面向移动与边缘设备的人工智能系统

AITime分享 徐梦炜

2020年8月21日

自我介绍 – 徐梦炜

北京大学 PEKING UNIVERSITY

- **博士 – 计算机系，北京大学**
  - 指导老师：黄罡 教授，刘譞哲 副教授
- **本科 – 计算机系，北京大学**
- **联合培养博士 – 普渡大学**
  - 指导老师：Prof. Felix Lin
- **访问学生 – 系统组，微软亚洲研究院**
  - Mentor: 刘云新 研究员

2015年9月 – 2020年6月

2011年9月 – 2015年6月

2018年11月 – 2019 年11月

2015年3月 – 2016年3月

- **研究方向:** 移动与边缘计算
  - 移动与边缘设备上的人工智能系统
  - 主页：https://xumengwei.github.io/

Background

- The increasing attention on **AI systems**

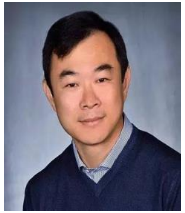
Ion Stoica
Berkeley

Michael I.J.
Berkeley

Jeff Dean
Google

Fei-Fei Li
Stanford

Yann Lecun
Facebook



Eric Xing
CMU

The 1st SysML conference in 2018

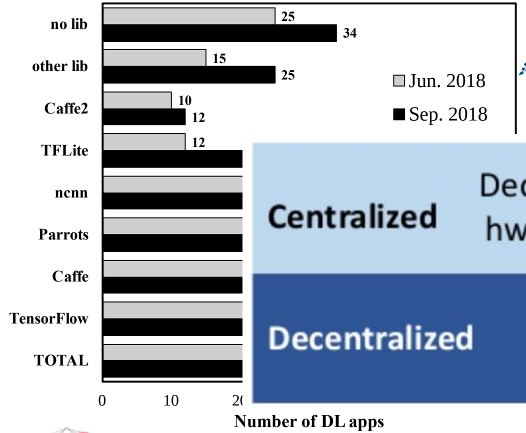
A Berkeley View of Systems Challenges for AI

Ion Stoica, Dawn Song, Raluca Ada Popa, David Patterson, Michael W. Mahoney, Randy Katz, Anthony D. Joseph, Michael Jordan, Joseph M. Hellerstein, Joseph Gonzalez, Ken Goldberg, Ali Ghodsi, David Culler, Pieter Abbeel*

The opportunities and challenges in the new AI era for systems

Background

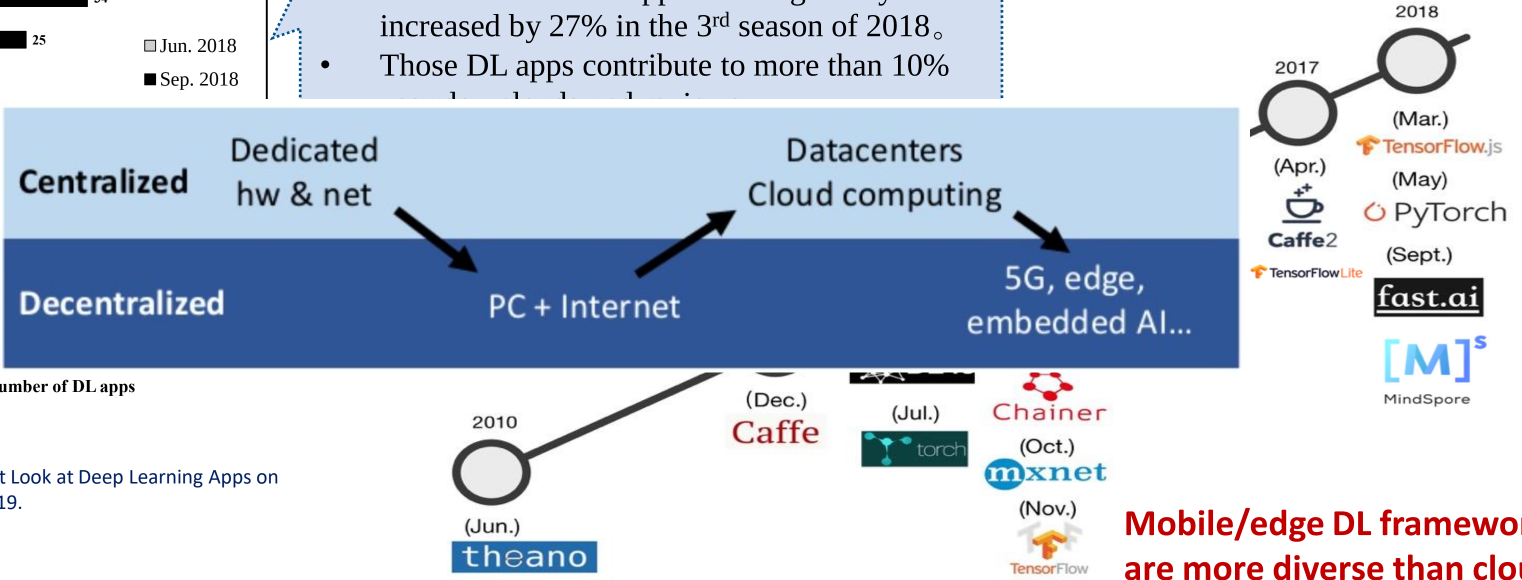
- The increasing attention on (mobile and edge) **AI systems**



THE WEB CONFERENCE

M. Xu, X. Liu, et al. A First Look at Deep Learning Apps on Smartphones. WWW 2019.

- The number of DL apps on Google Play has increased by 27% in the 3rd season of 2018。
- Those DL apps contribute to more than 10%



Mobile/edge DL frameworks are more diverse than cloud!

