当下人工智能技术正加速发展，渗透到云、边、端和应用的各个层面，与海量IoT设备进行深度融合，不断拓展应用场景。然而在AIoT场景中，嵌入式设备往往算力有限，难以承载庞大的AI模型。如何在资源有限的终端场景实现 AI模型的有效部署，是加速AI落地的重要问题。AI 工程师们研发了各种试图缩小模型大小并保持性能的办法，例如量化和蒸馏。其中，模型量化是将浮点计算转成低比特定点计算的一种模型压缩技术，可以有效减少模型算力消耗并提升计算速度，当前已经在工业界发展比较成熟。

目前相对成熟的模型量化方案是 INT8 量化。以ResNet-50 模型为例，原本需要用 float 32 表示的权重，量化后只需要使用 INT8表示，通过这样的处理，模型体积可以减少到原来的1/2，再加上 TensorCore 的加持，还会有近 8 倍的网络加速。而如果更进一步，将模型用INT4表示，可以带来更多的速度提升。

为了推动低比特量化技术的发展，旷视天元MegEngine 团队开源了 INT4 的源码实现，这也让MegEngine成为首个开源 CUDA INT4 源码实现的深度学习框架。MegEngine采用均匀线性量化方案，实现了非对称量化和对称量化两种INT4的数据类型，同时通过算子融合优化、kernel优化等方法，使得量化后的模型可以依然保持较高的精度以及良好的运行速度。同样以ResNet-50为例，INT4 相比 INT8 有 1.3倍的加速。

具体代码实现可以访问GitHub链接（https://github.com/MegEngine/examples）了解详情。

随着 CUDA INT4 的开源，目前MegEngine 框架不仅支持浮点数 FP32 和 FP16，而且支持 INT8 和 INT4 的对称和非对称量化推理。此外，MegEngine框架开发了诸多工具，帮助用户提升模型推理性能、简化部署流程，包括自动代码裁剪功能，支持用户全自动的针对算子进行代码裁剪；TracedModule 方案以及 MegEngine Lite，基于旷视海量业务打磨出的模型推理最佳实践，化解模型转换部署难题；流程管理工具FastRun,可以为每个计算自动选择最快的算法，从而保证整个网络的运行时间最短，让 MegEngine 用户运行不同的网络时都能收获最好性能。

自开源以来，MegEngine不断优化，已先后发布29个版本，推出一系列实用功能，降低AI算法生产门槛，助力AI应用快速落地。未来，旷视将继续支持和拥抱开源，并将自身在开源领域积累的技术和经验与业界共享，推动人工智能技术创新和行业发展。

本文源自中国经济网