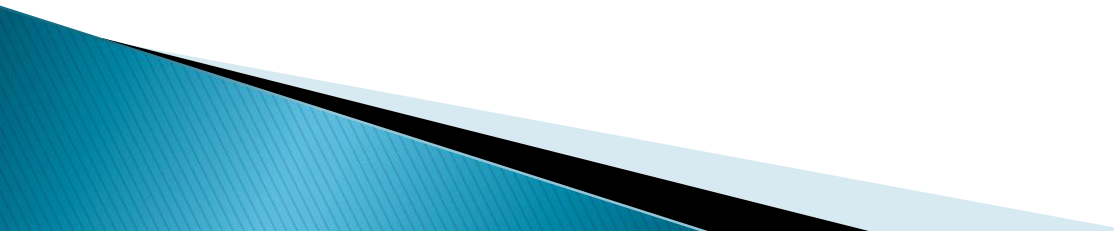


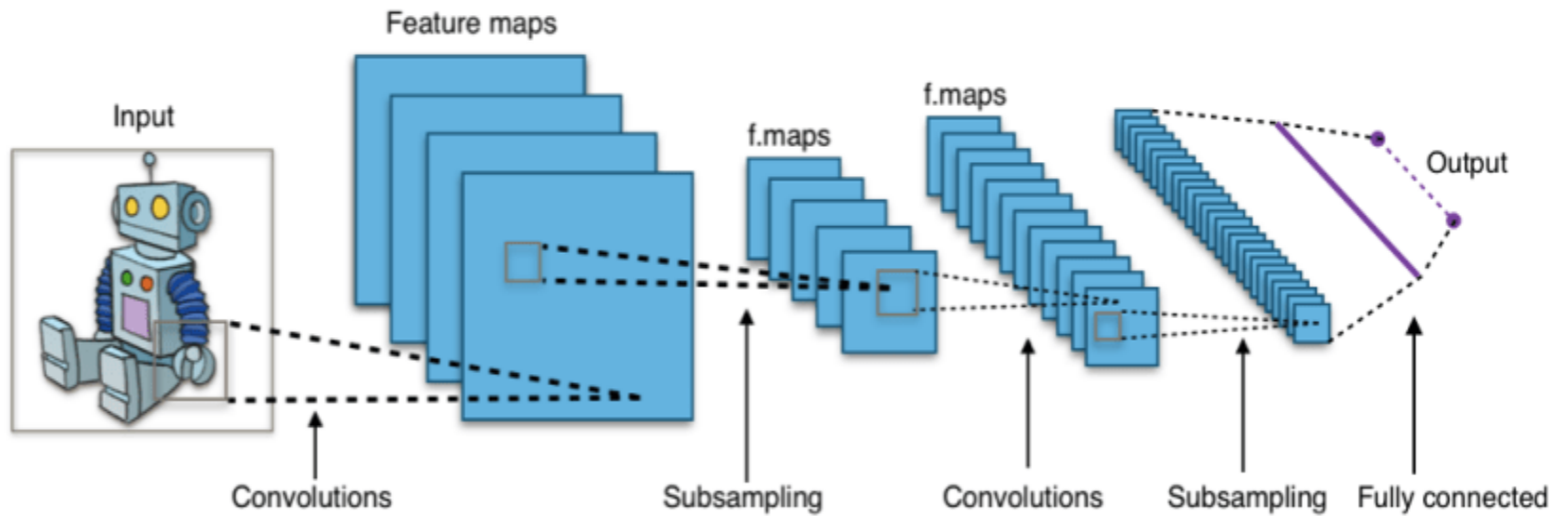
NoScope: Optimizing Neural Network Queries over Video at Scale

Prabhjot Singh

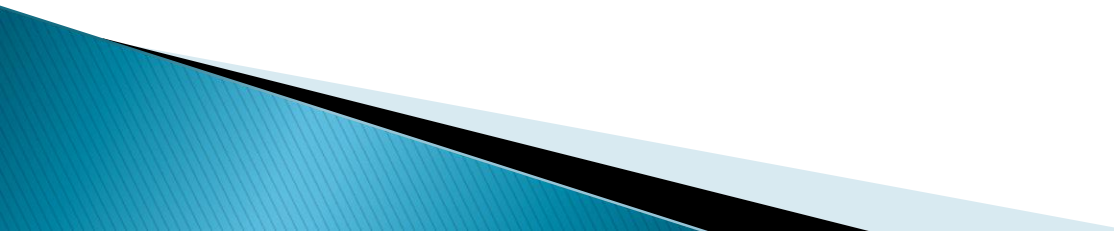
MOTIVATION

- ▶ Video represents a rich source of high-value, high-volume data
 - ▶ We can leverage this video data to answer queries about the physical world, our lives and relationships, and our evolving society.
 - ▶ Unfortunately, applying NNs to video data is prohibitively expensive at scale.
 - ▶ In response NoScope was proposed to solve this problem.
- 

