Deluceva: Delta-Based Neural Network Inference for Fast Video Analytics

Jingjing Wang and Magdalena Balazinska

University of Washington

UW DB

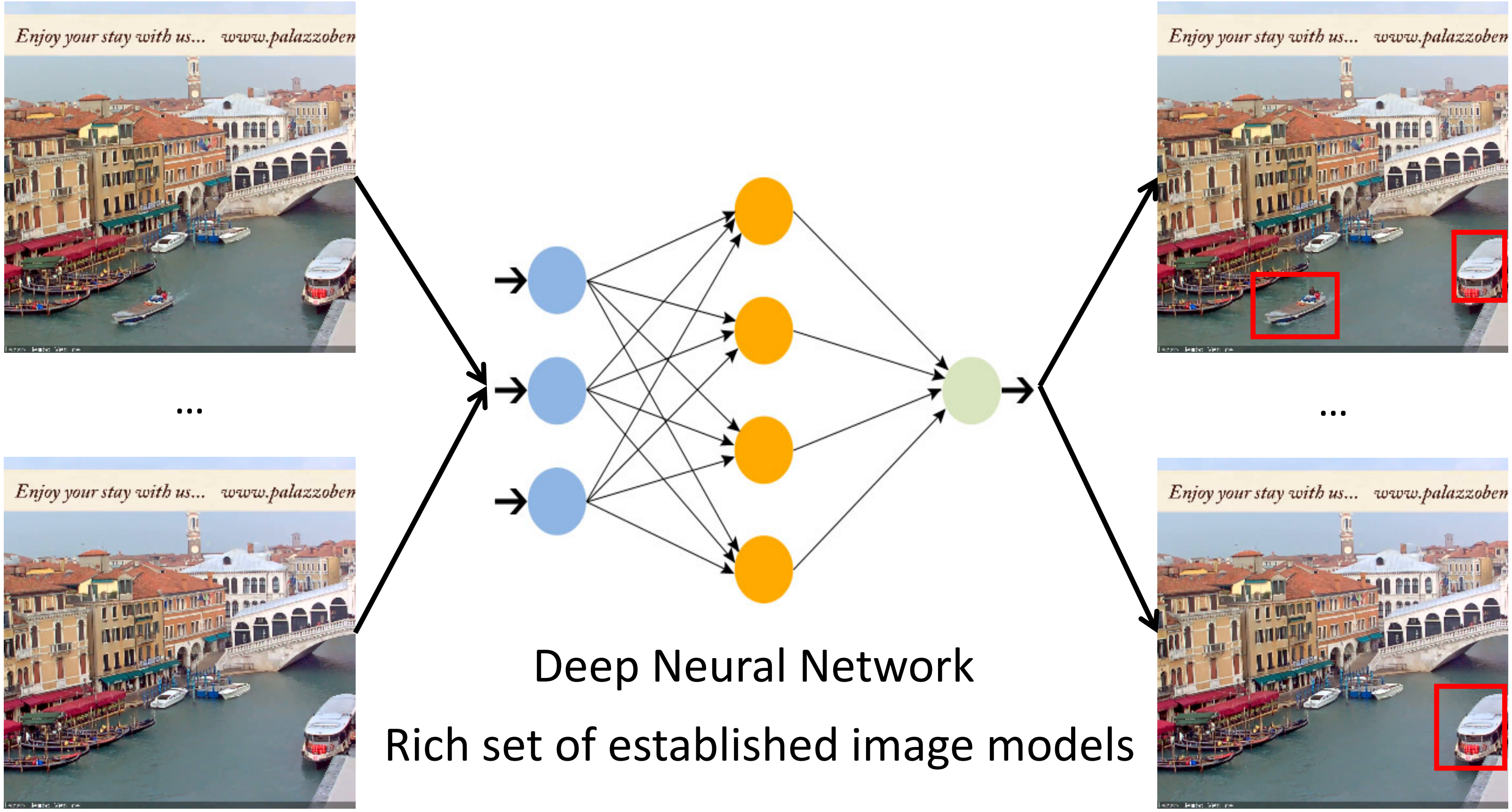
PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

Deluceva: Delta-Based Neural Network Inference for Fast Video Analytics

- Large volume of images/videos with valuable information
- A large set of neural network models for images
  - Object classification, detection, …

- Next step: **video analytics**
  - Larger volume
  - Efficiency critical, live output

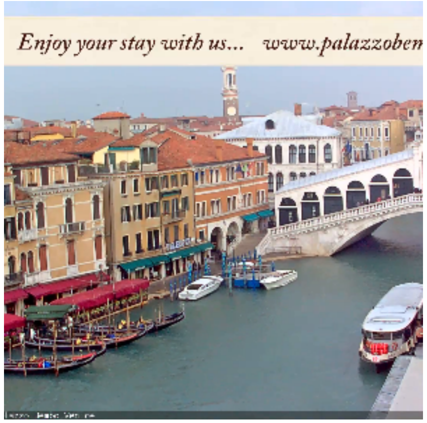
Video Analytics Using Neural Networks


Enjoy your stay with us... www.palazzoben
…
Enjoy your stay with us... www.palazzoben
Enjoy your stay with us... www.palazzoben
…
Enjoy your stay with us... www.palazzoben

