AdaMML: Adaptive Multi-Modal Learning for Efficient Video Recognition

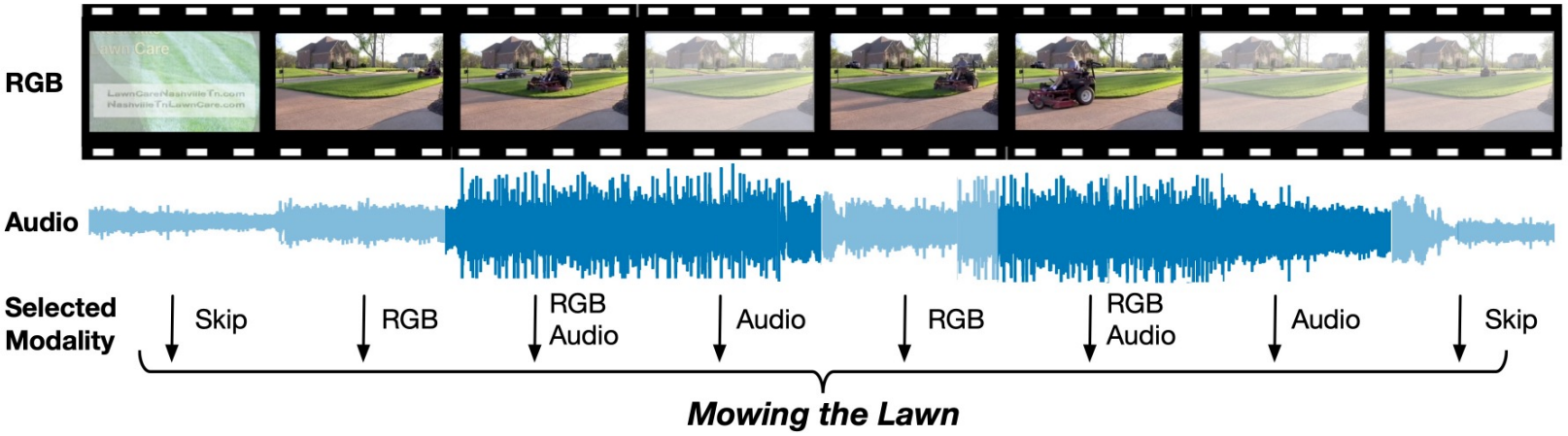
Rameswar Panda[1*], Chun-Fu (Richard) Chen[1*], Quanfu Fan[1],
Ximeng Sun[2], Kate Saenko[1,2], Aude Oliva[1,3], Rogerio Feris[1]

[1] MIT-IBM Watson AI Lab, [2] Boston University, [3] MIT
*: Equal Contribution

Project page: https://rpand002.github.io/adamml.html

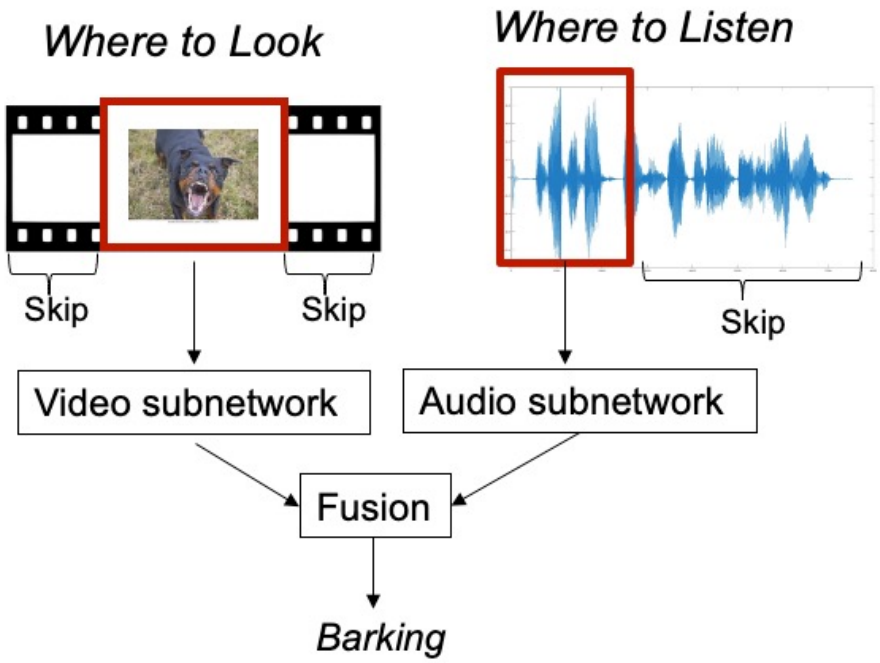
Motivation

Most of the multi-modal video recognition methods are computationally expensive, as they usually process all the data (including redundant/irrelevant parts)



Do all the segments require both RGB and audio stream to recognize the action as “Mowing the Lawn” in this video?