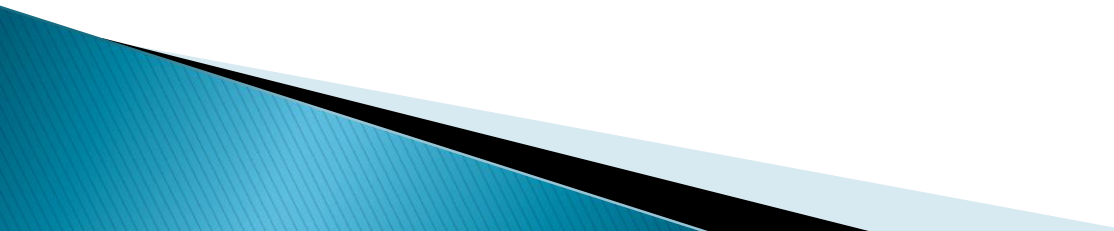
CNN



NN LIMITATIONS

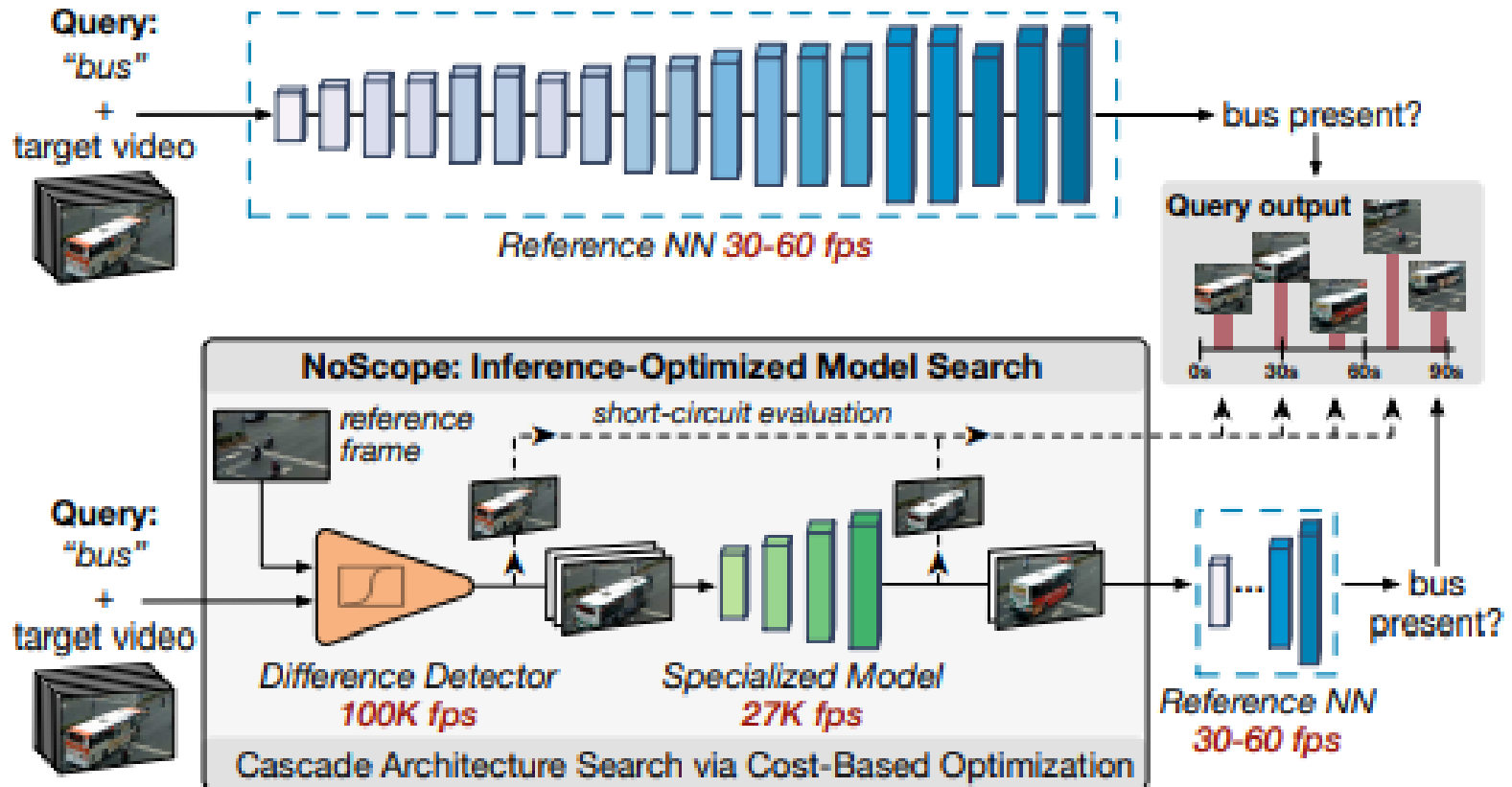
- ▶ Training NNs consists of fitting appropriate weights to a given architecture and this process is computationally expensive.
 - ▶ Only metric of interest in NNs is accuracy, not inference speed.
 - ▶ CNNs that are optimized for inference time primarily use “real time” (i.e., 30 fps) as a target , aiming to evaluate one video in real time on a GPU.
- 