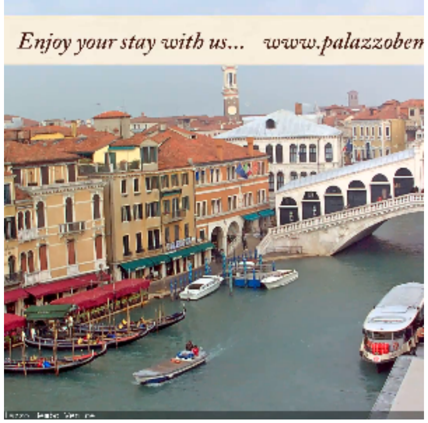
Deep Neural Network

Rich set of established image models

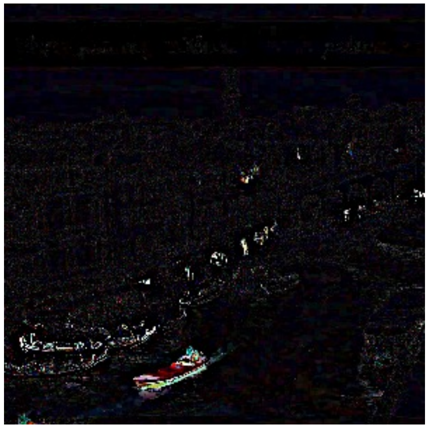
Key Observation:
Temporal Redundancy



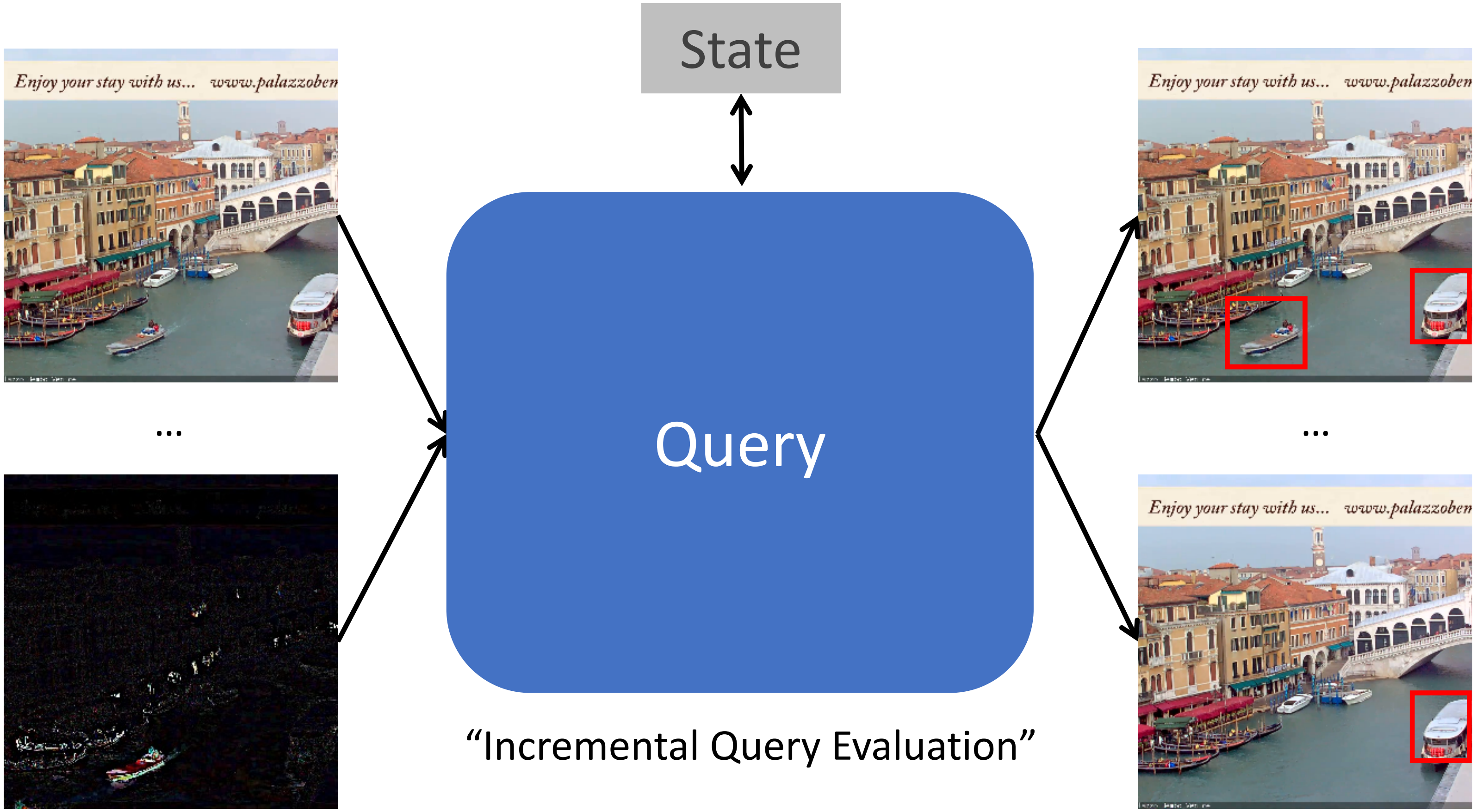
-



=

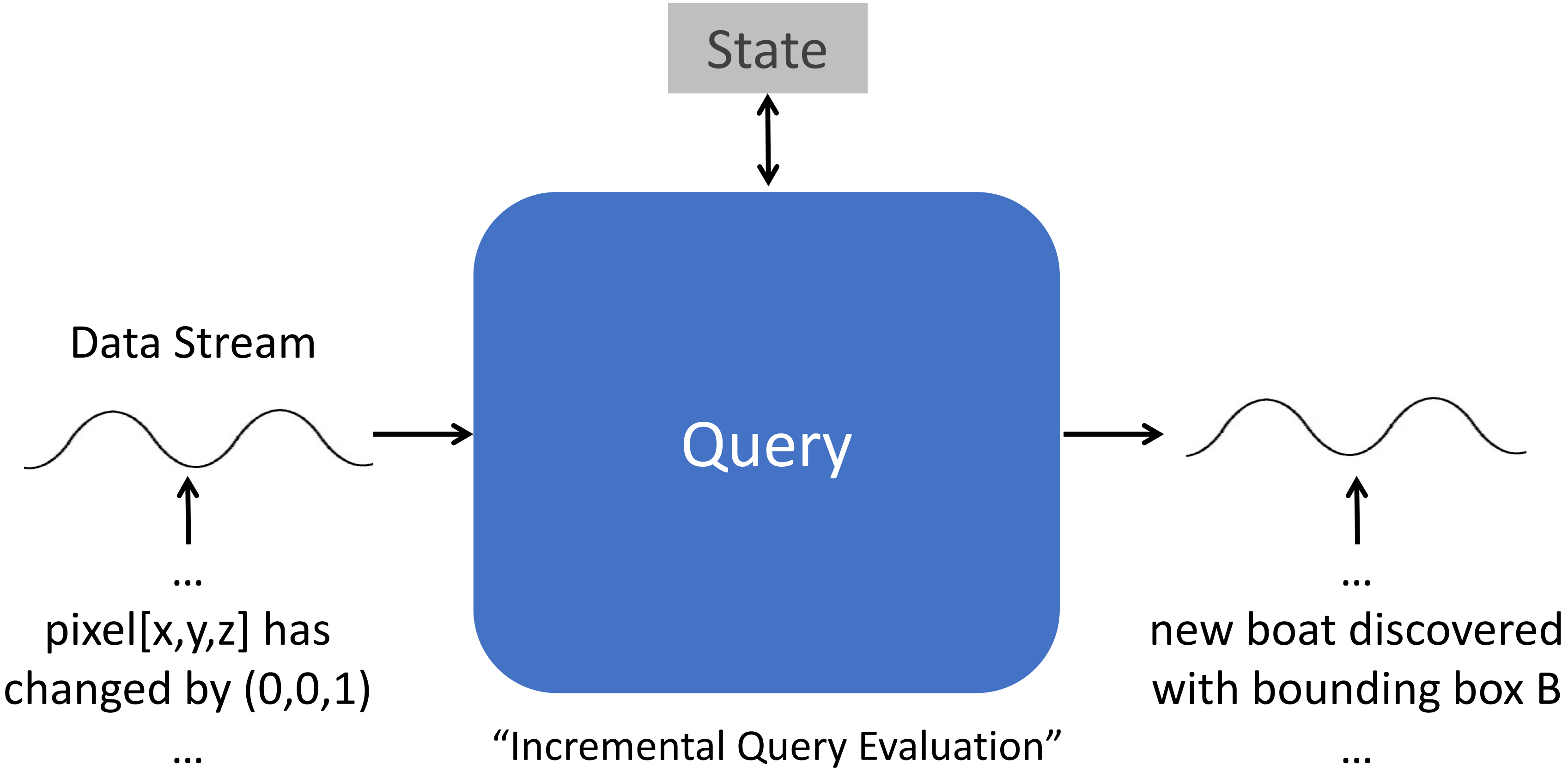


Process Deltas Instead of Full Frames



“Incremental Query Evaluation”

Process Deltas Instead of Full Frames



State

Data Stream

Query

…
