Adaptive Multimodal Learning for Efficient Video Understanding

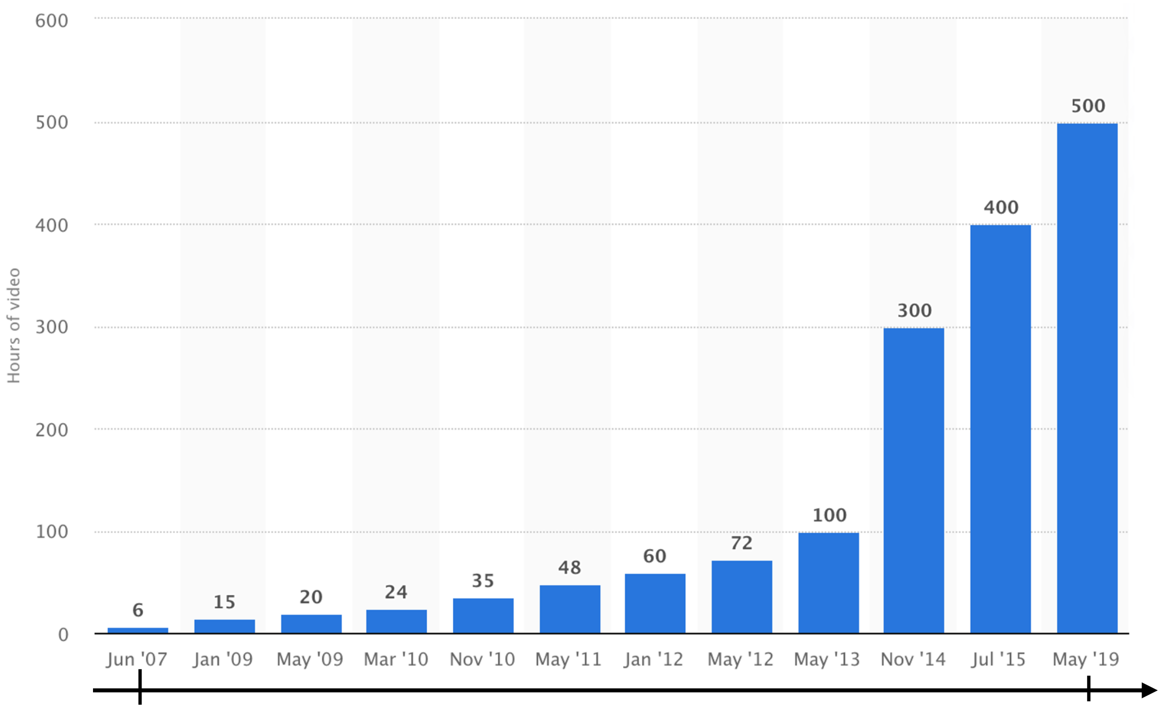
Rogerio Schmidt Feris

Principal Scientist and Manager

MIT-IBM Watson AI Lab

Huge Growth of Multimodal Video Data

500+ hours of video are uploaded to YouTube every minute

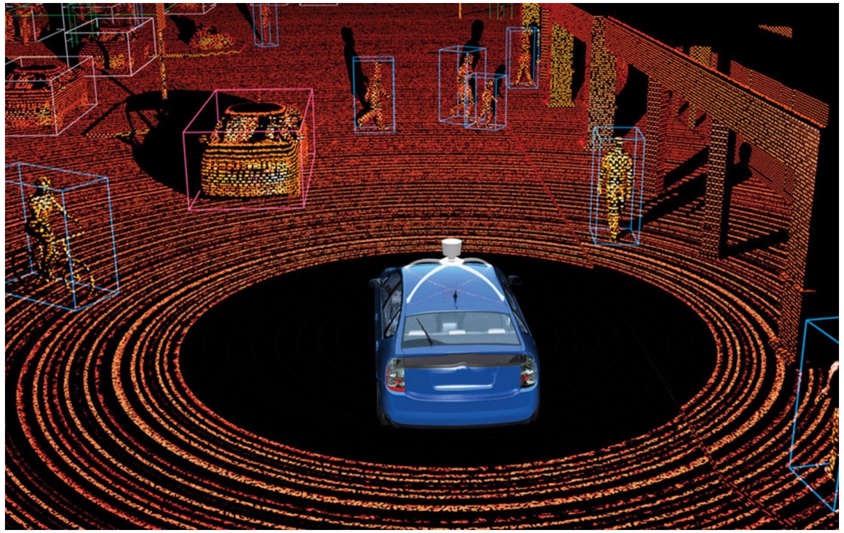


2007 2019

More than a billion hours of video are watched every day in youtube

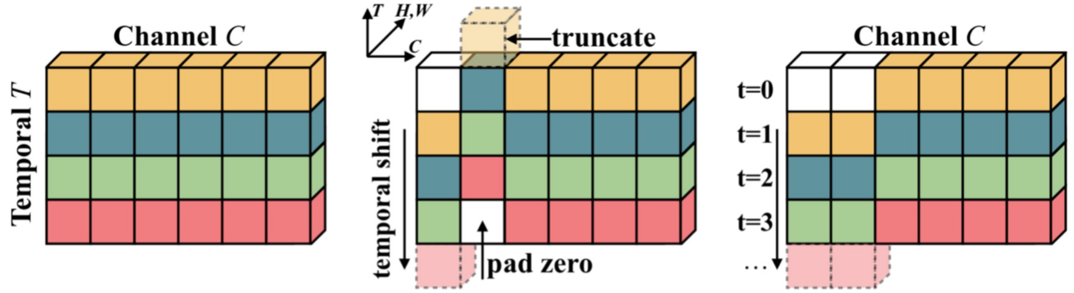
Despacito – 7+ Billion views

Huge Growth of Multimodal Video Data

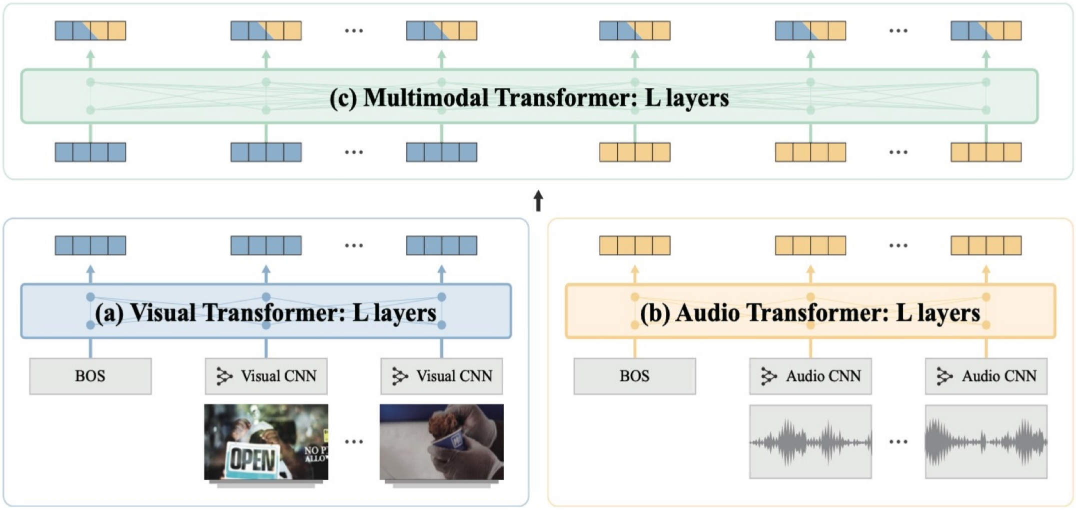


Video Model Compression and Acceleration

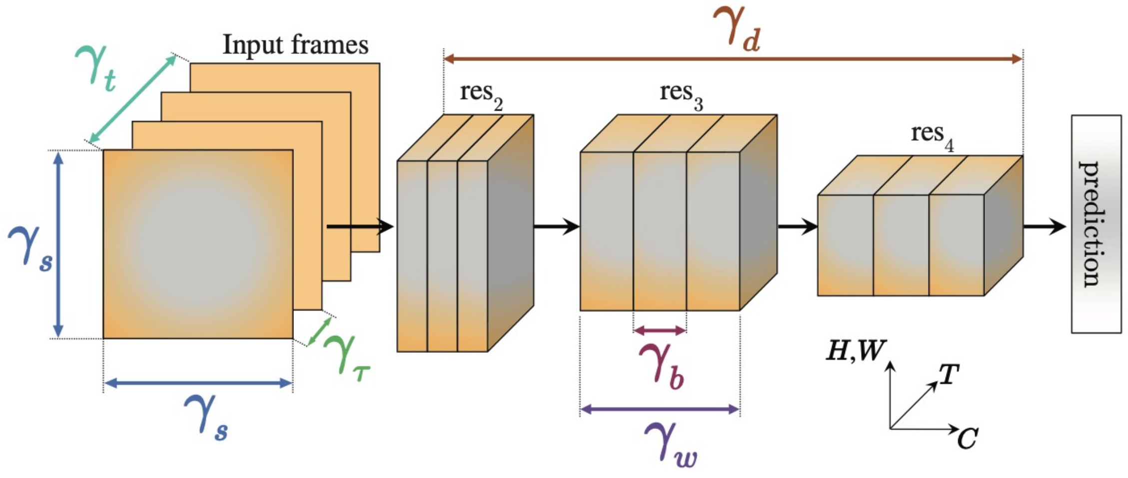
TSM [Lin et al, 2019]



AVBert [Lee et al, 2021]



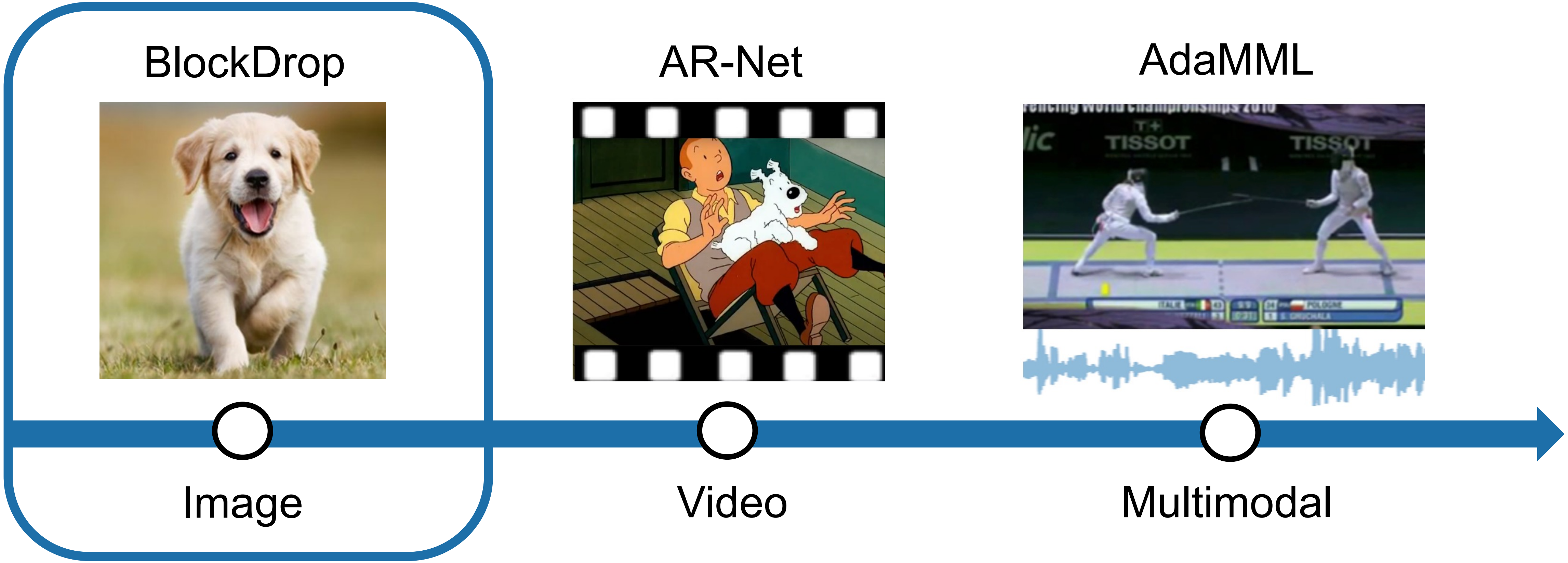
X3D [Feichtenhofer, 2020]



