

当前位置： 首页 > 新闻动态

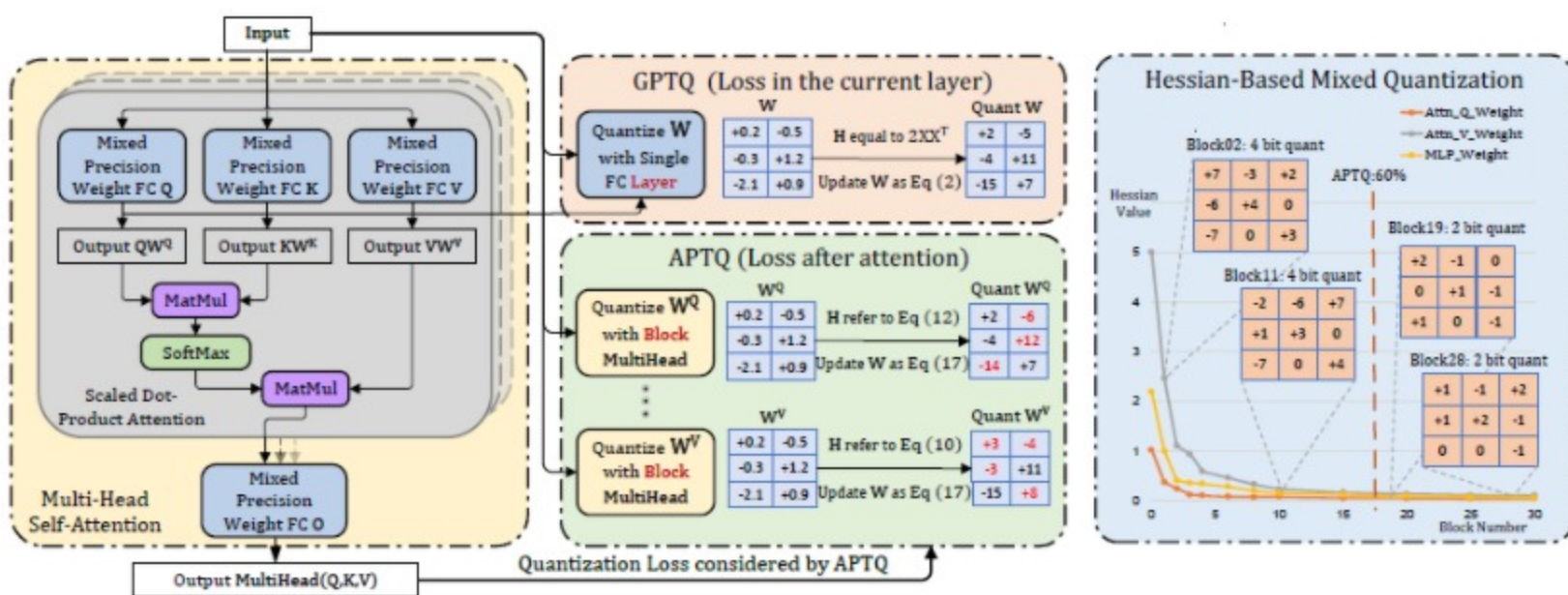
深港微电子学院交流博士生管子义在EDA顶级会议DAC上发表大语言模型量化工作

2024-02-28 综合新闻 浏览量: 2715

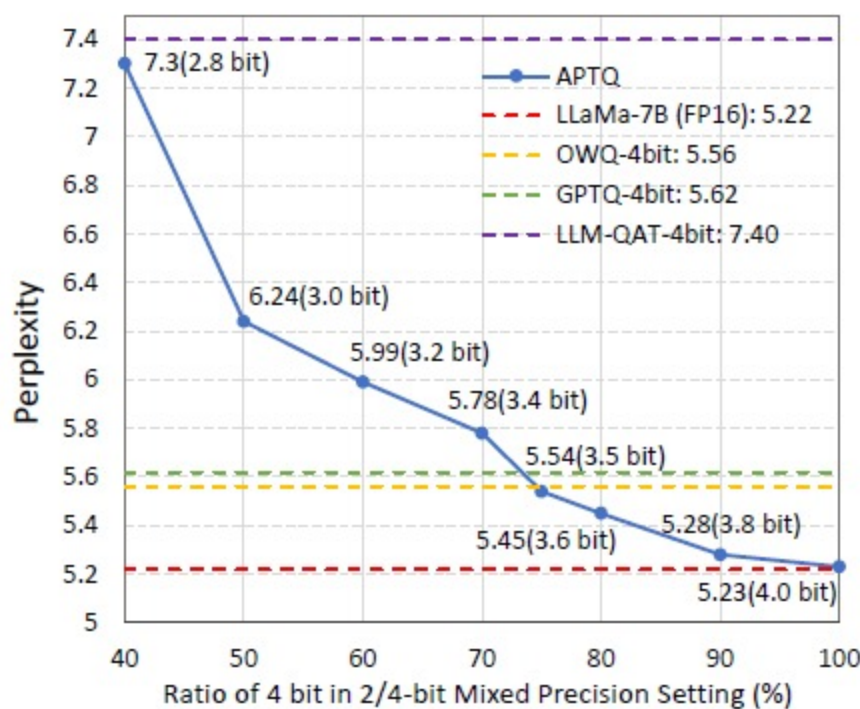
近日，南方科技大学深港微电子学院访问博士生，香港大学博士生管子义在EDA顶级会议DAC上发表了题为“APTQ: Attention-aware Post-Training Mixed-Precision Quantization for Large Language Models”的论文，该论文被EDA领域最顶级的会议国际设计自动化会议(Design Automation Conference, DAC 2024)接受。

DAC会议是集成电路辅助设计(EDA)和嵌入式系统领域的国际最权威会议，迄今已有超过50年的历史，主要关注芯片、电路以及系统设计的新工具和新方法。是中国计算机学会(CCF)推荐的计算机体系结构与高性能计算方向的A类顶级国际学术会议，该会议近两年的论文接收率为20%左右。

为了解决大语言模型LLM参数量过大难以部署的问题，本研究提出了APTQ(注意力感知的混合精度训练后量化技术)，这是一种创新的大型语言模型量化策略。APTQ技术不仅综合考虑了模型层权重的二阶信息，还首次纳入了注意力机制输出对模型整体的非线性影响，利用Hessian迹作为混合精度敏感度的评估指标，以实现在维持模型性能的同时对模型进行压缩量化。实验证明，APTQ在保持模型性能的基础上实现了平均4位宽度量化，并在C4数据集上达到了与全精度模型几乎相当的困惑度，同时在LLaMa-7B和LLaMa-13B模型上实现了在常识推理零样本学习任务中的高性能，证明了APTQ在量化大语言模型的鲁棒性。



