ASCNet: Self-supervised Video Representation Learning with Appearance-Speed Consistency

Deng Huang\textsuperscript{1,3*}, Wenhao Wu\textsuperscript{2*}, Weiwen Hu\textsuperscript{1}, Xu Liu\textsuperscript{1}, Dongliang He\textsuperscript{2}, Zhihua Wu\textsuperscript{2}, Xiangmiao Wu\textsuperscript{1}, Mingkui Tan\textsuperscript{1,4†}, Errui Ding\textsuperscript{2}

\textsuperscript{1}South China University of Technology \quad \textsuperscript{2}Baidu Inc. \quad \textsuperscript{3}Pazhou Laboratory \quad \textsuperscript{4}Key Laboratory of Big Data and Intelligent Robot Ministry of Education
Self-supervised video representation learning aims to learn video features from unlabeled video. The learned video representation can be used for downstream tasks, such as action recognition.
Challenges

1. Videos contain unstructured and noisy visual information.
   - It is hard to learn all information with single task.

2. Videos are unlabeled.
   - It is hard to find sufficient supervision for model training.

Green arrows: motion features
Masks: appearance features
- background, floor and human
Existing methods design pretext tasks to obtain supervision signal from the untrimmed video for representation learning.

- Future prediction task;
- Temporal order sorting task;
- Playback speed prediction task;
- Etc...

3D ST-puzzle (Kim et al. 2019)

CoCLR (Han et al. 2020)
Limitations

Limitations of existing pretext tasks

1. Some of the approaches rely on pre-computed motion information (e.g., optical flow), which is computationally heavy, particularly when the dataset is scaled up.

2. While negative samples play important roles in instance discrimination tasks, it is hard to maintain their quality and quantity. Moreover, same-class negative samples can be harmful to the representations used in downstream tasks.
Our method

Learn robust video representation from consistency between positive samples

- Appearance Consistency Perception (ACP)
- Speed Consistency Perception (SCP)
Appearance Consistency Perception (ACP) Task: minimize the representation distance between two augmented clips from the same video.

Motivation

Different data augmentations or playback speeds do not change the content of the clip.
Speed Consistency Perception (SCP) Task: minimize the distance between two clips with the same playback speed while the appearance can be different.

Motivations

- Temporal information is crucial for the downstream tasks;
- Changes of some motion may be not obvious under different playback speeds: we only minimize distance between the same playback speed.
Experiments

- Experimental results:
  - comparison to the state-of-the-art methods on action recognition;
  - comparison to the state-of-the-art methods on video retrieval.

- Datasets:
  - **Kinetics-400**: ~240K training videos, 400 human action classes;
  - **UCF-101**: 13,320 videos, 101 realistic action categories;
  - **HMDB-51**: 6,849 videos, 51 action classes.
Experimental results

- Comparison with SOTA on action recognition.

Table 2: Comparison with SOTA self-supervised learning methods on the UCF-101 and HMDB-51 datasets.

| Method       | Date  | Dataset (duration) | Backbone | Frames | Res. | Single-Mod | UCF | HMDB |
|--------------|-------|--------------------|----------|--------|------|------------|-----|------|
| Shuffle&Learn [25] | 2016  | UCF (1d)           | CaffeNet | -      | 224  |            | 50.2 | 18.1 |
| OPN [23]     | 2017  | UCF (1d)           | CaffeNet | -      | 224  | ✓          | 56.3 | 22.1 |
| CMC [30]     | 2019  | UCF (1d)           | CaffeNet | -      | 224  | ✓          | 59.1 | 26.7 |
| MAS [33]     | 2019  | UCF (1d)           | C3D      | 16     | 112  |            | 58.8 | 32.6 |
| VCP [24]     | 2020  | UCF (1d)           | C3D      | 16     | 112  | ✓          | 68.5 | 32.5 |
| ClipOrder [39] | 2019  | UCF (1d)           | R(2+1)D | 16     | 112  | ✓          | 72.4 | 30.9 |
| PRP [40]     | 2020  | UCF (1d)           | R(2+1)D | 16     | 112  | ✓          | 72.1 | 35.0 |
| PSP [9]      | 2020  | UCF (1d)           | R(2+1)D | 16     | 112  | ✓          | 74.8 | 36.8 |
| MAS [33]     | 2019  | K400 (28d)         | C3D      | 16     | 112  |            | 61.2 | 33.4 |
| 3D-RotNet [19] | 2018  | K400 (28d)         | 3D R18   | 16     | 112  | ✓          | 62.9 | 33.7 |
| ST-Puzzle [21] | 2019  | K400 (28d)         | 3D R18   | 48     | 224  |            | 65.8 | 33.7 |
| DPC [13]     | 2019  | K400 (28d)         | 3D R18   | 64     | 128  | ✓          | 68.2 | 34.5 |
| CTB [29]     | 2019  | K600+(273d)        | S3D-G    | -      | 112  | ✓          | 79.5 | 44.6 |
| SpeedNet [2] | 2020  | K400 (28d)         | S3D-G    | 64     | 224  | ✓          | 81.1 | 48.8 |
| Pace [34]    | 2020  | K400 (28d)         | S3D-G    | 64     | 224  | ✓          | 87.1 | 52.6 |
| CoCLR-RGB [15] | 2020  | K400 (28d)         | S3D-G    | 32     | 128  | ✓          | 87.9 | 54.6 |
| RSPNet [5]   | 2021  | K400 (28d)         | S3D-G    | 64     | 224  | ✓          | 89.9 | 59.6 |

