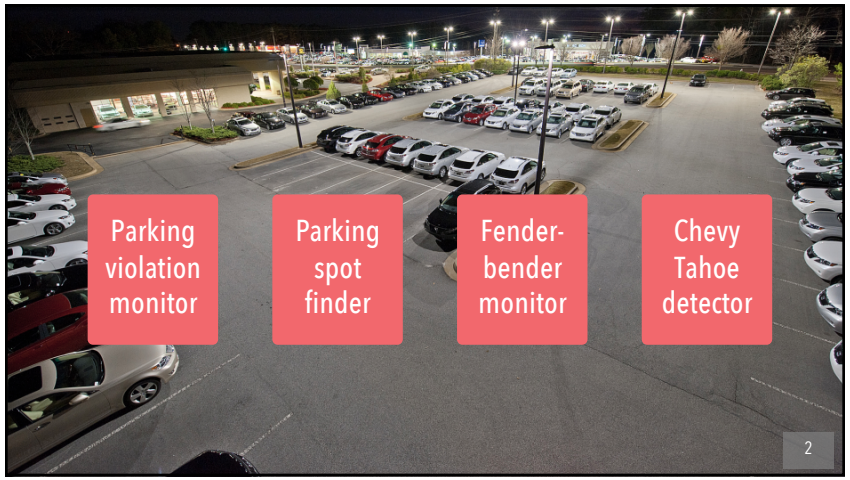
MAINSTREAM: DYNAMIC STEM-SHARING FOR MULTI-TENANT VIDEO PROCESSING

Angela Jiang, Daniel L.-K. Wong, Christopher Canel,
Lilia Tang, Ishan Misra, Michael Kaminsky,
Michael Kozuch, Babu Pillai,
David G. Andersen, Greg Ganger

1



2

MAINSTREAM CHALLENGES AND SOLUTIONS

❗ Concurrent DNN inference is slow

✓ Mainstream shares redundant computation

❗ Deployed apps are tuned in isolation and offline

✓ Mainstream dynamically tunes DNNs at runtime

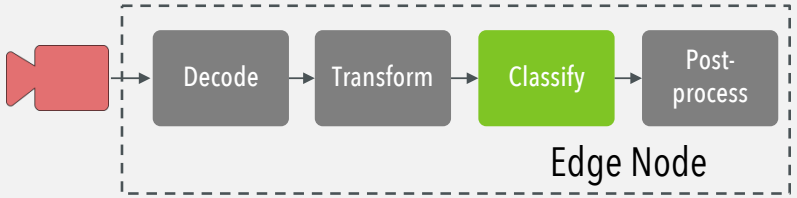
4

MAINSTREAM TAKEAWAYS

1. Enables efficient processing of set of apps
   By sharing redundant computation
2. Dynamically tune degree of specialization at runtime
3. Up to 87x improvement in F1-score
   Compared to No Sharing
4. Up to 47% improvement in F1-score
   Compared to Max Sharing

5

EXAMPLE VIDEO ANALYSIS PIPELINE



8

DNN IS PRIMARY COST OF PIPELINE

9

DNN IS PRIMARY COST OF PIPELINE

10

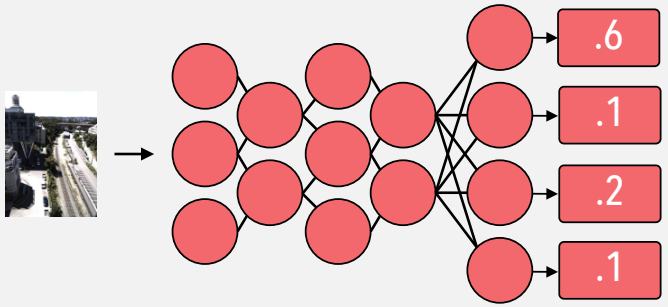
IN MAINSTREAM: PARTIAL SHARING OF 3 DNNS

11

HOW TO FIND REDUNDANCY BETWEEN DNNS?

