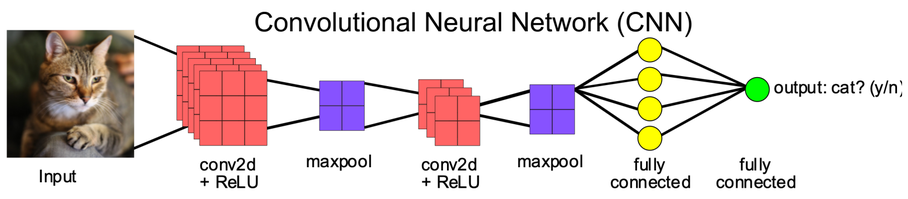
DeepFind: Sensor-driven Inference Acceleration for Continuous Deep Mobile Vision Applications

Chungkuk Yoo[1], Saiyma Sarmin[2], Inseok Hwang[1], Eric Rozner[2], Minsik Cho[1]
[1]IBM [2]University of Colorado Boulder

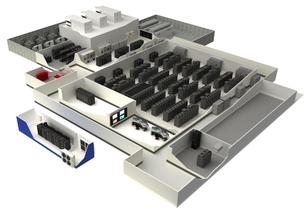
Problem and Goal

- Continuous vision enables smart environments

- Deep learning CNNs obtain human-scale accuracy



- Problem: CNN inference computationally expensive



- Move data to cloud?
  - Privacy concerns
  - Network cost
- Move computation to edge?
  - Fewer resources than cloud (e.g., energy, computation)

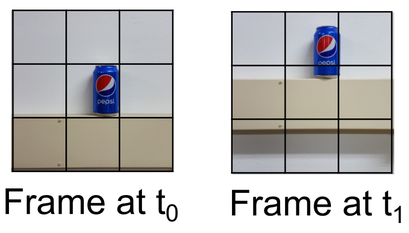
Goal: enable deep learning vision to run continuously and efficiently on mobile and embedded devices.

Approach

- Consecutive frames enable caching opportunities

- How to determine cacheable regions?

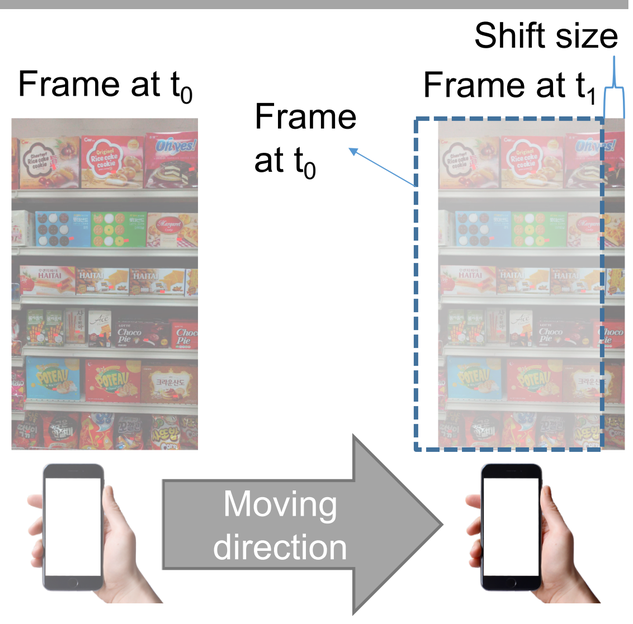
DeepMon



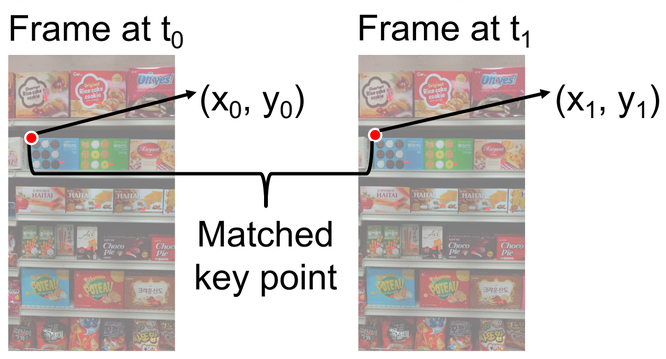
DeepMon: previous scheme

Expensive image-based analysis across input frames

Our Scheme



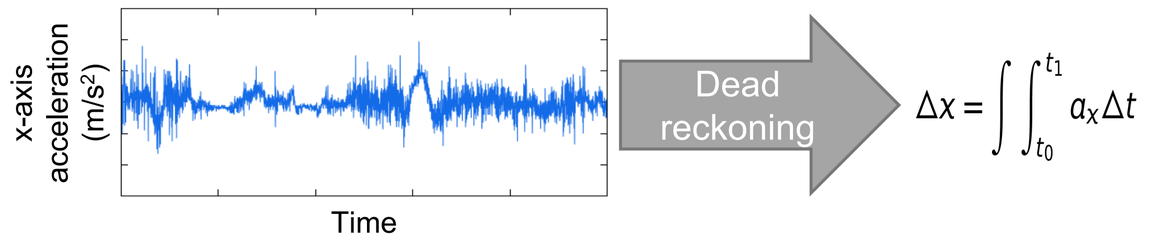
- Converting spatial distance Δx to pixel distance Δp



