Classification Matters: Improving Video Action Detection with Class-Specific Attention

Jinsung Lee, Taeoh Kim, Inwoong Lee, Minho Shim, Dongyoon Wee, Minsu Cho, Suha Kwak

POSTECH
NAVER Cloud
ECCV

Video Action Detection (VAD)

Motivation

- Original mAP
- When GT box is provided
- When GT class is provided

TubeR (CVPR '22)   EVAD (ICCV '23)   STMixer (CVPR '23)

TubeR
EVAD
Ours

talk to (e.g., self, a person, a group) | listen to (a person) | carry/hold (an object) | answer phone | listen (e.g., to music)

talk to | listen to | smoke | carry/hold | watch | touch | sit
put down | read | bend/bow | lift/pick up | sing to | press | give/serve | RESULT
talk to | listen to | read | carry/hold | watch | stand | RESULT

RESULT (orange):
"sit"
"carry/hold (an object)"
"smoke"
"listen to (a person)"

RESULT (green):
"stand"
"carry/hold (an object)"
"read"
"talk to (a person)"

In video action detection, you better

PAY MORE ATTENTION TO EVERY SINGLE ACTION CLASS!

Model

- ⊕ addition
- ⊘ Sine positional embedding
- ▭ Variables
- ⬭ Modules

Long story short, **our model creates an attention map** **for each action class** using *class queries*!

Performance

On standard VAD benchmarks (AVA, JHMDB, UCF):

Model	Input	Backbone	Pre-train	mAP
AIA	32×2	SF-R101	K700	32.3
VideoMAE	16×4	ViT-B	K400	31.8
WOO	32×2	SF-R101-NL	K600	28.3
TubeR	32×2	CSN-152	IG65M, K400	31.1
STMixer	32×2	CSN-152	IG65M, K400	32.8
STMixer	16×4	ViT-B	K400	32.6
STMixer	16×4	ViT-B	K710, K400	36.1
EVAD	16×4	ViT-B	K400	32.3
EVAD	16×4	ViT-B	K710, K400	37.7
Ours	32×2	CSN-152	IG65M, K400	**33.5**
Ours	16×4	ViT-B	K400	**32.9**
Ours	16×4	ViT-B	K710, K400	**38.4**

Model	Backbone	f-mAP / v-mAP JHMDB	f-mAP / v-mAP UCF
AVA	I3D-VGG	73.3 / 78.6	76.3 / 59.9
ACRN	SF-R101	77.9 / 80.1	- / -
CA-RCNN	R50-NL	79.2 / -	- / -
YOWO	3DResNext-101	74.4 / 85.7	80.4 / 48.8
WOO	SF-R101-NL	80.5 / -	- / -
AIA	R50-C2D	- / -	78.8 / -
ACAR	SF-R50	- / -	84.3 / -
TubeR	CSN-152	- / 82.3	83.2 / 58.4
STMixer	SF-R101-NL	86.7 / -	83.7 / -
EVAD	ViT-B	**90.2** / 77.8	85.1 / 58.8
Ours	ViT-B	86.9 / **88.5**	**85.9 / 61.7**

Our model achieves much greater efficiency:

- TubeR (CVPR '22)
- STMixer (CVPR '23)
- EVAD (ICCV '23)
- Ours

Params (M) | TFLOPs | Inference time (ms) | # Actor candidates | f-mAP | v-mAP

Capturing contexts dedicated to each action class helps achieving great performance with <u>fewer actor candidates</u>, leading to a significant improvement in efficiency.

Qualitative results

stand | take a photo | carry/hold (an object)
talk to (e.g., self, a person, a group) | sail boat | listen to (a person)
watch (e.g., TV) | watch (e.g., TV) | watch (e.g., TV)