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]

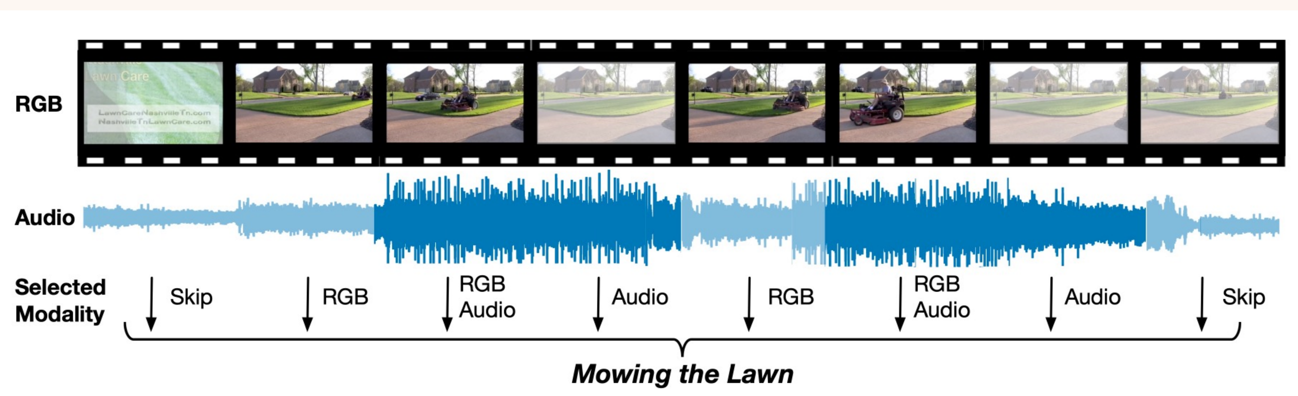
[1] MIT-IBM Watson AI Lab, [2] Boston University, [3] MIT (*: Equal Contribution)

MIT-IBM Watson AI Lab | BOSTON UNIVERSITY | MIT | 2021 ICCV OCTOBER 11-17 VIRTUAL

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).
- ❑ Utilizing information from all the input modalities may be counterproductive as informative modalities are often overwhelmed by uninformative ones in long videos.
- ❑ Some modalities require more computation than others and hence selecting the cheaper modality with good performance can significantly save computation leading to more efficient recognition .

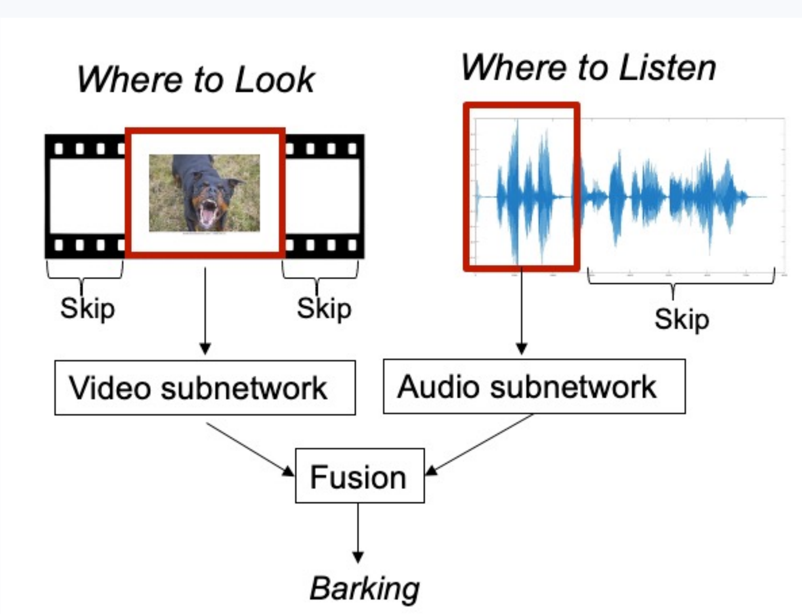


Do all the segments require both RGB and audio stream to recognize the action as "Mowing the Lawn" in this video?

No, we need both RGB and audio streams for only third and sixth video segments to improve the model confidence for recognizing the correct action, while the rest of the segments can be processed with only one modality or even skipped (e.g., the first and last video segment) without losing any accuracy.