pixel[x,y,z] has
changed by (0,0,1)
…

“Incremental Query Evaluation”

…
new boat discovered
with bounding box B
…

Delta-Based Inference for Videos

- Problem:
  - Input: a video stream, a reference model
  - Output: similar to the reference model’s result
- Approach:
  - Accelerate model inference by performing less computation

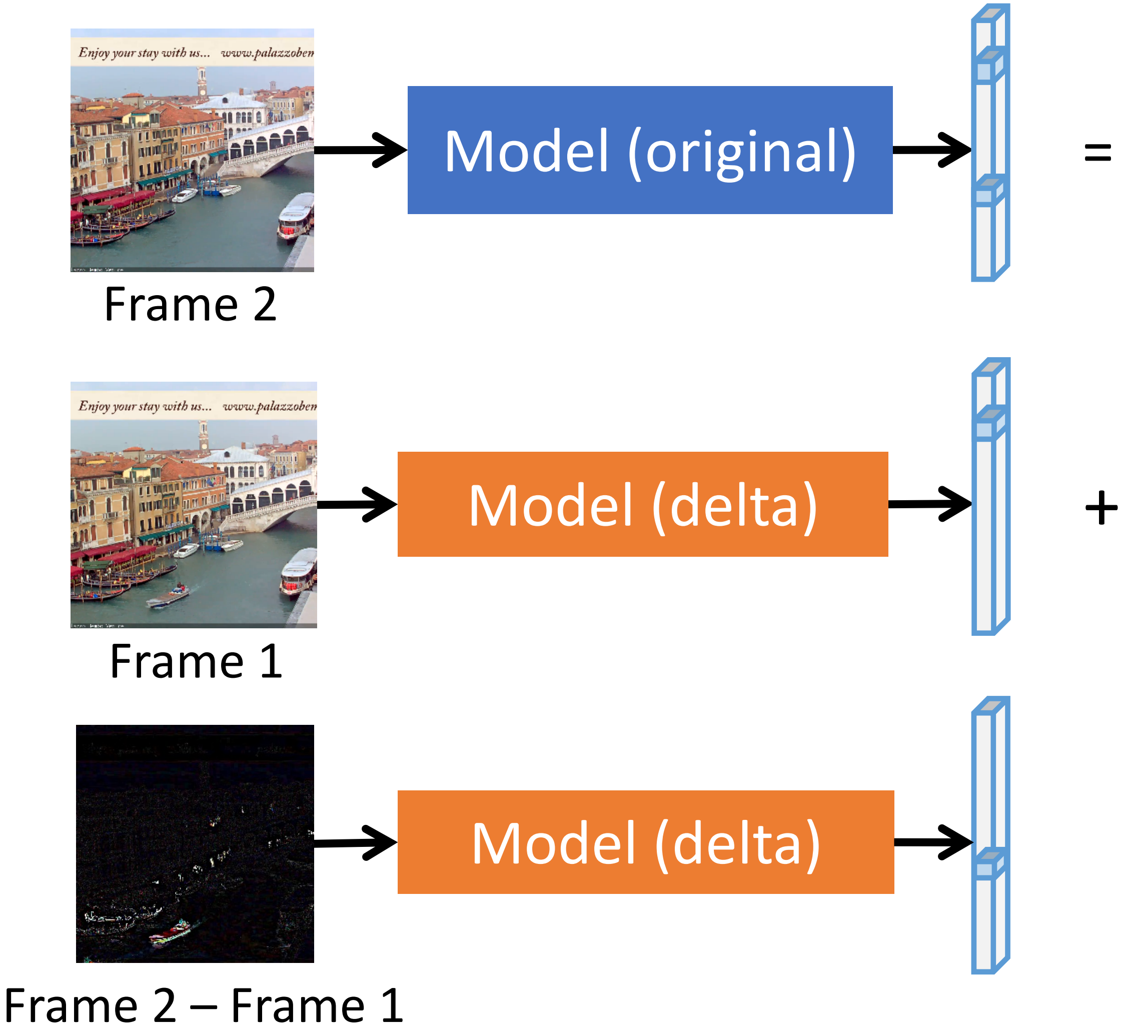
Delta-Based Inference for Videos: Overview

