Review and Examination

ECO: Efficient Convolutional Network for Online Video Understanding

by Zolfaghari et al

Video Understanding

- The ability to associate a logical description with a series of consecutive or non-consecutive frames that represents a unique action in the real world.
- Deep Learning as a provider of powerful classifiers.
- Large Datasets adding diversity and realism

Problem

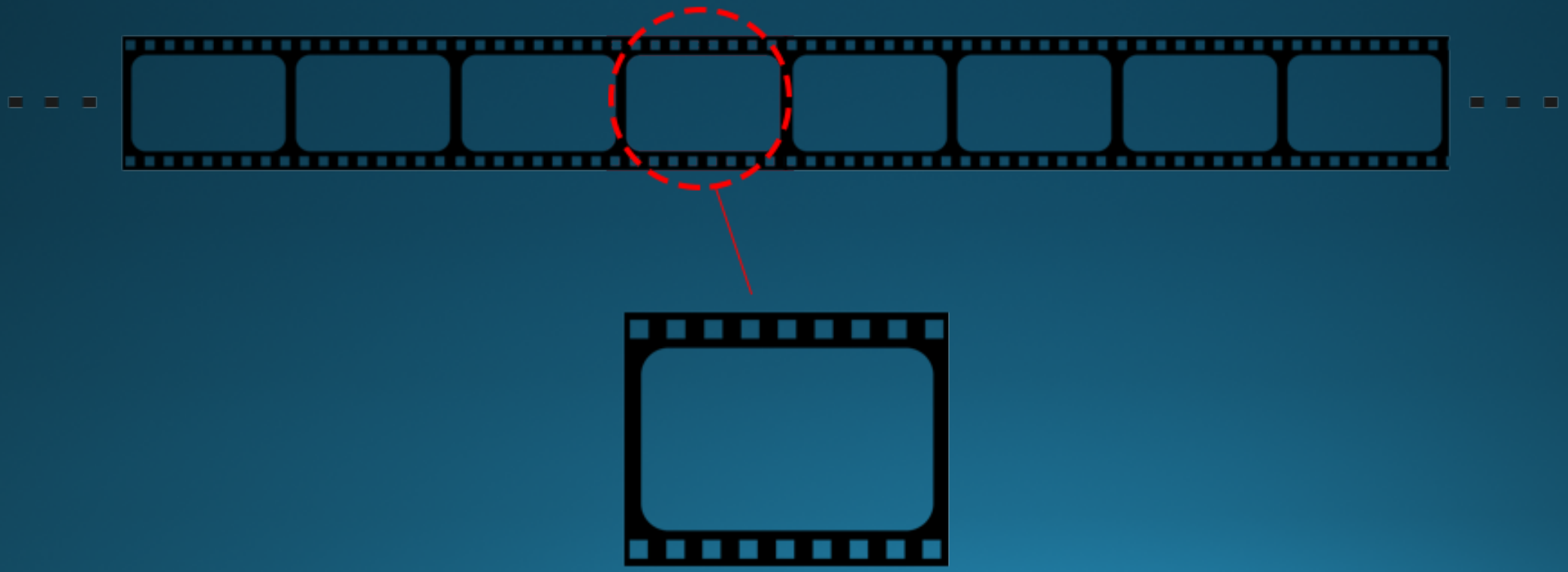
- Fast action-detection methods suffer from missing activities that span several seconds.
- Methods for understanding the temporal relationship between frames cover small spans due to their computational expense (post-hoc fusion of scores for small window weights).
- Frame Processing vs. Video Processing

Solution

- An aggregation mechanism.
- Use a 2D convolutional architecture to build a good initial belief about the action from a single frame within a temporal pool.
- Use a 3D convolutional architecture to bridge the contents of frames from different temporal pools to reinforce the initial belief obtained from the single frame.

Sampling Strategy

- Frames from within a temporal pool contain largely redundant information



Architecture…1: ECO Lite

- Divide the video into N equal segments.
- Select one frame randomly from each segment.
- Process each selected frame with a 2D convolutional network.
- Process feature representations from all the frames with a 3D convolutional network.