Key Idea

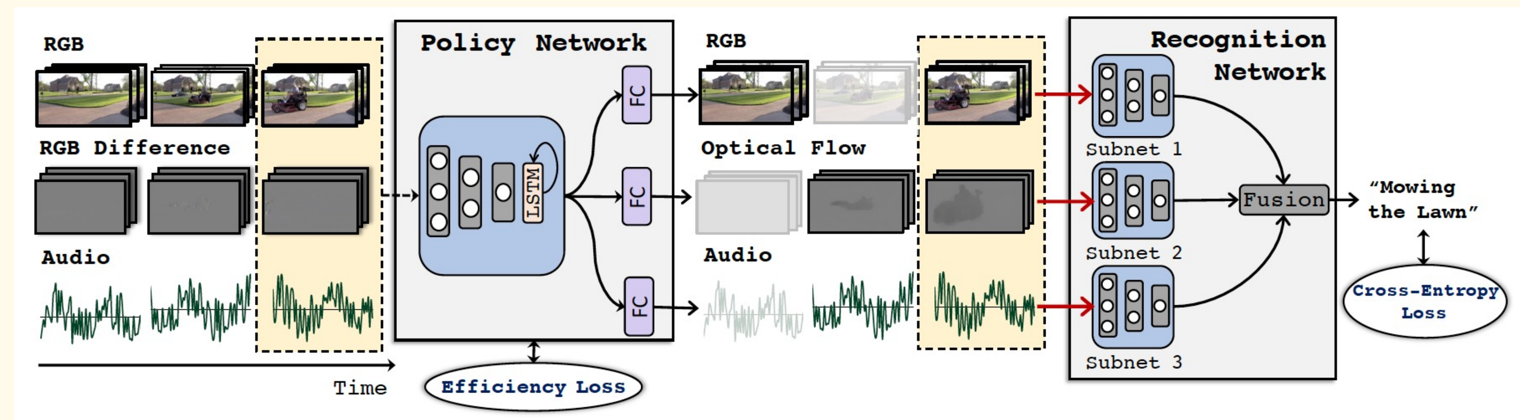
- ❑ **AdaMML:** A novel and differentiable approach to learn a decision policy that selects optimal modalities conditioned on the inputs for efficient video recognition.
- ❑ This is in sharp contrast to current multi-modal learning approaches that utilizes all the input modalities without considering their relevance to the video recognition.
- ❑ Learn a lightweight model (referred to as the multi-modal policy network) that outputs the posterior probabilities of all the binary decisions for using or skipping each modality on a per segment basis.



A new perspective for efficient video recognition by adaptively selecting input modalities, on a per segment basis, for recognizing complex actions.

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

Framework



- ❑ **AdaMML** 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.
- ❑ **Policy network** decides what modalities to use per segment to achieve the best accuracy and efficiency in video recognition.
- ❑ **In training**, policies are sampled from a Gumbel-Softmax distribution, which allows us to optimize the policy network via standard backpropagation.
- ❑ **During inference**, an input segment is first fed into policy network and then selected modalities are routed to recognition network to generate segment-level predictions. **Finally**, the network averages all segment-level predictions to obtain video-level prediction.

Multi-Modal Policy Network

The policy network contains a lightweight joint feature extractor and an LSTM module for modeling the causality across different time steps in a video.

$$h_t, o_t = \mathrm{LSTM}(f_t, h_{t-1}, o_{t-1})$$

Given the hidden state, the policy network estimates a policy distribution for each modality and then samples binary decisions indicating whether to select a modality at time step via Gumbel-Softmax operation.

Training using Gumbel-Softmax Sampling

An an effective way to replace the non-differentiable sample from a discrete distribution with a differentiable sample from a corresponding Gumbel-Softmax distribution.

$$\hat{P}_k = \underset{i\in\{0,1\}}{\operatorname{argmax}}(\log z_{i,k} + G_{i,k}), \quad k \in [1, ..., K]$$

$$P_{i,k} = \frac{\exp((\log z_{i,k} + G_{i,k})/\tau)}{\sum_{j\in\{0,1\}} \exp((\log z_{j,k} + G_{j,k})/\tau)}$$

During forward pass, we sample the policy and during the backward pass, we approximate the gradient of the discrete samples by computing the gradient of the continuous softmax relaxation.

Loss Function

$$\mathbb{E}_{(V,y)\sim\mathcal{D}_{train}}\left[-y\log(\mathcal{P}(V;\Theta)) + \sum_{k=1}^{K}\lambda_k\mathcal{C}_k\right] \quad \mathcal{C}_k = \begin{cases} (\frac{|U_k|_0}{C})^2 & \text{if correct} \\ \gamma & \text{otherwise} \end{cases}$$

First part: standard cross-entropy loss to measure the classification quality;
Second part: drives the network to learn a policy that favors selection of modality that is computationally more efficient in recognizing videos.

