工程与管理科学
Engineering and Management Science
第 7 卷◆第 7 期◆版本 1.0◆2025 年
文章类型：论文|刊号（ISSN）：2705-0637(P) / 2705-0645(O)

AI 大模型时代云计算与智能化融合路径

孔繁秋
宝信软件（云南）有限公司 650302

DOI:10.12238/ems.v7i7.14318

[摘 要] 为了更好地推动 AI 大模型在新时代背景下的训练与部署，满足复杂 AI 模型对计算资源和存储能力的高需求，我们需要深入挖掘云计算的核心优势，并将其与实际应用场景紧密结合。通过灵活调整资源规模，云计算不仅能够显著降低硬件设备部署方面的成本投入，还能有效解决行业发展中面临的各类业务挑战。本文立足于 AI 夤大模型与云计算的前沿技术领域，结合当前 AI 大模型训练与部署的实际需求，全面剖析了云计算与智能化深度融合的技术路径及其潜在价值。

[关键词]AI 大模型；云计算；智能化

一、AI 大模型

近年来，随着人工智能技术的迅猛发展，AI 大模型的崛起已成为不可逆转的趋势。特别是以 Transformer 架构为核心的深度学习模型，如 T5 和 BERT，凭借其庞大的参数规模，在语音合成、图像识别以及自然语言处理等多个领域展现出了卓越的性能表现。这些成就不仅重新定义了当前 AI 技术的发展格局，也为未来的创新应用奠定了坚实的基础。 从实际应用效果来看，AI 大模型相较于传统模型展现出更强大的泛化能力和学习能力。无论是在复杂任务还是多样化场景中，大模型均表现出显著的性能优势。例如，大规模图像模型能够大幅提升图像识别与生成的精度，而大语言模型则在文本生成方面实现了高度连贯性和自然性，为用户提供更加贴近真实人类表达的交互体验。进一步深入分析可知，AI 大模型的成功离不开科学的训练与部署策略。这一过程通常分为两个关键阶段：预训练和微调。AI 大模型以其卓越的技术特性和广泛的应用潜力，正在深刻改变我们的生活与工作方式。其背后的理论与实践成果无疑为未来 AI 技术的发展指明了方向，并将继续推动这一领域的边界不断拓展。通过利用 AI 大模型的构建形式，不仅能够显著减少模型开发过程中对海量数据的依赖，还能在小样本场景下展现出卓越性能。同时，在使用大量敏感数据进行训练时，还可能引发隐私泄露和安全隐患，这些问题对模型的实际部署提出了更高的要求。以一个简单的神经网络模型为例，我们可以更具体地探讨 AI 大模型的构建过程及其技术细节。首先，在模型结构设计中，分别设置输入层、隐藏层和输出层的神经元数量为 3 个、4 个和 2 个，并通过随机初始化权重和偏置参数完成初步配置。其次，在前向传播过程中，假设输入层的输入为 x，隐藏层的输出为 h，需要构建从输入层到隐藏层的权重矩阵，并确保该矩阵的维度与神经元数量相匹配。再次，针对输出层，设定其输入与输出的关系，进一步计算隐藏层到输出层的权重矩阵，同时调整隐藏层与输出层之间的偏置向量。这一系列步骤不仅体现了 AI 大模型的基本构建逻辑，也为后续优化和改进提供了理论基础。这种细致入微的设计方法不仅增强了模型的可解释性，还为解决实际应用中的复杂问题提供了有力支持。然而，这也提醒我们，在追求技术突破的同时，必须充分考虑资源消耗、隐私保护以及安全性等多方面因素，以实现 AI 大模型的可持续发展。假设输入层与隐藏层、隐藏层与输出层之间的激活函数分别为 ReLU 和 Softmax，则可分别用 $f(x)=\max(0,x)$ 与 $g(x)=\frac{e^{x}}{\sum_{m}^{n} e^{x}}$ 表示。最后，结合损失与学习率完成对输入层和隐藏层梯度下降的科学计算。

二、云计算

云计算是一种依托大量分布式计算机协同完成计算任务的先进技术，通过互联网提供动态、可扩展的计算资源与服务，构建灵活高效的网络计算环境。它能够对基础设施和应用程序进行虚拟化处理，并根据用户的具体需求量身定制相应的服务。这种计算模式不仅具备按需付费和高度灵活的特性，还为资源管理与商业模式创新提供了强有力的支持。作为一种革命性的技术手段，云计算不仅显著提升了服务质量与计算效率，还能以更低的服务器资源消耗和网络带宽占用，实现业务的高效部署与运营。其核心优势在于通过优化资源配置，大幅降低企业成本，同时支持业务的快速扩展与调整，从而助力产业转型升级。云计算主要涵盖三种服务模式：基