- Modify neural network to take deltas as inputs
- Decide which deltas are significant enough to process
- Generate a network of mixed-type (dense or delta-based) operators

Delta-Based Inference for Videos: Overview

- Modify neural network to take deltas as inputs
- Decide which deltas are significant enough to process
- Generate a network of mixed-type (dense or delta-based) operators

Neural Network with Sparse Deltas

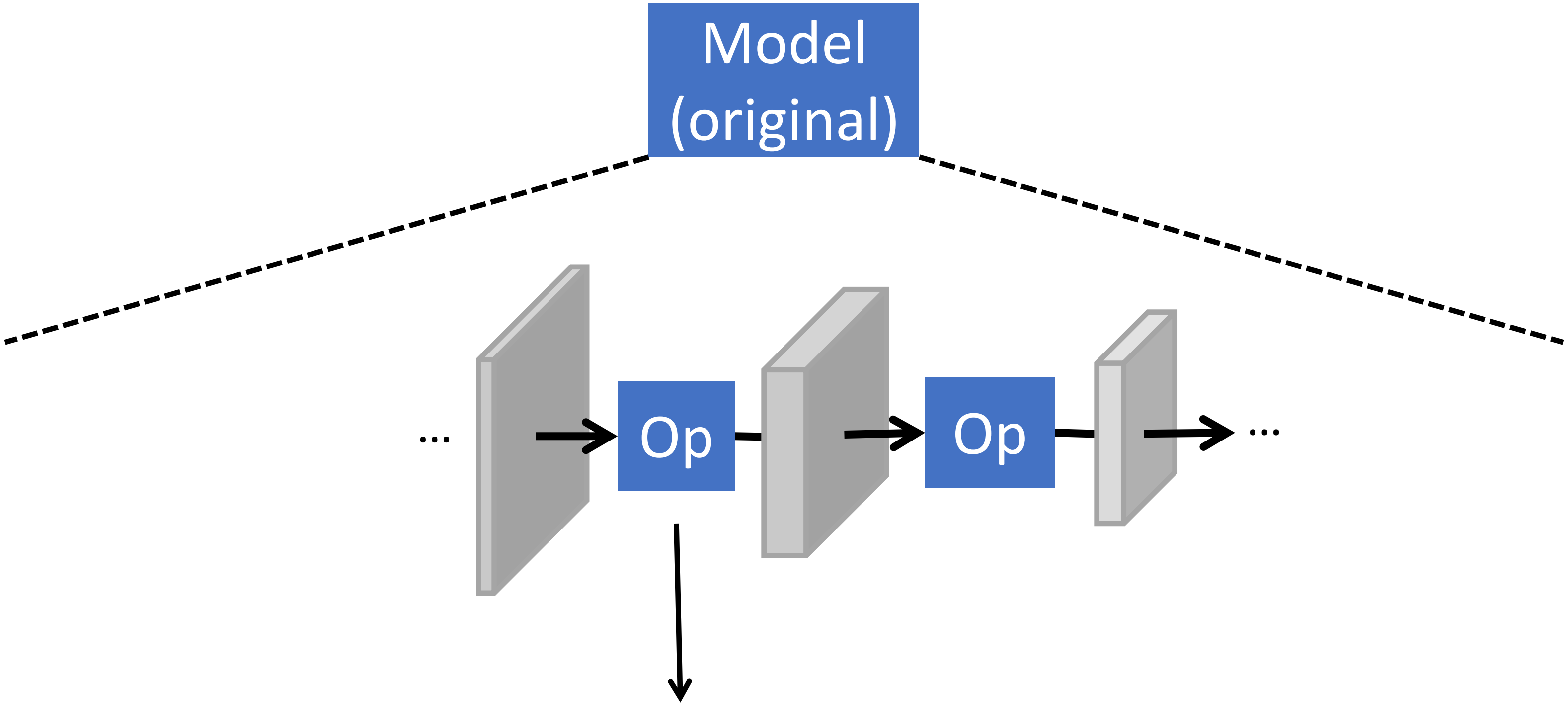


Frame 2 → Model (original) → =

Frame 1 → Model (delta) → +

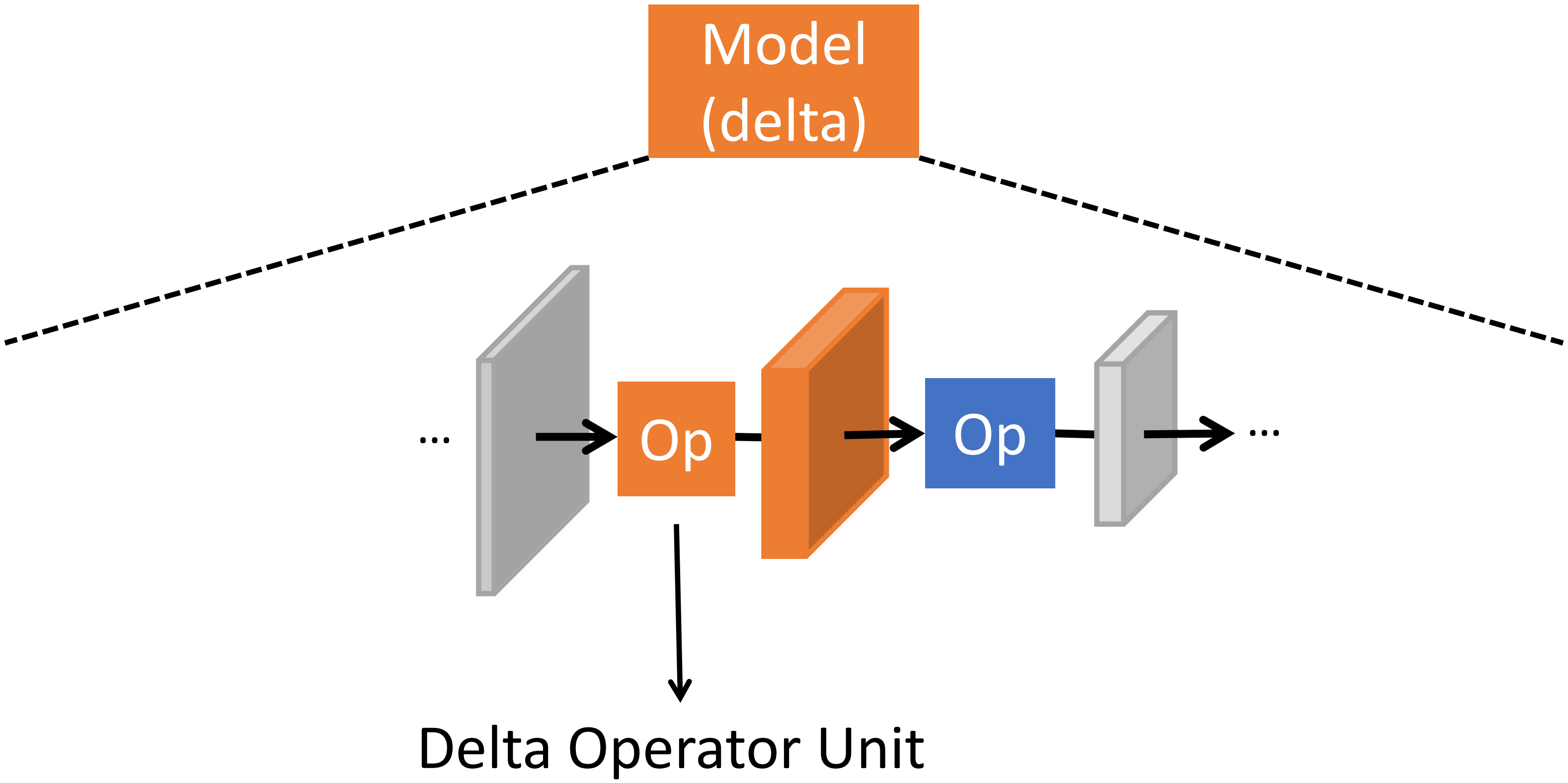
Frame 2 – Frame 1 → Model (delta) →

Example Neural Network Model



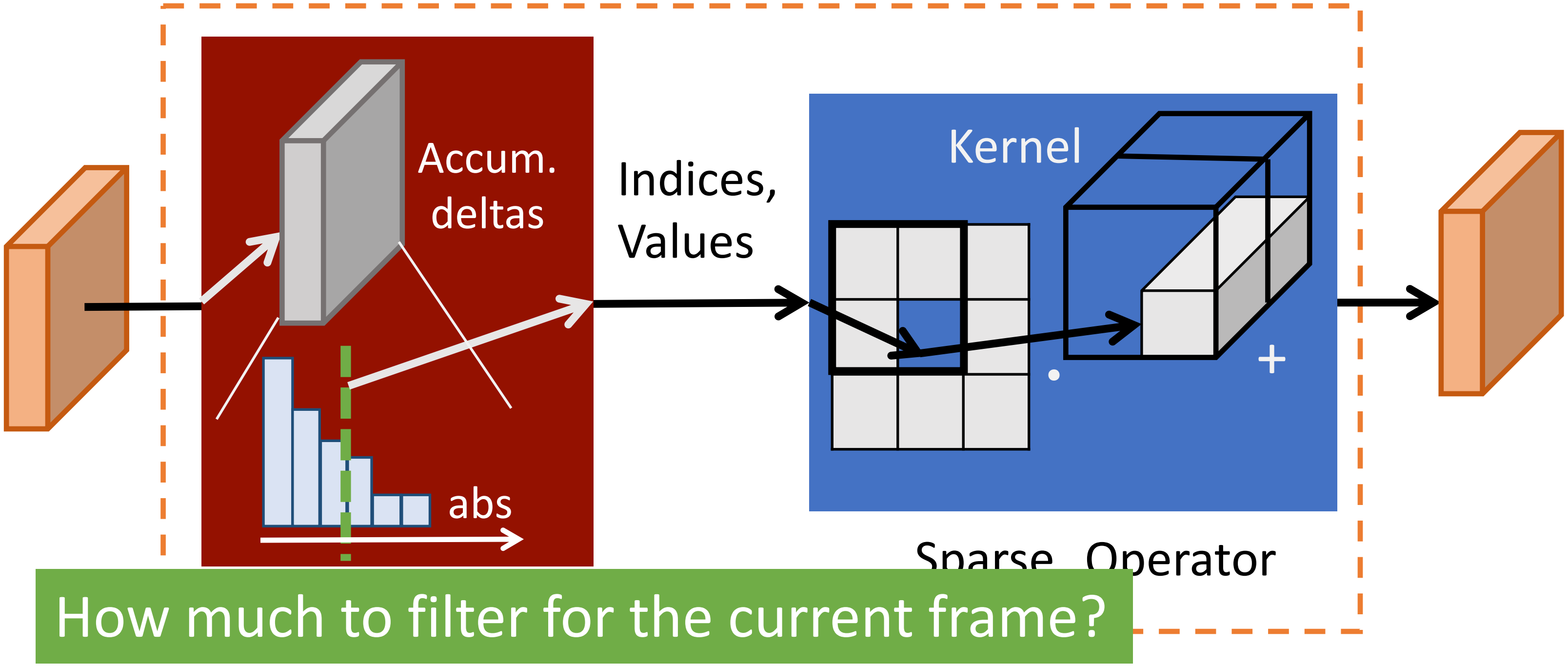
- Example operators: convolution, max pooling, ReLU, …