Datasets and Settings

- ❑ Datasets
  - ❑ Kinetics-Sounds: Training: 22,521 videos – Testing: 1532 videos – 31 classes
  - ❑ ActivityNet-v1.3: Training: 10,024 videos – Testing: 4926 videos – 200 classes
  - ❑ FCVID: Training: 45,611 videos – Testing: 45,612 – 239 classes
  - ❑ Mini-Sports1M: Training: 14,610 videos – Testing: 4870 videos – 487 classes
- ❑ Tasks: (I) RGB + Audio, (II) RGB + Flow, and (III) RGB + Flow + Audio
- ❑ Model Architectures
  - ❑ Policy Network: MobileNetV2; Recognition Network: TSN-like ResNet-50
- ❑ Evaluation Metrics: video-level mAP or top-1 accuracy, GFLOPS, Selection Rate

Results

Dataset	Kinetics-Sounds				ActivityNet			
		Selection Rate (%)				Selection Rate (%)		
Method	Acc. (%)	RGB	Audio	GFLOPs	mAP (%)	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%)

Video recognition results with RGB + Audio modalities on Kinetics-Sounds and ActivityNet

		Selection Rate (%)		
Method	Acc. (%)	RGB	Flow	GFLOPs
RGB	82.85	100	–	141.36
Flow	75.73	–	100	163.39
Weighted Fusion	83.47	100	100	304.75
AdaMML-Flow	83.82	56.04	36.39	151.54 (-50.3%)
AdaMML-RGBDiff	**84.36**	44.61	37.40	**137.03** (-55.0%)

RGB + Flow on Kinetics-Sounds

		Selection Rate (%)			
Method	Acc. (%)	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 on Kinetics-Sounds

	ActivityNet		FCVID	
Method	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

	Kinetics-Sounds		Mini-Sports1M	
Method	Acc. (%)	GFLOPs	mAP (%)	GFLOPs
LiteEval	72.02	104.06	43.64	151.83
AdaMML	**88.17**	76.45	**46.08**	138.32

Comparison with State-of-the-art Methods

	Network			
Method	RGB	Audio	mAP (%)	GFLOPs
ListenToLook	ResNet-18	ResNet-18	76.61	112.65
AdaMML $_{112\times112}$	ResNet-18	ResNet-18	79.48	70.87
AdaMML $_{224\times224}$	ResNet-18	ResNet-18	80.05	82.33
AdaMML $_{224\times224}$	ResNet-50	MobileNetV2	84.73	110.14
AdaMML $_{224\times224}$	EfficientNet-b3	EfficientNet-b0	85.62	30.55

ActivityNet

	RGB + Audio		RGB + Flow		RGB + Flow + Audio	
Method	Acc. (%)	GFLOPs	Acc. (%)	GFLOPs	Acc. (%)	GFLOPs
Average Fusion	88.15	145.17	83.30	304.75	88.18	308.56
Class-wise Weighted Fusion	87.86	145.17	83.82	304.75	87.75	308.56
Max Fusion	86.49	145.17	83.47	304.75	88.06	308.56
FC2 Fusion*	87.73	145.17	83.30	304.75	87.84	308.56
Weighted Fusion	87.86	145.17	83.47	304.75	88.25	308.56
AdaMML	**88.17**	**76.45**	**84.36**	**137.03**	**89.06**	**141.97**

*: concatenating feature vectors from all modalities and add two addition fully-connected layers to fuse features.

Comparison with fusion strategies on Kinetics-Sounds

- ❑ 35%-55% reduction in FLOPS while improving accuracy over weighted fusion baseline
- ❑ Significantly outperforms state-of-the-art methods on ActivityNet and FCVID
- ❑ Outperforms LiteEval on all datasets (~16% in Kinetics-Sounds)
- ❑ With same setting, AdaMML outperforms ListenToLook on ActivityNet (~3% mAP)
- ❑ New SOTA result on ActivityNet with EfficientNet
- ❑ Consistently outperforms all hand-designed fusion strategies (~50% FLOPS savings)
- ❑ Significantly better than random policy variants

Visualizations

Qualitative examples showing the effectiveness of AdaMML in selecting the right modalities per video segment (marked by green borders). Our adaptive approach focuses on right modalities to use per segment for correctly classifying videos while taking efficiency into account.