NoScope

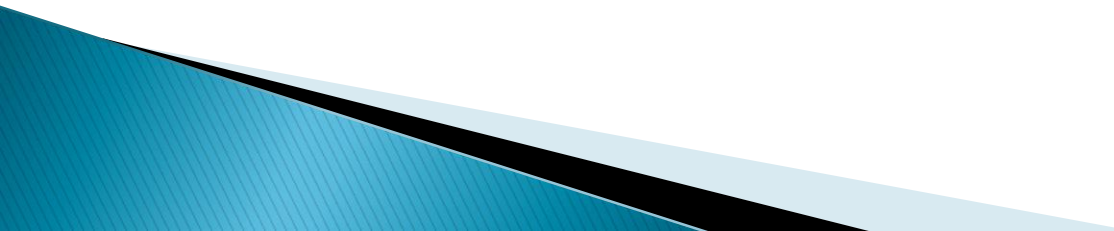
- ▶ A system for accelerating neural network analysis over videos via inference-optimized model search.
 - ▶ Can reduce the cost of neural network video analysis by up to three orders of magnitude.
 - ▶ Automatically searches for and trains a sequence, or cascade, of models.
 - ▶ Preserves the accuracy of the reference network but is specialized to the target video.
- 

CNN vs NoScope

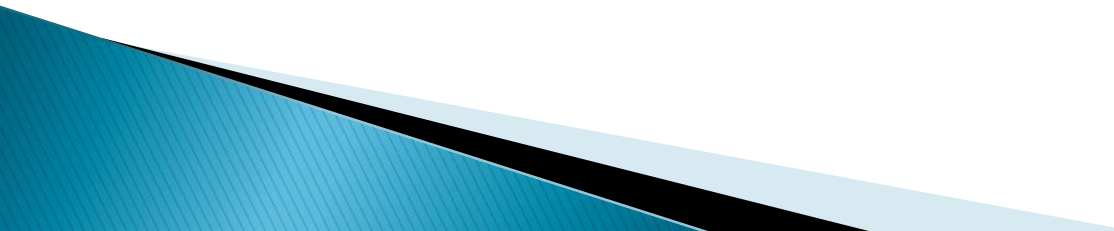
Traditional Deep Neural Network Inference (Frame by Frame)



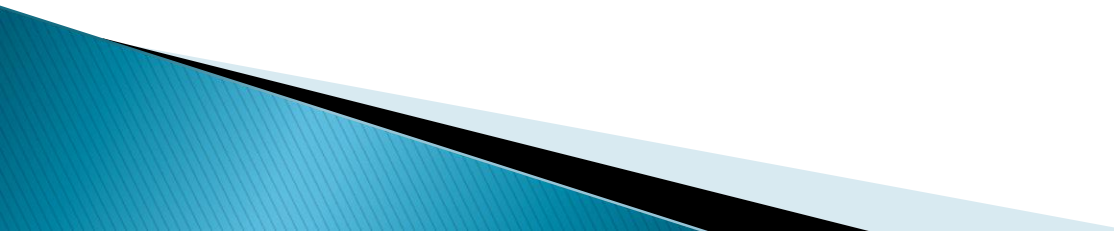
NoScope ARCHITECTURE

- ▶ NOSCOPE is comprised of three components:
 - specialized models
 - difference detectors
 - an inference-optimized cost-based optimizer.
 - ▶ Applies the reference model to a subset of the video, generating labeled examples.
 - ▶ Using these examples, searches for and learns a cascade of cheaper models to accelerate query on the video.
- 