Architecture…2: ECO Lite

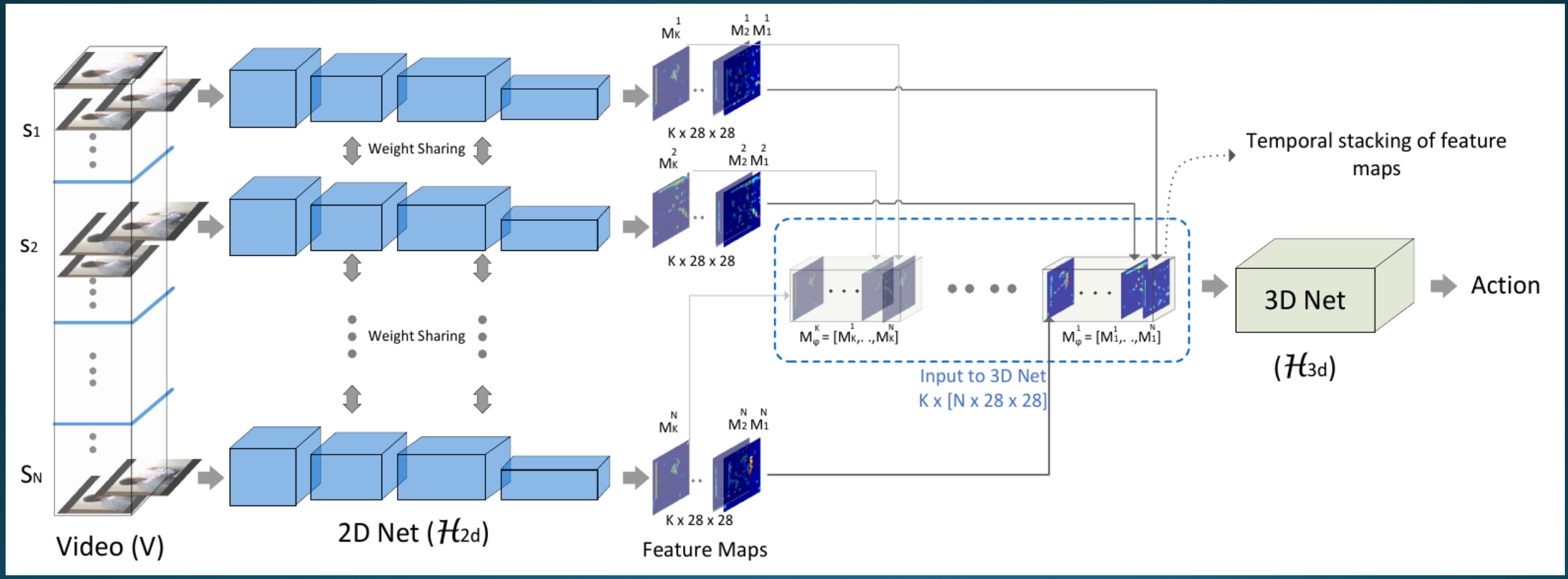


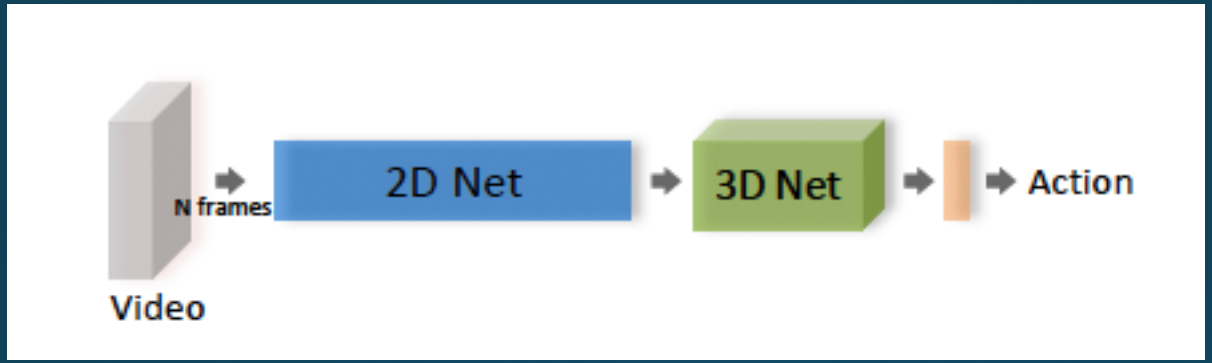
Image Source: Zolfaghari, M., Singh, K., Brox, T.: ECO: Efficient Convolutional Network for Online Video Understanding.In: European Conference on Computer Vision (ECCV) (2018), https://github.com/mzolfaghari/ECO-efficient-video-understanding/tree/master/doc_files