Research Overview

- **Supporting DL on smartphones**
  - CNN Cache to reduce inference time/energy (MobiCom 2018)
  - On-device training for input personalization (UbiComp 2018)
  - The first empirical study on smartphone DL apps (WWW 2019)
  - Adaptive Local Offloading for On-Wearable DL (TMC 2019)

Research Overview

- **Supporting DL on smartphones**
  - CNN Cache to reduce inference time/energy (MobiCom 2018)
  - On-device training for input personalization (UbiComp 2018)
  - The first empirical study on smartphone DL apps (WWW 2019)
  - Adaptive Local Offloading for On-Wearable DL (TMC 2019)

- **More efficient federated learning**
  - Heterogeneity-aware, automatic architecture search (arxivs)

Research Overview

- **Supporting DL on smartphones**
  - CNN Cache to reduce inference time/energy (MobiCom 2018)
  - On-device training for input personalization (UbiComp 2018)
  - The first empirical study on smartphone DL apps (WWW 2019)
  - Adaptive Local Offloading for On-Wearable DL (TMC 2019)

- **More efficient federated learning**
  - Heterogeneity-aware, automatic architecture search (arxivs)
- **Camera-centric Video Analytics**
  - Enabling video query on autonomous cameras (MobiSys'20)
  - Approximate video query on zero-streaming cameras (arxiv)

Research Overview

- **Supporting DL on smartphones**
  - CNN Cache to reduce inference time/energy (MobiCom 2018)
  - On-device training for input personalization (UbiComp 2018)
  - The first empirical study on smartphone DL apps (WWW 2019)
  - Adaptive Local Offloading for On-Wearable DL (TMC 2019)

- **More efficient federated learning**
  - Heterogeneity-aware, automatic architecture search (arxivs)
- **Camera-centric Video Analytics**
  - Enabling video query on autonomous cameras (MobiSys’20)
  - Approximate video query on zero-streaming cameras (arxiv)

Video Analytics is a Killer App 🔥

北京大學 PEKING UNIVERSITY

- Busy cross roads
- Retailing store
- Sports stadium
- Parking lots
- …

Video Analytics is a Killer App

北京大學 PEKING UNIVERSITY

- Busy cross roads
- Retailing store
- Sports stadium
- Parking lots
- …

Urban, residential areas

✓ Wired electricity
✓ Good internet

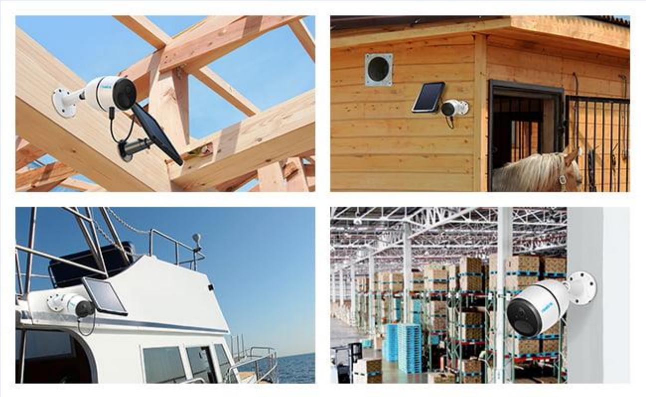
Video Analytics is a Killer App

- Busy cross roads
- Retailing store
- Sports stadium
- Parking lots
- …




Urban, residential areas

?



- Construction sites
- Cattle farms
- Highways
- Wildlifes
- …

Rural, off-grid areas

Autonomous Camera

- **Energy-independent** and **Compute-independent**

Autonomous Camera

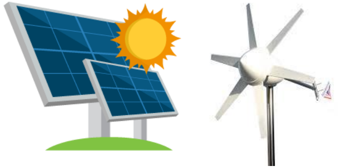
- **Energy-independent** and **Compute-independent**

Commodity SoCs, RPI-like, chargeable battery

Autonomous Camera

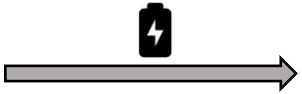
北京大学 PEKING UNIVERSITY

- **Energy-independent** and **Compute-independent**



Small-sized
energy harvester

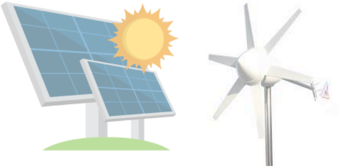
e.g., “10Wh today”



Commodity SoCs, RPI-
like, chargeable battery

Autonomous Camera

- **Energy-independent** and **Compute-independent**



Small-sized
energy harvester

e.g., “10Wh today”

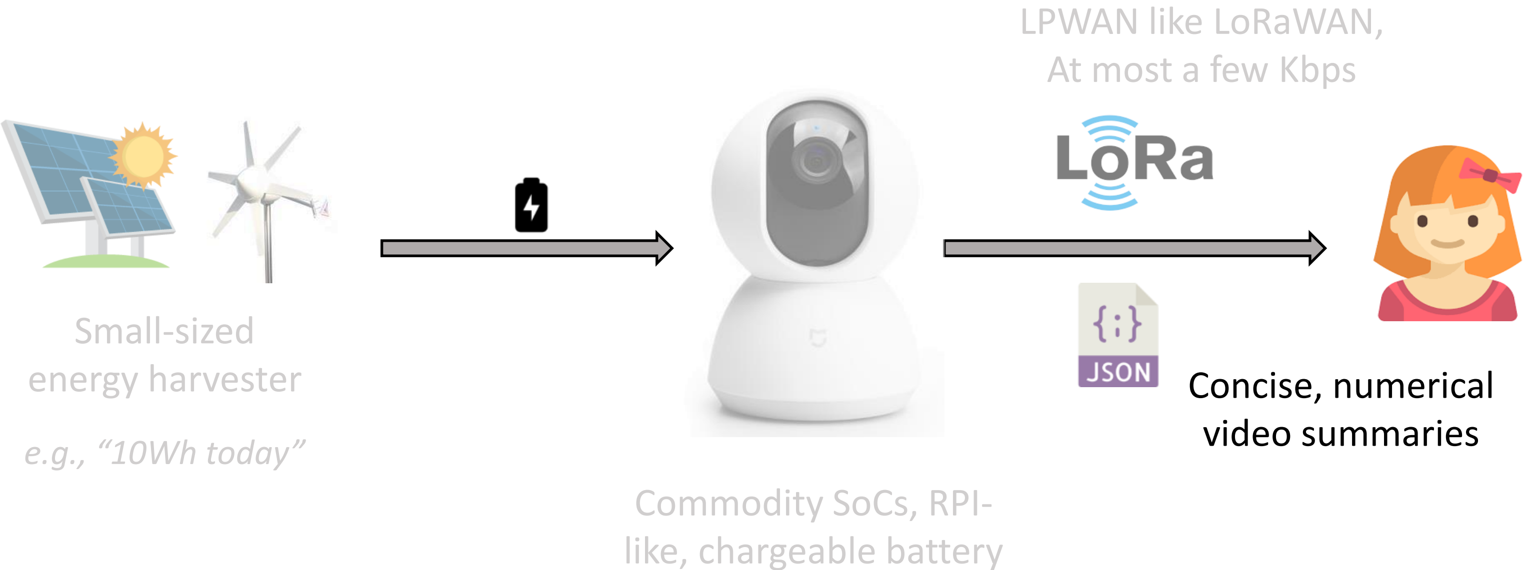
Commodity SoCs, RPI-
like, chargeable battery