1. MODEL SPECIALIZATION

- ▶ Generic NNs can detect thousands of classes, generality of these methods leads to costly inference.
 - ▶ NOSCOPE performs model specialization by applying a larger, reference model to a target video and uses output of the larger model to train a smaller, specialized model.
 - ▶ Mimics the reference model on the video while requiring fewer computational resources
- 

2. DIFFERENCE DETECTION

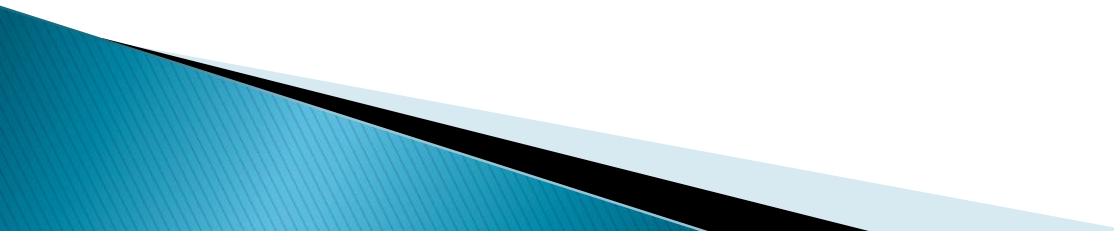
- ▶ Given a labeled video frame and an unlabeled frame, determines whether the unlabeled frame has the same or different label as the labeled frame.
 - ▶ NOSCOPE can quickly determine when video contents have changed
 - ▶ In videos where the frame rate is much higher than the label change rate, can provide up to $90\times$ speedups at inference time.
- 

3. COST-BASED MODEL SEARCH

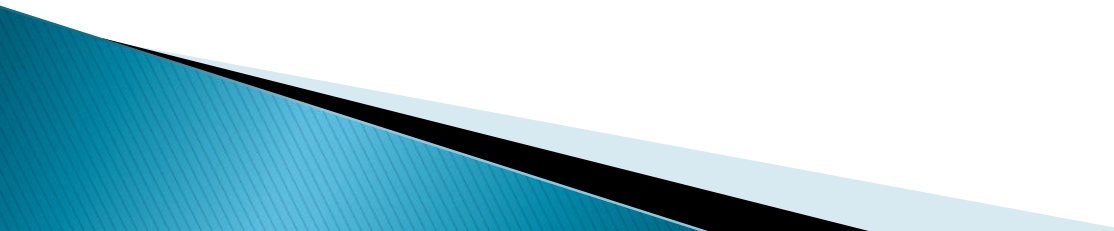
- ▶ NOSCOPE combines model specialization and difference detectors using inference-optimized model search.
- ▶ Takes as input a video as well as target accuracy values, FP^* and FN^*
- ▶ solves the following problem:

	maximize $E(\text{throughput})$
subject to	false positive rate $< FN^*$
and	false negative rate

IMPLEMENTATION

- ▶ Model Search Implementation – written in Python, calls C++ code for difference detectors. YOLOv2 is used as reference model.
 - ▶ Difference Detectors – OpenCV and C++
 - ▶ Specialized Model – TensorFlow framework
- 

LIMITATIONS

- ▶ Fixed-Angle Video – current prototype designed in fixed-angle video.
 - ▶ Model Drift – Assumes that training data obtained is from the same distribution as the subsequent video observed.
 - ▶ Image Batching – implementation batches video frames for greater efficiency.
- 

