张泽欢

通信地址: 英国伦敦帝国理工学院南肯辛顿校区赫胥黎楼 347
邮箱: zehuan.zhang22@imperial.ac.uk

教育背景

帝国理工学院 | 伦敦 | 英国 04/2023 –on
专业：定制化计算研究
导师： Prof. Wayne Luk
课题组主页: http://cc.doc.ic.ac.uk

天津大学 | 天津 | 中国 08/2019 – 03/2022
专业：微电子学与固体电子学 | 硕士学位
导师：刘强教授

天津大学 | 天津 | 中国 09/2015 – 07/2019
专业：集成电路设计与集成系统 | 学士学位

研究方向

- 面向 AI 的高效硬件系统设计:
  - 可信 AI 算法的加速器设计
  - 复值神经网络的可重构硬件架构设计
  - 大语言模型的异构加速方法
- 面向高效硬件系统设计的 AI 算法设计：
  - 可重构加速器的软硬件协同设计
  - 基于大语言模型的硬件设计自动化

论文

First-author

- Hardware-Aware Neural Dropout Search for Reliable Uncertainty Prediction on FPGA
  Zehuan Zhang, Hongxiang Fan, Hao Mark Chen, Lukasz Dudziak, Wayne Luk.. ACM/IEEE Design Automation Conference (**DAC**).
- Accelerating MRI Uncertainty Estimation with Mask-based Bayesian Neural Network
  Zehuan Zhang, Matej Genci, Hongxiang Fan, Andreas Wetscherek, Wayne Luk.. IEEE International Conference on Application-specific Systems, Architectures and Processors (**ASAP**).
- Harnessing Heterogeneous Sparsity and Adaptable Acceleration for Complex-Valued Neural Networks
  Zehuan Zhang, Zhengyan Liu, Qiang Liu, Wayne Luk.. (Under Review)
- Extending Dropout-based Bayesian Approximation to Complex Domains: Mapping Bayesian Complex-Valued Neural Networks on FPGA
  Zehuan Zhang, Hao Mark Chen, He Li, Wayne Luk.. (Under Review)

Co-authored

- Advancing AI-assisted Hardware Design with Hierarchical Decentralized Training and Personalized Inference-Time Optimization
  Hao Mark Chen, **Zehuan Zhang**, Wanru Zhao, Nicholas Lane, Hongxiang Fan.. (Under Review)
- CODESCA: Co-Design for Spectral Clustering Acceleration
  Zhengyan Liu, Ce Guo, **Zehuan Zhang**, Wayne Luk.. (Under Review)

科研经历

推测解码异构架构加速研究 01/2025 – 至今

- （系统）提出一种新型异构架构，将基于 FPGA 的草稿生成与基于 GPU 的验证相结合，用于奖励引导的推测解码。该架构充分利用 FPGA 的可重构性与低功耗特性，同时发挥 GPU 的高计算能力。
- （算法）将回溯机制融入奖励引导推测解码中，能够重新审视早期预测结果，并优化逐步生成的中间结果，进而提升推理准确性。
- （硬件）通过高级综合（HLS）设计，将小规模草稿模型映射到 AMD V80 FPGA 上。所开发的加速器可同时处理“被接受” 与 “被拒绝” 的草稿候选结果；针对被拒绝的情况，硬件设计采用渐进式处理方案，从而减少大量冗余计算。

贝叶斯复值神经网络加速器的软硬协同设计 09/2024 – 11/2024

- 提出一种新型基于 Dropout 的贝叶斯复值神经网络，可为复值应用提供不确定性估计，增强模型预测的可靠性。所设计的复值 Dropout 方法可作为即插即用模块，展示高通用性，可广泛用于现有的复值神经网络模型。
- 设计了一种自动化搜索方法，能够在兼顾算法目标与硬件约束的前提下，有效地为复值神经网络的实部与虚部选择最优 Dropout 配置。与传统贝叶斯神经网络相比，贝叶斯复值神经网络的设计空间呈指数级扩展。所提出的搜索方法可高效地探索并确定最优配置。
- 设计一系列用于生成贝叶斯复值神经网络的基础硬件模块，并提出贝叶斯复值神经网络定制化加速器的设计框架，以实现加速器的自动化设计，并提高加速器的硬件性能。
- 生成的加速器性能与 GPU 的实现结果相比，运行速度提高 3 倍，能效提升 23.5 倍。