Example Neural Network Model



- Modify operators to operate on deltas

Delta Operator Unit



How much to filter for the current frame?

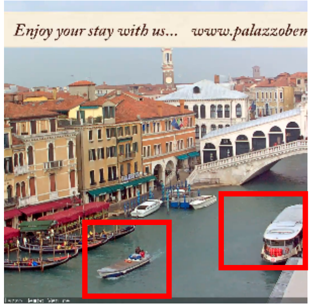
- Sparse operator: takes sparse deltas & outputs delta
  - Saves # of FLOPs by processing delta scalars only
- Filter: send only significant deltas to the operator
  - Builds histogram, keeps small deltas & outputs large deltas

Delta-Based Inference for Videos

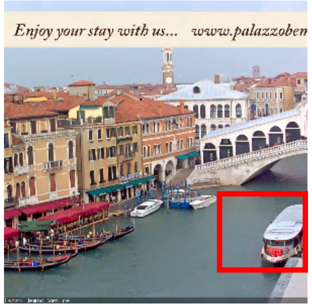
- Modify neural network to take deltas as inputs
- Decide which deltas are significant enough to process
- Generate a network of mixed-type (dense or delta-based) operators

Delta Impacts Output Quality

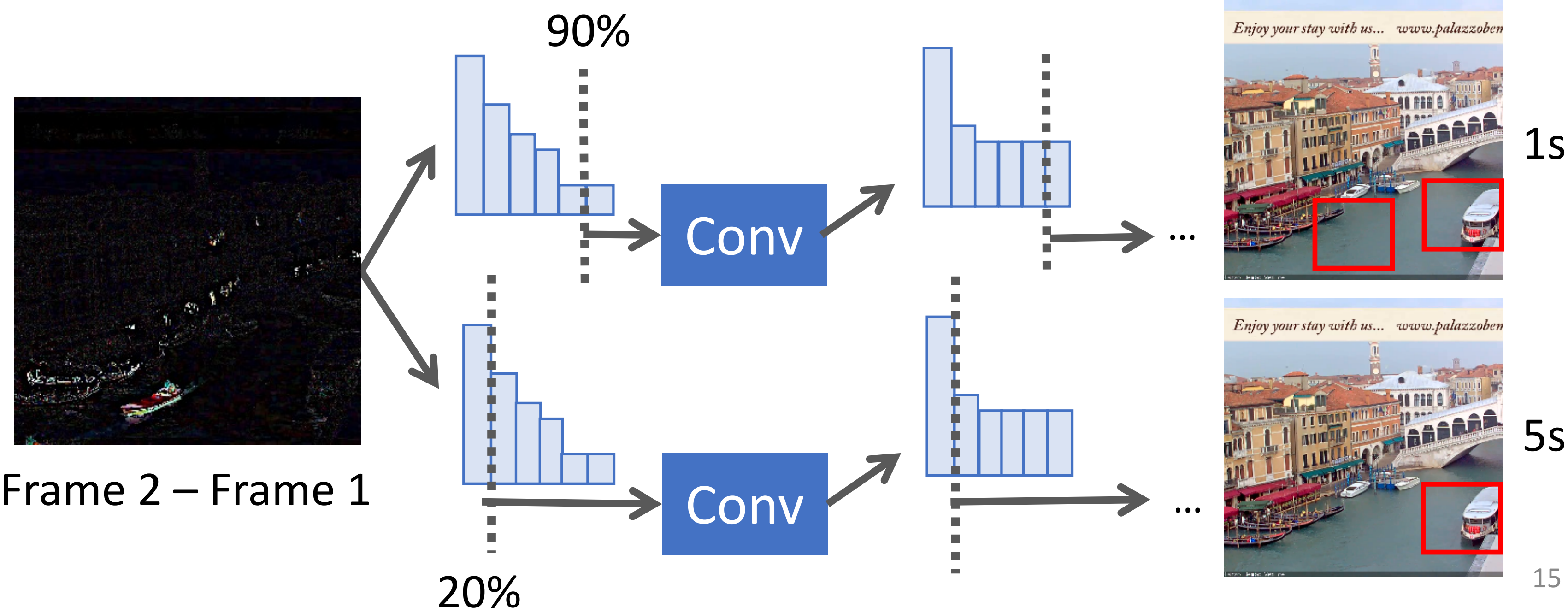
Ground truths:



