facebook Artificial Intelligence Research

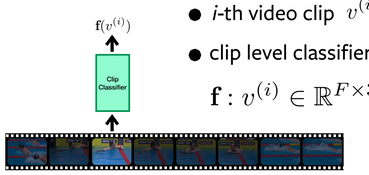
SCSAMPLER: SAMPLING SALIENT CLIPS FROM VIDEO FOR EFFICIENT ACTION RECOGNITION

Bruno Korbar
Du Tran
Lorenzo Torresani

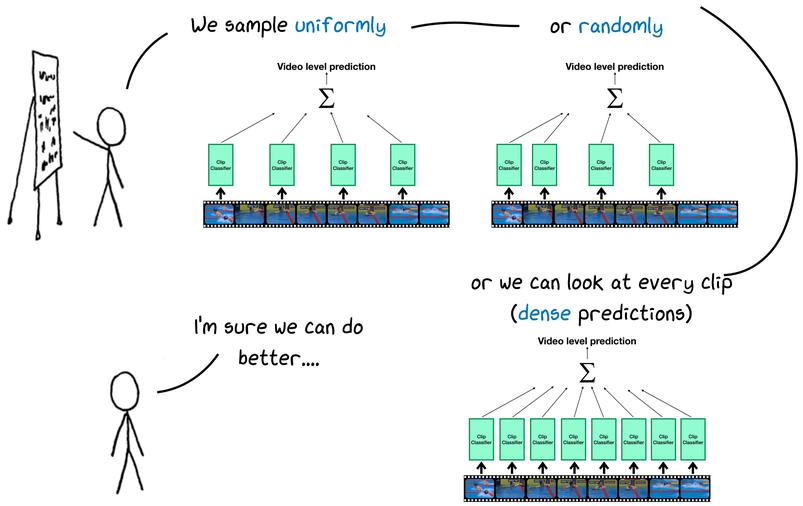
Background

- video $v \in \mathbb{R}^{L \times 3 \times H \times W}$
- *i*-th video clip $v^{(i)} \in \mathbb{R}^{F \times 3 \times H \times W}$
- clip level classifier :

$$\mathbf{f} : v^{(i)} \in \mathbb{R}^{F \times 3 \times H \times W} \rightarrow [0,1]^C$$

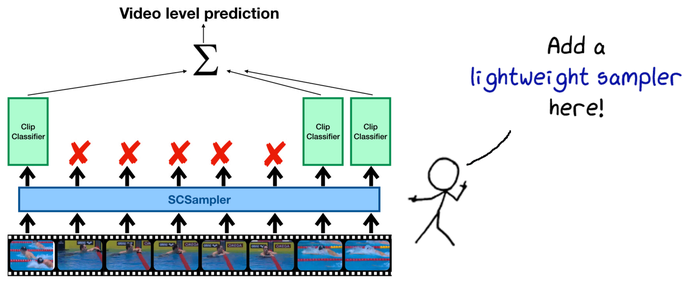


So how do we get predictions for the entire video?



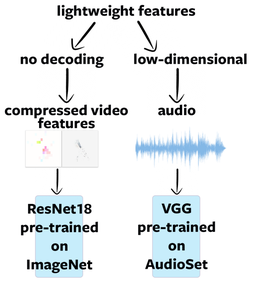
Not all clips in a video are equal!

Introducing a novel sampling method capable of identifying the most salient clips for video-level action recognition.



Approach overview:

- visual-only sampler using compressed visual features [1]: motion displacements (MD), RGB-residuals (RGB-R), and I-frames (I)
- audio-only sampler using lightweight VGG network on top of MEL-Spectrograms



Technical approach

Objective function

Action Classifier (AC)

1. train a lightweight action classifier $\hat{S}^{AC}(v^{(i)}) \in [0,1]^C$
2. define clip saliency by max pooling the result of the classifier $S^{AC}(v^{(i)}) = \max_c \hat{S}^{AC}_c(v^{(i)})$

Saliency Classifier (SAL-CL) and Saliency Ranker (SAL-RANK)

- assign saliency score to each clip of a video $\hat{S}^{SAL}(v^{(i)}) \in [0,1]$

