中科院计算所夏春伟论文信息：

1. **HOPE: A Heterogeneity-Oriented Parallel Execution Engine for Inference on Mobiles**

论文详情：**第一作者**，目前已投稿至**CCF-B类会议（CGO21）**，正在评审中 针对DNN模型中普遍存在的分支并行的特性，我们提出了面向移动设备的异构并行执行引擎，可以在多个异构设备自动的并行执行DNN的模型以加速DNN的推理。实验结果相对目前最好的工作能达到最高1.6倍加速比。

1. **DNNTune: Automatic Benchmarking DNN Models for Mobile-cloud Computing**

论文详情：**第一作者**，**CCF-B类会议+期刊（TACO +HiPEAC20），**已经录取 面向端云协同的AI推理新方法，针对如下几个问题：（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，提出了（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。最后得出端云协同最多1.66x加速比、节省15%能耗

1. **On Retargeting the AI Programming Framework to New Hardwares**

论文详情：**第四作者**，**CCF-C类会议（NPC18），**已经录取 针对多样化的AI设备，我们提出了一种将AI编程框架移植到不同的AI加速器的方法。我们以TensorFlow为例，将其迁移到了神威超算平台和FPGA上，实验结果证明了我们方法的有效性。

1. **Characterizing DNN Models for Edge-Cloud Computing**

**第一作者，IEEE收录（IISWC poster）** benchmark领域旗舰会议， 我们刻画了各种DNN（包括CNN、MLP、LSTM）在移动端和云端平台的性能和功耗。

1. **Graph-Based Web Query Classification**

**第一作者，EI检索** 基于百度搜索引擎数据集提出了图模型的Web查询分类算法