础设施即服务（IaaS）、平台即服务（PaaS）以及软件即服务（SaaS）。这三种模式各有侧重，但共同构成了一个全面、灵活的服务体系，为企业数字化转型和业务创新发展提供了坚实的技术支撑与战略保障。无论是在技术创新、资源整合还是商业模式优化方面，云计算都展现出了无可比拟的价值与潜力。基础设施即服务（IaaS）是一种创新的资源供给模式，它将网络、存储和计算等基础架构资源转化为一种公共产品。用户可以通过租用这些基础设施来完成软件测试、系统部署以及软件开发等一系列任务。与此同时，这种服务为用户的数据要素（Data Elements）提供了灵活的按需付费机制，并通过自动化与虚拟化管理手段，最大限度地满足用户的资源需求，从而显著降低硬件投入的成本。平台即服务（PaaS）则进一步将数据库、操作系统、编程语言及相关平台资源整合为一种服务形式。它以提供计算资源池为核心，帮助企业实现资源的高效调度与管理，从而为应用程序的开发、运行和部署提供强大的支持，极大地简化了企业的技术栈管理和运维负担。

三、云计算与智能化在 AI 大模型时代的融合路径

（一）高性能算力支撑

为了充分发挥云计算平台在训练与部署 AI 大模型方面的独特优势，深入践行“模型即服务”（MaaS）的理念，我们应充分利用 AI 大模型的能力，驱动产品和技术设计的革新，并构建完善的服务体系。在此基础上，结合 AI 大模型预训练的大规模特性及其对能力泛化的决定性影响，以行业需求和具体场景为导向，推动模型层次化结构的创新设计，开发面向多模态场景的高性能预训练大模型。这一举措将显著降低 AI 大模型在各行业领域的应用门槛，促进技术普惠。同时，以基础大模型为核心，进一步强化垂直领域模型的研发与优化，实现多端、高效、灵活的模型服务部署。通过构建全方位、立体化的服务机制，确保 AI 大模型的能力能够在不同终端设备上无缝运行，为用户提供一致且优质的体验。这种多层次、多维度的服务体系不仅能够满足多样化的需求，还将为社会各行业的数字化转型提供强有力的技术支撑，从而展现 AI 大模型在实际应用中的巨大潜力与价值。在人工智能大模型时代，云计算与智能化的深度融合已成为不可逆转的趋势。面对 AI 大模型通用化的发展方向及其对计算和存储资源的巨大需求，我们需要从平衡性能与成本的角度出发，进一步强化对模型压缩、模型剪枝以及模型蒸馏技术的应用。通过结合低比特量化、权重量化等方法，显著降低模型的计算复杂度与参数规模，从而大幅提升 AI 大模型训练与部署的效率。具体而言，可以通过在云端完成大模型的训练，再将经过蒸馏优化的小模型部署到边缘设备上运行，从而在保证性能稳定的同时显著减轻计算负担。研究表明，在不影响模型性能的前提下，AI 模型压缩与优化技术能够将模型服务所需的计算资源降至原来的 1/8，大幅节省了硬件成本与能耗。然而，构建支持大模型的高性能计算集群仍是一项极具挑战性的任务。为应对这一难题，可以通过提升计算系统总线性能以增强计算能力，并结合高效的存储技术确保数据交换的流畅性。通过综合运用数据缓存加速、多级存储以及 RDMA 网络技术，成功将端到端时延缩短至微秒级别，为云计算与智能化融合提供了强大的算力支撑。综上所述，通过技术创新与优化手段，我们不仅能够满足 AI 大模型对算力的需求，还能实现性能与成本之间的最佳平衡，为未来的智能化应用奠定坚实的基础。

（二）分布式模型训练

以高性能算力为基础，在推动云计算与智能化深度融合的过程中，需进一步强化对底层能力的充分利用，充分释放硬件计算潜力。在此背景下，借助云计算平台在大规模人工智能（AI）模型训练中的独特优势，将人工智能平台（Platform for AI，PAI）与高性能计算集群相结合，构建一套高效、灵活的分布式模型训练系统。在具体实施过程中，通过云计算平台为大规模 AI 模型训练提供强大的计算资源支持，将完整的模型训练任务分解为多个子任务，并分配至多个计算节点进行并行处理，从而实现任务的分布式执行。与此同时，同步协调各节点间的训练进程，确保整体任务的高效完成。在分布式模型训练的执行机制中，动态监控训练过程中的计算、内存及网络资源变化，实时调整执行策略。根据不同计算节点的性能特点和负载情况，合理分配数据样本，优化局部数据训练效率。最终，通过汇总各节点的训练结果，形成全局最优解，以高度优化的分布式训练方案高效处理超大规模数据集，显著缩短训练时间，大幅提升训练速度与效率。随着相关技术的不断创新与积累，通过对大量模型训练的实际运行考验，该系统已成功实现对具备 10 万亿参数规模的超大模型的高效训练，达到万卡级别的单任务分布式训练能力。其分布式加速比接近线性，展现出卓越的扩展性和性能表现，为未来更复杂的智能计算需求提供了坚实的技术保障。这一