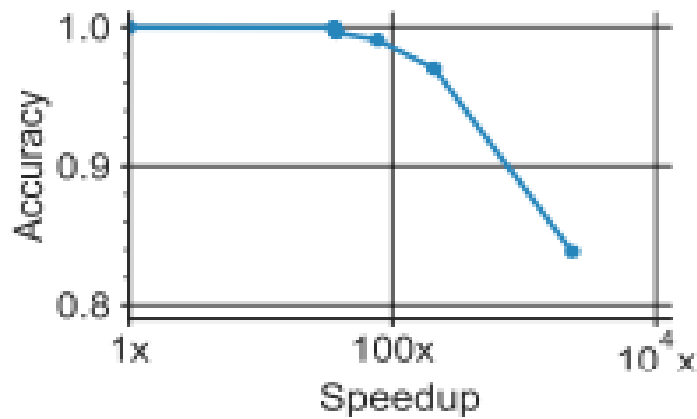
EVALUATION

- ▶ Experimental setup:

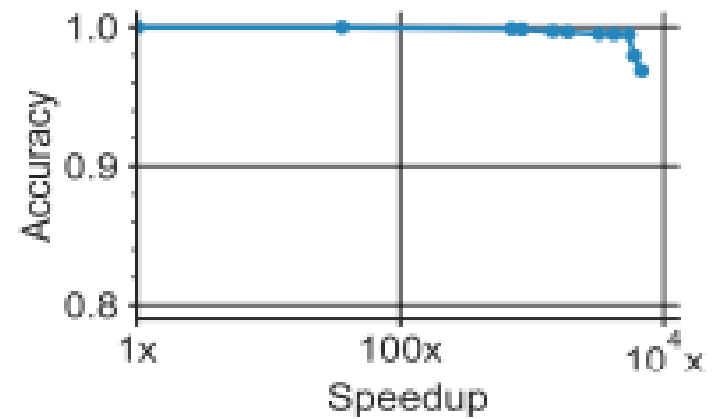
Table 1: Video streams and object labels queried in our evaluation.

Video Name	Object	Resolution	FPS	# Eval frames	Length (hrs)
taipei	bus	1000x570	30	1296k	12.0
coral	person	1280x720	30	1188k	11.0
amsterdam	car	680x420	30	1296k	12.0
night-street	car	1000x530	30	918k	8.5
store	person	1170x1080	30	559k	5.2
elevator	person	640x480	30	592k	5.5
roundabout	car	1280x720	25	731k	8.1

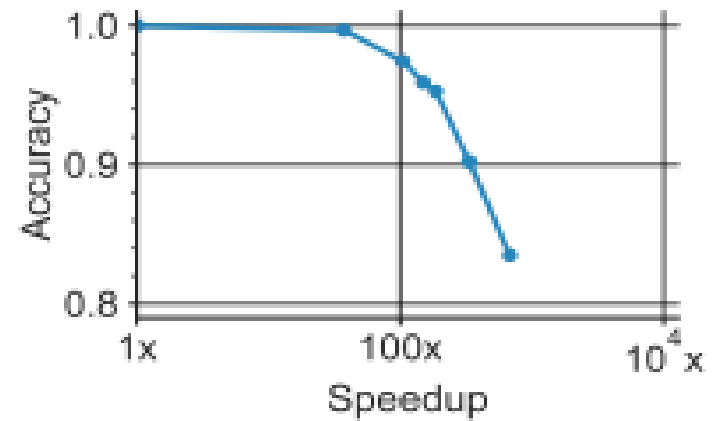
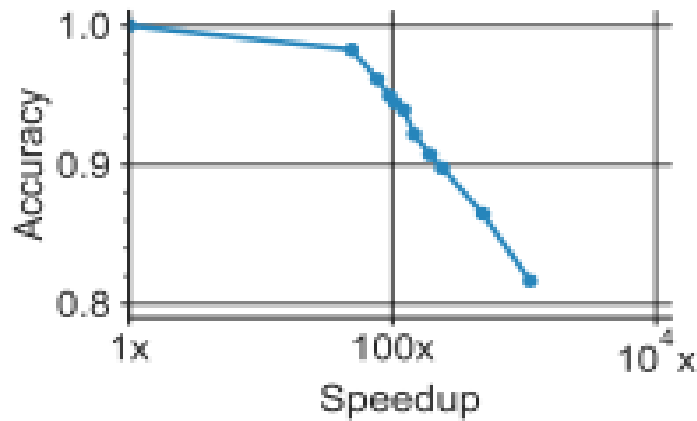
Accuracy vs Speedup