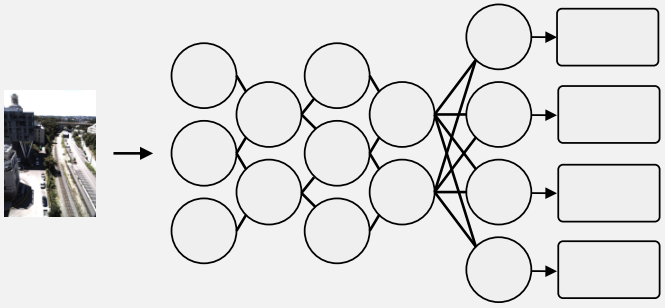
12

BASIC DNN



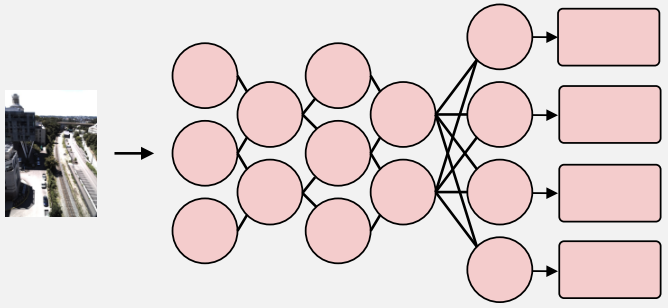
13

TRAINING A DNN



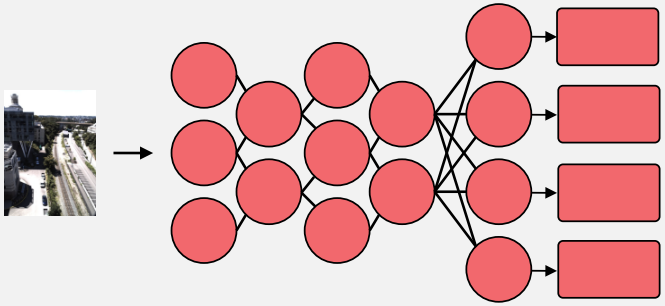
14

TRAINING A DNN



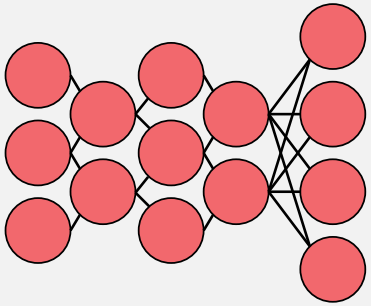
15

TRAINING A DNN



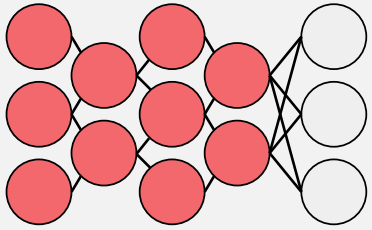
16

FINE-TUNING DNN #1



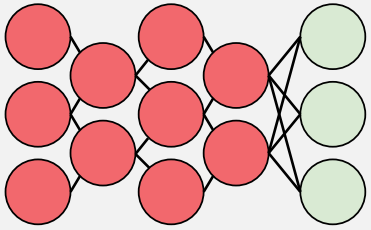
17

FINE-TUNING DNN #1



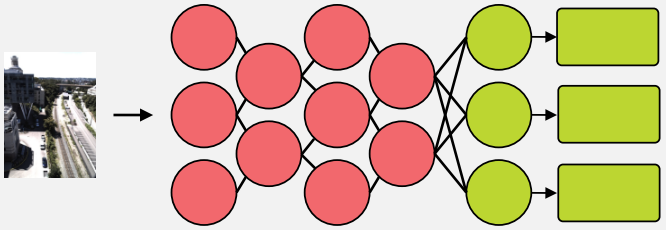
18

FINE-TUNING DNN #1



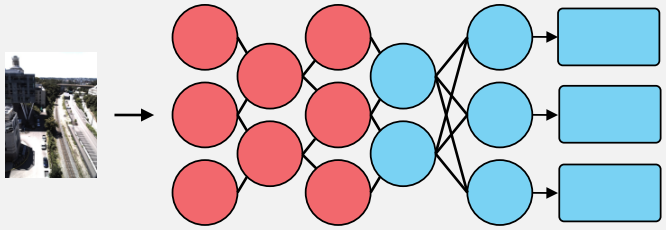
19

FINE-TUNING DNN #1



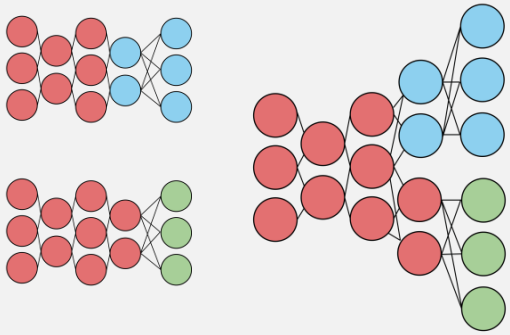
20

FINE-TUNING DNN #2



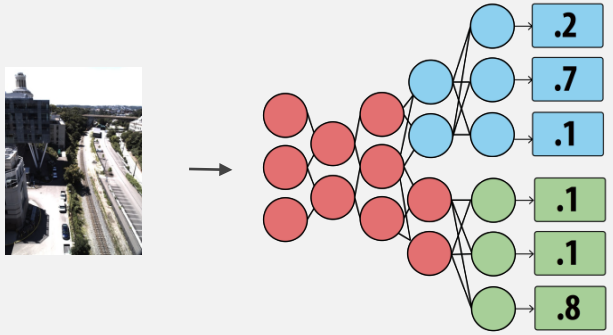
21

FINE-TUNING DNN #1



22

TWO DNNS AS RUN IN MAINSTREAM



23

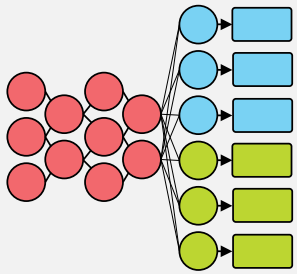
MAINSTREAM'S APPROACH

