更新于 2019 年 7 月 30 日

马凌霄

xysmlx@pku.edu.cn (邮件)
xysmlx@gmail.com (邮件)
博士四年级 · 北京大学分布式系统组　　北京市海淀区颐和园路 5 号北京大学 (地址)

研究方向：Systems for Machine Learning

机器学习系统， 大规模图计算， GPU 并行计算： 博士研究方向为利用现代高性能计算设备（例如：GPU）为大规模数据分析场景构建高性能并行计算系统，例如：深度学习、机器学习、大规模图计算

教育经历

- **北京大学**　信息科学技术学院 · 计算机系统结构 · 导师：代亚非 教授　2015.09 – 今　理学博士
- **北京师范大学**　信息科学与技术学院 · 计算机科学与技术　2011.09 – 2015.06　理学学士

实习经历

- **微软亚洲研究院（北京）– 系统组**　全职研究实习生　导师：薛继龙博士、伍鸣博士　2017.11 – 今

项目经历

- **NeuGraph (NGra)**　图神经网络计算系统 · 发表于 ***USENIX ATC'19*** · 第一作者　2017.12 – 今

  图神经网络 (Graph Neural Network, GNN) 具有对图节点之间依赖关系进行建模的强大功能，使得与图相关的研究领域取得了重大突破。由于图数据具有规模大、稀疏不规则的特性，GNN 超出了现有深度学习系统和图计算系统的设计，并导致在 GPU 上计算的 scalability 和 efficiency 的问题。NeuGraph 是面向大规模图神经网络的计算框架，它构建于现有深度学习系统 (如 TensorFlow) 之上，解决了 GNN 的 scalability 和 efficiency 的问题。
  - SAGA-NN 编程模型：以顶点为中心的数据流编程接口，让用户通过“像顶点一样思考”的方式来编写模型，使模型表达更加自然；其对图信息的描述为高效处理大图数据提供了可能
  - 通过“图感知”的 SAGA-NN 到子图粒度数据流图的翻译，和流式的子图粒度执行机制，解决 scalability 问题
  - 通过高性能图传播操作的 GPU kernel 以及 kernel fusion 机制，解决 efficiency 问题
  - 通过 chain 模式的多 GPU 流式执行机制，解决多 GPU 在 PCIe switch 上的带宽瓶颈问题
  - NeuGraph 与现有深度学习系统兼容，可自动将 SAGA-NN 程序翻译为现有深度学习系统可执行的数据流图
  - NeuGraph 在可放入 GPU 显存的小数据集上比 TensorFlow 最高快 5×；并可利用 GPU 直接处理大于 GPU 显存的大规模图数据集，单 GPU 比 TensorFlow-CPU 最高快 47×，并在多 GPU 上取得线性扩展性
  - 项目被新智元、机器之心等人工智能媒体报导; 项目在 2018 微软人工智能大会 (Microsoft AI Innovate) 和 2018 中国计算机大会 (CNCC) 进行了展示

- **Garaph**　*CPU-GPU* 异构图计算系统 · 发表于 ***USENIX ATC'17*** · 第一作者　2016.03 – 2017.03

  近年来，加速器和存储设备的快速发展为高效地在单个节点上处理大规模图数据提供了可能，而图数据的稀疏不规则特性使得在 CPU/GPU 上的图计算会产生线程冲突、随机访存、负载不均衡等问题。本项目提出 Garaph，一个面向 CPU/GPU 异构环境的大规模图计算系统，支持 CPU 和多 GPU 协同对大规模图数据进行高效处理。
  - GPU 图计算模块：通过保证显存的 coalescing memory access 提升显存访问速度，提出 replication 的策略解决 GPU 线程计算冲突和存储体冲突问题
  - CPU 图计算模块：基于边的平衡切分策略使得 CPU 端各个线程达到无锁计算和负载均衡，并针对不同图计算应用特性 (如：带有激活顶点特性的最短路算法) 综合切换双计算模式
  - CPU-GPU 动态调度策略：综合考虑任务特征和 CPU/GPU 的计算特征设计了动态任务调度策略
  - 对比基于 CPU 和基于 GPU 的大规模图计算系统，Garaph 比它们之中最快的系统最高快 5.36×
  - NASAC'17 受邀报告