Architecture…3: ECO (full)

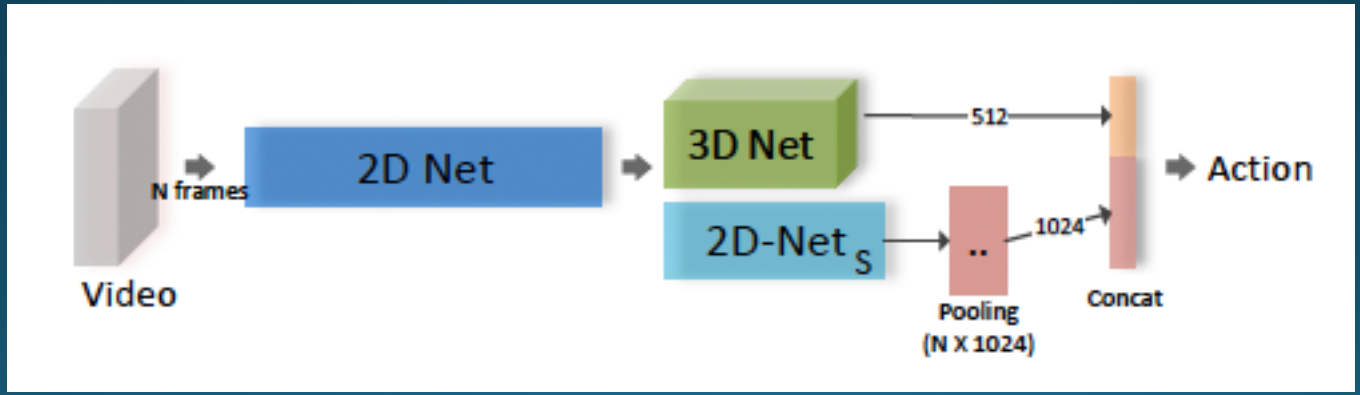
- The loss of capacity in the current architecture.
- Adding a 2D network parallel to the 3D.
- The new layer produces a feature vector
- Features from 2D Net and 3D Net are concatenated.
- Importance given to short term actions (in single frame)

