Signal, Image and Video Processing (2022) 16:1667–1672
https://doi.org/10.1007/s11760-021-02122-x

ORIGINAL PAPER

Study on temperature (τ) variation for SimCLR-based activity recognition

Pranjal Kumar[1] · **Siddhartha Chauhan**[1]

Received: 9 July 2021 / Revised: 8 December 2021 / Accepted: 12 December 2021 / Published online: 29 January 2022

Abstract
Human activity recognition (HAR) is a process to automatically detect human activities based on stream data generated from various sensors, including inertial sensors, physiological sensors, location sensors, cameras, time, and many others. Unsupervised contrastive learning has been excellent, while the contrastive loss mechanism is less studied. In this paper, we provide a temperature (τ) variance study affecting the loss of SimCLR model and ultimately full HAR evaluation results. We focus on understanding the implications of unsupervised contrastive loss in context of HAR data. In this work, also regulation of the temperature (τ) coefficient is incorporated for improving the HAR feature qualities and overall performance for downstream tasks in healthcare setting. Performance boost of 3.3% is observed compared to fully supervised models.

Keywords Contrastive learning · Activity recognition · Healthcare

1 Introduction

The purpose of human activity recognition (HAR), which consists of observations and analysis of human behavior and its environment, is to determine the current behavior and goals of the human body. HAR research has gained attention by its advantages in smart surveillance systems, healthcare systems, connections between virtual reality, smart homes, aberrant behavior detection and other areas and its capacity to support and connect with unique disciplines. One of the most widely discussed research areas is HAR [1] among academics from both academia and industry whose aim is the progress of all-round computing and human–computer interaction. The advances in deep learning have made the field a key component of the most smart systems and in the majority of computer vision tasks such as image classification, object detection, image separation and activity recognition, and natural languages processing (NLP). Due to the intensive work required by manually notifying millions of data samples, supervised strategy to learn features from labelled data has nearly been saturated. Though a plethora of information is available, researchers have been urged to find alternative ways of making use of it by lack of annotations. Unsupervised learning makes it possible for us to learn feature representations without the supervision of human beings. Contrastive learning has reached a state of the art in a variety of tasks, which was recently proposed as an unsupervised study [2–7]. The main difference from other techniques is that the data transformation and contrastive loss strategy are used. In short, most contrastive learning methods construct first a series of augmented data to build positive and negative pairs on an instance level. Similarity between positive pairs could then be maximized by different contrast losses, such as Triplet [8], NCE [9] and NT-Xent [3], while negative pairs could be minimized. Uniformity helps contrastive learning to learn distinguishable characteristics, but overpursuit of uniformity makes the contrastive loss unable to tolerate closely correlated samples which break down the underlying structure and damage downstream feature attributes [10, 11]. SimCLR is a basic visual learning framework for contrastive learning [3], and recently, it was incorporated for healthcare and HAR in particular for the first time [12]. The difficulty in collection of labelled data has been a major roadblock in using data-oriented methods for digital health. This is due to the limitations of HAR's labelled datasets, especially in healthcare-related contexts. Motivated from the work done by Tang et al. [12],

✉ Pranjal Kumar
pranjal@nith.ac.in

Siddhartha Chauhan
sid@nith.ac.in

[1] Computer Science & Engineering Department, NIT Hamirpur, Hamirpur, Himachal Pradesh (H.P) 177005, India

in this paper, we lay out a module that would deemed to be beneficial for HAR systems and other healthcare-related applications. Main contribution of this work is summarized below:

- We provide a study for understanding the behavior of contrastive learning (emphasizing on temperature coefficient,) in sensor data context for human activity recognition.
- We optimize the SimCLR module by regulating the temperature coefficient in order to enhance the quality of features for downstream tasks.
- Improved performance for overall model [12].

2 Related work

Many studies have examined the identification of human activities from diverse points of view. These include by specialized approach [13]; by predictions [25]. Different sensors such as video cameras, ambient temperature sensors, relative humidity, light, pressure and wearable sensors are used. The major forms of wearable sensors are generally algorithm type [14], sensor type [1, 15, 16]; fuse type [17] or device type [18], although other analyses have been carried out more generally by the HAR categories [19, 20]. HAR accomplished five primary tasks, namely the recognition of the fundamental activities [21], the recognition of everyday activities [22], uncommon events [23], biometric subjects [24], and energy expenditure integrated smartphone sensors or sensors incorporated into wearable devices. Dong and Biawas [26] introduced a wearable sensor network designed to monitor human activity. In a similar study, Curone et al. have used wearable triaxial accelerometers to monitor activity [27].

Progress in deep learning has made the field a central part of the smart systems. The ability to learn rich patterns from today's vast amount of data makes the use of deep neural networks (DNNs) an important approach in HAR. The amount of annotated training data available is very reliant on traditional supervised learning approaches. Self-supervised learning methods have recently integrated both generative [28] and contrastive [3] approaches that have been able to use unlabeled data to understand the underlying representations. In recent studies [7, 29–34] on unsupervised feature representation for images, concept known as contrastive learning was incorporated [5]. Contrastive learning (CL) is a discriminatory approach that aims to group similar samples closer and far away. The results after application of contrastive learning are astounding: for example, SimCLR [3] reduces the gap between unsupervised and supervised pre-training representations in linear classification performance.

