SKIP-CLIP: SELF-SUPERVISED SPATIOTEMPORAL REPRESENTATION LEARNING BY FUTURE CLIP ORDER RANKING

Alaaeldin El-Nouby*[,1,3], Shuangfei Zhai[2], Graham W. Taylor[1,3,4], Joshua M. Susskind[2]

*work was performed during an internship with Apple Inc.
[1]University of Guelph [2]Apple Inc. [3]Vector Institute [4]CIFAR

1 - MOTIVATION

- The study of the video domain in computer vision is of great importance. However, providing annotations for videos is particularly difficult due to its temporal dimension.
- Some efforts utilize the temporal order signal to design a Self-Supervision task [1, 2, 3]. These efforts disambiguate or sort video frames using no context outside the given frames, which can lead to ambiguous samples as the one shown below.

Ordered Frames

Reversed Frames

- Future prediction is a task that can be often used for learning representations. However, it is challenging since the future is not uni-modal and reconstructing the input can be expensive for videos.

2 - SKIP-CLIP

- We propose a method that alleviates the shortcomings of sorting and future frame prediction approaches by combining both ideas in a single method.
- We sparsely sample a set of frames to be sorted as well as a set of contiguous surrounding frames as a context. The model is trained to predict the correct relative position of every frame given the context using a ranking objective.
- Training the model to rank future frames given a context is a softer, more controlled instantiation of future prediction in latent space.
- To demonstrate the quality of the learned representations, we provide strong results for the downstream task of action recognition using the UCF101 dataset.

3 - NOTATION

Video frames:

$$V = \{v_1, v_2, \cdots, v_N\}$$

K context frames sampled starting time-step t:

$$c = \{v_t, v_{t+1}, \cdots, v_{t+K}\}$$

M target frames subsequent to c:

$$T = \{x_1, x_2, \cdots, x_M\}$$

M Negative frames (sampled from different videos):

$$Q = \{\overline{x}_1, \overline{x}_2, \cdots, \overline{x}_M\}$$

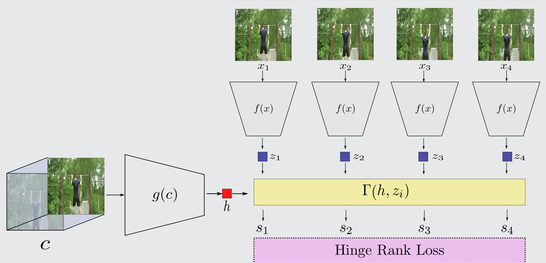
Context encoding:

$$h = g(c)$$

Targets encoding:

$$z_i = f(x_i), \quad \overline{z}_i = f(\overline{x}_i)$$

4 - METHODOLOGY



Objective Functions:

$$L_{\text{rank}} = \sum_{i=0}^{M-1} \sum_{j=i+1}^{M} \max(0, -\Gamma(h, z_i) + \Gamma(h, z_j) + \delta_{rank}).$$

$$L_{\text{contrastive}} = \sum_{i=0}^{M} \max(0, -\Gamma(h, z_i) + \mathbb{E}_{\overline{z} \in Q}[\Gamma(h, \overline{z})] + \delta_{neg}).$$

Scoring Function:

$$\Gamma(h, z_i) = \frac{1}{H * W} \sum_{m=0}^{H} \sum_{n=0}^{W} \frac{h^{m,n} \cdot z_i^{m,n}}{||h^{m,n}|| \cdot ||z_i^{m,n}||}$$

- An auxiliary rotation prediction task is used as a regularization for the encoder $f(x)$.
- $g(c)$ encoder parameters are trained for the pretext task of future frames ranking. The parameters are then fine-tuned for the downstream task of action recognition. Both pre-training and fine-tuning are performed using the UCF-101 dataset.

5- RESULTS

Model	UCF101
Skip-Clip	59.5
Skip-Clip + rotation	63.1
Skip-Clip + rotation + negative sampling	**64.4**

Table 1: Ablation Study comparing the base model to models with additional auxiliary objectives.