Architecture…4

ECO Lite



ECO

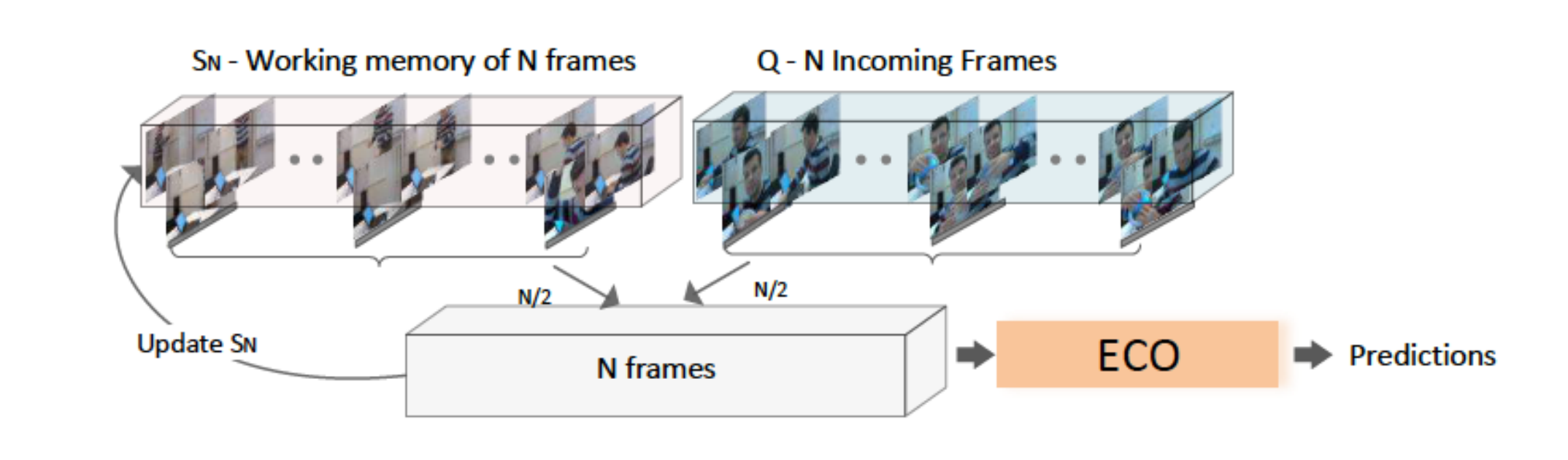


Images Source: Zolfaghari, M., Singh, K., Brox, T.: ECO: Efficient Convolutional Network for Online Video Understanding.In: European Conference on Computer Vision (ECCV) (2018), https://github.com/mzolfaghari/ECO-efficient-video-understanding/tree/master/doc_files

Streaming Video Understanding...1

- Change to sampling mechanism with no change to ECO architecture.
- Uses working memory to store N frames and update it with incoming frames.
- Efficient as it keeps N frames at any one time.

Streaming Video Understanding…2



Images Source: Zolfaghari, M., Singh, K., Brox, T.: ECO: Efficient Convolutional Network for Online Video Understanding.In: European Conference on Computer Vision (ECCV) (2018), https://github.com/mzolfaghari/ECO-efficient-video-understanding/tree/master/doc_files

Benchmark: ECO vs state-of-the-art

Method	Inference speed (VPS)	UCF101 (%)	HMDB51 (%)
Res3D	<2	85.8	54.9
TSN	21	87.7	51
EMV	15.6	86.4	-
I3D	0.9	95.6	74.8
ARTNet	2.9	93.5	67.6
$ECO_{Lite-4F}$	**237.3**	87.4	58.1
ECO_{4F}	**163.4**	90.3	61.7
ECO_{12F}	**52.6**	92.4	68.3
ECO_{20F}	**32.9**	93.0	69.0
ECO_{24F}	**28.2**	93.6	68.4

