Large-Scale Self-Supervised Human Activity Recognition

Mohammad Zaki Zadeh
University of Texas at Arlington
Arlington, Texas, USA

Ashish Jaiswal
University of Texas at Arlington
Arlington, Texas, USA

Hamza Reza Pavel
University of Texas at Arlington
Arlington, Texas, USA

Aref Hebri
University of Texas at Arlington
Arlington, Texas, USA

Rithik Kapoor
University of Texas at Arlington
Arlington, Texas, USA

Fillia Makedon
University of Texas at Arlington
Arlington, Texas, USA

ABSTRACT

In this paper, a self-supervised approach is used to obtain an effective human activity representation using a limited set of annotated data. This research is aimed on acquiring human activity representation in order to improve the accuracy of classifying videos of human activities in the NTU RGB+D 120 dataset. The effectiveness of various self-supervised approaches, as well as a supervised method, is studied. The results reveal that when the training set gets smaller, the performance of supervised learning approaches diminishes, whereas self-supervised methods maintain their performance by utilizing unlabeled data.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**.

KEYWORDS

computer vision, deep learning, self-supervised learning

ACM Reference Format:
Mohammad Zaki Zadeh, Ashish Jaiswal, Hamza Reza Pavel, Aref Hebri, Rithik Kapoor, and Fillia Makedon. 2022. Large-Scale Self-Supervised Human Activity Recognition. In *The15th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '22), June 29-July 1, 2022, Corfu, Greece.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3529190.3534720

1 INTRODUCTION

The supervised method of learning features from annotated data has nearly hit its limit due to the enormous work required to manually annotate millions of data samples. Self-supervised approaches [4] are at the forefront of efforts to adapt deep learning methods to learn feature representations without the need front expensive annotations.

The goal of this study (Figure 1) is to recognize human activity in the NTU RGB+D 120 dataset [5], which contains 120 different action types such as daily, mutual, and health-related activities. Similar to our earlier work [6], three distinct state-of-the-art self-supervised learning approaches including MoCo [3], SimSiam [2], and VICReg [1] were used to pre-train the model in a self-supervised way and their results were compared to a supervised approach.

Table 1: Different methods' top-1 classification accuracy.

Method	Cross-Subject			Cross-Setup		
	50%	25%	10%	50%	25%	10%
Sup.	**57.77**	37.01	25.28	**57.10**	21.10	20.28
MoCo	42.16	30.54	26.40	39.63	29.32	18.29
SimSiam	44.03	30.99	29.43	42.79	28.38	19.74
VICReg	50.92	**39.50**	**34.67**	49.68	**31.48**	**25.18**

2 METHODOLOGY AND RESULTS

Contrastive learning is one of the most popular self-supervised methodologies (CL). CL aims to group similar (positive) samples together, while separating different (negative) samples. Because the amount of negative samples has an impact on the performance of CL techniques, various strategies have been developed to address this issue [4]. The momentum encoder method (MoCo) (Figure 2 left) creates a dictionary in the form of a queue of encoded samples, with the current mini-batch enqueued and the oldest mini-batch dequeued [3]. The momentum encoder shares the same parameters as the query encoder (θ_q) and its parameters (θ_k) are updated based on the parameters of the query encoder ($\theta_k = m\theta_k + (1-m)\theta_q, m \in [0, 1)$: momentum coefficient).

The second method, called SimSiam [2], avoids collapsing solutions by maximizing the similarity of an image's two views directly, without utilizing negative pairs or a momentum encoder. Stop-gradient action, according to the authors, is crucial in preventing collapsing solutions. The SimSiam method architecture is depicted in Figure 2 (middle).

Another self-supervised method for dealing with the collapsing solutions problem is VICReg [1] (Variance-Invariance-Covariance Regularization). The VICReg architecture illustrated in 2 (right) is symmetric and is based on three simple principles: variance, invariance, and covariance. By regulating the variance of the representations along each dimension independently, the variance principle is a simple yet effective technique for preventing collapse. The invariance principle learns invariance to diverse viewpoints of a picture using a conventional mean-squared Euclidean distance without requiring any negative pairs. Finally, the covariance principle employs the covariance criteria, which decorrelates the various dimensions of learnt representations in order to disseminate information across dimensions and avoid dimension collapse.

