夏 **春伟**



求职意向：面向移动端的深度学习优化 · 深度学习编程与编译

北京市海淀区中关村科学院南路 *6* 号，中科院计算所

(+86) 17600100408 | [xiachunwei@ict.ac.cn](mailto:xiachunwei@ict.ac.cn) | 研究方向：面向深度学习的端云协同推理编程与优化

# Education

## 中国科学院 计算技术研究所 计算机体系结构国家重点实验室 北京

专业：计算机系统结构 课题组：编译与编程 导师：冯晓兵、崔慧敏 博士在读 *2016.09 -* 至今

* 2019.10 所级奖励 华为奖学金 | 2018.10 校级奖励 中国科学院大学三好学生

**天津大学** 天津

专业：计算机科学与技术 学士 · 英语:CET4(606), CET6(519) *2012.09 - 2016.07*

* 2015.10 天津大学三好学生 | 2014.10 国家励志奖学金 | 2013.10 天津大学三好学生 | 2013.10 天津大学优秀学生干事

**Skills**

**科研** 了解常用DNN 模型，熟悉典型移动端算法优化与系统算法，熟悉国内国际端云协同方法与移动端benchmark

**开发** 熟悉典型移动端AI 推理框架（TF-Lite、MNN、MACE）与AI 加速器硬件架构（Mobile GPU、Edge TPU、Cambricon NPU 等）

**编程** 熟练使用C++, JAVA, Python, LaTeX，熟悉OpenCL 并行编程，熟悉Android Native 开发与优化

# Research

## 异构融合的移动端AI 推理加速 北京

优化难：自动子图映射与调度 *2019.04-*至今

* 问题：移动端集成多种异构加速芯片，如何充分利用异构芯片在功耗敏感的情况下加速AI 推理
* 难点：（1）移动设备巨大性能差异（2）日趋复杂的 DNN 负载（3）轻量级异构融合的并行执行
* 设计：（1）基于权值的子图划分（2）基于线性规划和贪心的算子分配与调度（3）Cache-aware 的优化和轻量级调度
* 结果：典型DNN 模型（Inception-V3/V4，Pnasnet-large）等相比State-of-Art 最多 1.65× 加速（2 计算单元）
* 专利一篇：在异构处理单元上执行深度神经网络的方法，申请号：202010493830.8
* **论文 1：《HOPE: A Heterogeneity-Oriented Parallel Execution Engine for Inference on Mobiles》投稿至CGO21（CCF-B)**
* **论文 2：《On Re-targeting the AI Programming Framework to New Hardwares》，NPC18（CCF-C，第四作者)**

## 面向端云协同AI 推理的DNN 调优框架 北京

编程难：自动选择最优配置 *2017.06 - 2019.04*

* 问题：DNN 在移动端和云端协同的推理成为DNN 部署新方法
* 难点：（1）如何准确的测量延迟和功耗（2）如何切分 DNN 负载最小化延迟和功耗（3）如何选择最优的软硬件配置
* 设计：（1）动态的 layer-wise profiling 机制（2）自动调优框架：DNNTune 自动发现最优切分点和最佳软硬件配置
* 结果：（1）端云协同最多 1.66× 加速比、节省 15% 能耗
* 结果：（2）刻画各种 DNN（CNN/LSTM/MLP) 在各种平台（高中低、CPU/GPU/AI、手机/嵌入式）各种配置的性能
* **论文 3：《DNNTune: Automatic Benchmarking DNN Models for Mobile-cloud Computing》ACM-TACO (CCF-B)，HiPEAC 会议报告**
* **DNNTune 成为过去一年ACM TACO 下载量最高的论文(截止目前）。**
* **论文 4：《Characterizing DNN Models for Edge-Cloud Computing》(poster)，发表在IISWC-2018（Benchmark 旗舰会议）**

## Web 查询的分类算法 天津

CIKM-2015 COMPETITION TOP-20 *2014.10 - 2015.10*

* **论文 5：基于百度搜索引擎数据集的图模型的Web 查询分类算法,《Graph-Based Web Query Classification》WISA-2015（EI 检索）**

# Projects

## 面向异构融合数据流加速器的编程模型及编译器优化-国家重点研发计划子课题 北京

主要参与人员，负责DNN 模型在异构融合加速器的算子映射和任务调度 *2017.12 - 2019.12*

## 张量计算异构处理框架验证项目 北京

主要参与人员，负责将原本在CPU 上使用MKL 加速的FFT 应用，迁移到NVIDIA JETSON 嵌入式开发板，验证ARM

CPU+ NVIDIA GPU 的异构加速效果

*2018.04-2018.06*

## 5G RCA 应用在大数据处理框架的评估建模 北京

主要参与人员，通过在SPARK 框架中代码插桩，来测量RCA 应用在SPARK 上各个阶段的延迟 *2016.02-2016.08*

## 国家大学生创新创业训练项目（国家级）：基于信息融合的短时交通流速度预测 天津

项目负责人，基于天津市真实的出租车交通流数据，预测在短时间内的道路车辆运行速度与拥堵状况 *2014.05-2015.05*

SEPTEMBER 18, 2020 夏春伟个人简历 1