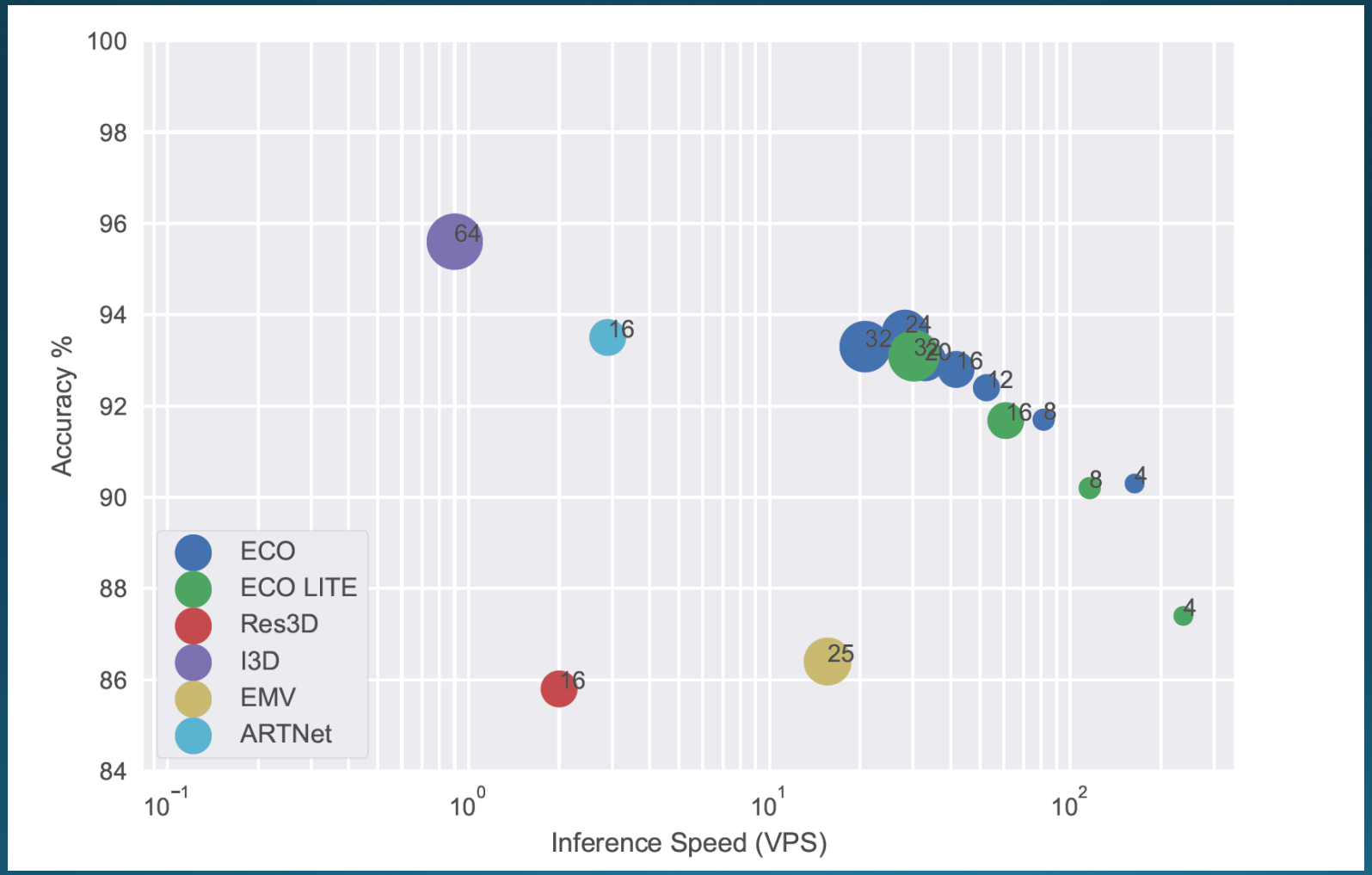
Modified table Source: Zolfaghari, M., Singh, K., Brox, T.: ECO: Efficient Convolutional Network for Online Video Understanding.In: European Conference on Computer Vision (ECCV) (2018), https://github.com/mzolfaghari/ECO-efficient-video-understanding/tree/master/doc_files

Accuracy: ECO vs. ECO Lite

Model	Sampled Frames	Speed (VPS)		Accuracy (%)			
		Titan X	Tesla P100	UCF101	HMDB51	Kinetics	Someth.
ECO	4	99.2	163.4	90.3	61.7	66.2	–
	8	49.5	81.5	91.7	65.6	67.8	39.6
	16	24.5	41.7	92.8	68.5	69.0	41.4
	32	12.3	20.8	93.3	68.7	67.8	–
ECO Lite	4	142.9	237.3	87.4	58.1	57.9	–
	8	71.1	115.9	90.2	63.3	–	38.7
	16	35.3	61.0	91.6	68.2	64.4	42.2
	32	18.2	30.2	93.1	68.3	–	41.3