LPWAN like LoRaWAN,
At most a few Kbps

LoRa

Autonomous Camera

- **Energy-independent** and **Compute-independent**



Elf for Autonomous Cameras

- Target video query: **object counting**

Elf for Autonomous Cameras

北京大学 PEKING UNIVERSITY

- Target video query: **object counting**

Query: (car, 30 mins)

Install

Elf for Autonomous Cameras

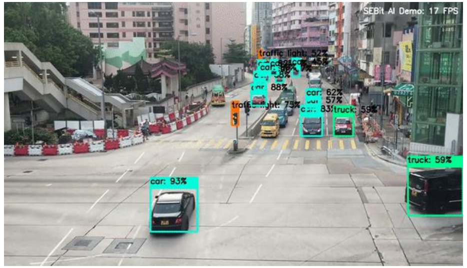
北京大學
PEKING UNIVERSITY

- Target video query: **object counting**

Query: (car, 30 mins)

Install

Sample & capture



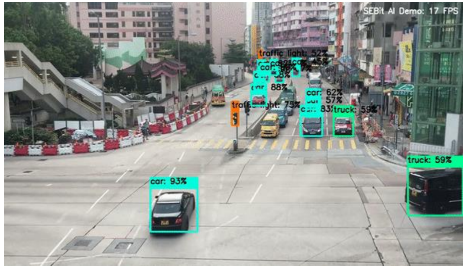
PEKING UNIVERSITY

- Target video query: **object counting**

Query: (car, 30 mins)

Install

Sample & capture



7:00AM-7:30AM	[500 ± 100] Cars
7:30AM-8:00AM	[700 ± 140] Cars
8:00AM-8:30AM	[800 ± 180] Cars
8:30AM-9:00AM	[400 ± 100] Cars
9:30AM-10:00AM	[200 ± 80] Cars

Elf for Autonomous Cameras

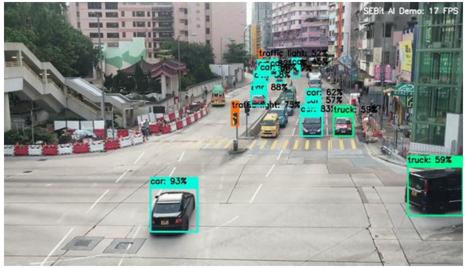
北京大學 PEKING UNIVERSITY

- Target video query: **object counting** with confidence interval (CI)

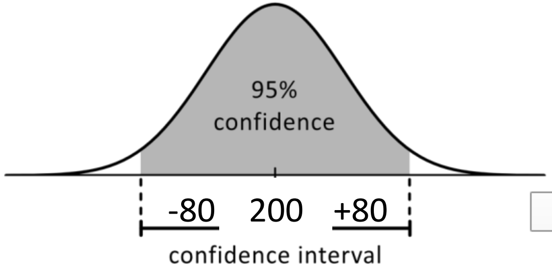
Query: (car, 30 mins)

Install

Sample & capture



7:00AM-7:30AM	[500 ± 100] Cars
7:30AM-8:00AM	[700 ± 140] Cars
8:00AM-8:30AM	[800 ± 180] Cars
8:30AM-9:00AM	[400 ± 100] Cars
9:30AM-10:00AM	[200 ± 80] Cars



Elf for Autonomous Cameras

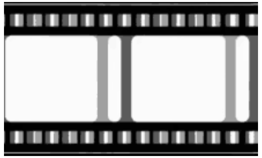
北京大學 PEKING UNIVERSITY

- Target video query: **object counting** with confidence interval (CI)

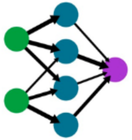
- The central problem: **planning constrained energy for counting**
  - Energy model: a budget that cannot be exceeded in a horizon (e.g., 24 hrs)
  - Trade-offs: frame sampling and NN selection
  - Target: smallest mean CI widths across all (30-min) windows in a horizon