1. Merge video stream processing of concurrent apps
   By sharing redundant layers
2. Dynamically tune degree of specialization at runtime
   Based on available resources, other applications
   Maximize application effectiveness

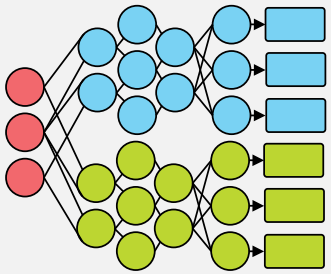
24

PERFORMANCE TRADE-OFFS FROM SPECIALIZATION

25



More sharing = Higher throughput



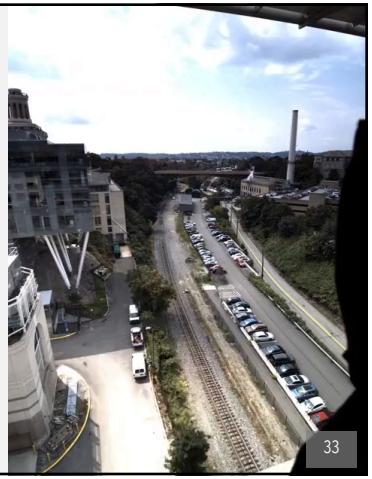
More specialization = Higher per-frame acc.

26

HOW DO WE MEASURE APPLICATION PERFORMANCE?

32

APPLICATION EXAMPLE: EVENT DETECTION



33

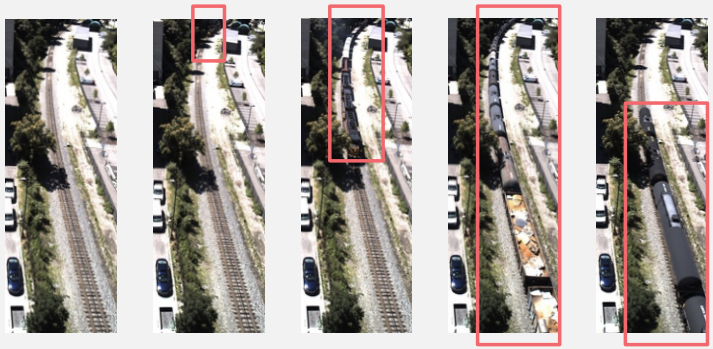
GOAL: MAXIMIZE EVENT F1-SCORE

Event-Recall How many relevant events are detected?
Event-Precision How many detected events are relevant?
Event-F1 score Harmonic mean between precision and recall