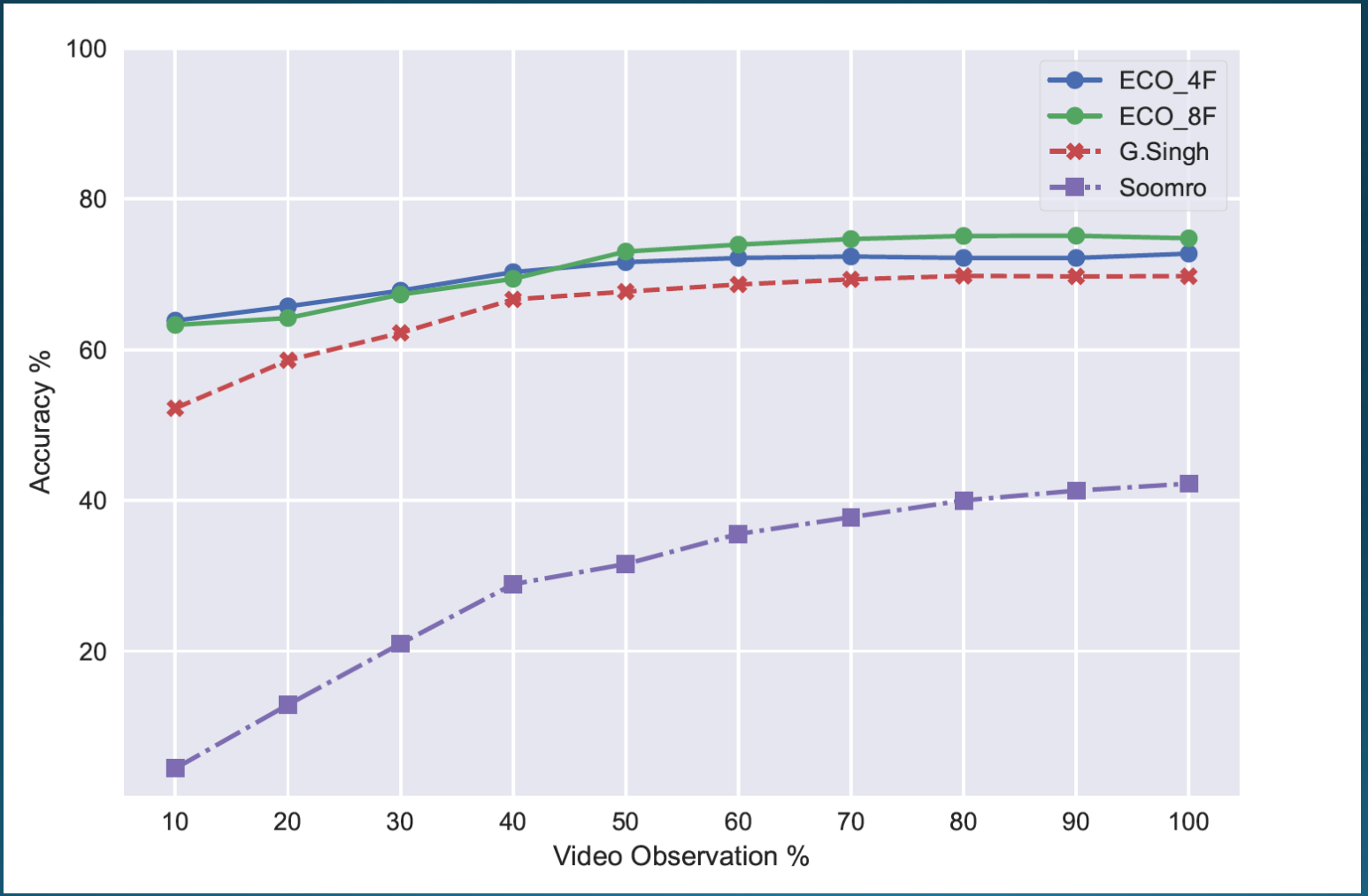
Table Source: Zolfaghari, M., Singh, K., Brox, T.: ECO: Efficient Convolutional Network for Online Video Understanding.In: European Conference on Computer Vision (ECCV) (2018), https://github.com/mzolfaghari/ECO-efficient-video-understanding/tree/master/doc_files

Accuracy vs. Runtime: on UCF101



Images Source: Zolfaghari, M., Singh, K., Brox, T.: ECO: Efficient Convolutional Network for Online Video Understanding.In: European Conference on Computer Vision (ECCV) (2018), https://github.com/mzolfaghari/ECO-efficient-video-understanding/tree/master/doc_files

