“Knights” CRCV VIPriors21 Action Recognition submission

Ishan Dave[1⋆], Naman Biyani[2⋆], Brandon Clark[1], Rohit Gupta[1], Yogesh Rawat[1] and Mubarak Shah[1]

[1] Center for Research in Computer Vision (CRCV), University of Central Florida, Orlando, Florida, USA
[2] Indian Institute of Technology, Kanpur, India,
{ishandave, brandonclark314, rohitg}@knights.ucf.edu, namanb@iitk.ac.in {yogesh, shah}@crcv.ucf.edu

Abstract. This technical report presents our approach *“Knights”* to solve the action recognition task on a small subset of Kinetics-400 i.e. *Kinetics400ViPriors* without using any extra-data. Our approach has 3 main components: state-of-the-art self-supervised pretraining, video transformer models, and optical flow modality. Along with the use of standard test-time augmentation, our proposed solution achieves **73%** on Kinetics400ViPriors test set, which is the best among all of the other entries *Visual Inductive Priors for Data-Efficient Computer Vision*’s Action Recognition Challenge, ICCV 2021.

1 Introduction

Deep learning has enabled progress in video understanding tasks like action recognition [13,16,7], action detection [26,11,6,25] temporal action localization [29,28] etc. Most of the advancement in the video understanding due to deep network is built upon existence of the large scale labeled data like Kinetics [2], HACS [32], LSHVU [9] etc.

Recent works in video self-supervised learning [5,30,22,18,24,19] show that spatio-temporal representations learned from the self-supervised learning on the same dataset also helps in improving results of the video encoder by a significant margin over the training from scratch. In our proposed solution we opt for the Temporal Contrastive Learning for video Representation (TCLR) method [5] which gets the maximum gain among all methods while pretraining from the same dataset. On UCF101, without using any additional labeled or unlabeled data, TCLR pretrained model results in a boost of **20%** Top-1 accuracy over the baseline model.

Recently, transformers have been applied to key computer vision tasks such as image classification after the introduction of Vision Transformer (ViT) [10]. The impressive performance of transformers in the image domain led to investigation of Transformer-based architectures for video-based classification tasks.

⋆ equal contribution

Video transformers have lead to state of the art performances on Kinetics-400 [3], SSv2 [14] and Charades [27]. Adding temporal attention encoder on top of ViT [10](Pretrained) was proposed in VTN [23] which led to good performance on video action recognition. A factorized spacetime attention based approach was proposed in TimeSformer [1] after analysis of various variants of space-time attention based on compute-accuracy tradeoff. Video Swin Transformer [21] investigated spatiotemporal locality and showed that an inductive bias of locality a better speed-accuracy trade-off compared to other approaches which use global self-attention. In our proposed solution we adopt MViT [12] transformers which shows state-of-the-art performance on Kinetics-400 without requiring any pretraining checkpoint unlike other video transformers architectures which requires ImageNet [8] pretraining. Apart from eliminating the pretraining requirement, another advantage of using MViT is low computational requirement due to its pooling attention for spatiotemporal modeling.

While learning from scratch, it is difficult to optimize the parameters of a 3D ConvNet based architecture with a single stream of RGB video frames from relatively smaller datasets such as Kinetics400ViPriors as compared to Kinetics-400 [3] as the given dataset has same number of classes as Kinetics400 but roughly ∼20% of the number of videos in Kinetics-400. Carreira et al. [3] show that two-stream-based 3D ConvNet approaches significantly surpass single stream RGB video-based 3D ConvNet approaches; there is a ∼30% improvement for the task of action recognition on both UCF101 and HMDB when no pretrained weights are used. Also, [4] shows that optical flow is a powerful prior for modeling motion information while learning from scratch.

We use an ensemble of various TCLR self-supervised pretrained 3D ConvNets and video transformers in the RGB stream and an ensemble of various 3D ConvNets in the optical flow streams. This helps in mitigating common generalization errors as well as decreasing the variance in neural network predictions.