成果不仅验证了系统的高效性与可靠性，更为人工智能领域的持续发展奠定了重要基础。

（三）低延时模型推理

在完成大规模人工智能模型的训练任务后，通过采用与训练阶段相同的分布式推理技术，可以实现大模型在云计算平台上的高效部署。这一过程从两方面着手：首先，将推理任务分解为多个子任务,并将其分布到多个计算节点上执行，从而显著提升推理效率；其次，充分利用云计算平台对大模型训练和推理的强大支持能力，结合分布式存储技术，实现模型参数和数据的快速读取与传输。以阿里云为例，其低延时推理模型和服务平台的应用展现了云计算与智能化技术深度融合的潜力。随着这两项技术的不断发展，阿里云充分挖掘了云计算的核心优势，在全球范围内建立了 29 个数据中心。在进行大模型推理时，阿里云能够根据业务用户的具体地理位置，智能选择最近的数据中心区域完成模型部署，从而确保服务的低延迟和高效率。此外，阿里云还为模型推理提供了自动扩缩容功能，通过实时监测低延时模型推理服务平台的负载变化，动态调整资源分配，进一步优化推理效率。值得一提的是,阿里云基于灵机模型服务和人工智能平台打造的低延时模型推理与服务平台，具备一键式模型部署的能力。这不仅突破了传统技术体系中寻找底层资源、上传模型等繁琐操作的限制,还让用户仅需一行代码即可轻松实现模型在云端的高效部署。这种简洁而强大的功能设计，不仅大幅降低了开发门槛，也为企业和个人开发者带来了前所未有的便利性，彰显了阿里云在人工智能领域的领先地位和技术实力。

（四）开放式服务平台

为了更好地满足不同业务的实际需求，并切实解决各类业务难题，我们以“模型即服务”（MaaS）理念为指导，结合云计算与智能化的深度融合，以优化用户体验为核心目标，构建了一个开放式的创新服务平台。通过这一平台，用户能够借助一站式模型社区，迅速筛选并获取适用于各类问题的高质量模型。例如，2022 年发布的魔搭社区便是这种开放式模型服务平台的典范。该社区不仅提供了涵盖自然语言处理、语音识别、图像分析等领域的预训练模型，还为开发者提供了丰富的开放性数据集，用于模型的进一步训练与调试。自发布以来，魔搭社区迅速吸引了大量 AI 模型开发者的关注，其模型下载量在短时间内突破了 1700 万次,充分证明了这些模型在实际开发过程中的高效应用能力。这些模型不仅在技术层面实现了突破，更在各行各业的实际场景中发挥了重要作用，真正实现了技术的价值辐射。随着魔搭社区内模型数量的持续增长，它已经成为推动云计算与智能化融合的重要引擎，为行业提供了大量优质的人工智能模型。同时，这一平台也为所有模型开发者提供了一个展示和验证成果的舞台。开发者可以将他们的研究成果上传至魔搭社区，并通过实际应用和技术评估，深入探索模型的技术潜力、商业化路径以及具体应用场景。这不仅促进了技术的交流与进步，也为人工智能的广泛应用开辟了新的可能性。通过魔搭社区的成功实践，我们可以清晰地看到，开放式服务平台不仅能够汇聚全球开发者的智慧，还能有效推动技术创新与产业应用的深度融合，为未来智能化社会的建设奠定坚实基础。

四、结束语

在人工智能大模型领域，通过深度融合智能化技术，显著提升了云计算的自适应能力和智能化水平。这种提升不仅强化了云计算对大模型在计算资源与存储能力方面的支撑作用，还为全球数字技术的快速发展注入了更强大的动力。这一进程充分证明，智能化技术与云计算的结合是推动数字时代创新突破的关键力量,也为未来技术演进奠定了坚实基础。

[参考文献]

[1]DanClement.人工智能和云计算[J].世界电子元器件，2022（7）：30-32.

[2]赵文丹.人工智能在大数据时代计算机网络技术中的应用[J].中文科技期刊数据库(全文版)自然科学,2022(10)：47-49.

[3]施巍松，孙辉，曹杰，张权，刘伟.边缘计算：万物互联时代新型计算模型[J].计算机研究与发展,2017,54(5)：907-924.

[4]林永青.人工智能起源处的“群星”[J].金融博览，2017，0（9）：46-47.

[5]施巍松，张星洲，王一帆，张庆阳.边缘计算：现状与展望[J].计算机研究与发展，2019，56（1）：69-89.

[6]崔雍浩，商聪，陈锶奇，郝建业.人工智能综述：AI 的发展[J].无线电通信技术，2019，45（3）：225-231.

[7]王鑫，韩振东，严斌峰.基于安全隔离度的 5G 专网部署模式[J].移动通信，2020，44（1）：44-47.

[8]方琰崴，李立平，陈亚权.5G 2B 专网解决方案和关键技术[J].移动通信，2020，44（8）：1-6.