Elf Overview

Check our paper for details



Sampled frames



Selected Neural Net

Aggregator with error Integration

Object counts with CIs

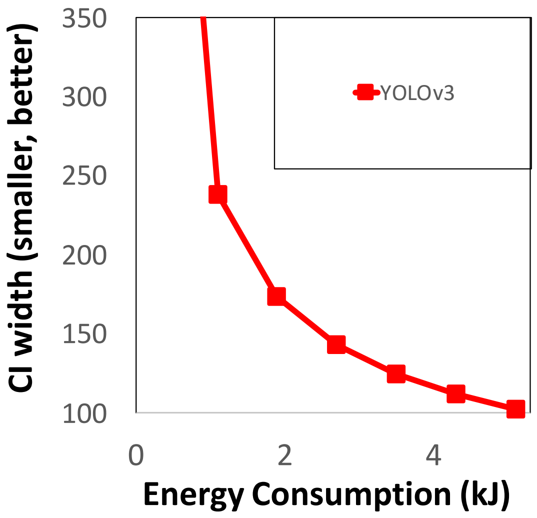
Reinforcement Learning Planner

Focus of the talk

Camera Operating System

Elf tech #1: per-window energy/CI fronts

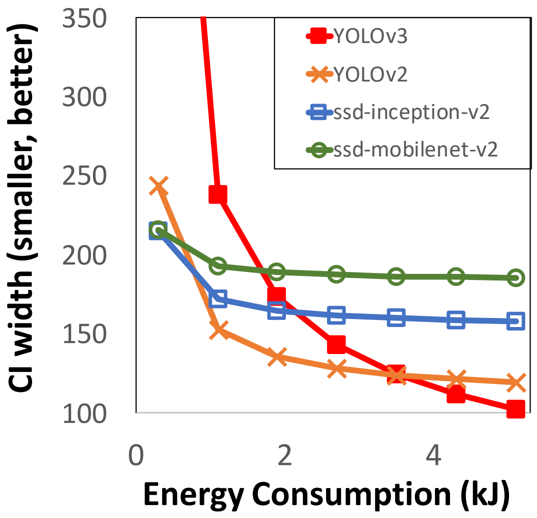
- What's the best count action for a window?
  - A *count action*: determining (1) an NN and (2) # of frames to process



Energy Consumption = E(NN) * frame_num

Elf tech #1: per-window energy/CI fronts

- What's the best count action for a window?
  - A *count action*: determining (1) an NN and (2) # of frames to process

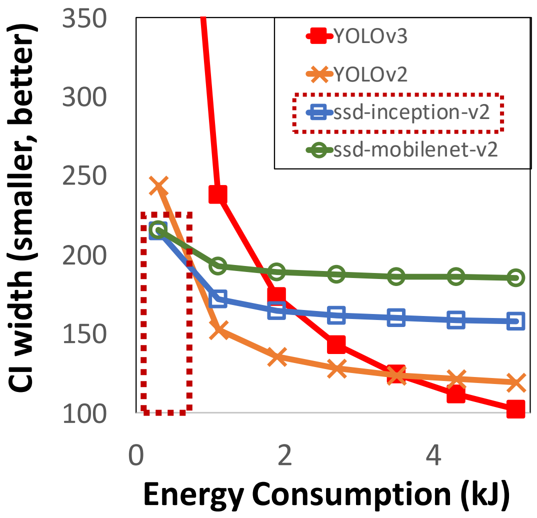


Energy Consumption = E(NN) * frame_num

NN Counters	Input	mAP	Energy
YOLOv3 (Golden, GT) [85]	608x608	33.0	1.00
YOLOv2 [84]	416x416	21.6	0.22
faster rcnn inception-v2 [86]	300x300	28.0	0.40
ssd inception-v2 [68]	300x300	24.0	0.08
ssd mobilenet-v2 [88]	300x300	22.0	0.05
ssdlite mobilenet-v2 [88]	300x300	22.0	0.04

Elf tech #1: per-window energy/CI fronts

- What's the best count action for a window? No silver bullet.
  - A *count action*: determining (1) an NN and (2) # of frames to process

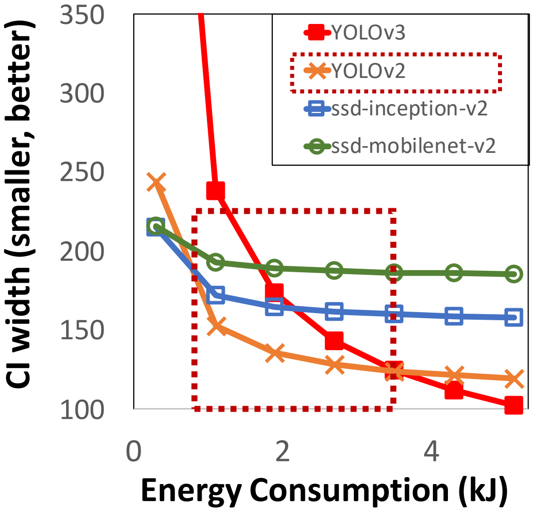


When energy is low: cheaper NNs win

- **Bottlenecked by sampling error (frame quantity)**

Elf tech #1: per-window energy/CI fronts

- What's the best count action for a window? No silver bullet.
  - A *count action*: determining (1) an NN and (2) # of frames to process



When energy is low: cheaper NNs win

- **Bottlenecked by sampling error (frame quantity)**

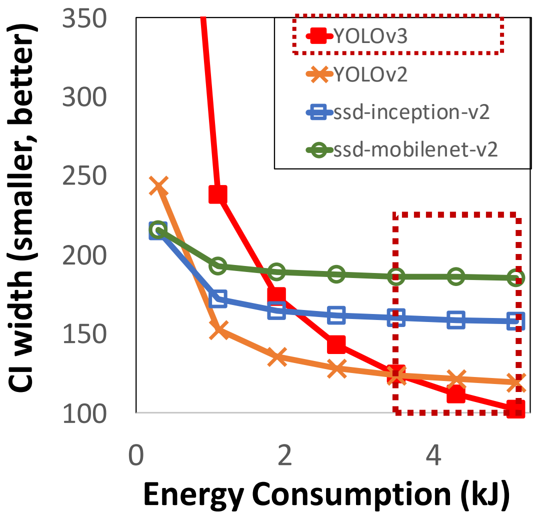
When energy is low: more accurate NNs win