Most methods rely on **one-size-fits-all networks** that require the same fixed set of features to be extracted for all inputs, no matter their complexity

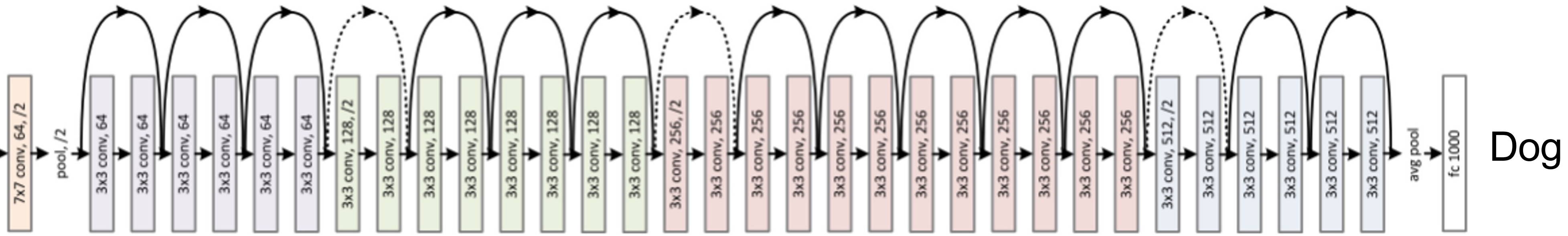
This talk: Dynamic (Adaptive) Neural Networks for Efficient Inference

- Networks models that are dynamically reconfigured **depending on the input**



- Conditional Computation [Bengio et al, 2013/2016]

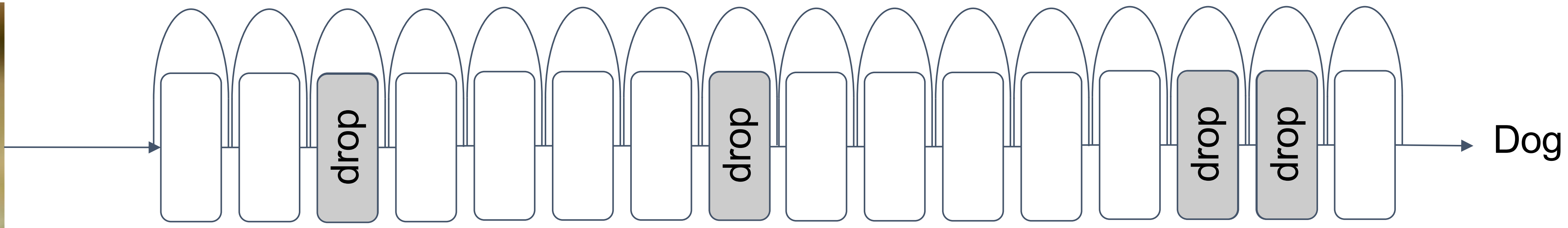
BlockDrop: Dynamic Inference Paths in Residual Networks



Do we really need to run 100+ layers / residual blocks of a neural network if we have an "easy" input image?

[Wu & Nagarajan et al, CVPR 2018]

BlockDrop: Dynamic Inference Paths in Residual Networks



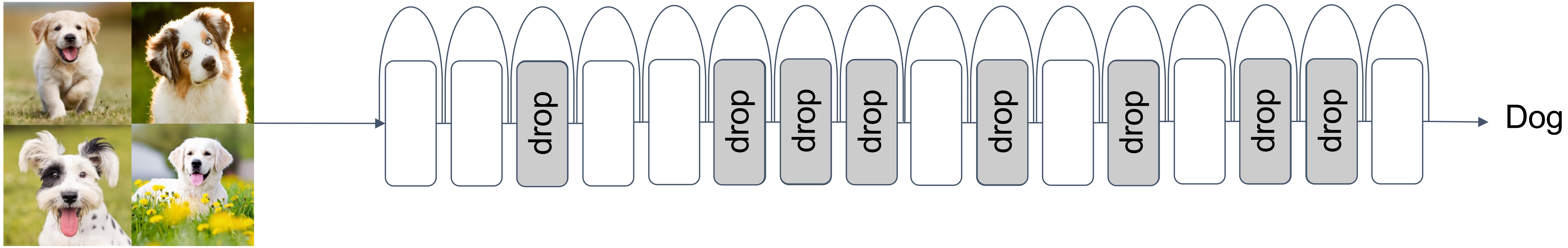
"Dropping some blocks during testing doesn't hurt performance much"

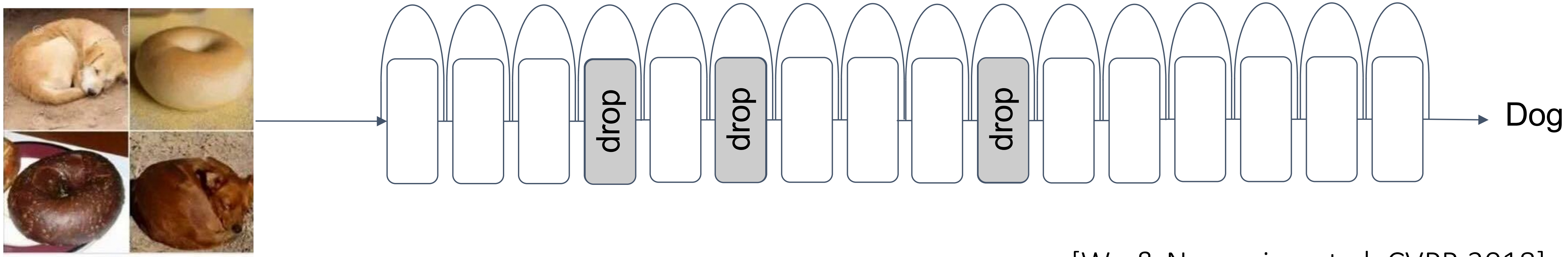
(Veit et al., NIPS 16)

[Wu & Nagarajan et al, CVPR 2018]

BlockDrop: Dynamic Inference Paths in Residual Networks

How to determine which blocks to drop depending on the input image?





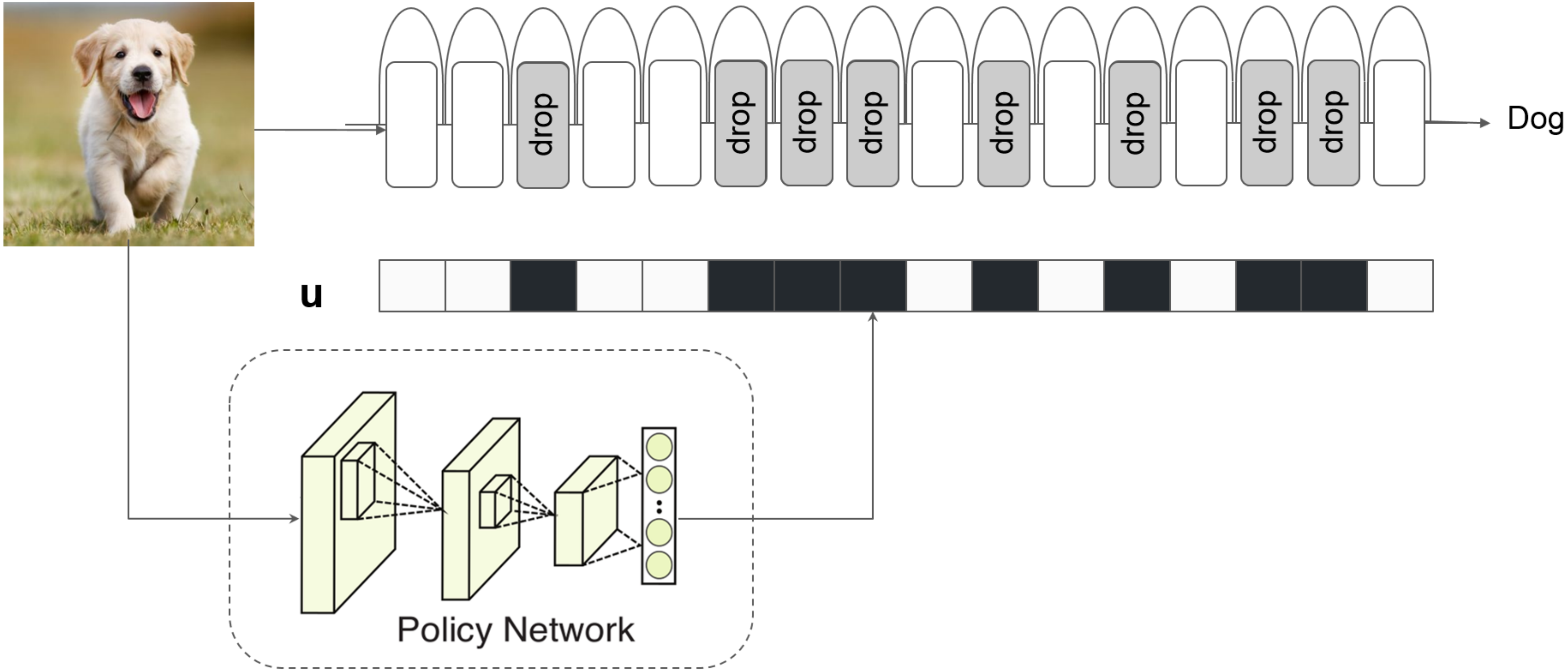
[Wu & Nagarajan et al, CVPR 2018]

Our Idea: BlockDrop

Predict which blocks to drop conditioned on the input image, in one shot, without compromising accuracy

[Wu & Nagarajan et al, CVPR 2018]