Action Recognition: Online Video



Images Source: Zolfaghari, M., Singh, K., Brox, T.: ECO: Efficient Convolutional Network for Online Video Understanding.In: European Conference on Computer Vision (ECCV) (2018), https://github.com/mzolfaghari/ECO-efficient-video-understanding/tree/master/doc_files

Example: using somethingsomething dataset

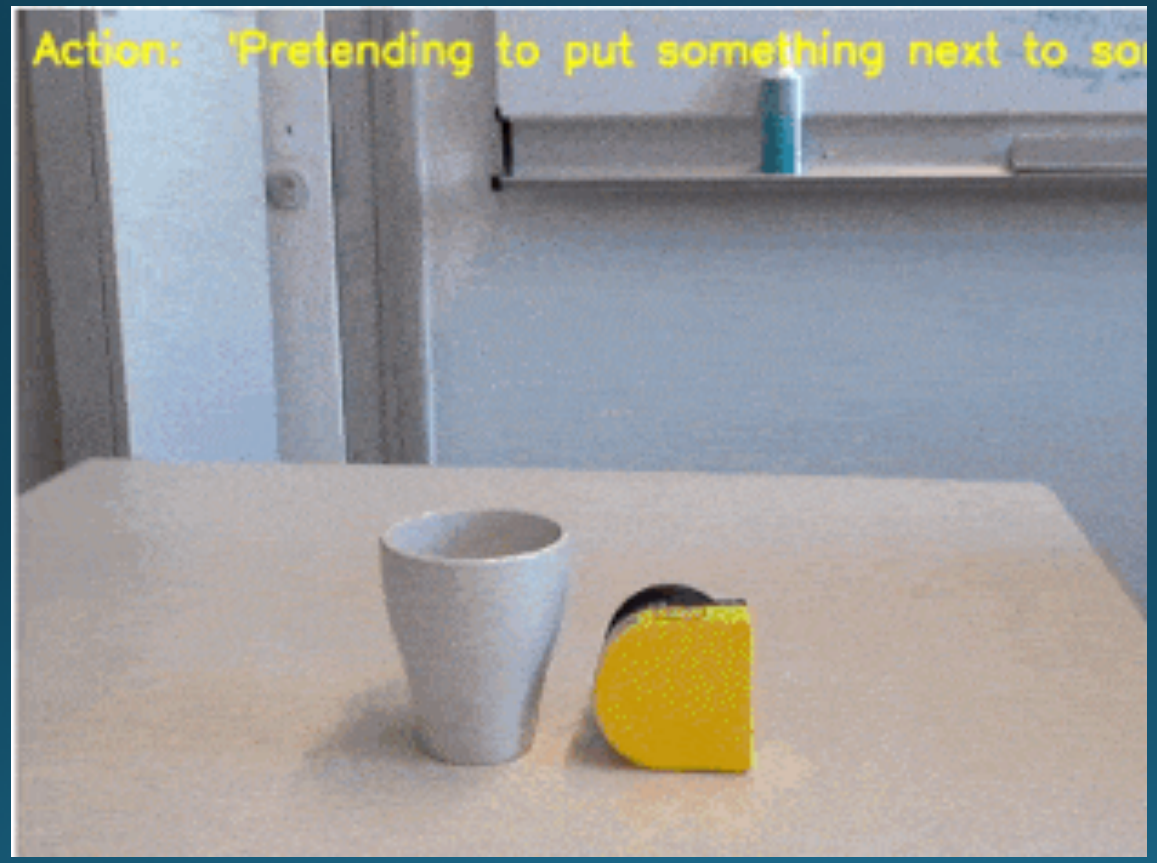


Image Source: Zolfaghari, M., Singh, K., Brox, T.: ECO: Efficient Convolutional Network for Online Video Understanding.In: European Conference on Computer Vision (ECCV) (2018), https://github.com/mzolfaghari/ECO-efficient-video-understanding/tree/master/doc_files