- **Bottlenecked by NN error (frame quality)**

Elf tech #1: per-window energy/CI fronts

北京大學 PEKING UNIVERSITY

- What's the best count action for a window? No silver bullet.
  - A *count action*: determining (1) an NN and (2) # of frames to process



When energy is low: cheaper NNs win

- **Bottlenecked by sampling error (frame quantity)**

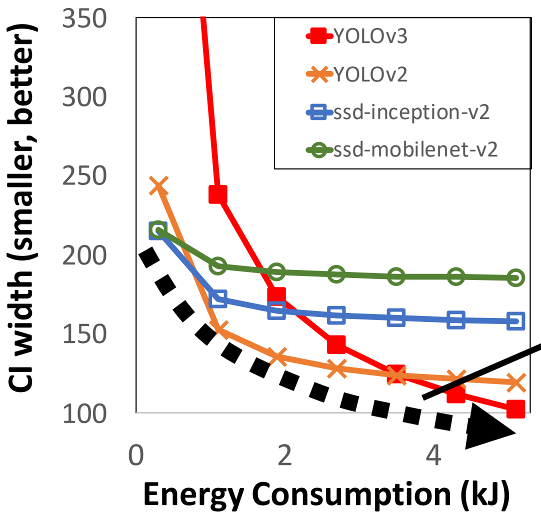
When energy is low: more accurate NNs win

- **Bottlenecked by NN error (frame quality)**

Elf tech #1: per-window energy/CI fronts

PEKING UNIVERSITY

- What's the best count action for a window? No silver bullet.
  - A *count action*: determining (1) an NN and (2) # of frames to process



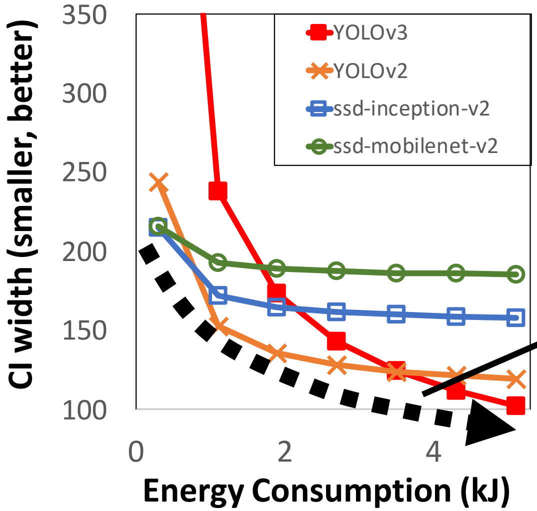
When energy is low: cheaper NNs win
When energy is low: more accurate NNs win

***Energy/CI front:** the combination of all "optimal" count actions with varied energy*

Elf tech #1: per-window energy/CI fronts

北京大學 PEKING UNIVERSITY

- What's the best count action for a window? No silver bullet.
  - A *count action*: determining (1) an NN and (2) # of frames to process



When energy is low: cheaper NNs win
When energy is low: more accurate NNs win

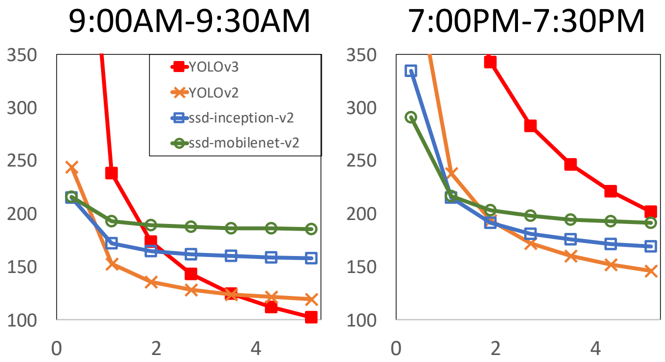
***Energy/CI front:** the combination of all "optimal" count actions with varied energy*

- How to construct? Error integration
- Depends on the video characteristics

Elf tech #1: per-window energy/CI fronts

北京大學 PEKING UNIVERSITY

- What's the best count action for a window? No silver bullet.
  - A *count action*: determining (1) an NN and (2) # of frames to process



Different windows have different energy/CI fronts

When energy is low: cheaper NNs win
When energy is low: more accurate NNs win

Energy/CI front: *the combination of all "optimal" count actions with varied energy*

- How to construct? Error integration
- Depends on the video characteristics

Elf tech #2: across-window joint planning

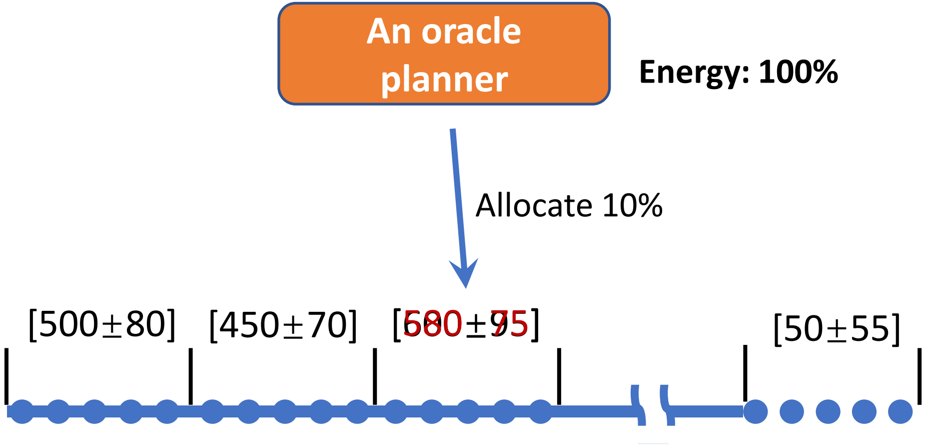
北京大學 PEKING UNIVERSITY

- An Oracle Planner: best performance but unrealistic
  - knows all energy/CI fronts

