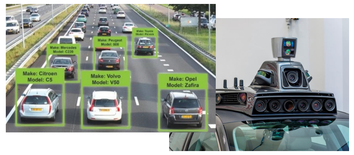
Ekya: Continuous Learning of Video Analytics Models on Edge Compute Servers

Romil Bhardwaj, Zhengxu Xia, Ganesh Ananthanarayanan, Junchen Jiang, Nikolaos Karianakis, Yuanchao Shu, Kevin Hsieh, Victor Bahl, Ion Stoica

romil.bhardwaj@berkeley.edu; http://aka.ms/ekya

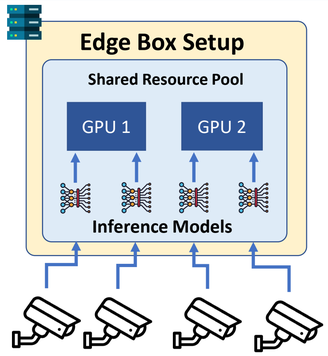
riselab UC Berkeley · Microsoft · The University of Chicago

Video Analytics at the Edge



Why edge analytics?

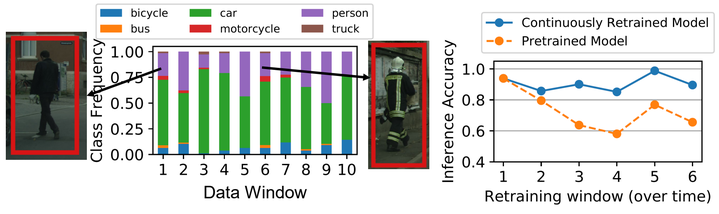
- Privacy
- No network required



Data Drift and Continuous Learning

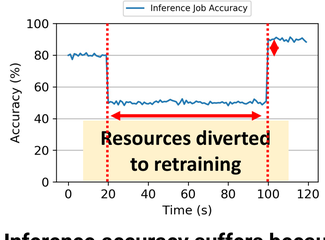
Compressed Models **don't generalize** and are sensitive to **data drift**

Data Drift Example – Class Definition Shifts

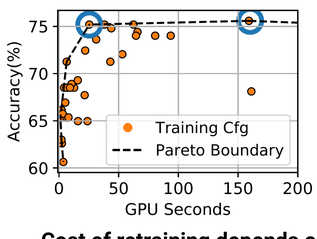


The performance of pretrained models varies as the incoming data distribution and class definitions change

Continuous Learning addresses data but steals GPU-time from inference and has variable **costs**



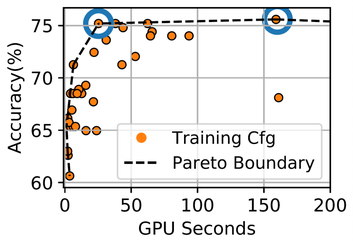
Inference accuracy suffers because GPUs are allocated to training

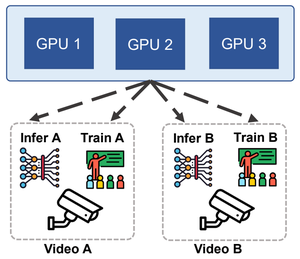


Cost of retraining depends on retraining hyperparameters

Joint Hyperparameter Selection and Resource Scheduling – an NP-hard problem

Pick Hyperparameters to retrain ⟷ **Allocate resources between Training and Inference across streams**





Scheduler Objective

Maximize mean inference accuracy across all video streams

Constraints

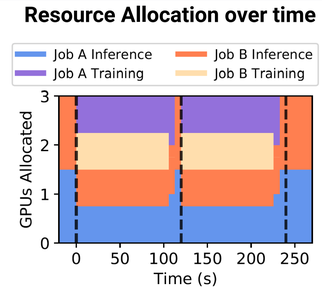
Resource capacity constraints (do not oversubscribe)

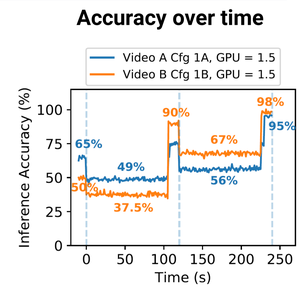
Minimum inference accuracy constraint (ensure min. accuracy)

Example of smart joint selection and resource allocation

Fair scheduler

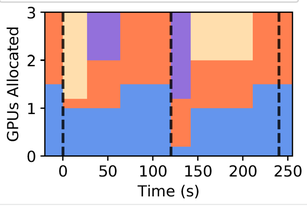
- Allocate resources equally
- Pick configs with highest accuracy

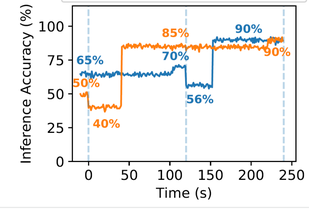




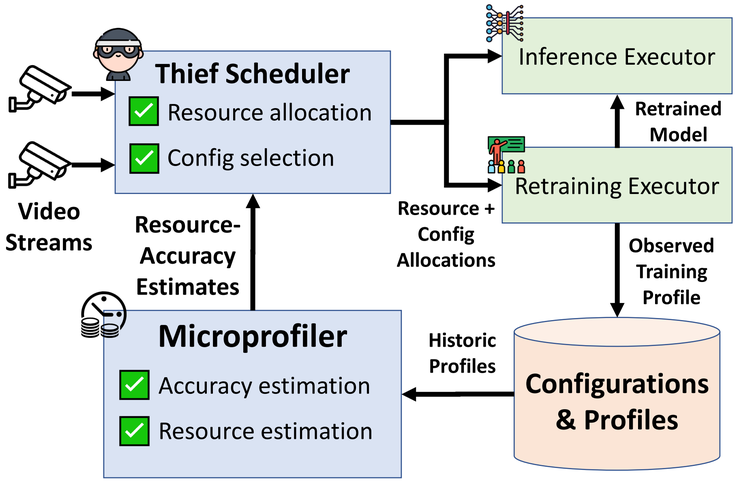
Smarter scheduler

- Prioritize retraining one job
- Pick configs with max accuracy gain per unit resource





Ekya at a Glance



Microprofiler produces accuracy and resource demand estimates, informing the thief scheduler's heuristics to maximize accuracy

Evaluation and (new!) Datasets

- We release and evaluate Ekya on two new datasets and on two public datasets – Waymo and Cityscapes
- Ekya saves upto 4.3x resources and achieves 29% higher accuracy than baselines

Comparing accuracy with varying number of video streams