Ours

K400 (28d) 3D R18 16 112 ✓ 80.5 52.3

Table 3: Performance of different evaluation protocols on UCF-101 dataset. The models are pre-trained on Kinetics-400.

| Arch. | Res. | #Frames | Crop Type | Top-1 |
|-------|------|---------|-----------|-------|
| S3D-G | 224  | 64      | Center-crop | 90.77% |
| 3D R18 | 112  | 16      | Three-crop  | 90.88% |
|       | 128  | 32      | Ten-crop   | 87.31% |
|       | 128  | 16      | Center-crop | 80.52% |
|       | 128  | 16      | Three-crop  | 80.73% |
|       | 128  | 16      | Three-crop  | 80.99% |

Table 4: Performance of different pre-training epochs on UCF-101 dataset.
The model uses a pre-trained 3D ResNet-18 as the backbone.

| Epochs | Top-1 (%) |
|--------|-----------|
| 100    | 76.34     |
| 200    | 80.52     |
| 300    | 81.31     |
| 400    | 81.50     |
Experimental results

- Comparison with SOTA on nearest neighbor video retrieval.

Table 5: Comparison with SOTA methods on the UCF-101 dataset.

| Method       | Architecture | Top-k       |
|--------------|--------------|-------------|
|              |              | $k = 1$ | $k = 5$ | $k = 10$ | $k = 20$ | $k = 50$ |
| OPN [23]     | CaffeNet     | 19.9    | 28.7   | 34.0     | 40.6     | 51.6     |
| Buchler et al. [3] | CaffeNet | 25.7    | 36.2   | 42.2     | 49.2     | 59.5     |
| ClipOrder [39] | 3D R18    | 14.1    | 30.3   | 40.0     | 51.1     | 66.5     |
| SpeedNet [2]  | S3D-G        | 13.0    | 28.1   | 37.5     | 49.5     | 65.0     |
| VCP [24]      | 3D R18       | 18.6    | 33.6   | 42.5     | 53.5     | 68.1     |
|               | R(2+1)D      | 19.9    | 33.7   | 42.0     | 50.5     | 64.4     |
| Pace [34]     | 3D R18       | 23.8    | 38.1   | 46.4     | 56.6     | 69.8     |
|               | C3D          | 31.9    | 49.7   | 59.2     | 68.9     | 80.2     |
| RSPNet [5]    | C3D          | 36.0    | 56.7   | 66.5     | 76.3     | 87.7     |
|               | 3D R18       | 41.1    | 59.4   | 68.4     | 77.8     | 88.7     |
| Ours          | 3D R18       | **58.9**| **76.3**| **82.2**| **87.5**| **93.4**|

Results

- Our ASCNet outperforms other methods on nearest neighbor video retrieval task.
Conclusions

Contributions

- We propose the ACP and SCP tasks for unsupervised video representation learning.
- We propose the appearance-based feature retrieval strategy to select the more effective positive sample for speed consistency perception.
- We verify the effectiveness of ACP and SCP tasks for learning meaningful video representations on two downstream tasks and two datasets.
Thank you!