Elf tech #2: across-window joint planning

北京大學 PEKING UNIVERSITY

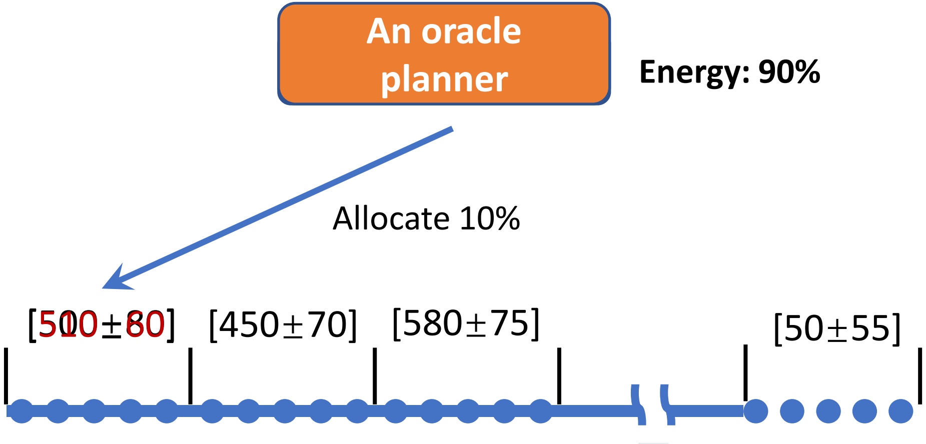
- An Oracle Planner: best performance but unrealistic
  - knows all energy/CI fronts



A greedy approach: giving energy to the window with the most benefit (i.e., CI width reduction).

Elf tech #2: across-window joint planning

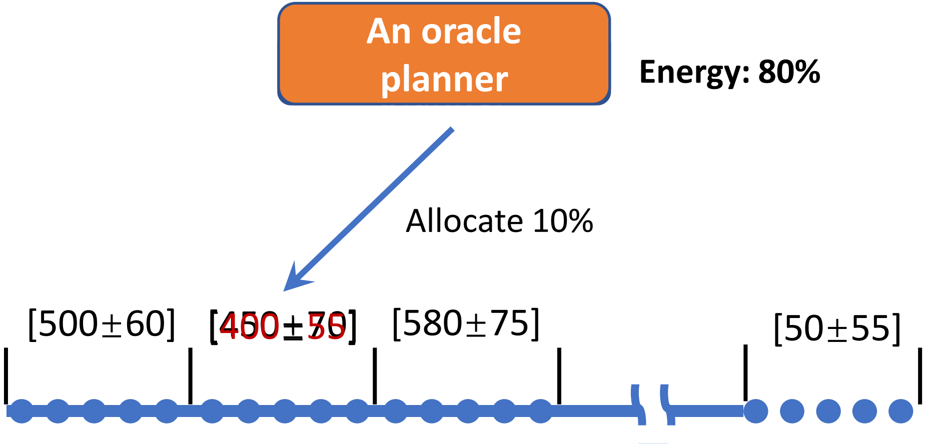
- An Oracle Planner: best performance but unrealistic
  - knows all energy/CI fronts



A greedy approach: giving energy to the window with the most benefit (CI width reduction)

Elf tech #2: across-window joint planning

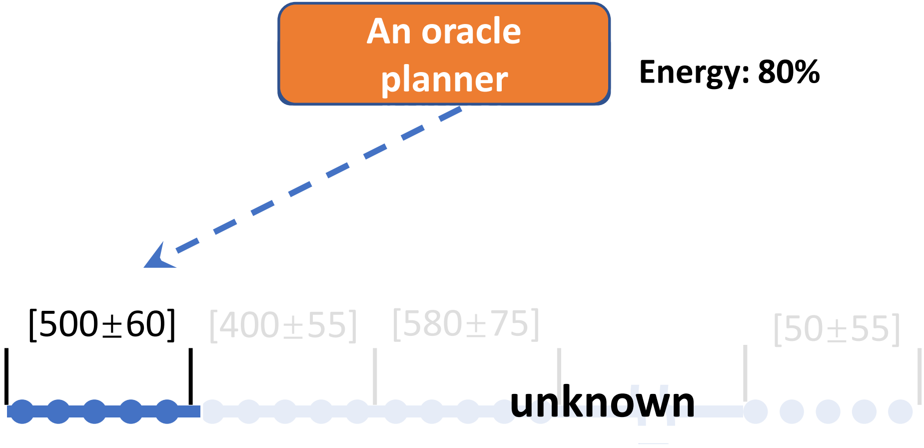
- An Oracle Planner: best performance but unrealistic
  - knows all energy/CI fronts



A greedy approach: giving energy to the window with the most benefit (CI width reduction)

Elf tech #2: across-window joint planning

- An Oracle Planner: best performance but unrealistic
  - knows all energy/CI fronts

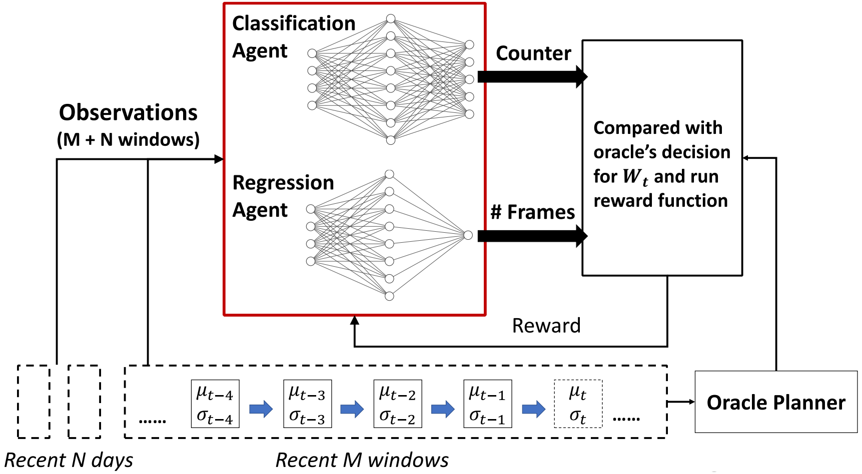


A greedy approach: giving energy to the window with the most benefit (CI width reduction)