2 Proposed Method

2.1 Self-supervised pretraining- TCLR

TCLR self-supervised framework explicitly encourages the learning of temporally distinct video representations. TCLR framework consists of mainly three components:

Instance Contrastive Loss In a mini-batch of N different video instances, 2 clips are taken from a video instance and stochastically transformed using various geometric and appearance based transforms. Following the instance discrimation objective the 2 differently augmented clips are brought together in the representation space whereas the clips from the different instances are pushed further apart using Instance Contrastive Loss ($\mathcal{L}_{IC}$)

$$\mathcal{L}_{IC}^{i} = -\log \frac{h\left(G_i, G_i'\right)}{\sum_{j=1}^{N}[\mathbb{1}_{[j\neq i]} h(G_i, G_j) + h(G_i, G_j')]}, \tag{1}$$

where, (G_i, G'_i) are two clip representations from same instance i, $h(u, v) = \exp\left(u^T v/(\|u\|\|v\|\tau)\right)$ is used to compute the similarity between u and v vectors with an adjustable parameter temperature, τ. $\not\Vdash_{[j \neq i]} \in \{0, 1\}$ is an indicator function which equals 1 iff $j \neq i$.

Local-Local Temporal Contrastive Loss For this loss, we treat non-overlapping clips sampled from different temporal segments of the same video instance as negative pairs, and randomly transformed versions of the same clip as a positive pair. This allows the model to learn differences between timestamps of a video. The loss is defined as

$$\mathcal{L}^i_{LL} = -\sum_{p=1}^{N_T} \log \frac{h\left(G_{i,p}, G'_{i,p}\right)}{\sum_{q=1}^{N_T} [\not\Vdash_{[q \neq p]} h(G_{i,p}, G_{i,q}) + h(G_{i,p}, G'_{i,q})]}. \tag{2}$$

where, $G_{i,p}$ represents a clip from instance i at time p and $G'_{i,p}$ represents the transformed version of the same clip. The positive pairs for this loss are formed by two clips from the same instance i and the same timestamp p (e.g. $G_{i,p}$ and $G'_{i,p}$). Any two clips from different timestamps of an instance i are treated as a negative pairs (e.g. $G_{i,p}$ and $G_{i,q}$ form a negative pair). A given video instance i is divided into N_T non-overlapping clips. Hence, for every positive pair, the local-local contrastive loss has $2 \times N_T - 2$ negative pairs.

Global-Local Temporal Contrastive Loss The purpose behind global-local temporal contrastive loss is to encourage the model to learn features that represent the temporal locality of the input clip across the temporal dimension of the feature map.

$$\mathcal{L}^i_{GL_k} = \log \frac{h\left(L_{i,k}, G_{i,k}\right)}{\sum_{q=1}^{N_T} h(L_{i,k}, G_{i,q})} + \log \frac{h\left(G_{i,k}, L_{i,k}\right)}{\sum_{q=1}^{N_T} h(G_{i,k}, L_{i,q})}, \tag{3}$$

$$\mathcal{L}^i_{GL} = -\sum_{k=1}^{N_T} \mathcal{L}^i_{GL_k}. \tag{4}$$

Where, $L_{i,k}$ represents the local clip from instance i for timestamp k, and $G_{i,k}$ represents the features that are pooled from a global clip of i but represent the same timestamp k. Hence, there are two separate ways to represent the timestamp k of any instance i. This loss has two sets of reciprocal terms, with $G_{i,k}$ and $L_{i,k}$ serving as the anchor for each term and creating a positive pair. The negative pairs are supplied by matching the anchors with representations corresponding to other non-overlapping local clips. Note that similar to our local-local temporal contrastive loss we do not use negatives from other video instances for calculating this loss.

2.2 Multiscale Vision Transformer

MViT[12] is a multi-scale vision transformer which is trained from scratch. Contrary to the typical Vision Transformer [10] models which use a constant feature