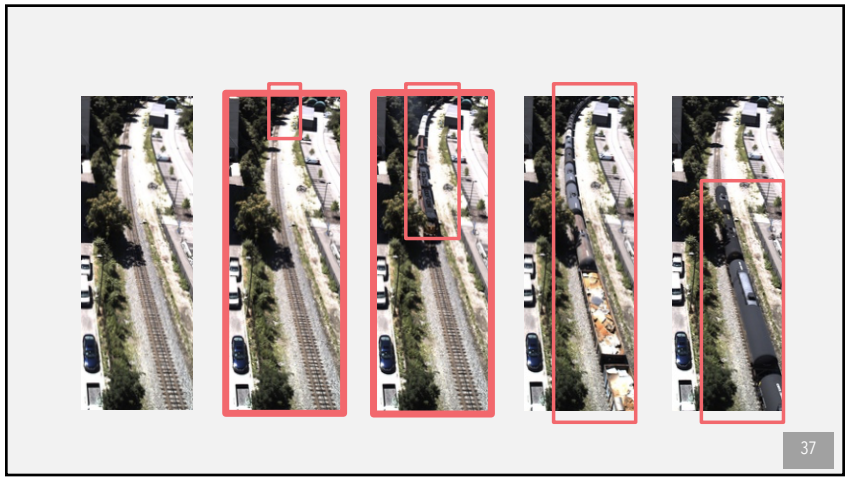
34

EFFECT OF ACCURACY AND FPS ON F1-SCORE

35

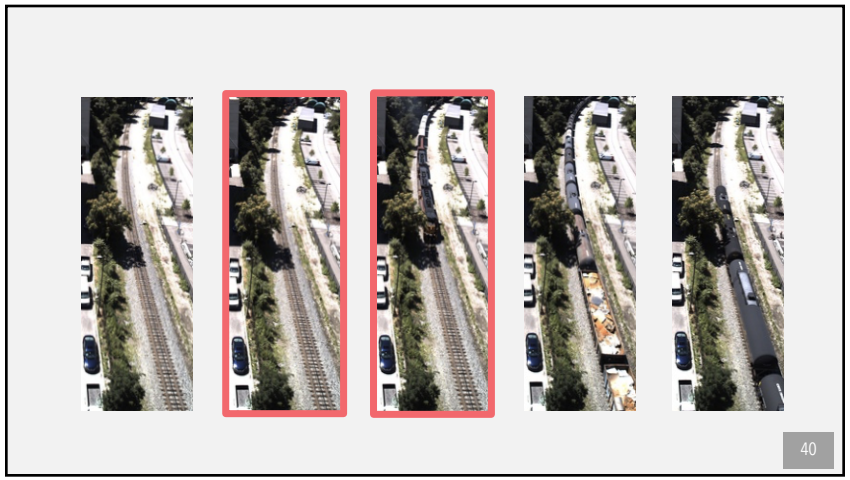


36


37


38

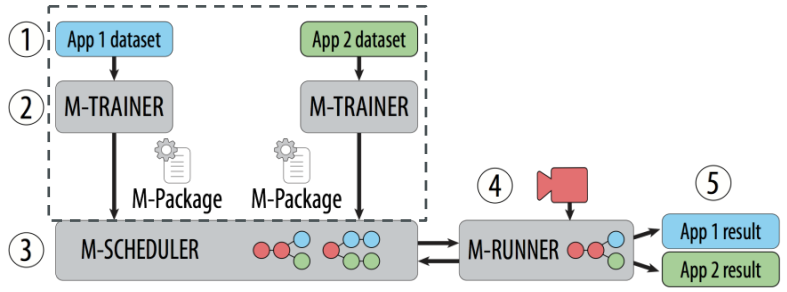

39


40

MAINSTREAM ARCHITECTURE

45

MAINSTREAM ARCHITECTURE



46

MAINSTREAM RESULTS

50

EXPERIMENTAL SETUP

Run concurrent image classification pipelines

Using MobileNets-224 DNNs

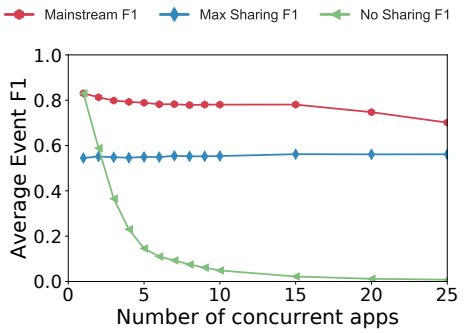
Tasks: Pedestrian, Bus, Red Car, Scramble, Schoolbus, Trains, Car

