Low-Latency Proactive Continuous Vision

Yiming Gan
Department of Computer Science, University of Rochester

with
Yuxian Qiu, Shanghai Jiao Tong University
Lele Chen, University of Rochester
Jingwen Leng, Shanghai Jiao Tong University
Yuhao Zhu, University of Rochester
Continuous Vision: Long Frame Latency
Bottleneck: Serialization

Light -> Sensor -> Raw Pixels
Bottleneck: Serialization

Light → Raw Pixels → RGB
Bottleneck: Serialization

Light → Raw Pixels → RGB → Results
Proactive Pipeline

Frame 1

Sensing  Imaging  Vision
Proactive Pipeline

Latency

Frame 1
- Sensing
- Imaging
- Vision

Frame 2
- Sensing
- Imaging

Frame 3
- Sensing
- Imaging
Proactive Pipeline

Frame 1
- Sensing
- Imaging
- Vision

Frame 2
- Sensing
- Imaging
- Chek

Frame 3
- Sensing
- Imaging
- Chek
Proactive Pipeline

Frame 1
- Sensing
- Imaging
- Vision

Frame 2 Fail Check
- Sensing
- Imaging
- Chek
- Vision

Frame 3
- Sensing
- Imaging
- Chek
Proactive Pipeline

Frame 1
- Sensing
- Imaging
- Vision

Frame 2  Fail Check
- Sensing
- Imaging
- Chek
- Vision

Frame 3  Pass Check
- Sensing
- Imaging
- Chek
Challenges

Frame 1
Sensing | Imaging | Vision

Frame 2
Fail Check
Sensing | Imaging | Chek | Vision

Frame 3
Pass Check
Sensing | Imaging | Chek

Latency
Resource Contention
Solutions

Snapdragon 675

CPU
4th gen Kryo CPU
Performance @ 2.0GHz
Efficiency @ 1.7GHz

Artificial Intelligence
3rd gen Qualcomm® AI Engine

Modem
Snapdragon X12
Cat 12 DL, up to 600 Mbps

Audio
Qualcomm Aqstic™ audio
Qualcomm® aptX™ audio

Display
Up to FHD+ display

GPU
6th gen Adreno GPU

DSP
6th gen Hexagon DSP
DSP Security

Camera
2nd gen Qualcomm Spectra ISP
Up to 25 Megapixels @ 30fps ZSL
48 Megapixels snapshot

Charging
Qualcomm® Quick Charge™ 4+ technology

Commercial devices expected in Q1 2019

Qualcomm AI Engine and Qualcomm Quick Charge are products of Qualcomm Technologies, Inc. and/or its subsidiaries.
Solutions

Snapdragon 675

CPU
4th gen Kryo CPU
Performance @ 2.0GHz
Efficiency @ 1.7GHz

Artificial Intelligence
3rd gen Qualcomm® AI Engine

Modem
Snapdragon X12
Cat 12 DL, up to 600 Mbps

Audio
Qualcomm Aqstic™ audio
Qualcomm® aptX™ audio

Display
Up to FHD+ display

GPU
6th gen Adreno GPU

DSP
6th gen Hexagon DSP
DSP Security

Camera
2nd gen Qualcomm Spectra ISP
Up to 25 Megapixels @ 30fps ZSL
48 Megapixels snapshot

Charging
Qualcomm® Quick Charge™ 4+ technology

Commercial devices expected in Q1 2019
Challenges

Latency

Frame 1: Sensing, Imaging, Vision

Energy Wasting

Pred, Vision

Vision

Frame 2: Fail Check, Sensing, Imaging, Chek, Vision

Frame 3: Pass Check, Sensing, Imaging, Chek
Solutions

- Relaxing Checking Criterion (Threshold $T$)
Solutions

- Relaxing Checking Criterion (Threshold T)
- Relaxing Checking Frequency (Degree K)
Frames Sequence

Precise Frames  Unchecked-Predicted Frames  Checked-Predicted Frames

Predicted Sequence

Time
PVF Framework

Static

Vision Apps

Similarity T

Degree K

Dynamic

SoC

CPU

NPU

DSP

GPU

ISP

Memory

Sensor

BUS

Accuracy Target

Similarity Metric etc.
PVF Framework

Static

Vision Apps

Similarity T

Degree K

Dynamic

SoC

CPU

NPU

DSP

GPU

ISP

Predictor

BUS

Memory

Sensor
PVF Framework

Static

Vision Apps

Similarity $T$

Degree $K$

Dynamic

SoC

Runtime

CPU

NPU

DSP

GPU

ISP

Predictor

Memory

Sensor

BUS
PVF Framework

Static

Vision Apps

Similarity T

Degree K

Dynamic

Runtime

Scheduler

Control

Checking

SoC

CPU

NPU

DSP

GPU

ISP

Predictor

Memory

Sensor

BUS
Experimental Setup

I. In house simulator modeling state-of-the art SoCs
   • Real measurement of latency and energy on different IPs.
Experimental Setup

I. In house simulator modeling state-of-the art SoCs
   • Real measurement of latency and energy on different IPs.

II. RTL Implementations for NPU and Predictor
    • 20x20 Systolic Array for NPU, 10x10 Systolic Array for Predictor
Experimental Setup

I. In house simulator modeling state-of-the art SoCs
   • Real measurement of latency and energy on different IPs.

II. RTL Implementations for NPU and Predictor
    • 20x20 Systolic Array for NPU, 10x10 Systolic Array for Predictor

III. Evaluate on Object Detection and Tracking
    • KITTI dataset for object detection, VOT-challange for tracking.
Experimental Setup

I. In house simulator modeling state-of-the art SoCs
   • Real measurement of latency and energy on different IPs.

II. RTL Implementations for NPU and Predictor
    • 20x20 Systolic Array for NPU, 10x10 Systolic Array for Predictor

III. Evaluate on Object Detection and Tracking
     • KITTI dataset for object detection, VOT-challange for tracking.

IV. Different Input Resolutions
Baselines

I. Base
   • Baseline with traditional execution pipeline

II. BO
   • Baseline with optimized back-end

III. FCFS
   • Traditional pipeline with multiple hardware IPs
Results

Latency Reduction (%)

Energy Budget (mJ)
Results

Latency Reduction (%)

Energy Budget (mJ)

PVF

Better
Results

Latency Reduction (%)

Better

Energy Budget (mJ)
Conclusion

I. Long Latency Bottleneck Continuous Vision

II. Proactive Execution Pipeline
   1) Leveraging Heterogeneities in Mobile SoCs
   2) Relaxed Checking

III. Non-mission-critical System
Collaborators

Yuxian Qiu  Jingwen Leng  Lele Chen  Yuhao Zhu
Questions