倪培洹

北京市 海淀区 中关村东路 95 号 中国科学院自动化研究所

158 0375 7631 ◇ nipeihuan24@mails.ucas.ac.cn ◇ peihuanni.github.io

教育经历

吉林大学 学士 2020 年 9 月 - 2024 年 6 月

电子科学与工程学院, 微电子科学与工程

CET4: 603, CET6: 553

中国科学院大学 硕士（研二） 2024 年 9 月 - 2027 年 6 月

人工智能学院, 人工智能，导师：李钢、程健

GPA: 3.81

发表论文

目前以一作身份发表论文 1 篇，在投 1 篇；共同作者身份发表论文 1 篇，在投 1 篇。

一作论文

Peihuan Ni, Zitao Mo, Tielong Liu, Hongli Wen, Zeyu Zhu, Minnan Pei, Junwen Si, Weifan Guan, Peisong Wang, Qinghao Hu, Gang Li and Jian Cheng. APEX: Integer-only Non-linear Function Approximation for Efficient Cross-Modal Inference. In **DATE**, 2026. **(CCF-B, EDA 顶会)**

- 我们发现不同模态的模型由于量化后数据分布的差异，导致其对非线性算子近似的敏感度不同，并且强制保留精度会导致位宽过高，提高硬件开销。
- 提出了一个高效的统一计算架构，融合了各种常用的非线性算子的计算图，并且通过 bit-level 剪枝的方式减少了计算位宽同时保留了精度。
- 精度方面：在语言模型上精度提高 0.7%，在视觉模型上精度提高 1.3%，并且在多模态模型中精度近乎无损。架构方面：相比于之前的方法，我们减少了 1.73-8.71× 的面积，同时使功耗降低 1.21-10.83×。

CoRL, 2027. **(在投，CCF-A，具身智能顶会)**

共同作者

Zeyu Zhu, Gang Li, Minnan Pei, Zitao Mo, **Peihuan Ni**, Peisong Wang, Tielong Liu and Jian Cheng. **KL-MoE: A Hierarchical MoE Pruning Framework Exploiting KL Divergence**. In **DAC**, 2026. **(CCF-A，EDA 顶会)**

In **TCAD**. **(在投，CCF-A，电子设计顶刊)**

In **ASPLOS**, 2026. **(在投，CCF-A，体系结构顶会)**

In **MICRO**, 2026. **(在投，CCF-A，体系结构顶会)**

项目经历

VLA 量化压缩和 FPGA 端侧部署 **主要完成人**

- 为了实现模型在片上的高效计算，负责设计了一种 W4A8（KV Cache 使用 8 bit 量化）的量化方法，并且使用纯整数的非线性近似方法来降低计算量，并且对精度近乎无损。

- 为了尽可能将所有运算在片上执行，减少数据搬运的开销，实现了基于 FP16-INT8 MAC 的量化反量化融合算子，在不牺牲精度的前提下尽可能保证激活值存在片上，从而提高推理效率。
- 为了提高数据复用和算子复用，并且与整体架构位宽对齐，复用 FP16-INT8 MAC，并且改变部分 RoPE 的计算顺序从而保证数据可以全部得到复用。
- 由于权重为 4 bit 量化，KV Cache 为 8 bit 量化，为了实现混合精度的矩阵乘法，设计了比特序列化的矩阵乘法器。同时为了配合比特序列化的设计，实现了高效的，同时支持大小模型不同注意力头维度的 KV 重排序模块以及 BRAM 到计算单元的 CrossBar 高效数据映射。
- 为了减少非线性单元的硬件按的开销，提高计算速度，设计了一种统一的纯整数非线性计算单元。并且在这个过程中发现了一些难题并提出相应的高效计算架构来解决，论文发表在 DATE 2026。
- 为了配合 VLIW 指令集，参与设计编译器的实现，并负责设计并实现高效的片上数据流以及片上的指令处理器和全部算子的控制模块。并且为了减少完成一次模型推理的指令数量，设计了各种支持地址偏移和偏移模式的控制器逻辑，将指令数量减少 32× 以上。
- 负责在 GPU 上实现了 Bit-Accuracy 的仿真器，保证片上数据的正确性和一致性。与 Vivado 上进行一个 Block 的仿真对齐，并且、与片上实际整体运行结果对齐。
- 负责调试 DMA 访存通路，高效利用 HBM 带宽，缓解 Decoding 过程中的 Memory-Bound 问题。
- 调试 VHK158 FPGA 开发板的数据通路调试，时序优化，petalinux 系统制作等工作，并将模型实际部署在片上。

奖项荣誉

竞赛获奖： 英特尔杯全国大学生电子设计竞赛邀请赛全国三等奖

荣誉奖项： 中国科学院大学三好学生