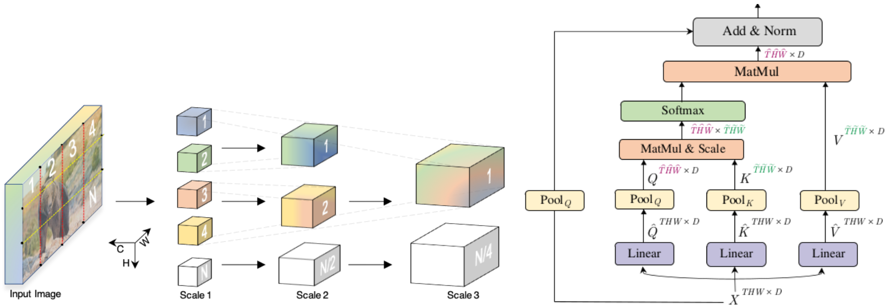

Fig. 1. The left picture shows an overview of MViT's approach to learn from dense simple features to courser and complex features as the number of channels increase and resolution decreases. The right picture shows an overview of the flexible Pooling Attention mechanism of MViT[12]

dimension and resolution in all layers and an attention mechanism to determine which previous tokens to focus, MViT [12] proposes a flexible Multi Head Pooling Attention mechanism that pools the projected query, key, and value vectors, enabling reduction of the visual resolution. The right of Figure 1 gives overview of Multi Head Pooling Attention approach proposed by MViT [12].

This pooling attention is combined with an increase in the channel dimension with the idea of hierarchical feature construction from simple features which have high visual resolution and lower dimensions to more complex features having higher dimension features with lower resolution. The left of Figure 1 gives overview of hierarchical feature construction approach of MViT.

2.3 Optical Flow

Optical flow has shown to increase performance on several video related tasks. In our experiments we calculate TVL-1 optical flow [31] for the dataset and use those features for training. The TVL-1 algorithm is based on a minimization of the energy function:

$$E(\mathbf{u}) = \int_{\Omega} |\nabla u_1| + |\nabla u_2| + \lambda |\rho(\mathbf{u})| \tag{5}$$

3 Experiments

This section covers details of the challenge dataset, implementation details, and results.

3.1 Dataset

We use the provided Kinetics400ViPriors dataset, a modification of the official Kinetics400[20] dataset. The challenge dataset consists of 40k videos for training, 10k videos for validation, and 20k videos for testing. We use TV-L1 optical flow computed by the [31] method.

3.2 Implementation Details

We perform TCLR self-supervised pretraining on train+val set without using any labels. The pretraining is performed for 400 epochs for R3D-18 and R3D-50 architectures following the learning rate schedule of [5]. After the pretraining, each model is finetuned for the 150 epochs. More details input clip during training and inference setting is given in Table 1.

We have used the MViT-B with a convolutional pooling function as described in [12] which consists of 4 scale stages. The model takes 16 frames of 224×224 resolution as input with a skip rate of 4. We used the code provided in MViT [12] paper[3]. We have followed truncated normal distribution initialization [15] and trained for 200 epochs with 2 repeated augmentation [17] repetitions as described in MViT [12].

3.3 Results

We performed our initial experiments on Kinetics400ViPriors validation set by training on just training set to observe the performance of our ensemble and model selection purposes, shown in Table 1. The best models from the validation set performance are later finetuned on the train+val set for 100 epochs, and submitted for the test set evaluation. The proposed method achieves a Top-1 accuracy of 73% on Kinetics400ViPriors test set without using additional data in our training.

Table 1. Overview of accuracies of approaches used on val set

#	Model	Resolution	Frames × skip	Test Time Crops (spatial × temporal)	Validation Accuracy
1	MViT	224	16 x 4	3 x 10	51.3%
2	I3D-OF	224	16 x 2	3 x 10	40.5%
3	I3D-OF	224	16 x 4	3 x 5	42.1%
4	R3D18-OF	112	16 x 2	3 x 10	31.9%
6	R3D-50 TCLR	224	16 x 2	3 x 10	61.1%
6	R3D-18	112	16 x 2	3 x 10	33.2%
7	R3D-18 TCLR	112	16 x 2	3 x 10	46.0%

[3] `https://github.com/facebookresearch/SlowFast`

4 Conclusion

In this technical report, we have shown that the self-supervised pretraining improves results significantly while learning from limited data without pretraining from additional labeled or unlabeled data. Combining the self-supervised 3D-CNNs with the state-of-the-art video transformer models and optical flow performs competitively on the test set. We believe our method can be further improved by pretraining the video transformer models in self-supervised manner on the same dataset.