PETRA '22, June 29-July 1, 2022, Corfu, Greece

ACM ISBN 978-1-4503-9631-8/22/06.
https://doi.org/10.1145/3529190.3534720

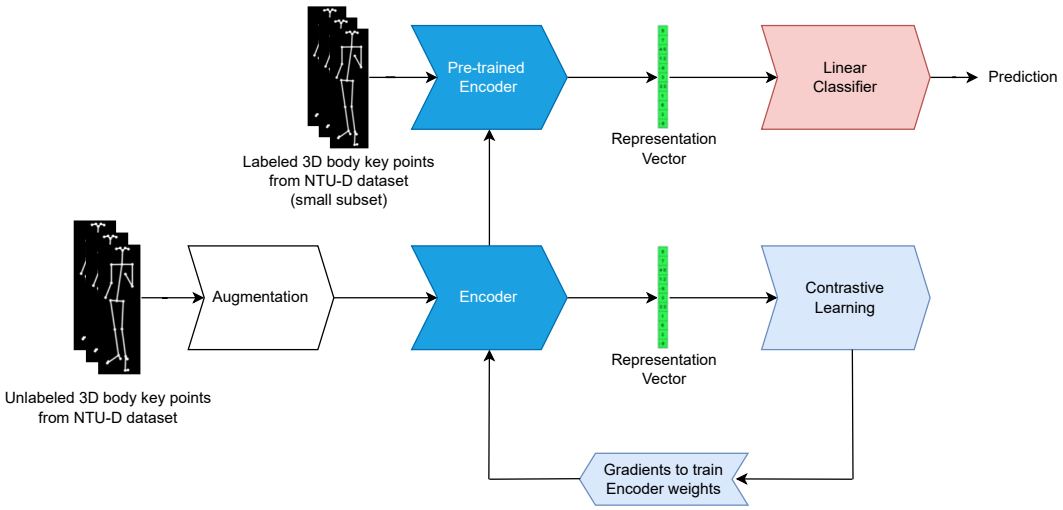

Figure 1: Proposed method architecture: top—supervised classification; bottom—self-supervised pre-training.

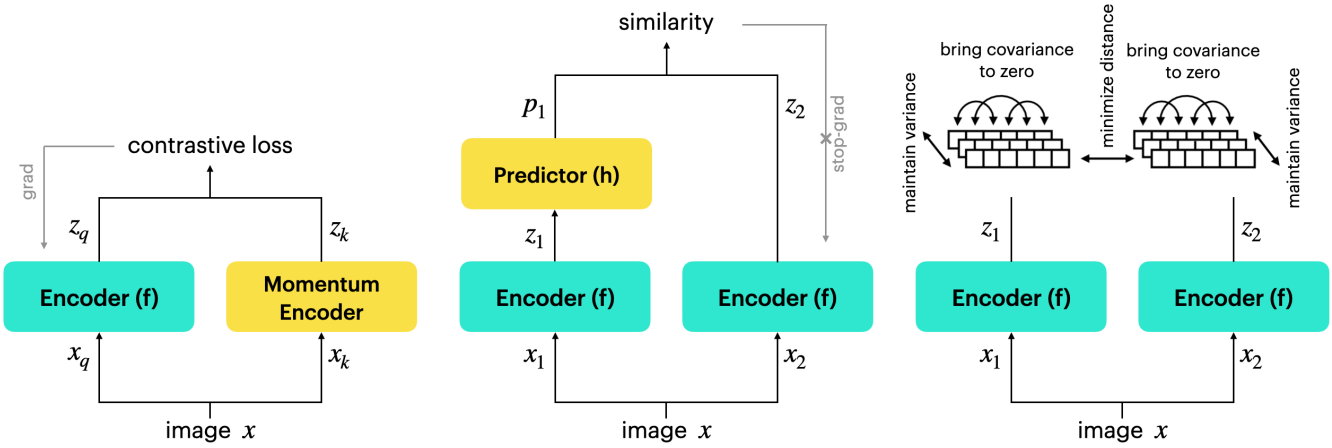

Figure 2: Different self-supervised learning architectures: left—MoCo [3]; middle—SimSiam [2]; right—VICReg [1].

The architecture of the proposed computer vision system is depicted in Figure 1. In order to pre-train the classifier model, the publicly available NTU-RGB+D 120 [5] was used without any labels. This dataset contains 120 action classes and 114,480 video samples. In this work, only 3D skeletal data were employed and a four-layer 1D convolutional neural network (CNN) with one penultimate transformer layer was used as the encoder network.

Three scenarios were created to evaluate the performance of the proposed methods in the case of a modest amount of annotated data. In the first instance, half of the data was utilized for training and half was used for testing. In the second instance, 25% of the data was used for training while the other 75% was used for testing. Finally, in the final case, 10% of the data was used for training while the other 90% was used for testing. The results show that when the training set gets smaller, the supervised method's classification accuracy declines, whereas the self-supervised techniques retain or even beat the supervised approach.

REFERENCES

[1] Adrien Bardes, Jean Ponce, and Yann LeCun. 2021. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. arXiv:2105.04906 [cs.CV]
[2] Xinlei Chen and Kaiming He. 2020. Exploring Simple Siamese Representation Learning. arXiv:2011.10566 [cs.CV]
[3] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2019. Momentum Contrast for Unsupervised Visual Representation Learning. arXiv:1911.05722 [cs.CV]
[4] Ashish Jaiswal, Ashwin ramesh babu, Mohammad Zadeh, Debapriya Banerjee, and Fillia Makedon. 2020. A Survey on Contrastive Self-Supervised Learning. *Technologies* 9 (12 2020), 2. https://doi.org/10.3390/technologies9010002
[5] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. 2019. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019). https://doi.org/10.1109/TPAMI.2019.2916873
[6] Mohammad Zaki Zadeh, Ashwin Ramesh Babu, Ashish Jaiswal, and Fillia Makedon. 2022. Self-Supervised Human Activity Representation for Embodied Cognition Assessment. *Technologies* 10, 1 (2022). https://doi.org/10.3390/technologies10010033