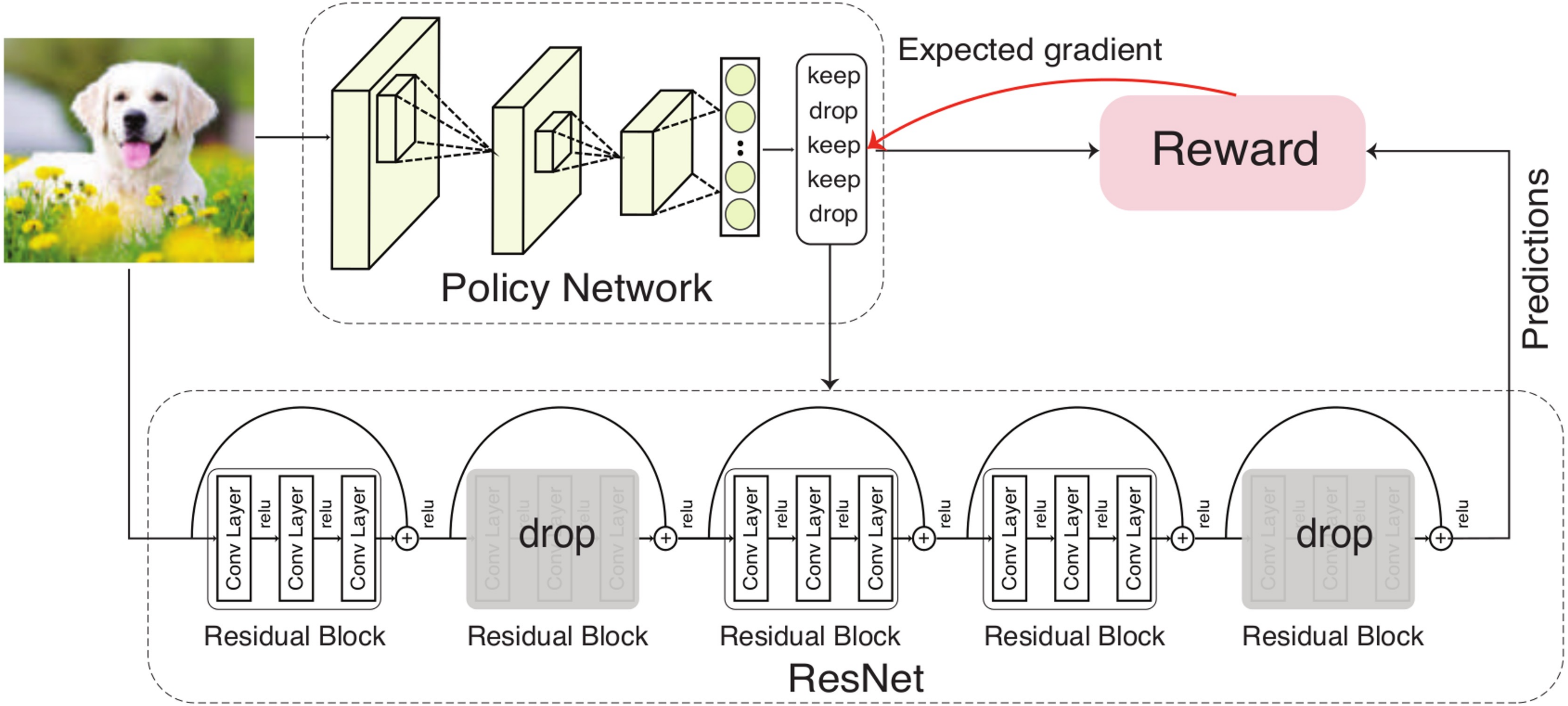
BlockDrop: Dynamic Inference Paths in Residual Networks



[Wu & Nagarajan et al, CVPR 2018]

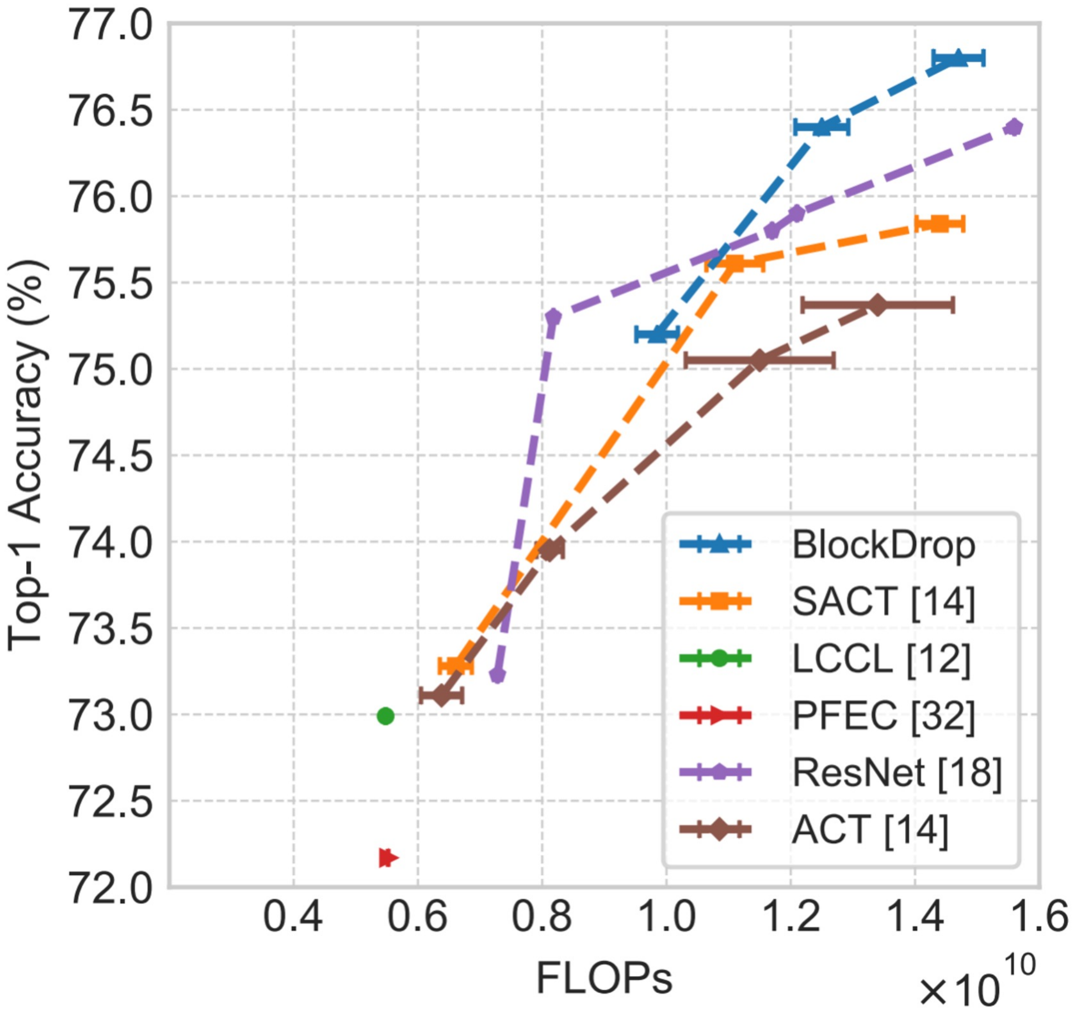
BlockDrop: Dynamic Inference Paths in Residual Networks

Policy Network Training through Reinforcement Learning



[Wu & Nagarajan et al, CVPR 2018]

BlockDrop: Dynamic Inference Paths in Residual Networks



Results on ImageNet:

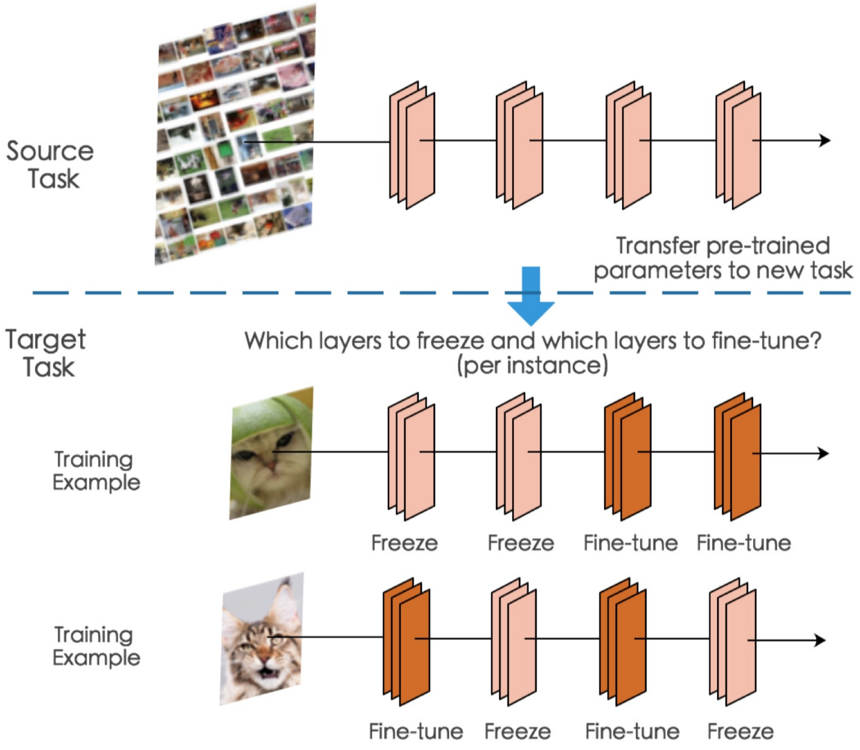
20% - 36% computational savings (FLOPs)

Complementary to other model compression techniques

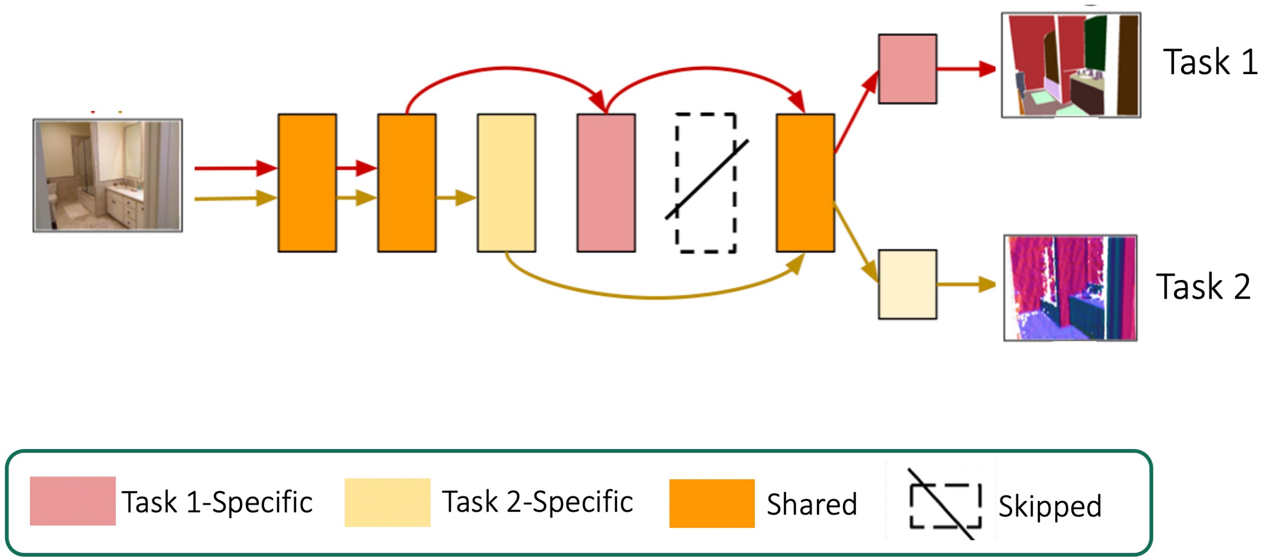
[Wu & Nagarajan et al, CVPR 2018]

See Also:

