功能简化，但却能完成所有你需要的任务。这就类似于小模型在保持一定准确度的基础上，能够高效地运行在手机、智能手表等计算资源有限的设备上。

2. 推理速度与能效

由于小模型参数较少，推理速度比大模型要快。在一些实时性要求极高的场景中，像自动驾驶系统、智能家居中的语音识别等，小模型能够迅速做出响应，减少延迟，保证用户体验。

比喻：小模型就像是节能灯泡

小模型的运行类似于节能灯泡，低能耗、高效率；而大模型则像是大功率的电暖器，消耗的电力要多得多。通过蒸馏，AI应用可以在性能不打折的情况下，大大降低能源消耗，尤其在电池驱动的设备上（如无人机、智能手表）尤为重要。

3. 实时决策与能效管理

在自动驾驶等领域，边缘计算设备需要实时处理大量数据。如果使用未经过蒸馏的大模型，推理速度可能不够快，导致无法及时响应。通过蒸馏，学生模型可以在有限的硬件资源下做出快速决策，保证行车安全。

四、总结

蒸馏技术在AI领域的应用，正如精妙的教学方法，它让“学生模型”能够在不具备大规模计算资源的情况下，模仿和学习“教师模型”的知识与推理能力，从而实现高效、快速的推理和决策。通过蒸馏技术，小模型能够在智能手机、智能手表、自动驾驶等多个应用场景中，提供流畅、高效的体验，同时降低计算资源和能耗的需求。

随着DeepSeek R1模型的发布，蒸馏技术又一次走到了前沿，成为国产AI技术创新的重要突破。在未来，蒸馏技术将持续推动AI应用的普及和创新，帮助更多设备和场景实现智能化和高效化。如果你对蒸馏技术或AI领域有更多疑问，欢迎留言讨论！



大模型 238

大模型 · 目录

上一篇

神秘的 DeepSeek：探索未知之境

下一篇

DeepSeek 带来的最大震撼：推理层面不再需要 SFT