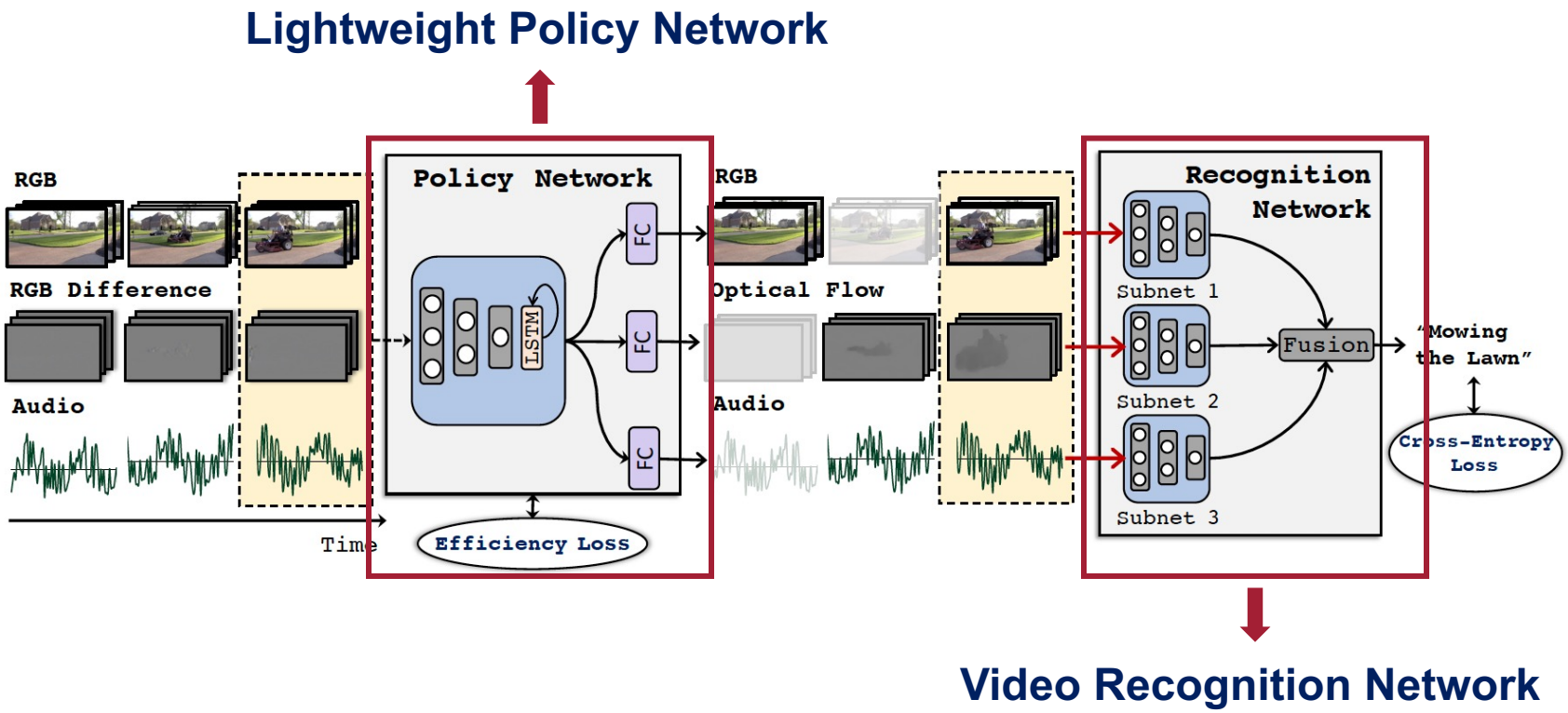
Key Idea



An **adaptive multi-modal learning framework**, that selects on-the-fly the optimal modalities for each segment conditioned on the input for efficient video recognition

To the best of our knowledge, this is the first work on data-dependent selection of different modalities for efficient video recognition.

AdaMML



Our approach consists of a policy network and a recognition network composed of different sub-networks that are trained jointly (via late fusion with learnable weights) for recognizing videos.

