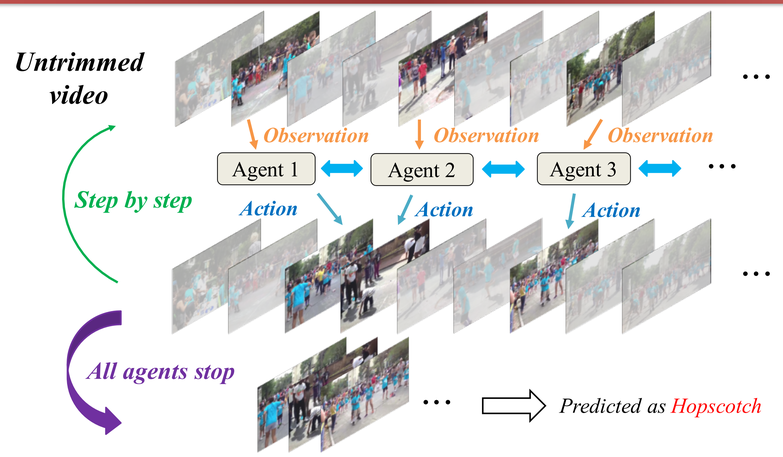
Multi-Agent Reinforcement Learning Based Frame Sampling for Effective Untrimmed Video Recognition

Wenhao Wu[1,2], Dongliang He[3], Xiao Tan[3], Shifeng Chen[1], Shilei Wen[3]
[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences, [3]Department of Computer Vision Technology (VIS), Baidu Inc.

ICCV 2019 Seoul, Korea

MOTIVATION



We focus on developing learning based sampler instead of exploring network architectures for effective untrimmed video recognition.

CONTRIBUTION

- We focus on a previously overlooked point, i.e., the frame sampling strategy, in improving untrimmed video classification performance and intuitively formulate it as a Markov decision process.
- Multi-agent reinforcement learning is adopted to solve the formulated sequential decision problems. A novel framework that takes both context information and historical environment states into consideration for decision making is designed.
- The proposed method can be effectively applied to various existing untrimmed video recognition models to improve the performance, which is well witnessed by the excellent experimental results.

METHOD

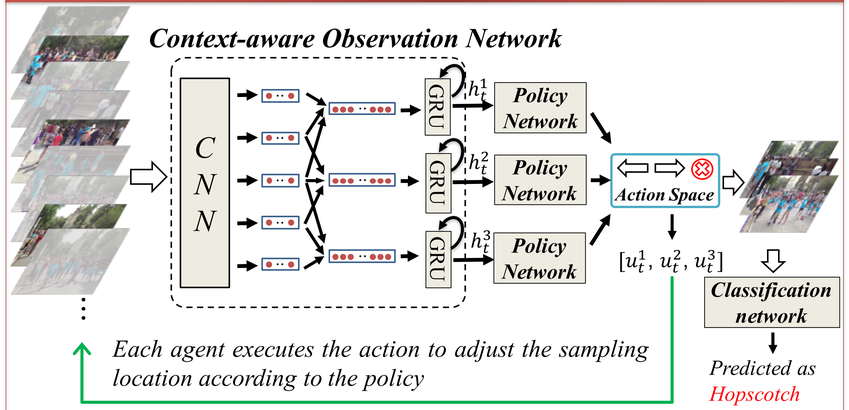


Figure 2. The proposed MARL framework is composed of a context-aware observation network for capturing environment state h_t^n , a policy network for estimating probabilistic distribution over action space and a classification network for video-level prediction. There are N (which is 3 for illustration) agents in our system. The agent interacts with the environment by taking action to adjust sampling locations iteratively. GRU, which integrates preceding and current states, is used to model the sequential decision making property.

EXPERIMENTS

- *Different number of agents*

	N=5	N=15	N=25	N=120	All
ResNet-152	80.19	82.99	83.81	83.72	82.53

Table 5. Impact of N on ActivityNet v1.3 val set in terms of mAP.