SpotTune, CVPR 2019



Adashare, NeurIPS 2020

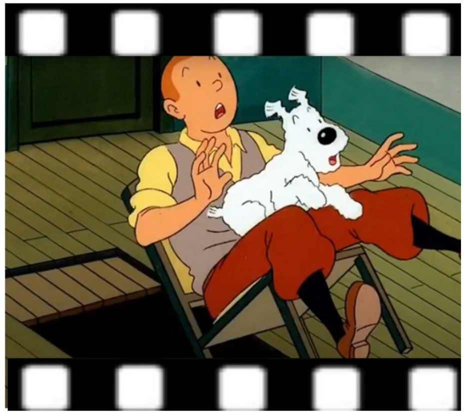


This talk: Dynamic (Adaptive) Neural Networks for Efficient Inference

- Networks models that are dynamically reconfigured **depending on the input**

BlockDrop

AR-Net



AdaMML

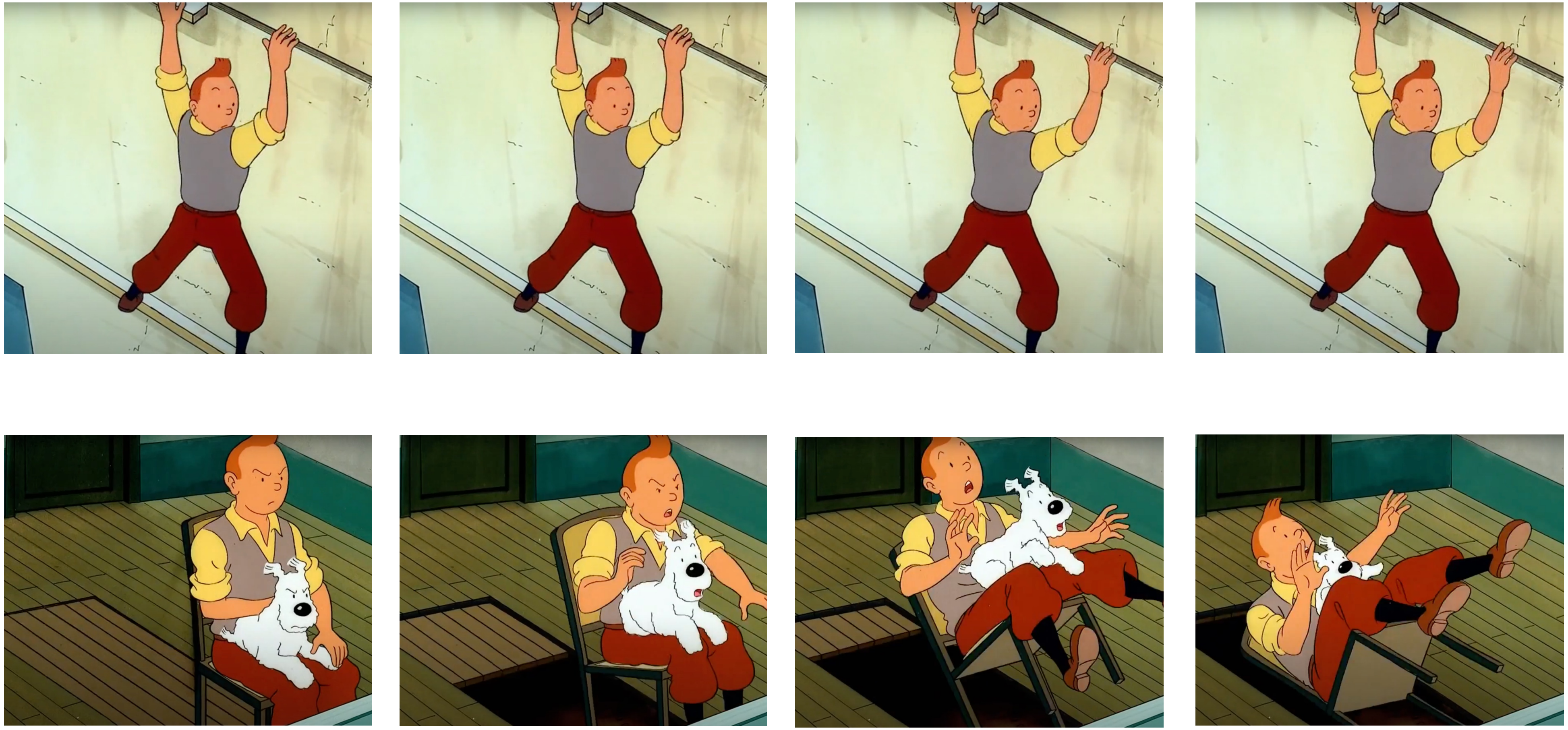
Image

Video

Multimodal

- Conditional Computation [Bengio et al, 2013/2016]

Videos are redundant.
Do we need all frames of a video to make a prediction?



Different video segments have different levels of redundancy

[Meng et al, ECCV 2020]

How about Spatial Resolution?

- Most methods process all video frames at the same resolution

High-Resolution → More Accuracy, Less Efficiency

Low-Resolution → Less Accuracy, More Efficiency

[Meng et al, ECCV 2020]

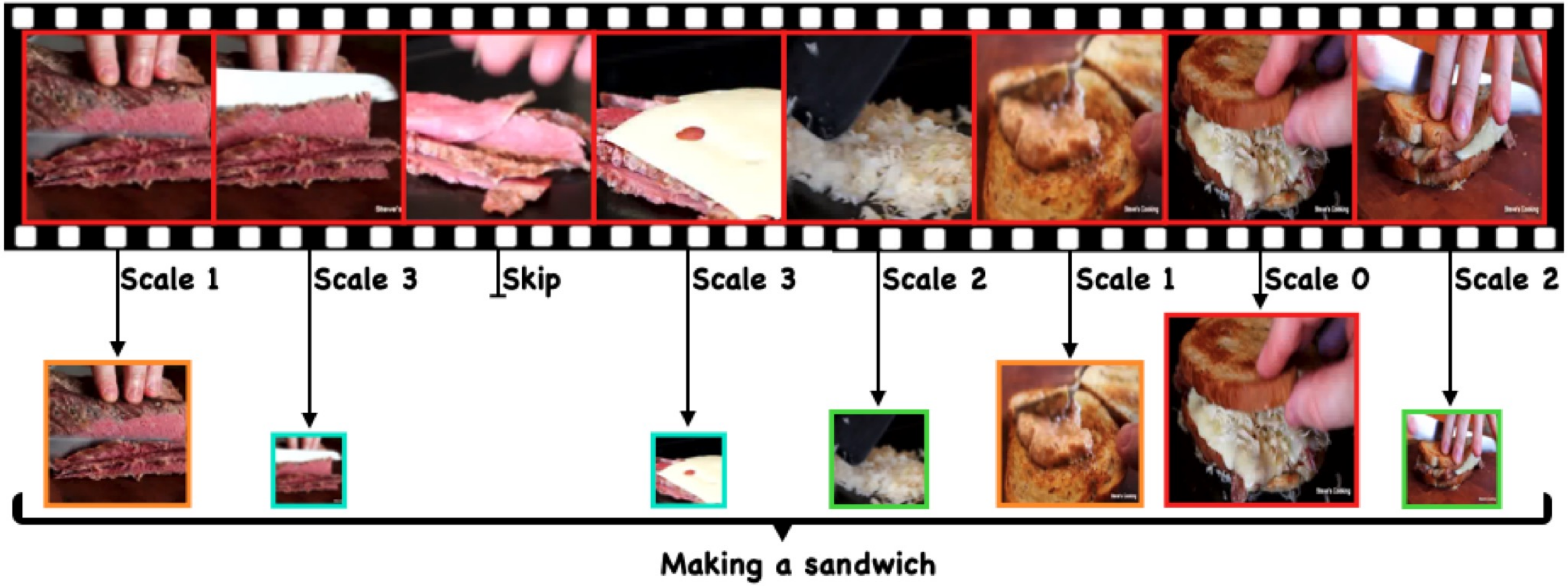
Our Idea: AR-Net

Adaptively select the right data, at the right level of detail (resolution), to make video recognition more efficient

[Meng et al, ECCV 2020]

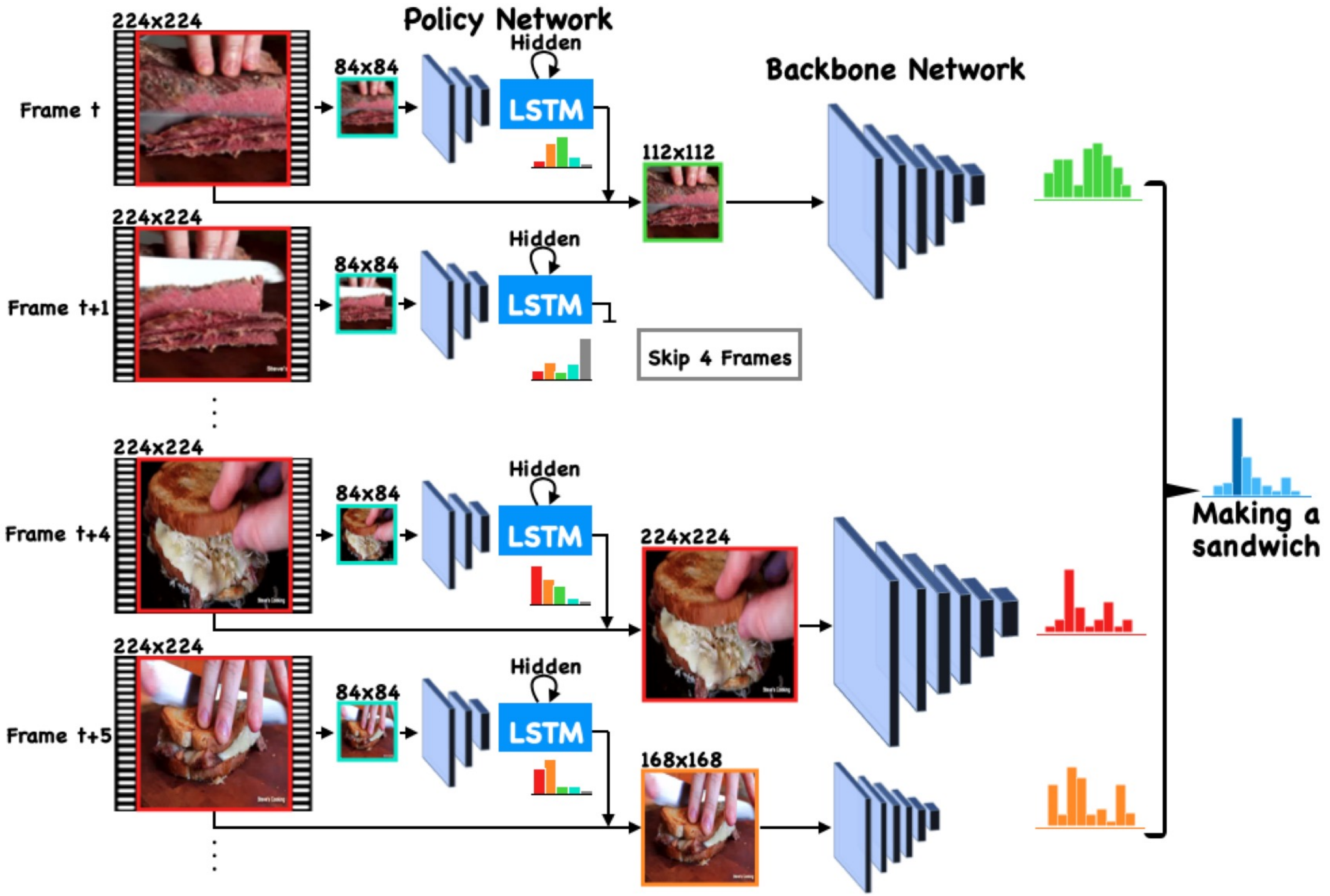
AR-Net: Adaptive Frame Resolution for Efficient Action Recognition

