On Temporal Granularity in Self-Supervised Video Representation Learning

Rui Qian, Yeqing Li, Liangzhe Yuan, Boqing Gong,
Ting Liu, Matthew Brown, Serge Belongie, Ming-Hsuan Yang, Hartwig Adam, Yin Cui

Research Motivation

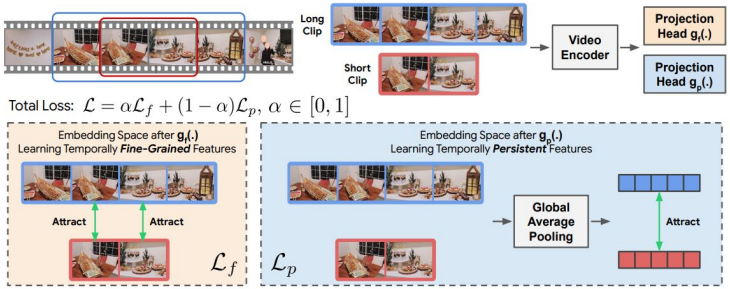
- **Common video representation learning methods mainly focus on temporally persistent learning**:
  - The features learned are persistent across all temporal locations of the whole video.
- **This objective is not always preferable:**
  - The strong performances are only on common video benchmarks of action recognition.
  - The feature learned ignores the changing nature in videos, may be not suitable for all video tasks.

Contribution

- **We propose the concept of temporal granularity.**

- **Given the same short video:**
  - Event boundary detection task calls for temporally fine-grained features to be aware of the temporal content shifts.
  - Video-level recognition task requires the model to robustly predict the target label based on some sampled clips from the video, preferring coarse-grained features.

Proposed Method



- **Temporal sampling:**
  - Long-short sampling of two clips.
- **Two embedding spaces:**
  - Fine-grained space using dense contrastive learning for temporal discriminative features.
  - Coarse-grained space using global contrastive learning for coarse features.

Evaluation and benchmarks

- **Two representation setting:**
  - TeG-PS: a = 0.0, only persistent
  - TeG-FG: a = 0.9, mostly fine-grained
- **Benchmarks:**
  - Prefer fine-grained feature:
    - Event classification
    - Event boundary detection
  - Prefer coarse-grained feature:
    - Scene-heavy action recognition
  - Interesting to explore:
    - Temporal-heavy action recognition
    - Spatio-temporal action localization on AVA-Kinetics

Results

method		backbone	acc.
Supervised	Train from scratch [62]	I3D	31.2
		I3D + NL	30.2
		R3D-50 + NL	33.1
		SlowFast + NL	32.6
Unsupervised	CVRL [58]	R3D-50	28.3
	TeG-PS	R3D-50	28.3
	TeG-FG	R3D-50	**31.1**

Event classification

method	external data	finetuning	F1 score
BMN [50]	IN + THUMOS	✗	18.6
SemeDet [12]	-	✗	27.5
PA [65]	IN	✗	39.6
BMN-SE [50]	IN + THUMOS	✗	49.1
TCN [44]	IN	✗	58.8
PC [65]	IN	✓	62.5
CVRL [58]	-	✓	69.1
TeG-PS	-	✓	69.9
TeG-FG	-	✓	**71.4**

Event boundary detection

method	backbone	pre-train dataset	pre-train epoch	pre-train frame	pre-train FLOPs (G)	acc.
SimCLR [58]	R3D-50	K400	-	-	-	46.8
VINCE [28]	R-50	K400	200	-	-	49.1
ImageNet [58]	R3D-50	IN	-	-	-	53.5
SeCo [83]	R-50	IN† + K400	400	-	-	61.9
ρSwAV (ρ=2) [21]	R3D-50	K400	800	8+8	83.5	63.2
CVRL [58]	R3D-50↓	K400	800	16+16 → 8+8	91.2	66.1
ρBYOL (ρ=2) [21]	R3D-50	K400	800	8+8	83.5	66.2
ρMoCo (ρ=2) [21]	R3D-50	K400	800	8+8	83.5	67.4
ρMoCo (ρ=2) [21]	R3D-50	K400	200	16+16	167.0	67.6
ρBYOL (ρ=4) [21]	R3D-50	K400	200	16+16+16+16	**334.0**	**71.4**
TeG-FG	R3D-50↓	K400	800	16+32 → 8+16	136.4	65.0
TeG-PS	R3D-50↓	K400	800	16+32 → 8+16	136.4	**67.8**

Scene-heavy action recognition