Hardware: Intel NUC

Video stream: 20FPS, 640x480

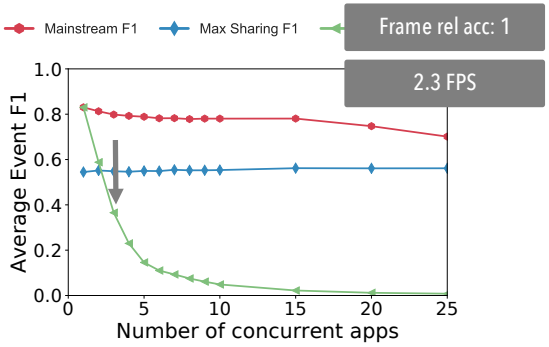
51

MAINSTREAM IMPROVES EVENT F1-SCORE BY 87X



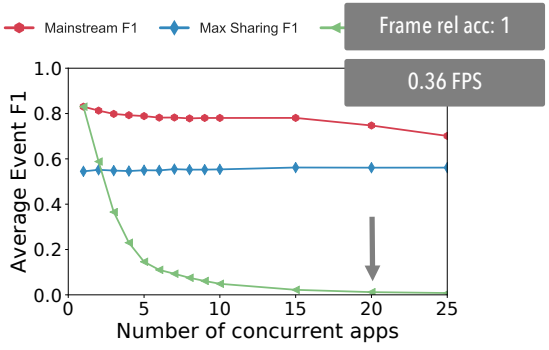
52

MAINSTREAM IMPROVES EVENT F1-SCORE BY 87X



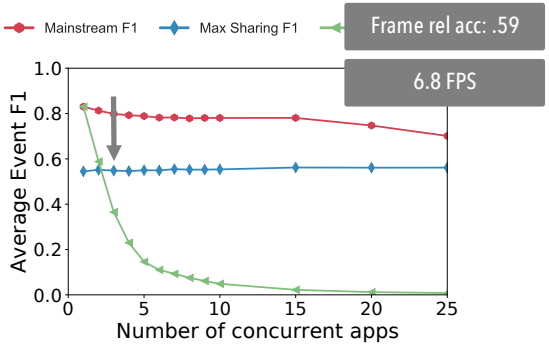
53

MAINSTREAM IMPROVES EVENT F1-SCORE BY 87X



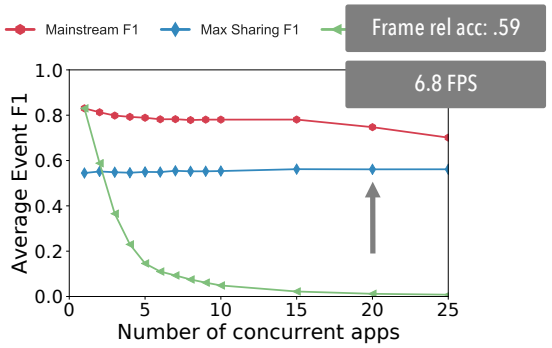
54

MAINSTREAM IMPROVES EVENT F1-SCORE BY 87X



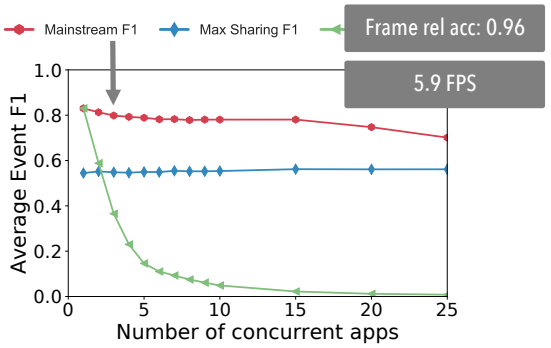
55

MAINSTREAM IMPROVES EVENT F1-SCORE BY 87X



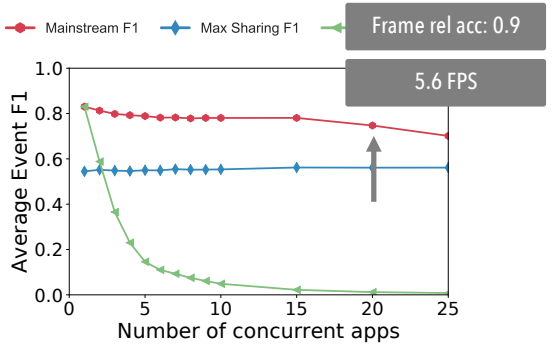
56

MAINSTREAM IMPROVES EVENT F1-SCORE BY 87X



57

MAINSTREAM IMPROVES EVENT F1-SCORE BY 87X



58

MAINSTREAM TAKEAWAYS

1. **Enables efficient processing of set of apps**
   By sharing redundant computation
2. **Dynamically tune degree of specialization at runtime**
3. **Up to 87x improvement in F1-score**
   Compared to No Sharing
4. **Up to 47% improvement in F1-score**
   Compared to Max Sharing

60