Our Idea: AR-Net



[Meng et al, ECCV 2020]

AR-Net: Adaptive Frame Resolution for Efficient Action Recognition

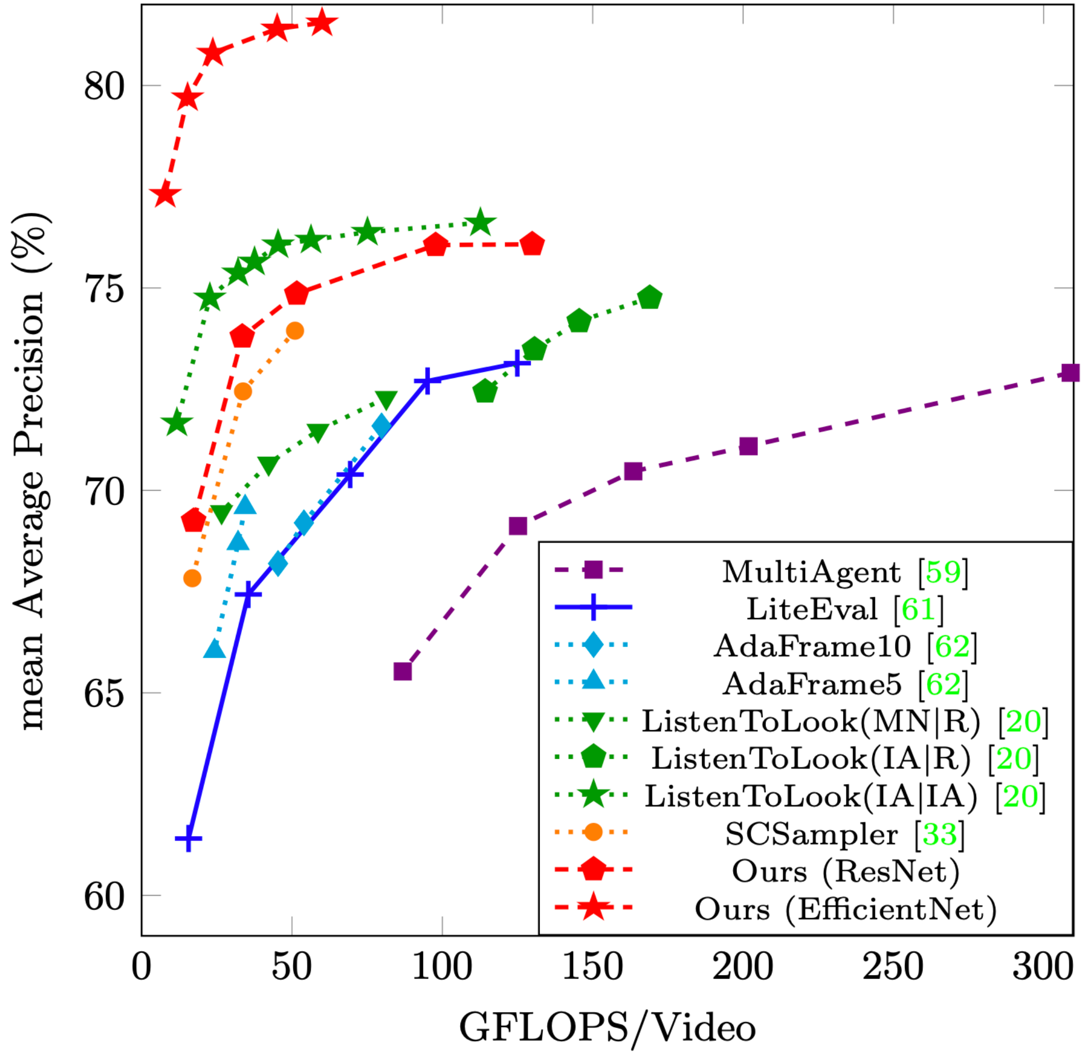


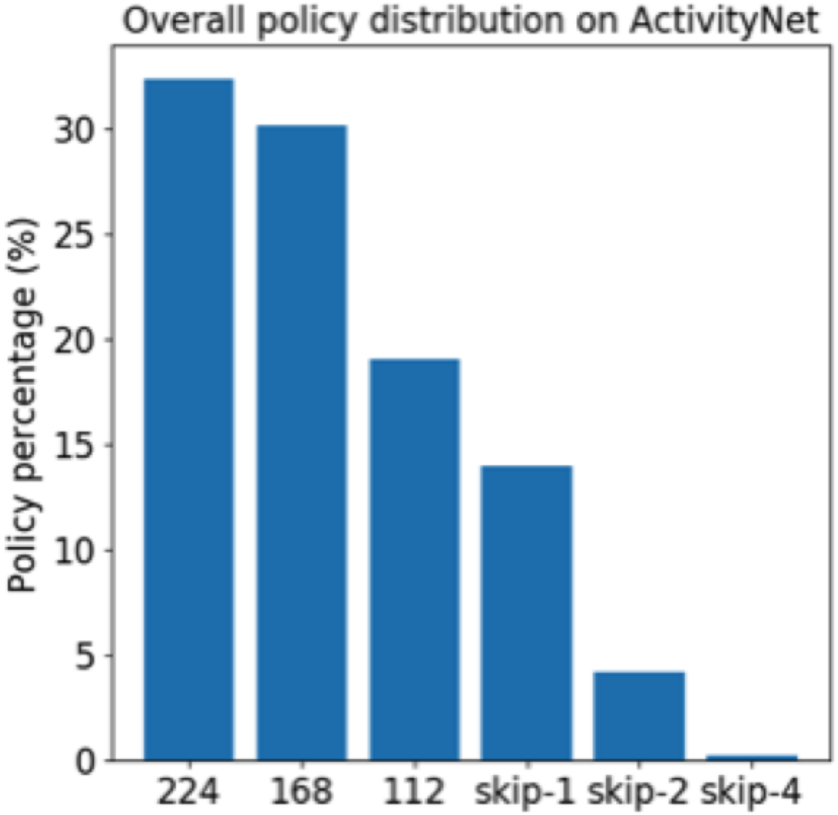
The policy network is trained using Gumbel Softmax sampling (instead of reinforcement learning)

[Meng et al, ECCV 2020]

AR-Net: Adaptive Frame Resolution for Efficient Action Recognition

Results - ActivityNet

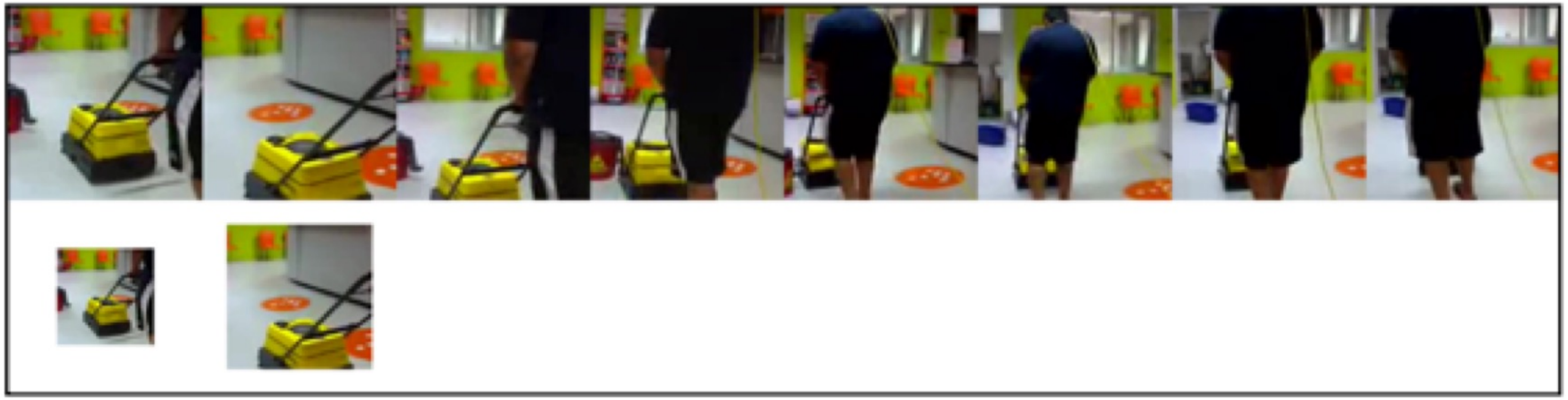




[Meng et al, ECCV 2020]

AR-Net: Adaptive Frame Resolution for Efficient Action Recognition

Qualitative Results



Cleaning Floor (easy)

Fireworks (easy)

[Meng et al, ECCV 2020]

AR-Net: Adaptive Frame Resolution for Efficient Action Recognition

Qualitative Results

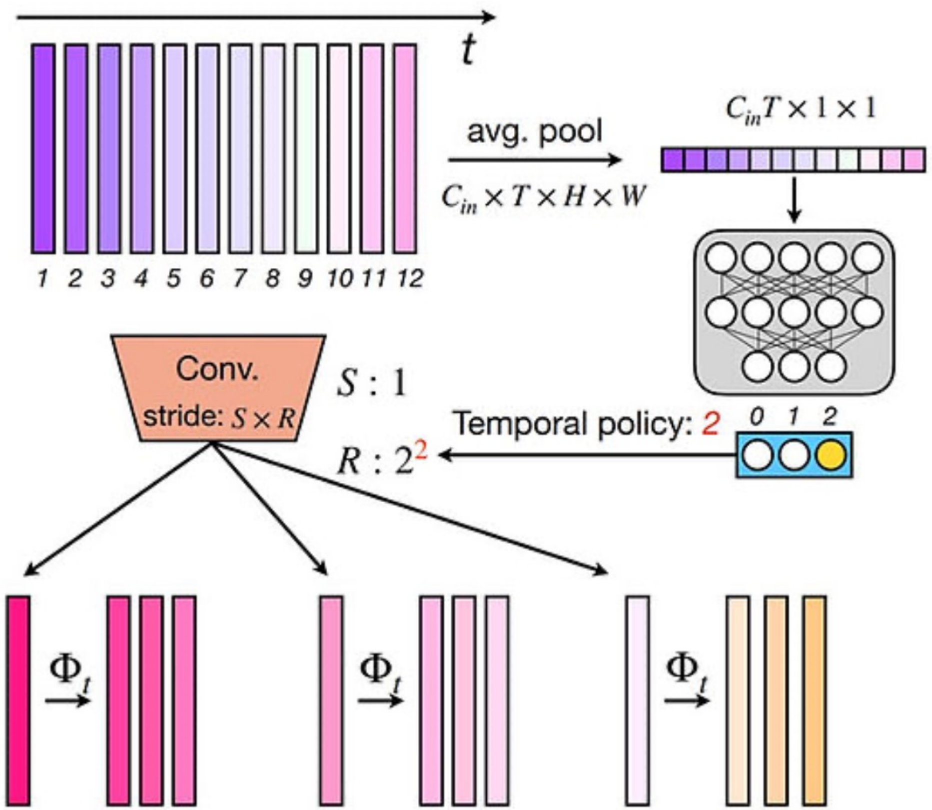
Making Salad (Medium)