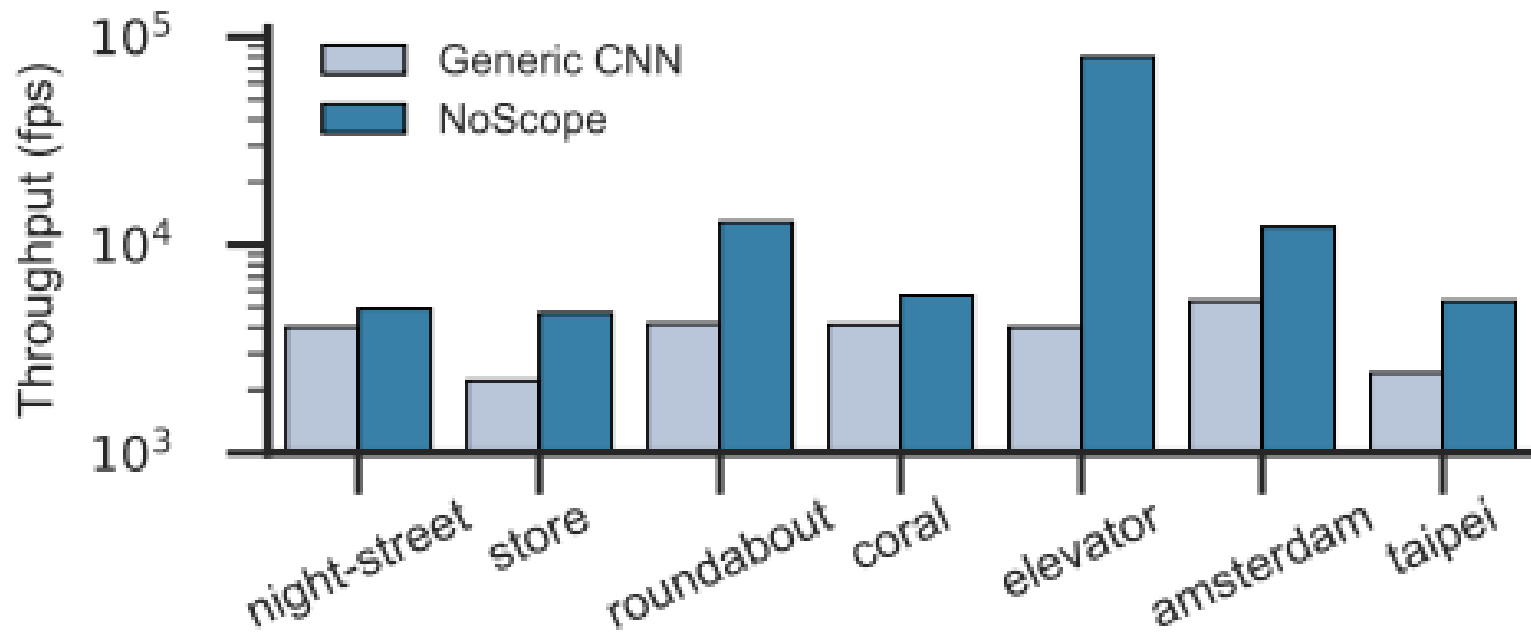
(a) taipei



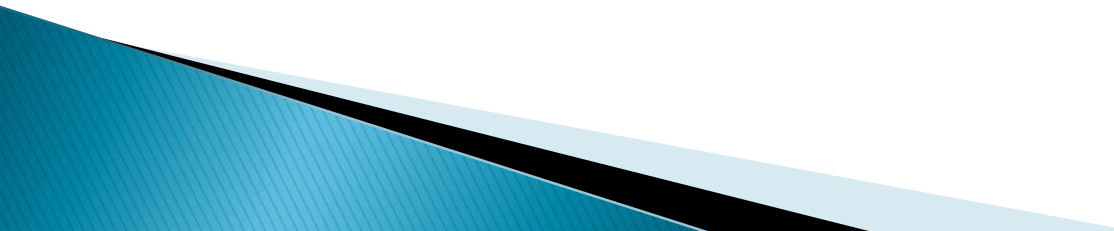
(b) coral



Throughput of Generic NN vs NoScope



CONCLUSION

- ▶ NNs have dramatically improved our ability to extract semantic information from video but to detect objects in video it is prohibitively expensive at scale, currently requiring a dedicated GPU to run at real-time.
 - ▶ NOSCOPE prototype demonstrates that by prioritizing inference time in model architecture search, NNs can be applied to large datasets with computational cost lower by orders of magnitude.
- 

Thank You