SAL-CL:
- learn to distinguish ✓ from ✗

SAL-RANK:
- learn to rank ✓ higher than ✗

Visual-only sampler

- requires no decoding
- omitting I-Frames makes it lightweight

Audio-only Sampler

- very low dimensional
- lightweight VGG-like architecture

Experimental results

Sports1M, *K*=10

Classifier	SCSampler $\mathcal{S}$ (K clips) accuracy (%)	SCSampler $\mathcal{S}$ (K clips) runtime (day)	Random / Uniform / Empirical (K clips) accuracy (%)	Random / Uniform / Empirical (K clips) runtime (day)	Dense (*all* clips) accuracy (%)	Dense (*all* clips) runtime (days)	Oracle $\mathcal{O}$ (K clips) accuracy (%)
MC3-18	72.8	0.8	64.5 / 64.8 / 65.3	0.4	66.6	12.9	85.1
R(2+1)D-18	73.9	0.8	63.0 / 63.2 / 63.9	0.4	68.7	13.1	87.0
R3D-18	70.2	0.8	59.8 / 59.9 / 60.3	0.4	65.6	13.3	85.0
R(2+1)D-34	78.0	0.9	71.2 / 71.5 / 72.0	0.6	70.9	14.2	88.4
ir-CSN-152	84.0	0.9	75.3 / 75.8 / 76.2	0.5	77.0	14.0	92.6

SCSampler improves a state-of-the-art model [2] by 8.5%

and it speeds up inference 15 times.

Kinetics, *K*=10

Classifier	SCSampler $\mathcal{S}$ (K clips) accuracy (%)	Random / Uniform / Empirical (K clips) accuracy (%)
R(2+1D)-34*	76.7	73.8 / 74.0 / 74.1
I3D-RGB**	75.1	71.9 / 71.8 / 71.9
ir-CSN-152*	80.2	77.8 / 78.5 / 79.2

*pre-trained on Sports1M
**pre-trained on Imagenet

Assessing design choices on miniSports dataset

Objective function

	AC	SAL-CL	SAL-RANK
Visual SCSampler	73.05	63.17	64.77
Audio SCSampler	66.37	58.73	67.82

train with AC loss

train with SAL-RANK loss

Feature selection for visual-only sampler

SCSampler features	SCSampler architecture	accuracy (%)	runtime (min)
MD	ResNet-18	63.5	19.8
RGB-R	ResNet-18	68.0	20.4
MD + RGB-R	ResNet-18	73.1	20.9
IF+MD+RGB-R	ResNet-18	74.9	27.3
MD + RGB-R	ShuffleNet-26	67.9	19.1
IF+MD+RGB-R	ShuffleNet-26	69.9	23.8

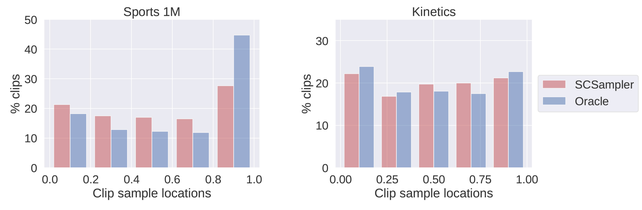
I-Frames are too expensive!

Combining visual and audio samplers

SCSampler Audio-Video Combination	accuracy (%)	runtime (min)
AV-union-list ($K' = 8$)	75.98	23.4
AV-joint-training	75.53	23.4
Visual SCSampler only	73.05	20.9
Audio SCSampler only	67.82	22.0

So the main takeaways are:
- visual signal dominant
- audio and video complementary
- joint training possible

Temporal localization of the most salient clips



References

[1] Wu, Chao-Yuan et al. (2018). "Compressed video action recognition." In *CVPR 2018*, pp. 6026-6035.
[2] Tran, Du, et al. (2019) "Video Classification with Channel-Separated Convolutional Networks", In *ICCV 2019*.
[3] Carreira, Joao, and Andrew Zisserman (2017). "Quo vadis, action recognition? a new model and the kinetics dataset." In *CVPR 2017*, pp. 6299-6308.
[4] Karpathy, Andrej et al. (2014). "Large-scale video classification with convolutional neural networks." In *CVPR 2014*, pp. 1725-1732.
[*all graphics*] xkcd.com