References

1. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? (2021)
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
4. Dave, I., Carter, K., Shah, M.: "kallis" crcv vipriors challenge submission
5. Dave, I., Gupta, R., Rizve, M.N., Shah, M.: Tclr: Temporal contrastive learning for video representation. arXiv preprint arXiv:2101.07974 (2021)
6. Dave, I., Scheffer, Z., Tirupattur, P., Rawat, Y., Shah, M.: Ucf-system: Activity detection in untrimmed videos (2020)
7. Demir, U., Rawat, Y.S., Shah, M.: Tinyvirat: low-resolution video action recognition. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 7387–7394. IEEE (2021)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
9. Diba, A., Fayyaz, M., Sharma, V., Paluri, M., Gall, J., Stiefelhagen, R., Van Gool, L.: Large scale holistic video understanding. In: European Conference on Computer Vision. pp. 593–610. Springer (2020)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021)
11. Duarte, K., Rawat, Y., Shah, M.: Videocapsulenet: A simplified network for action detection. Advances in Neural Information Processing Systems **31**, 7610–7619 (2018)
12. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C.: Multiscale vision transformers (2021)
13. Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 203–213 (2020)
14. Goyal, R., Kahou, S.E., Michalski, V., Materzyńska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., Memisevic, R.: The "something something" video database for learning and evaluating visual common sense (2017)

15. Hanin, B., Rolnick, D.: How to start training: The effect of initialization and architecture (2018)
16. Hara, K., Kataoka, H., Satoh, Y.: Towards good practice for action recognition with spatiotemporal 3d convolutions. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 2516–2521 (2018)
17. Hoffer, E., Ben-Nun, T., Hubara, I., Giladi, N., Hoefler, T., Soudry, D.: Augment your batch: Improving generalization through instance repetition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
18. Huo, Y., Ding, M., Lu, H., Lu, Z., Xiang, T., Wen, J.R., Huang, Z., Jiang, J., Zhang, S., Tang, M., Huang, S., Luo, P.: Self-supervised video representation learning with constrained spatiotemporal jigsaw (2021), `https://openreview.net/forum?id=4AWko4A35ss`
19. Jenni, S., Meishvili, G., Favaro, P.: Video representation learning by recognizing temporal transformations. In: The European Conference on Computer Vision (ECCV) (August 2020)
20. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The kinetics human action video dataset (2017)
21. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer (2021)
22. Luo, D., Liu, C., Zhou, Y., Yang, D., Ma, C., Ye, Q., Wang, W.: Video cloze procedure for self-supervised spatio-temporal learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 11701–11708 (2020)
23. Neimark, D., Bar, O., Zohar, M., Asselmann, D.: Video transformer network (2021)
24. Qian, R., Meng, T., Gong, B., Yang, M.H., Wang, H., Belongie, S., Cui, Y.: Spatiotemporal contrastive video representation learning. arXiv preprint arXiv:2008.03800 (2020)
25. Rana, A.J., Rawat, Y.S.: We don't need thousand proposals: Single shot actor-action detection in videos. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2960–2969 (January 2021)
26. Rizve, M.N., Demir, U., Tirupattur, P., Rana, A.J., Duarte, K., Dave, I.R., Rawat, Y.S., Shah, M.: Gabriella: An online system for real-time activity detection in untrimmed security videos. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 4237–4244. IEEE (2021)
27. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding (2016)
28. Swetha, S., Kuehne, H., Rawat, Y.S., Shah, M.: Unsupervised discriminative embedding for sub-action learning in complex activities. arXiv preprint arXiv:2105.00067 (2021)
29. Tirupattur, P., Duarte, K., Rawat, Y.S., Shah, M.: Modeling multi-label action dependencies for temporal action localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1460–1470 (June 2021)
30. Wang, J., Jiao, J., Liu, Y.H.: Self-supervised video representation learning by pace prediction. In: The European Conference on Computer Vision (ECCV) (August 2020)
31. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime tv-l 1 optical flow. In: Joint pattern recognition symposium. pp. 214–223. Springer (2007)

32. Zhao, H., Torralba, A., Torresani, L., Yan, Z.: Hacs: Human action clips and segments dataset for recognition and temporal localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8668–8678 (2019)