- *Different context range*

	U25	All	M=0	M=1	M=2	M=4
mAP	82.08	82.53	82.99	83.81	83.80	83.72

Table 6. Evaluation of context range on ActivityNet v1.3 val set using ResNet-152. U25 and All are hand-crafted strategies.

- *Policy network transferring*

U25	All	MARL	Birds' Sampler	Cars' Sampler	ANet1.2' Sampler	R101' Sampler
82.08	82.53	83.81	82.70	82.66	83.41	83.43

Table 7. The performance of different settings on ActivityNet v1.3 val set.

EXPERIMENTS

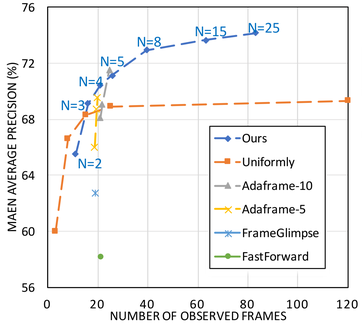


Figure 3. Mean average precision vs. number of observed frames.

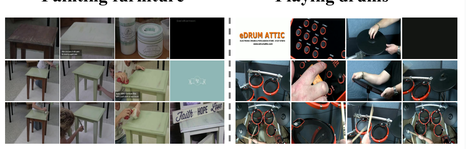


Figure 4. Visualization of the selected frame with different strategies on ActivityNet. The first row show frames from uniformly sampling, the second row depicts frames from our method without context-aware observation while the last row contains frames from our method.

Architecture	ActivityNet v1.2				ActivityNet v1.3			
	R25	U25	All	Ours	R25	U25	All	Ours
C3D	62.06	62.89	63.00	**64.13**	59.73	60.68	60.83	**62.00**
BN-Inception	78.76	80.02	80.50	**81.99**	75.08	76.48	77.33	**78.32**
ResNet-101	80.73	81.94	82.26	**83.76**	78.69	79.96	80.64	**81.54**
Inception-V3	81.90	82.66	83.25	**85.01**	79.27	80.33	80.86	**82.34**
ResNet-152	82.72	83.71	84.07	**85.70**	80.69	82.08	82.53	**83.81**

Table 1. Table 1. Performance comparison of different ConvNet architectures on the ActivityNet dataset. For different architectures, randomly sampling 25 frames (R25), uniformly sampling 25 frames (U25), using all frames (All) and using our method to sample 25 frames (Ours) are evaluated. All architectures are based on ImageNet pre-trained model, except C3D.

Method	Backbone	Pre-trained	top-1	mAP
IDT [43]	-	ImageNet	64.70	68.69
C3D [33]	-	Sports1M	65.80	67.68
TSN [45]*	BN-Inception	ImageNet	72.97	76.56
P3D [33]	ResNet-152	ImageNet	75.12	78.86
RRA [55]	ResNet-152	ImageNet	78.81	83.42
Ours	ResNet-152	ImageNet	**79.82**	**83.81**
TSN [45]*	BN-Inception	Kinetics	78.98	81.80
Ours	BN-Inception	Kinetics	80.22	83.52
C16	Ensemble	-	-	**90.9**
Ours	SEResNeXt152	Kinetics	**85.72**	90.05

Table 3. Comparing with methods on the ActivityNet v1.3 validation dataset **using RGB modality**. * indicates the results of our implementation. C16 denotes the champion submission of the ActivityNet 2016 challenge, it fuses multiple powerful models and multi-modal (RGB, optical flow and audio) results.

Strategies	Instance	Video
R25/U25/All	80.69/82.08/82.53	80.17/81.23/81.73
Ours	**83.81**	**82.98**

Table 2. Performance comparison between different supervision on ActivityNet v1.3 val set.

Method	YouTube Birds	YouTube Cars
BN-Inception*	60.13	61.96
I3D [3]*	40.68	40.92
TSN [45]*	72.361	74.340
RRA [55]*	73.205	77.625
U25/All/Ours	76.56/76.77/**79.01**	76.49/76.99/**79.77**

Table 4. Comparing with methods on YouTube Birds and YouTube Cars. * indicates the results of the method come from the latest project page of these datasets.