Elf tech #2: across-window joint planning

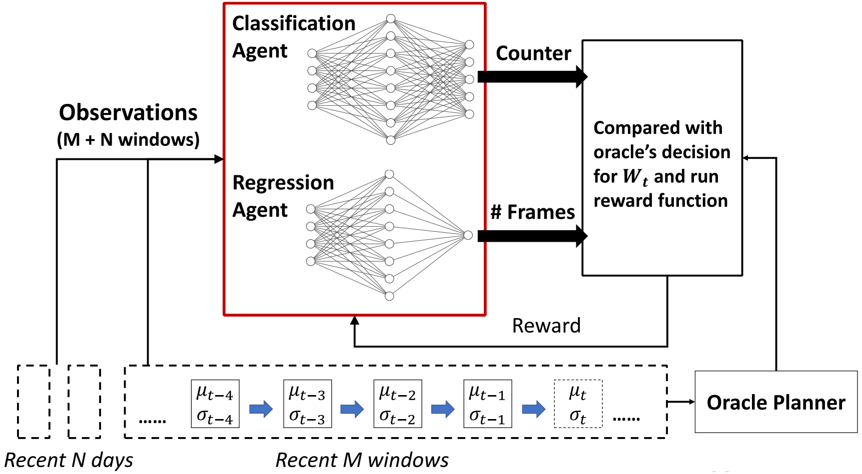
北京大學 PEKING UNIVERSITY

- An Oracle Planner: best performance but unrealistic
  - knows all energy/CI fronts
  - planned offline
- A learning-based planner: imitating the oracle planner
  - basis: reinforcement learning
  - rationale: daily and temporal patterns



Elf tech #2: across-window joint planning

- An Oracle Planner: best performance but unrealistic
  - knows all energy/CI fronts
  - planned offline
- A learning-based planner: imitating the oracle planner
  - basis: reinforcement learning
  - rationale: daily and temporal patterns
  - offline training -> online prediction
    - Two agents: NN selection and # of frames
    - Observations: knowledge of past windows
    - Penalty: deviation from oracle's decision



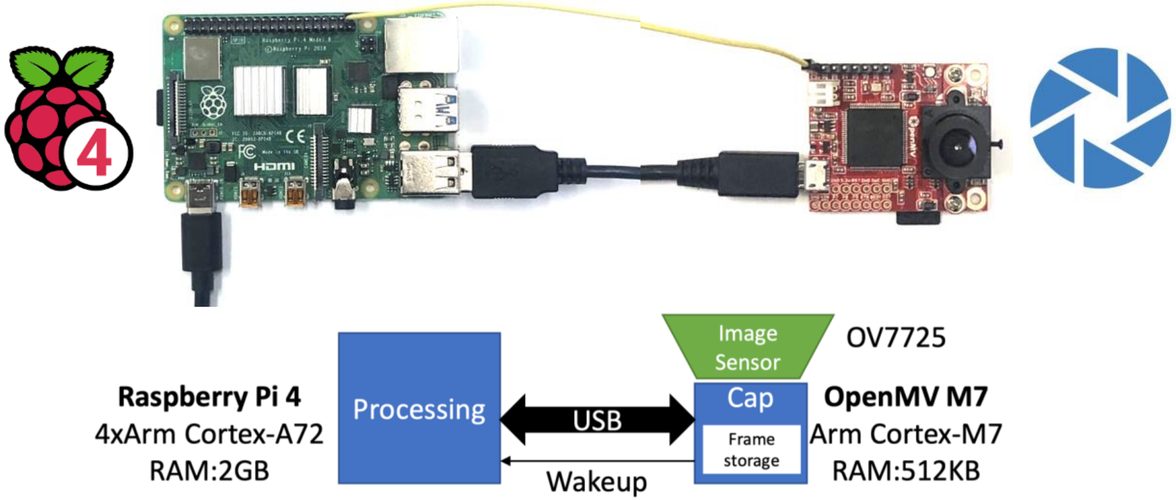
Elf tech #2: across-window joint planning

PEKING UNIVERSITY

- An Oracle Planner: best performance but unrealistic
  - knows all energy/CI fronts
  - planned offline
- A learning-based planner: imitating the oracle planner
  - basis: reinforcement learning
  - rationale: daily and temporal patterns
  - offline training -> online prediction
  - Enforce energy budget: make reservation for future windows
    - 30 frames to be statistically meaningful

Elf Implementation

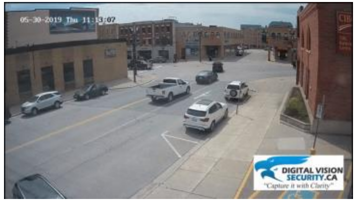
- Capture & processing decoupled for higher energy efficiency
  - Processing batched at the end of each window



Elf Evaluation

- Over 1,000-hr videos
  - Public, 2-week long each stream
- Baselines
  1. *GoldenNN*: most accurate NN
  2. *UniNN*: one fixed best NN
  3. *Oracle*: offline planned
- Small solar panel
  - 10Wh~30Wh per day