Example: action classification of an incoming stream

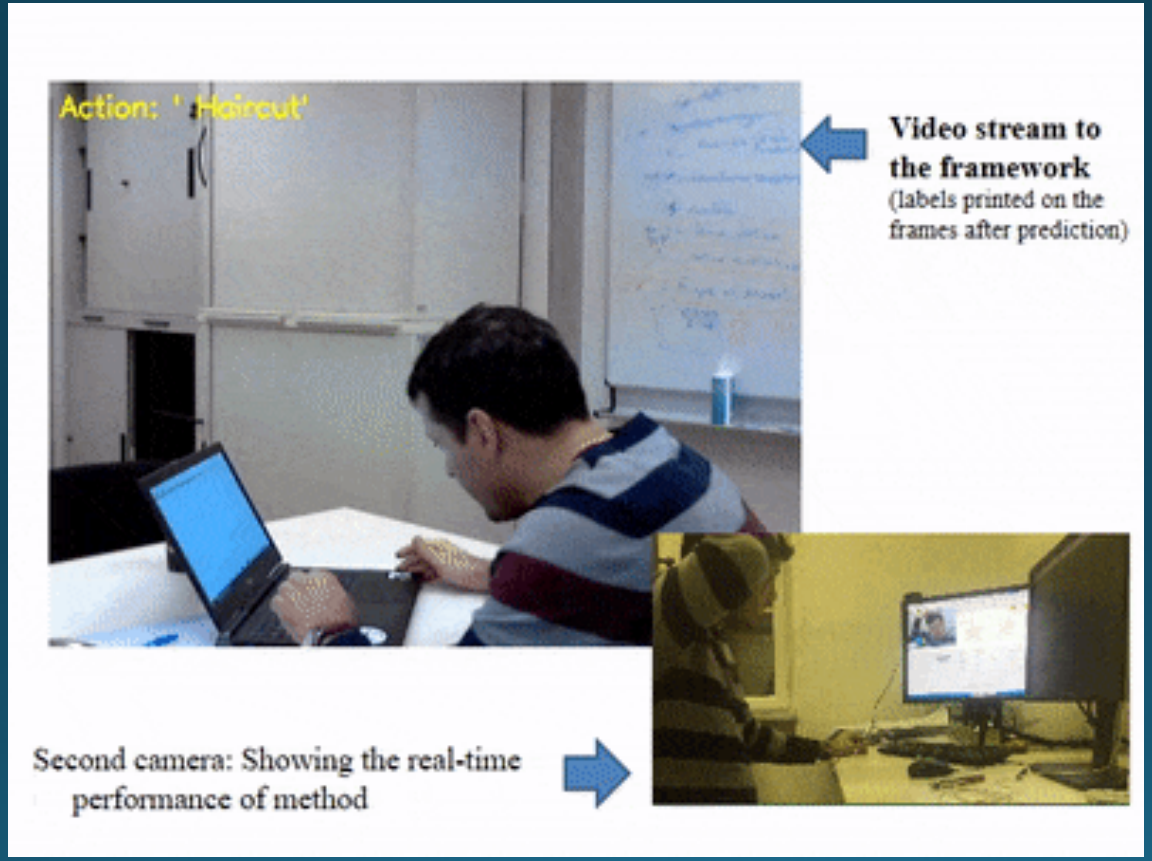


Image Source: Zolfaghari, M., Singh, K., Brox, T.: ECO: Efficient Convolutional Network for Online Video Understanding.In: European Conference on Computer Vision (ECCV) (2018), https://github.com/mzolfaghari/ECO-efficient-video-understanding/tree/master/doc_files

Future

- Contextual Searchability
- Preforming video understanding on low computing devices (i.e smartphones, home cameras, smart home hubs)

Investigation

- Capacity to run on mobile devices.
- Capacity to process live video.

References

Zolfaghari, M., Singh, K., Brox, T.: ECO: Efficient Convolutional Network for Online Video Understanding.In: European Conference on Computer Vision (ECCV) (2018)

Thank you

Questions?