Contributions

- Accelerate CNN on mobile and embedded devices
- A caching mechanism to reduce CNN inference time
  - Exploits spatial/temporal similarities in CNN inputs

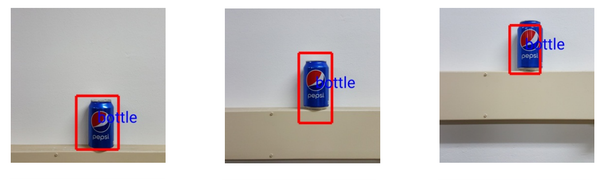
  - Utilizes mobile sensors to determine similarities



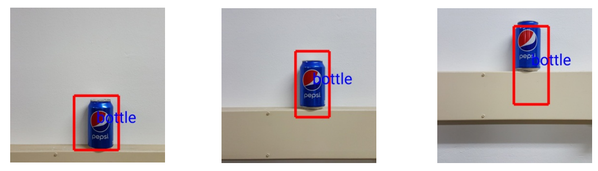
Evaluation

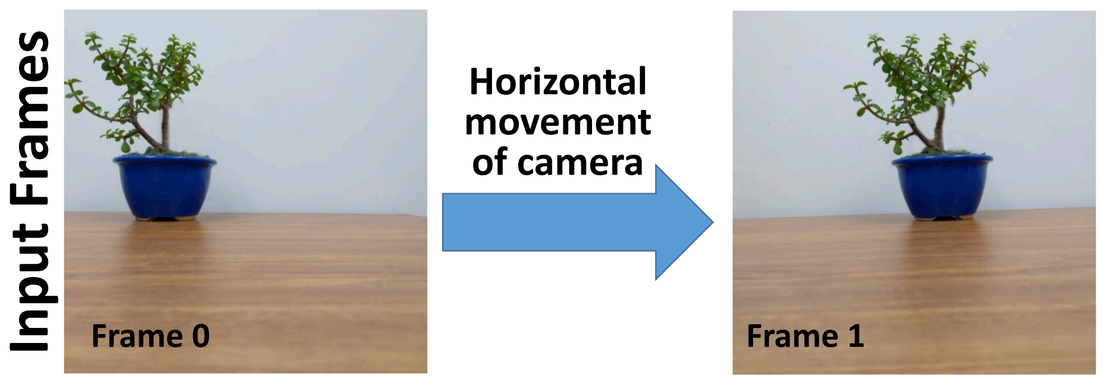
- Original Tiny-Yolo

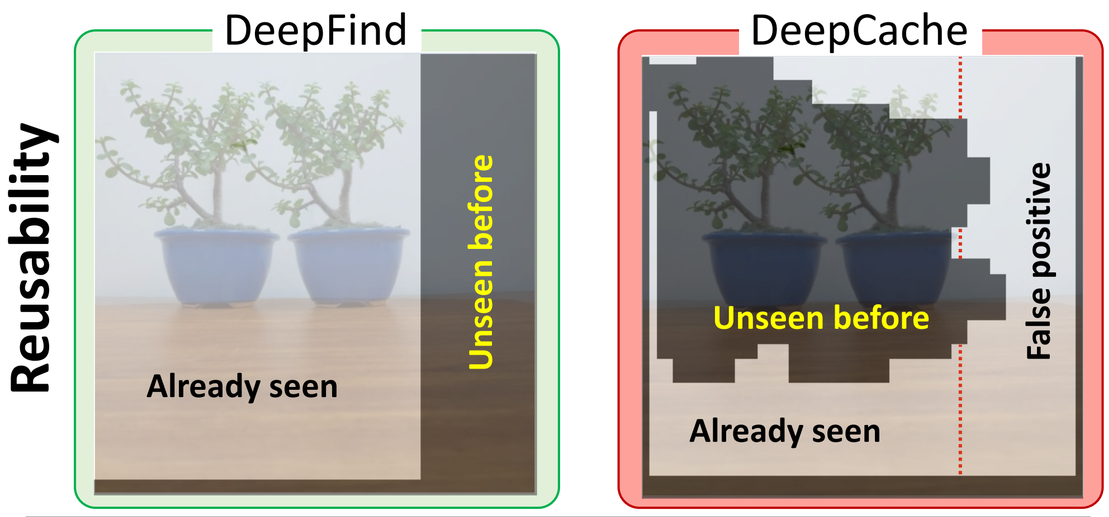
- DeepFind



- DeepMon







Time to determine cached region (per frame)		
DeepFind	DeepMon	DeepCache
0.42 ms	6.0 – 18 ms	11 – 30 ms

Summary

- Continuous mobile vision important
  - Visual info provides context of users and environments
- Current deep learning algorithms are too expensive
  - Edge devices have less power, energy than cloud
- Our work makes efficient continuous vision on mobile and embedded devices a reality
  - Allows personalized intelligence to become truly pervasive