Auburn, AL



Unknown

Hampton, NY

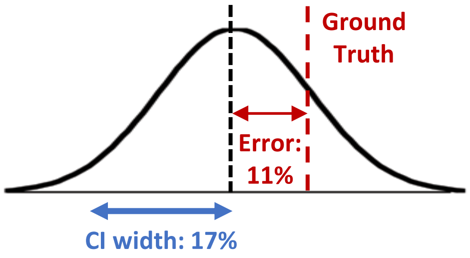
Jackson, WY

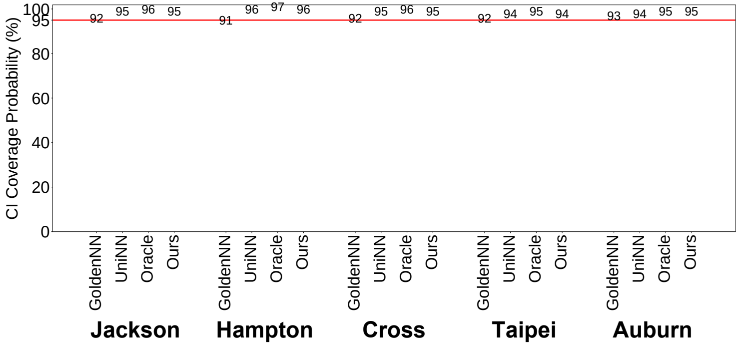
Taipei

Taipei

Elf Evaluation

- Average: 11% error, valid and 17%-width CI
  - 95% confidence level



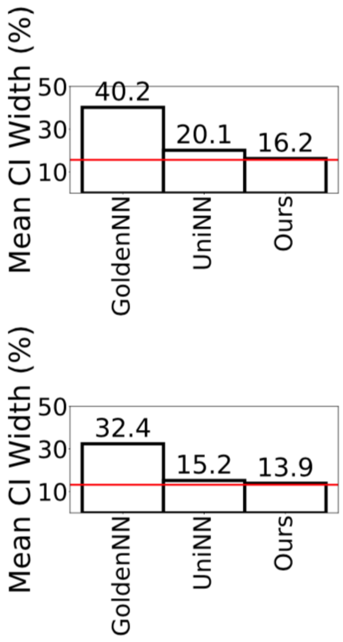


Cis cover ground truth with 95% probability (specified)

Elf Evaluation

- Average: 11% error, valid and 17%-width CI
- Significant improvements over baselines in CI widths
  - 66.6%, 59.8%, and 56.2% smaller over *GoldenNN* (up to 3.4x)
  - 41.1%, 16.6%, and 9.7% smaller over *UniNN*

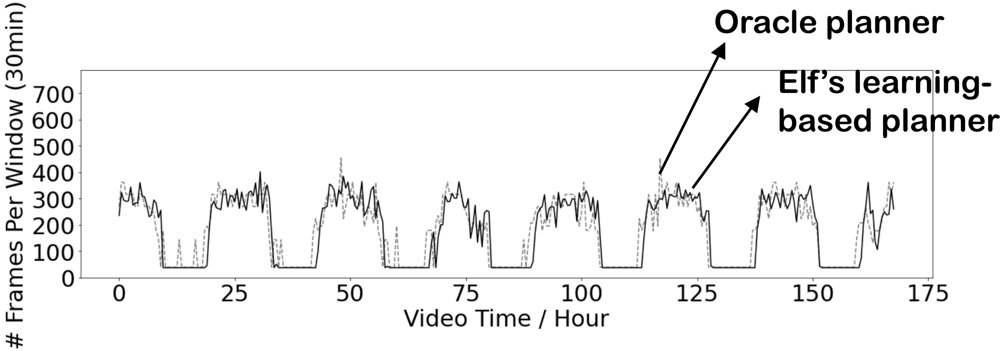
10Wh per day | 20Wh per day | 30Wh per day

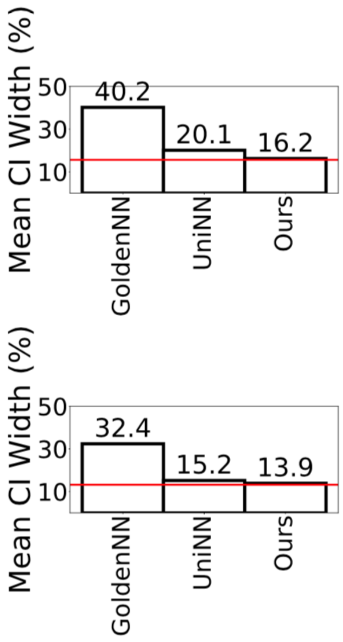


(a) Jackson

Elf Evaluation

- Average: 11% error, valid and 17%-width CI
- Significant improvements over baselines in CI widths
- Very close to *Oracle*
  - < 5% wider CI
  - Well imitating the oracle planner

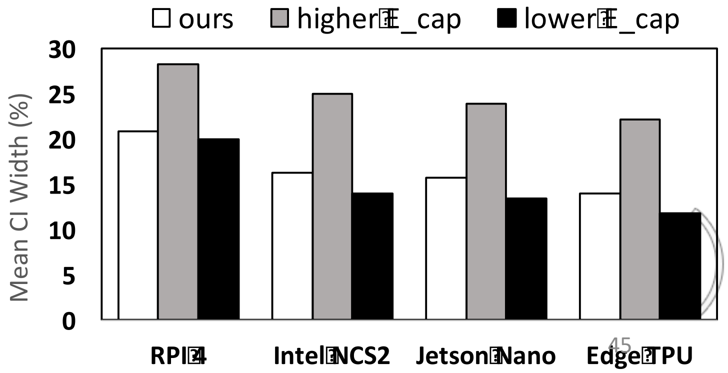




(a) Jackson

Elf Evaluation

- Average: 11% error, valid and 17%-width CI
- Significant improvements over baselines in CI widths
- Very close to *Oracle*
- What if we have AI accelerators?
  - CIs are reduced noticeably (by 22.1%–33.1%)
  - Still cannot process every frame (short of energy)



Summary

- Autonomous camera: expanding the geo-frontier of video analytics
  - Energy-independent and compute-independent
- Elf: the first runtime for autonomous camera
  - Target query: object counting
  - Key idea: count planning per- and across-windows
- Prototyped on heterogeneous hardware
- Evaluated on over 1,000-hr videos
  - 11% error, 17% CI width