图一展示了APTQ的总体架构，它将全面的Transformer注意力分析与特定层的Hessian量化结合起来，从而增强了对模型的理解。



图二比较了在C4数据集上使用APTQ的LLaMa-7B模型在不同4位比特量化下与其他量化方法的困惑度结果，证明了APTQ在混合精度设置中的优越性。

文章以南方科技大学为第一完成单位，深港微电子学院交流博士生管子义为第一作者。相关工作在深港微电子学院余浩教授课题组及香港大学黄毅教授课题组的指导下完成，并得到了深圳市科创委-深圳市高层次人才项目和深圳市科创委-基础研究重点项目的支持。

相关内容

[综合新闻](#) | [科研新闻](#) | [通知公告](#) | [学术讲座](#)

近年来，可穿戴生物传感器器件因其可动态、无创地监测表皮分泌物中的生物标志物引起了学界巨大的研究兴趣。...

闪耀暑期，深港微电子学院2024年集创赛捷报频传！

在这个充满挑战与机遇的暑期，深港微电子学院的同学们以卓越的创新能力和不懈的努力，在全国大学生集成电...

我院本科生曹宇恒在IEEE IWS 2024 荣获最佳学生论文奖

2024年 5月16-19日，无线通信领域重要国际学术会议IEEE 国际无线研讨会...

深港微电子学院方小虎课题组在2024年全国微波毫米波会议荣获最佳学生论文奖

2024年中国微波周于5月16日~19日在北京国际会议中心成功举办。本届微波周包括全国微波毫米波会议...

南方科技大学深港微电子学院-安森德半导体联合实验室揭牌仪式顺利举行

南方科技大学深港微电子学院成功举办首届深港澳实验室专业介绍会