Assembling a computer (Hard)

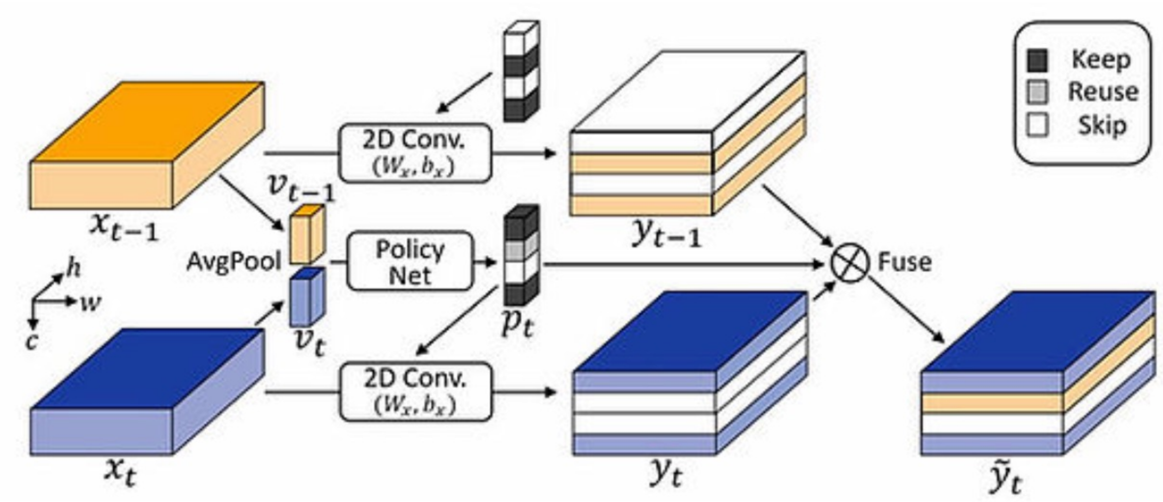
[Meng et al, ECCV 2020]

See Also:

VA-RED^2, ICLR 2021

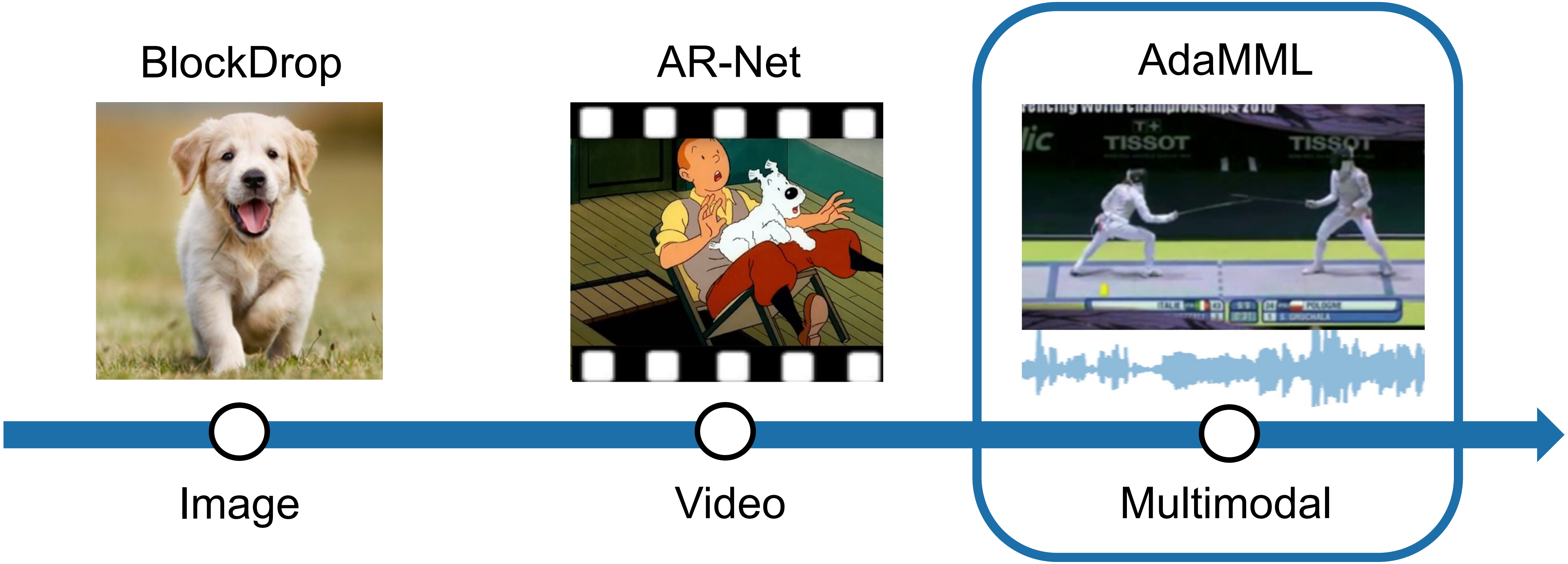


AdaFuse, ICLR 2021



This talk: Dynamic (Adaptive) Neural Networks for Efficient Inference

- Networks models that are dynamically reconfigured **depending on the input**



- Conditional Computation [Bengio et al, 2013/2016]

Observation: For a given video segment, not all modalities may be required or relevant for recognizing a particular action class

Action: Running

(Commentator talking about the weather)

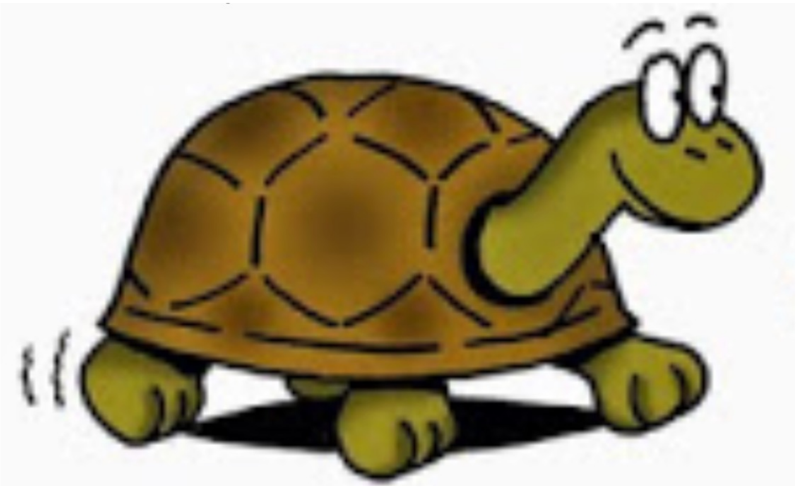
Video Frame

Audio

Some modalities require more computation than others

Audio (efficient)

Optical Flow
(expensive)

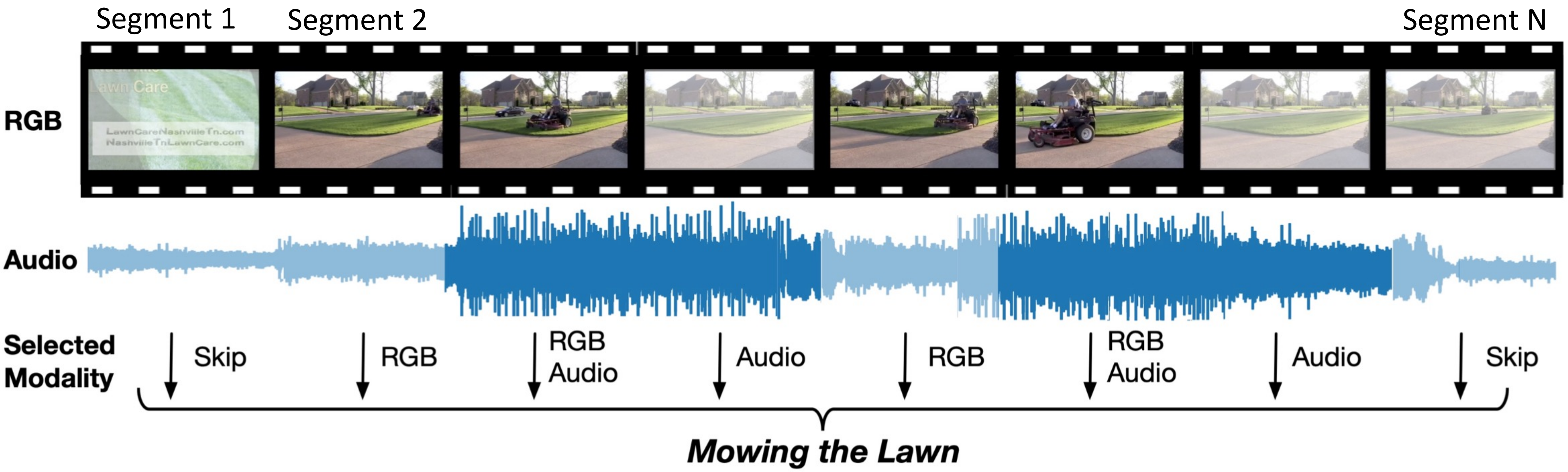


Our Idea: AdaMML

Predict which modality to use for each video segment (conditioned on the input) so as to maximize action recognition accuracy and efficiency

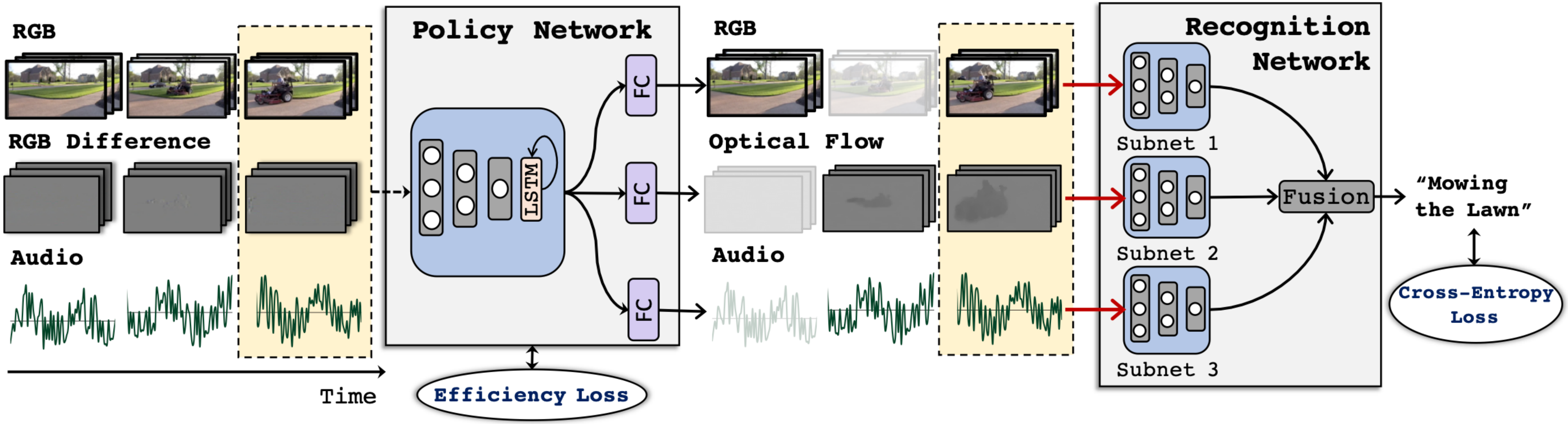
[Panda and Chen, Arxiv 2021]