- **DL Compiler**　深度学习 *Inference* 编译器 · 在投论文 · 第一作者　2018.10 – 今

  现有深度学习框架（如 TensorFlow）在 inference 场景中会存在很高且不稳定的 latency。我们针对该问题，设计了一个编译器，通过数据流图层面的计算优化取得低且稳定的 latency，在多个 CNN/RNN 模型上进行实验，与 TensorFlow 和 TensorRT 对比取得加速。

- **SeerNet** 稀疏卷积计算 · 发表于 ***CVPR'19*** · 第二作者 2018.10 – 2018.11

  SeerNet 关注卷积神经网络中输出特征图的稀疏性，例如，经过 ReLU 或 Max-pooling 层后，卷积层的大部分输出被置为零或丢弃。如果跳过这部分对应的先导卷积计算，则可以大大减少卷积层的计算量。
  - 利用低精度模型以极低的代价预测输出特征的稀疏性，然后通过稀疏的原精度计算加速原神经网络计算
  - SeerNet 可以直接应用于预训练好的模型中，无需对原始模型做任何修改或重训练
  - SeerNet 充分利用了现有硬件 (如 CPU、GPU) 对混合精度计算的支持
  - 基于 CPU 的实现在卷积层上取得了最高 5.79× 的加速，并在端到端的模型推理取得 1.2×∼1.4× 的加速

发表论文

[1] NeuGraph: Parallel Deep Neural Network Computation on Large Graphs
Lingxiao Ma, Zhi Yang, Youshan Miao, Jilong Xue, Ming Wu, Lidong Zhou, Yafei Dai
2019 USENIX Annual Technical Conference (**USENIX ATC'19**) (CCF A) (ATC'19 北大唯一)

[2] Garaph: Efficient GPU-accelerated Graph Processing on a Single Machine with Balanced Replication
Lingxiao Ma, Zhi Yang, Han Chen, Jilong Xue, Yafei Dai
2017 USENIX Annual Technical Conference (**USENIX ATC'17**) (CCF A) (ATC'17 北大第一单位唯一)

[3] SeerNet: Predicting Convolutional Neural Network Feature-Map Sparsity through Low-Bit Quantization
Shijie Cao, **Lingxiao Ma**, Wencong Xiao, Chen Zhang, Yunxin Liu, Lintao Zhang, Lanshun Nie, Zhi Yang
30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR'19**) (CCF A)

[4] CuWide: Towards Efficient GPU Training for Large-Scale Wide Model
Xupeng Miao, **Lingxiao Ma**, Yingxia Shao, Bin Cui, Zhi Yang, Jiawei Jiang, Lele Yu
Submitted to **VLDB'20** (single-blind peer-review, under one shot revision) (CCF A)

申请专利

[1] 一种基于自动选择副本因子模型的图计算方法. 201710533444.5

主要奖励

- **北京大学优秀科研奖** 2018.12, 2017.12
- **北京大学博士研究生校长奖学金** 2017.06
- **北京师范大学优秀毕业生** 2015.05
- **北京师范大学第 12 届励耘优秀本科生奖学金** 全校 *6* 人 2014.12
- **国家奖学金** 2014.10

竞赛获奖

- **第 39 届 ACM/ICPC 国际大学生程序设计竞赛亚洲区域赛鞍山站** 银牌 2014.10
- **美国数学建模竞赛** 一等奖 *(Meritorious Winner)* 2014.02
- **全国大学生数学建模竞赛北京赛区** 一等奖 2013.10

专业能力

- **编程语言**: C, C++, CUDA, Python, LaTeX, Markdown, Java, Shell
- **熟悉领域**: 面向 GPU 和多核环境的并行编程, 大规模图计算, 机器学习, 分布式系统, 数据结构与算法
- **熟悉系统**: TensorFlow, TVM, TensorRT, PyTorch, PowerGraph

更多个人信息请访问: https://xysmlx.github.io/