Frame 1



Frame 2


Enjoy your stay with us... www.palazzoben
Enjoy your stay with us... www.palazzoben

90%

Conv

…

1s

Frame 2 – Frame 1

20%

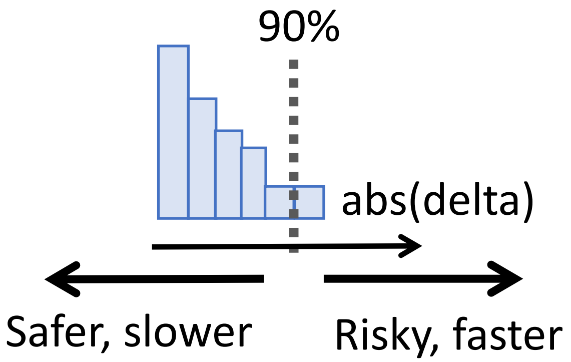
Conv

…

5s

Dynamic Filtering Percentage

- Filtering percentage
  - Higher is faster but risky
  - Lower is safer but slower



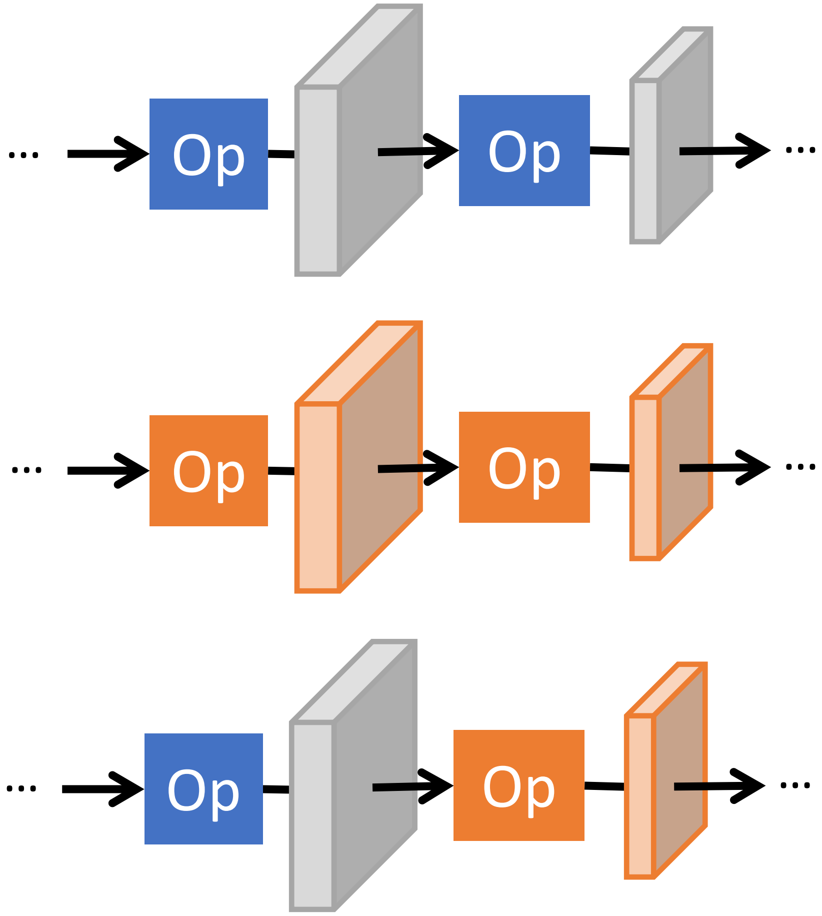
- Target filtering percentage: largest percentage that generates good result
  - Applies to all operators
  - Two approaches: PI controller / Machine learning

Delta-Based Inference for Videos

- Modify neural network to take deltas as inputs
- Decide which deltas are significant enough to process
- Generate a network of mixed-type (dense or delta-based) operators

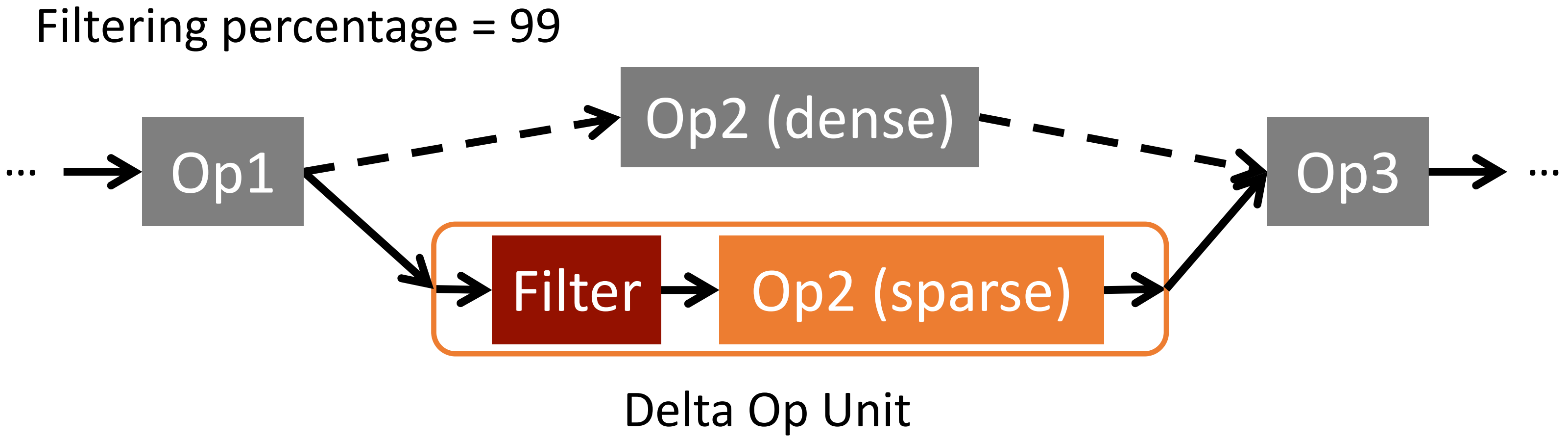
Mixed Network

Filtering percentage is:
low (e.g. 0)
high (e.g. 99)
medium (e.g. 50)