Our Idea: AdaMML



[Panda and Chen, Arxiv 2021]

Approach



[Panda and Chen, Arxiv 2021]

Policy Network

- RGB difference as an efficient proxy for optical flow
- Input Data is subsampled (both spatially and temporally)
- Lightweight Backbone (MobileNetV2)
- Gumbel Softmax Sampling

[Panda and Chen, Arxiv 2021]

Loss Function

$$\mathbb{E}_{(V,y)\sim\mathcal{D}_{train}}\left[-y\log(\mathcal{P}(V;\Theta))+\sum_{k=1}^{K}\lambda_k\mathcal{C}_k\right]$$

Cross-Entropy + Efficiency Loss

$$\mathcal{C}_k=\begin{cases}(\frac{|U_k|_0}{C})^2 & \text{if correct}\\ \gamma & \text{otherwise}\end{cases}$$

Percentage of used video segments per modality K

Penalty for misclassification

[Panda and Chen, Arxiv 2021]

RGB + Audio (Kinetics-Sounds)

Dataset	Kinetics-Sounds			
		Selection Rate (%)		
Method	Acc. (%)	RGB	Audio	GFLOPs
RGB	82.85	100	–	141.36
Audio	65.49	–	100	3.82
Weighted Fusion	87.86	100	100	145.17
AdaMML	**88.17**	46.47	94.15	**76.45** (-47.3%)

[Panda and Chen, Arxiv 2021]

RGB+Audio+Flow (Kinetics-Sounds)

Method	Acc. (%)	Selection Rate (%) RGB	Flow	Audio	GFLOPs
RGB	82.85	100	–	–	141.36
Flow	75.73	–	100	–	163.39
Audio	65.49	–	–	100	3.82
Weighted Fusion	88.25	100	100	100	308.56
`AdaMML-Flow`	88.54	56.13	20.31	97.49	**132.94** (**-56.9%**)
`AdaMML-RGBDiff`	**89.06**	55.06	26.82	95.12	141.97 (**-54.0%**)

[Panda and Chen, Arxiv 2021]

Qualitative Results

[Panda and Chen, Arxiv 2021]

Qualitative Results

[Panda and Chen, Arxiv 2021]

Qualitative Results

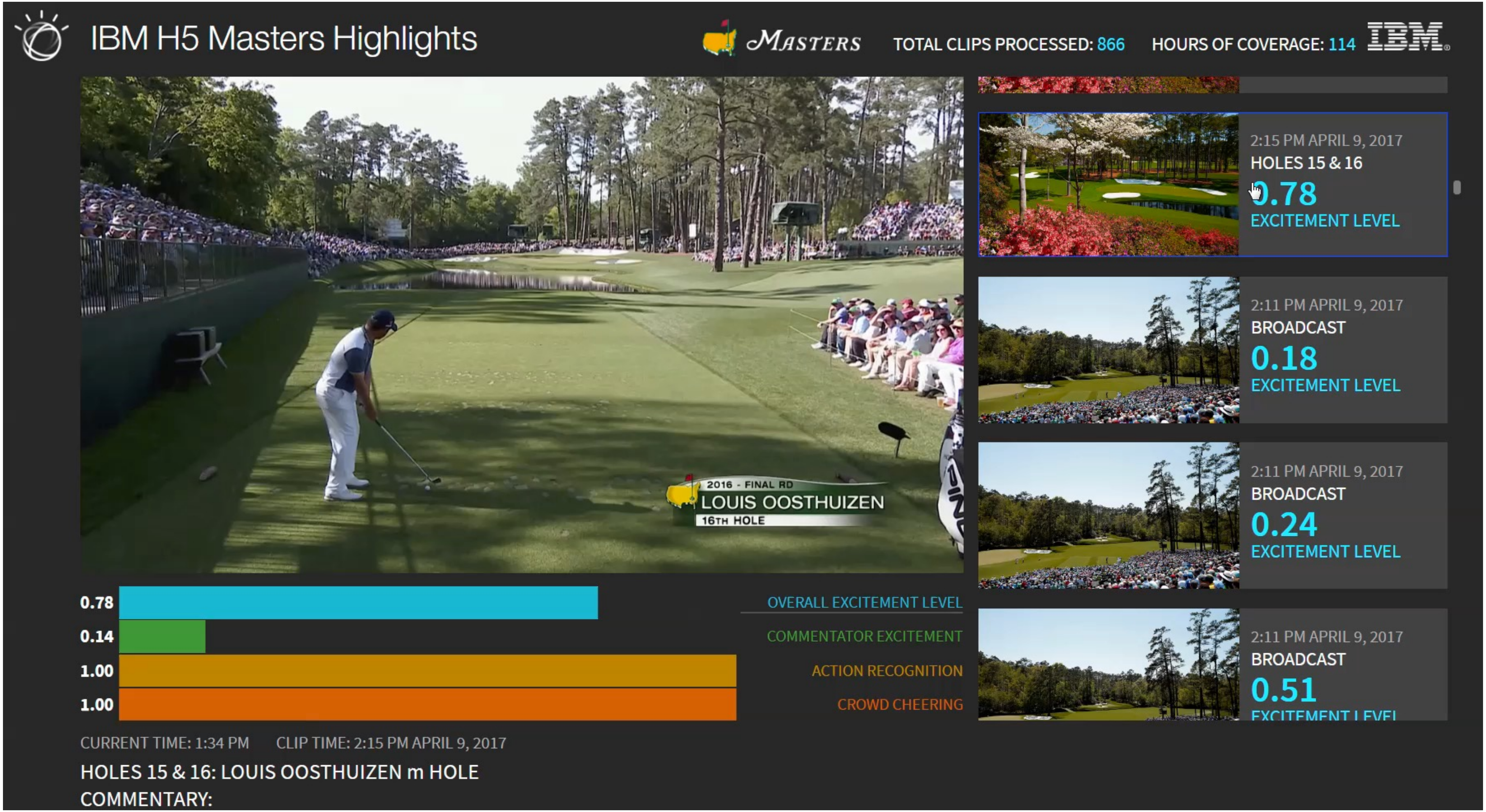
Action: Doing Fencing

[Panda and Chen, Arxiv 2021]

Qualitative Results

[Panda and Chen, Arxiv 2021]

Other Related Projects



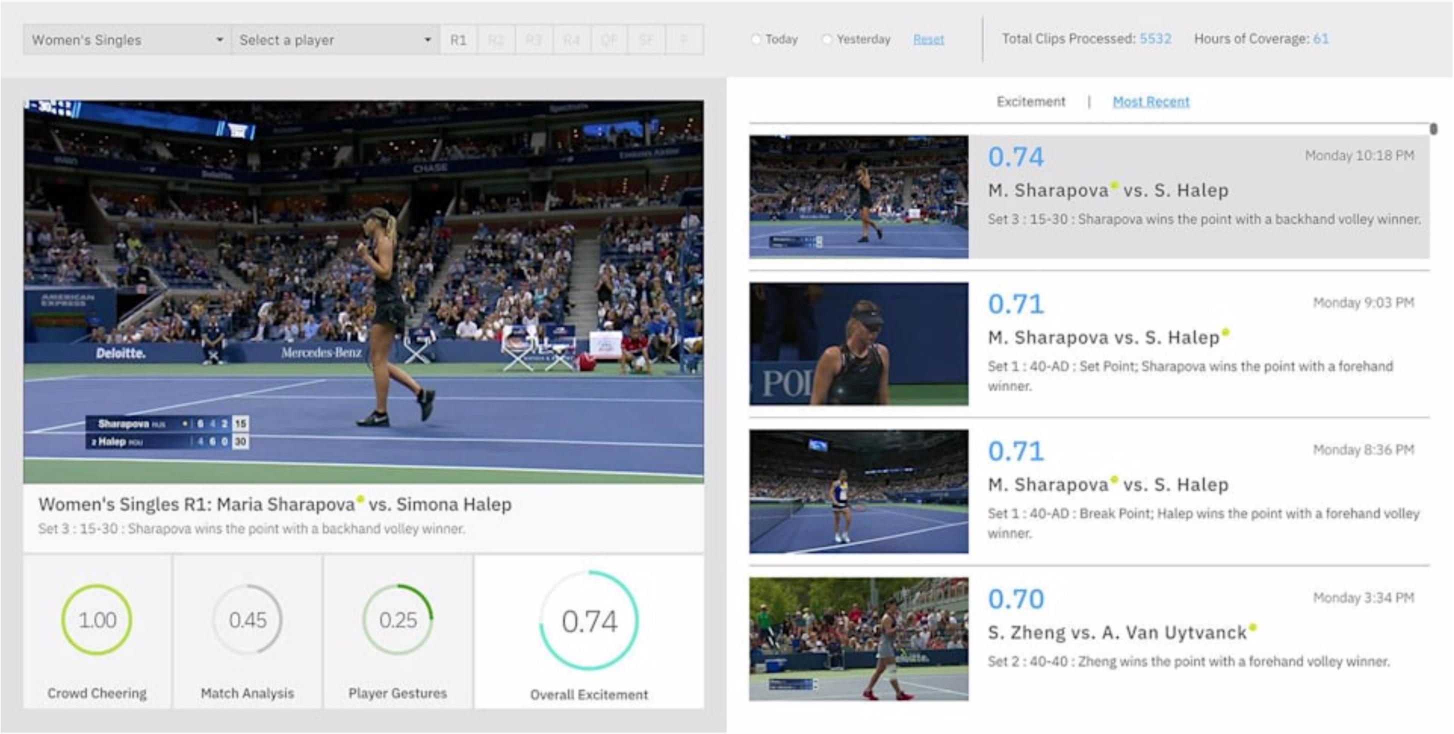
IBM H5 Masters Highlights
MASTERS
TOTAL CLIPS PROCESSED: 866
HOURS OF COVERAGE: 114
IBM
2:15 PM APRIL 9, 2017
HOLES 15 & 16
0.78
EXCITEMENT LEVEL
2:11 PM APRIL 9, 2017
BROADCAST
0.18
EXCITEMENT LEVEL
2:11 PM APRIL 9, 2017
BROADCAST
0.24
EXCITEMENT LEVEL
2:11 PM APRIL 9, 2017
BROADCAST
0.51
2016 - FINAL RD
LOUIS OOSTHUIZEN
16TH HOLE
0.78
OVERALL EXCITEMENT LEVEL
0.14
COMMENTATOR EXCITEMENT
1.00
ACTION RECOGNITION
1.00
CROWD CHEERING
CURRENT TIME: 1:34 PM
CLIP TIME: 2:15 PM APRIL 9, 2017
HOLES 15 & 16: LOUIS OOSTHUIZEN m HOLE
COMMENTARY:

IBM Highlights @ USOpen and Wimbledon

Watched by millions of fans worldwide

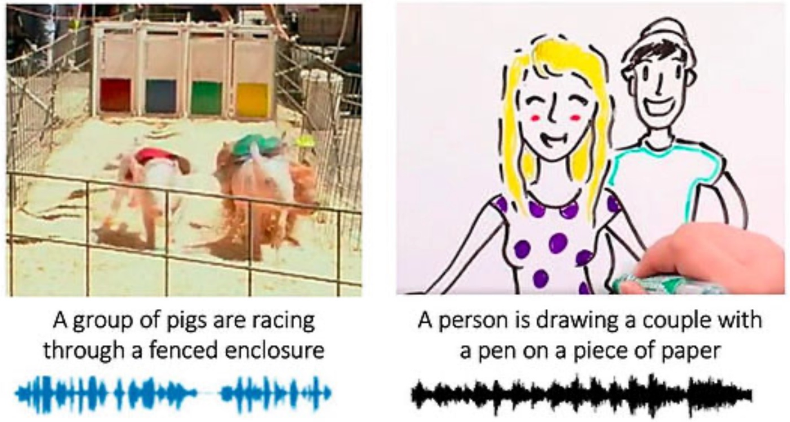
The New York Times

Enjoy Those U.S. Open Highlights. A Computer Picked Them for You.