复值神经网络的剪枝和加速方法研究 12/2023 – 08/2024

- 设计实验验证在复值神经网络中，实部与虚部均可进行剪枝，但不同剪枝策略会导致模型性能的差异。
- 提出一种新的面向复值神经网络的异构剪枝技术。对模型权重的实部与虚部采用异构化的剪枝策略，使得模型在保持算法性能的同时，模型的稀疏率也显著提升。
- 针对复值的运算特点，设计了复值量化融合等多种优化方案，并集成至定制化的加速器中，有效实现了对异构剪枝的复值神经网络的高效加速。
- 实验表明，所提出的异构剪枝及硬件设计方案，在模型精度仅下降 0.63%的情况下，FLOPs 减少 51.23%。所设计的可重构加速器相比于 GPU 和 CPU 的运行结果，实现 6.7 倍与 12.2 倍的速度提升，同时显著提高了推理能效，展示了其在资源受限环境中的良好应用潜力。

基于 Dropout 的贝叶斯神经网络自动搜索研究 04/2023 – 11/2023

- 提出一种新的 Dropout 搜索框架，结合一轮式超网络训练与进化算法，能够根据具体应用的需求与约束，自动优化 Dropout 贝叶斯神经网络及其定制化的 FPGA 加速器设计。
- 为基于 Dropout 的贝叶斯神经网络构建逐层式的搜索空间，支持异构 Dropout 层的组合优化，从而提升模型的算法性能及硬件实现效率。
- 设计并实现四类 Dropout 方法的 FPGA 加速方案，支持多种 Dropout 组合的贝叶斯神经网络的高效推理。
- 实验验证所提出框架能够有效搜索出符合帕累托最优的模型配置。所生成的 FPGA 加速器相比于采用先进工艺的 CPU 和 GPU，分别提升 65 倍和 33 倍的能效，与现有的最先进的贝叶斯神经网络 FPGA 加速器相比，所提出的自动搜索方法生成的模型和硬件加速器在算法性能与能效方面均实现显著提升。

具备可靠的不确定性估计的 MRI 分析研究 09/2022 – 03/2023

- 提出了一种算法–硬件协同优化流程，可将深度神经网络转换为基于 Mask 的贝叶斯神经网络。该转换后的模型具备良好的硬件友好性，便于实现高效的硬件设计
- 将该设计流程应用于医学模型 IVIM-NET，构建出可输出不确定性信息的模型 uIVIM-NET，用于磁共振成像（MRI）分析，使得模型的使用者同时得到预测结果及对结果的自信程度估计。
- 针对 uIVIM-NET 设计了定制化 FPGA 加速器，并采用了跳零机制和批处理级优化策略，提升硬件性能。
- 实验表明，相比于 GPU 和 CPU 运行结果，在 Xilinx VU13P FPGA 上设计的加速器分别实现了 7.5 倍和 32.5 倍的速度提升，并降低了 34.4 倍和 82.8 倍的功耗。同时，该协同设计方法满足 MRI 分析的不确定性需求，提高了模型分析的可靠性。

荣誉奖项

DAC 国际会议 Young Fellow	2024
国家留学基金委博士奖学金（CSC）	2023 – 2027
天津大学一等学业奖学金	2020 – 2021
天津大学研究生二等入学奖学金	2019 – 2020
天津大学优秀毕业生	2018 – 2019
第十八届“天津大学学生科技英才”	2018 – 2019
天津市人民政府奖学金	2016 – 2017
天津大学科技创新先进个人	2016 – 2017
天津大学三好学生	2015 – 2016, 2016 – 2017
天津大学成绩优异先进个人	2015 – 2016

技能

编程语言: Python/Matlab/Verilog/C++
硬件描述语言：Verilog, HLS
深度学习框架: PyTorch, TensorFlow, Keras