Mixed Network

- Logical plan: a DAG of operators
- Physical plan: choose between delta op unit / original dense implementation
  - Profile each operator with different filtering percentages
  - Pick the faster variant



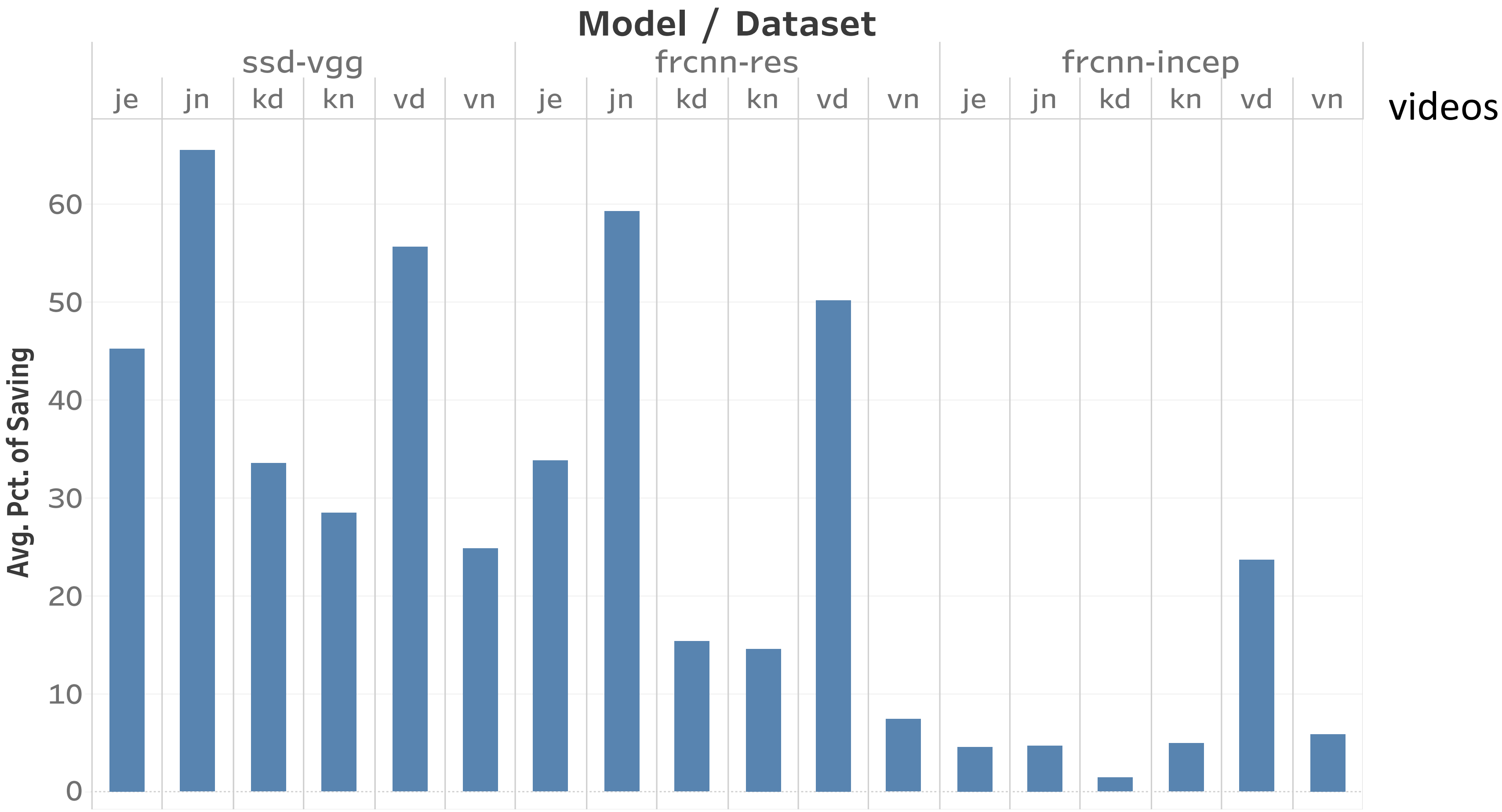
Evaluation: Setup

- Three object detection models

Model	Abbrv.	# of FLOPs	Time
SSD-VGG16	ssd-vgg	123B	3s
FRCNN-RESNET101	frcnn-res	550B	16s
FRCNN-INCEPTION-RESNET-V2	frcnn-incep	1395B	41s

- Six 10-minute videos from three YouTube live streams
  - Taken at different times (e.g. day/night) for each stream
  - Typical objects: people, cars, buses, boats, …
  - One frame per second
- TensorFlow, one CPU thread, Amazon EC2 r3.2xlarge

Evaluation: End-to-End Comparison



- Highest runtime savings by PI controller
  - When error less than a threshold

Deluceva: Conclusion

- Observe rich temporal redundancy in videos
- Accelerate model inference by processing significant deltas only
  - Modify NN models to consist of sparse & dense ops
  - Adjust the filtering granularity adaptively
  - Generate a network of mixed-type operators based on cost models
- Improve runtime up to 67% with low error
- Applies to convolutional neural network models

- Ongoing: GPU implementation, compare to other work