Method	Accuracy
Random Initialization [4]	42.4
ImageNet Inflation [5]	60.3
Skip-Clip	**64.4**

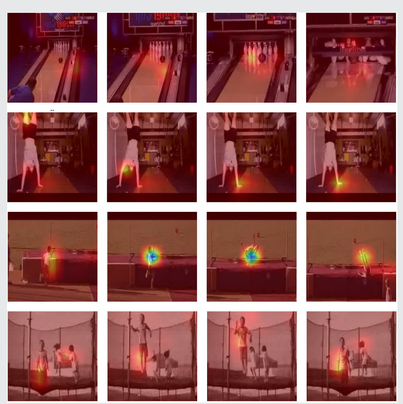
Table 2: Top-1 accuracy comparison to standard initialization baselines performance for action recognition task on UCF-101 dataset.

- Basic Skip-Clip model without the auxiliary rotation objective can be susceptible to learning trivial solutions to the ranking task, which does not necessarily transfer well to other semantic tasks like action recognition.

5 - RESULTS (CONTD.)

Method	Backbone	Source	UCF101
Shuffle and Learn [2]	AlexNet	UCF101	50.9
Arrow of time [1]	AlexNet	UCF101	55.3
OPN [3]	AlexNet	UCF101	56.3
VideoGAN [6]	C3D	UCF101	52.1
Motion & Appearance [7]	C3D	UCF101	58.8
Motion & Appearance [7]	C3D	Kinetics	61.2
3DRotNet [8]	3D ResNet-18	Kinetics	62.9
Video Clip Ordering [9]	R3D	UCF101	64.9
3DCubicPuzzles [5]	3D ResNet-18	Kinetics	**65.8**
Skip-Clip	3D ResNet-18	UCF101	64.4

Table 3: Top-1 Accuracy performance for action recognition task on UCF-101 dataset. Different backbones used for by the methods can account for some of the performance difference.



Visualizations of the cosine similarity per aligned cell between context and target representations that are part of computing the scores $\Gamma(h, z)$. We can observe that the highlighted regions correspond to salient motions in the frames.

6- REFERENCES

[1] D Wei et al. "Learning and Using the Arrow of Time". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
[2] I Misra et al. "Unsupervised Learning using Sequential Verification for Action Recognition". In: *CoRR* abs/1603.08561 (2016). arXiv: 1603.08561. URL: http://arxiv.org/abs/1603.08561.
[3] H Lee et al. "Unsupervised Representation Learning by Sorting Sequences". In: *CoRR* abs/1708.01246 (2017). arXiv: 1708.01246. URL: http://arxiv.org/abs/1708.01246.
[4] K Hara et al. "Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?" In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (June 2018). DOI: 10.1109/cvpr.2018.00685. URL: http://dx.doi.org/10.1109/CVPR.2018.00685.
[5] D Kim et al. "Self-Supervised Video Representation Learning with Space-Time Cubic Puzzles". In: *CoRR* abs/1811.09795 (2018). arXiv: 1811.09795. URL: http://arxiv.org/abs/1811.09795.
[6] C Vondrick et al. "Generating Videos with Scene Dynamics". In: *Advances in Neural Information Processing Systems 29*. Ed. by DD Lee et al. Curran Associates, Inc., 2016, pp. 613–621. URL: http://papers.nips.cc/paper/6194-generating-videos-with-scene-dynamics.pdf.
[7] J Wang et al. *Self-supervised Spatio-temporal Representation Learning for Videos by Predicting Motion and Appearance Statistics*. 2019. arXiv: 1904.03597 [cs.CV].
[8] L Jing et al. *Self-Supervised Spatiotemporal Feature Learning via Video Rotation Prediction*. 2018. arXiv: 1811.11387 [cs.CV].
[9] D Xu et al. "Self-Supervised Spatiotemporal Learning via Video Clip Order Prediction". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.