Results

Dataset	Kinetics-Sounds				ActivityNet			
Method	Acc. (%)	Selection Rate (%) RGB	Audio	GFLOPs	mAP (%)	Selection Rate (%) RGB	Audio	GFLOPs
RGB	82.85	100	–	141.36	73.24	100	–	141.36
Audio	65.49	–	100	3.82	13.88	–	100	3.82
Weighted Fusion	87.86	100	100	145.17	72.88	100	100	145.17
AdaMML	**88.17**	46.47	94.15	**76.45** (-47.3%)	**73.91**	76.25	56.35	**94.01** (-35.2%)

RGB + Audio

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%)

RGB + Flow + Audio

Comparison with Weighted Fusion Baseline

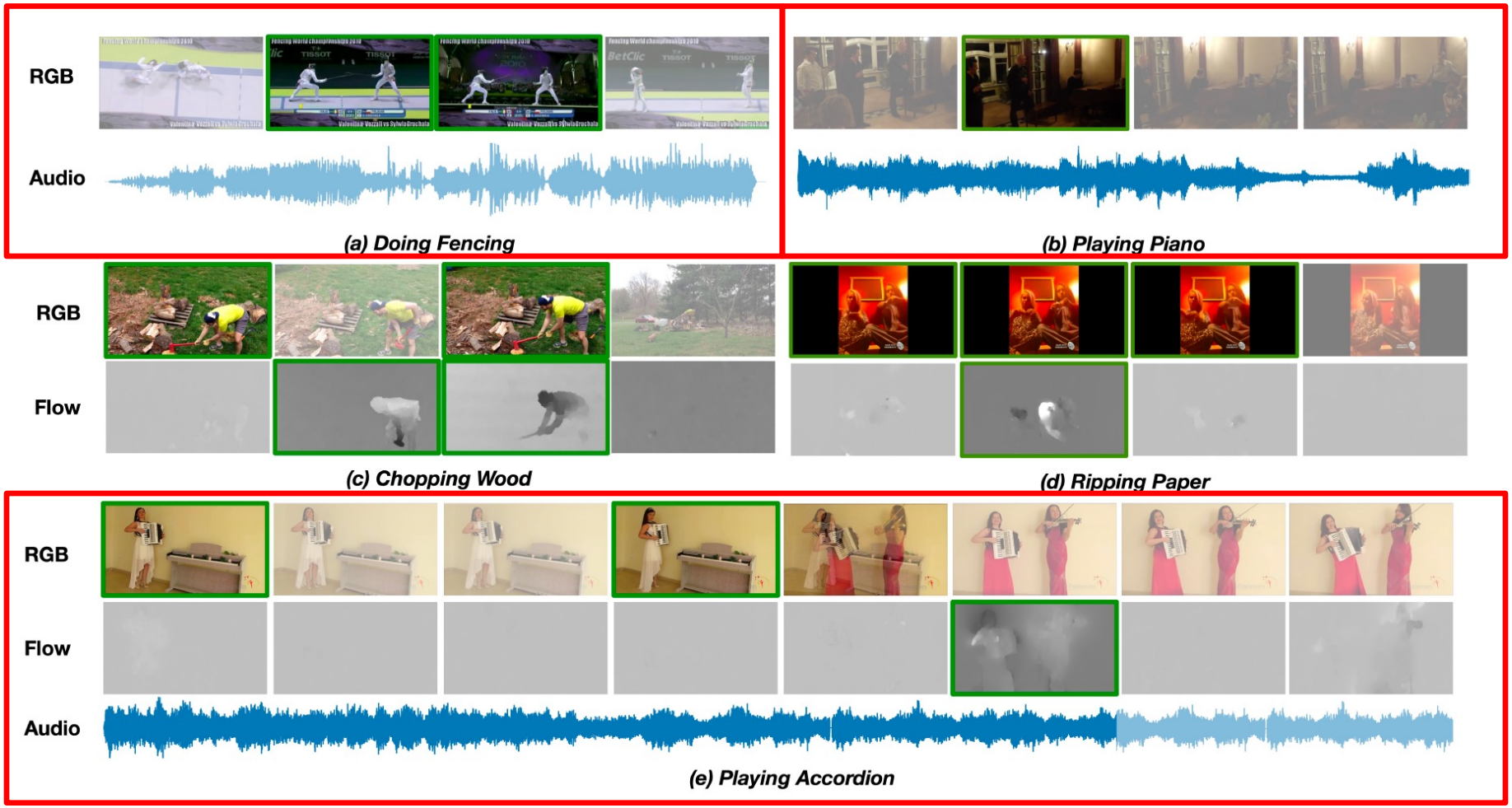
Results

Method	ActivityNet		FCVID	
	mAP (%)	GFLOPs	mAP (%)	GFLOPs
FrameGlimpse	60.14	33.33	67.55	30.10
FastForward	54.64	17.86	71.21	66.11
AdaFrame	71.5	78.69	80.2	75.13
LiteEval	72.7	95.1	80.0	94.3
AdaMML	**73.91**	94.01	**85.82**	93.86

Comparison with State-of-the-art Methods

New SOTA for efficient video recognition, improving prior best result in terms of accuracy, and computational efficiency

Results



Please refer to our paper for more detailed results and analysis

Thank you and welcome to our poster!

Project Page: https://rpand002.github.io/adamml.html