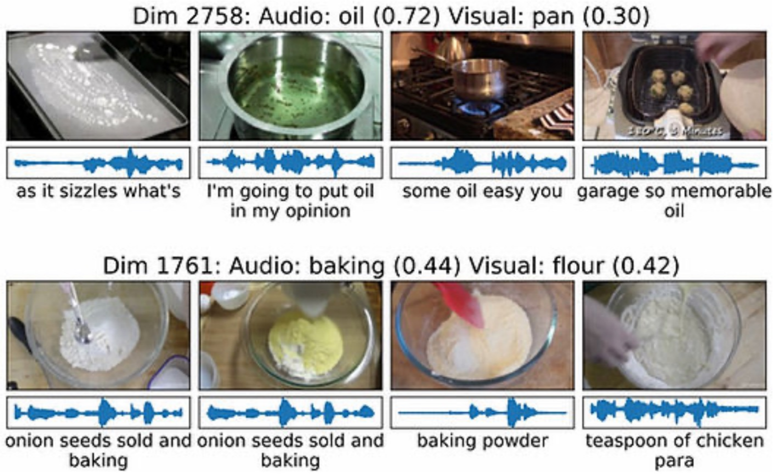


Grounding Spoken Words in Video (without supervision)

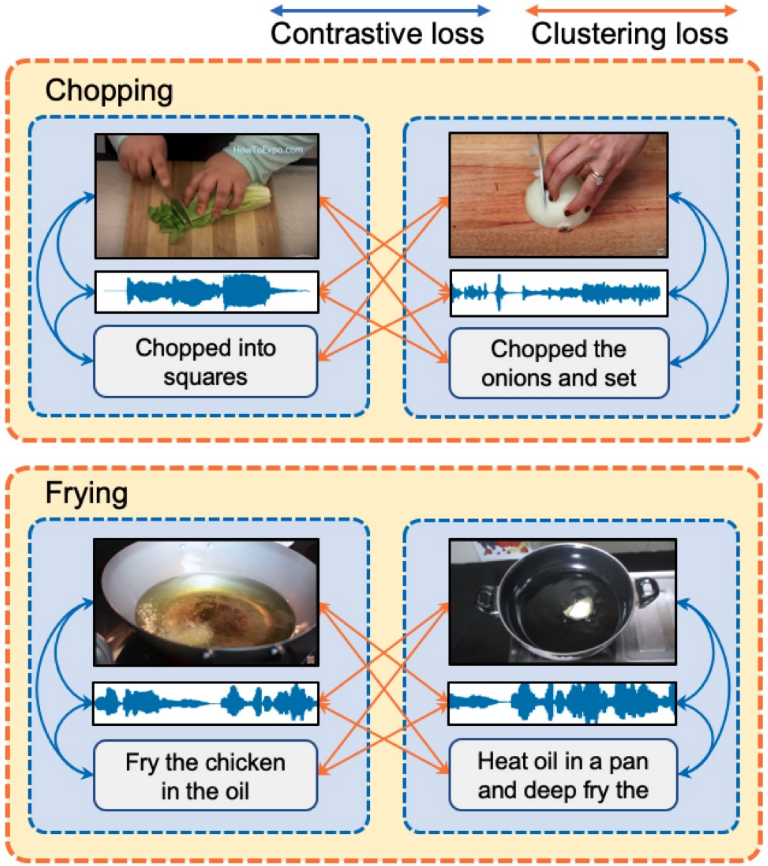
Spoken Moments, CVPR 2021



AVLNet, Interspeech 2021

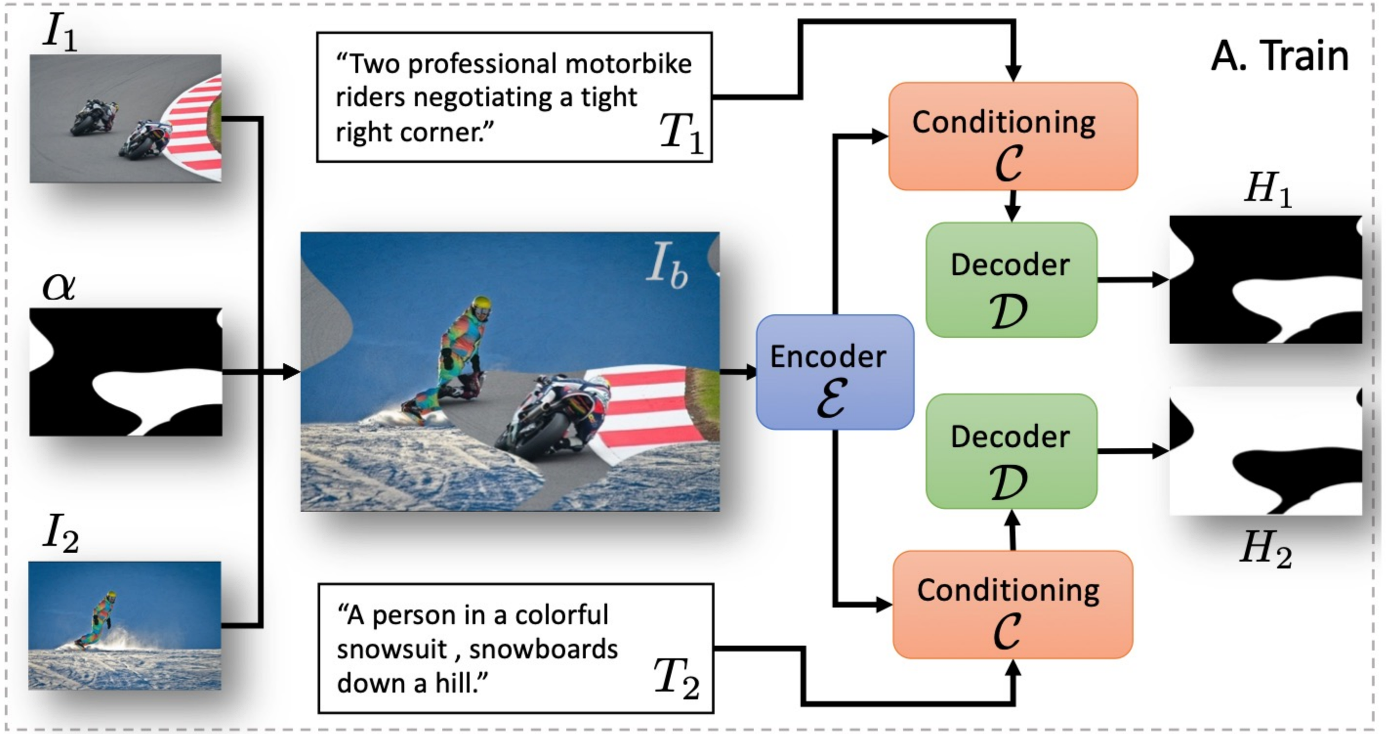


Multimodal Clustering Networks, Arxiv 2021

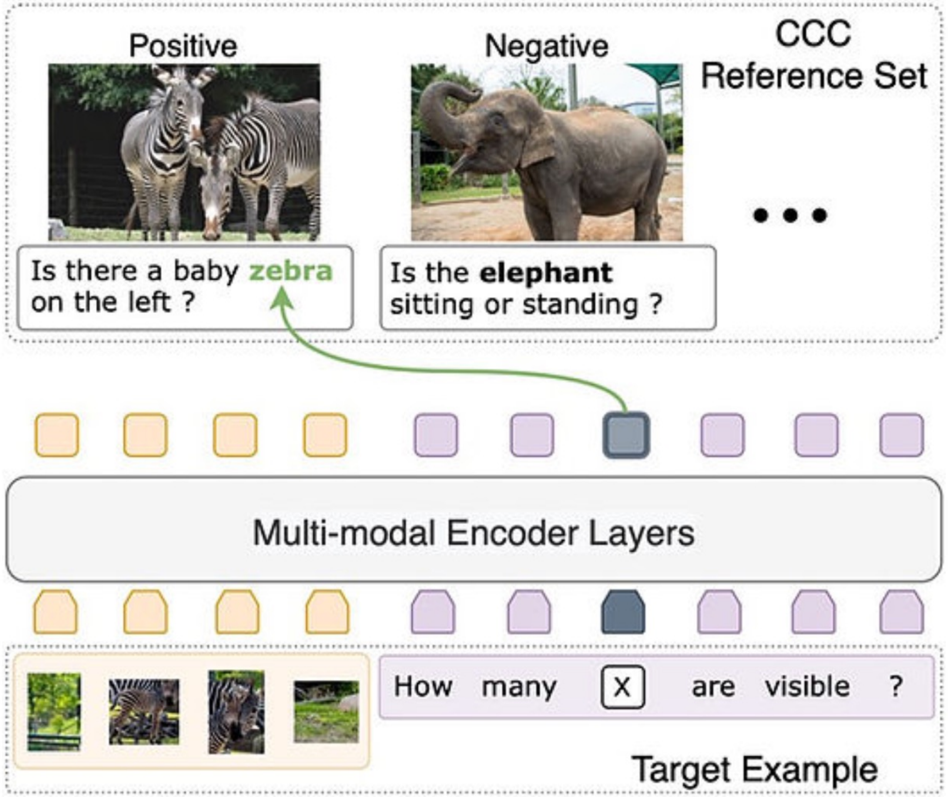


Grounding Text in Images (without supervision)

Grounding by separation, Arxiv 2021



Separating skills and concepts, CVPR 2021



Summary

Adaptive (dynamic) neural networks for efficient inference

- Blockdrop: dynamic selection of layers to execute for efficient image classification
- AR-Net: dynamic selection of frame resolution for efficient video recognition
- AdaMML: dynamic selection of modalities for efficient multimodal video understanding

References

- ZuxuanWu*, Tushar Nagarajan*, Abhishek Kumar, Steve Rennie, Larry Davis, Kristen Grauman, and Rogerio Feris. **BlockDrop: Dynamic Inference Paths in Residual Networks**. CVPR 2018

- Yue Meng, Chung-Ching Lin, Rameswar Panda, Prasanna Sattigeri, Leonid Karlinsky, Aude Oliva, Kate Saenko, and Rogerio Feris. **AR-Net: Adaptive Frame Resolution for Efficient Action Recognition**. ECCV 2020

- Rameswar Panda*, Richard Chen*, Quanfu Fan, Ximeng Sun, Kate Saenko, Aude Oliva, and Rogerio Feris. AdaMML: **Adaptive Multi-Modal Learning for Efficient Video Recognition**. Arxiv 2021

See more at http://rogerioferis.org

(* equal contribution)