3 Contrastive learning

The purpose of contrastive methodology is to comprehend a function that maps the input data features to the features on a hypersphere dimension. Wang et al. depicted that the contrastive loss for a given unlabeled training sample set $X = \{x_1, x_2 \ldots x_N\}$ is given as follows [11]:

$$L(x_i) = -\log\left[\frac{\exp\left(\frac{s_{i,i}}{\tau}\right)}{\sum_{k\neq i}\exp\left(\frac{s_{i,k}}{\tau}\right) + \exp\left(\frac{s_{i,i}}{\tau}\right)}\right] \tag{1}$$

where $s_{i,j} = f(x_i)^T g\left(x_j\right) f()$ is an extractor that maps pixel space corresponding images onto space in a hypersphere $g(\cdot)$ could serve the same purpose as of f [3]. τ is a temperature hyper-parameter that helps in distinguishing positive and negative samples. The contrastive loss attempts to attract positive key samples and separates the negative key samples. This goal can also be achieved with a simpler contrastive loss function as shown below [11]:

$$L_{simple}(x_i) = -s_{i,i} + \lambda \sum_{i\neq j} s_{i,j} \tag{2}$$

The goal of contrastive learning is to learn augmented data alignment and discriminatory embedding. The contrastive loss does not restrict negative sample distribution. The temperature contributes to the control of penalty strength on hard negative samples. Specifically, small contrastive losses tend to penalize much more the most severe negative samples in the form of a more separate local structure in a sample and a more uniform embedding distribution [11].

4 Methodology

By using a contrastive loss in the latent space, SimCLR[3] learns representations by maximizing agreement between views of the same data that have been augmented in different ways. SimCLR architecture consists of these primary modules.

- A data incrementation module that randomly transforms a given example of data leading to two correlated views on the same example.
- A network base neural encoder that extracts vectors from enhanced data examples.
- A neural network projection head maps the space where the contrast loss is applied.
- A loss function set to a contrastive prediction task.

The image augmentation operators have been replaced by 8 augmentation functions [12] (adding a random amount of

Gaussian noise, reversing signals, scrambling different sections of signal data, rearranging the different input channels, etc.) designed for time-series sensor data that mimic common sensor noises. By incorporating different pairs of functions in various orders, these functions are utilized to create up to 64 non-identical augmentation functions. With the adoption of the NT-Xent [3] (normalized temperature scaled cross entropy loss), the model is trained so that it has the greatest possible agreement between positive pairs. TPN [35], a lightweight neural network architecture consisting of three 1D convolution layers, has been used as the base encoder in this study. The projection head was made up of a three-layer MLP that was fully connected.

5 Results & discussion

5.1 Dataset

MotionSense [37] was used in our assessment as a publicly available dataset. This dataset comprises data from 24 individuals who carried an iPhone 6s in the front pocket of their pants and perform 6 different activities: walking downstairs, upstairs, walking, jogging, sitting and standing. In this study 6630 windows, each 400 timestamping and 50 percent overlap were used for data from a 50% triaxial accelerometer.

5.2 Experimental setup

Linear and fine tune evaluation was administered on NVIDIA TESLA V100 SXM2. During pre-entraining for 200 epochs and batch size 512, the SGD optimizer with a cosine decay of learning rate is used. TPN [35] is incorporated as a base encoder for the HAR systems to suit the needs of a comparatively lightweight neural network architecture. The projection head was utilized as a three-layer, fully connected MLP with a loss function of NT-Xent. The base decoder is composed of three temporal (1D) layers, each with 24, 16, 8 and 32, 64 and 96 kernel sizes. During preparation, the projection head is composed of 3 fully connected layers with 256, 128 and 50 units, and the grading head is composed of two fully connected layers of 1024 and 6 units in the fine-tuned evaluation. A 0.1 drop-off rate is used to activate the ReLU function. At the end, there is an additional global maximum pooling layer. The model is trained at the SGD optimizer for linear assessment for 50 epochs and a learning rate of 0.03. The model is perfectly tuned with Adam optimizer and a study rate of 0.001 for 50 epochs for a finely tuned assessment.

5.3 Quantitative results

In this section, we conduct extensive experimentation on the temperature coefficient, in order to understand the modeling relationship of the proposed network using activity prediction precision as the assessment metric. The effect of the temperature is assessed. In the first place, we try to determine whether the temperature precisely checks the severity of the penalties in severe negative samples. Numerical results are tabulated in Table 2.

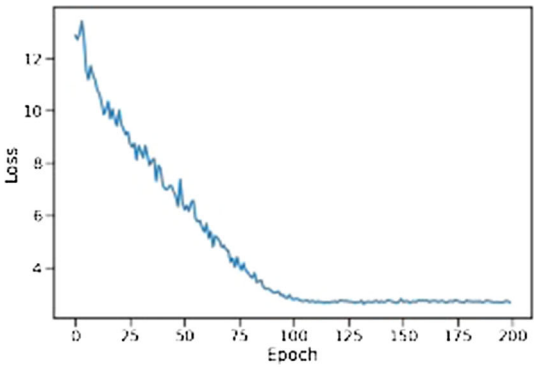
- When the temperature is 0.2 or 0.3, the model achieves the best results. Small or large temperature model achieve inadequate performance.
- The current model shows a 1.3% increase in performance than the previous [12].

As the τ value decreases, the relative penalty distribution becomes more uniform, leading to all the negative samples having the same magnitude of penalties. The relative penalty distribution becomes more concentrated in the prominently equivalent regions as the τ increases. As the temperature drops, the effective penalty interval gets shorter. The values $\tau = 0.07$ and $\tau = 1$ are on the low and high ends, respectively, while the rest are in the middle. The contrastive loss will only affect the nearest one or two samples at extremely low temperatures, severely degrading the performance. To avoid this, the temperatures in this paper are kept inside plausible interval [11] Figure 1.

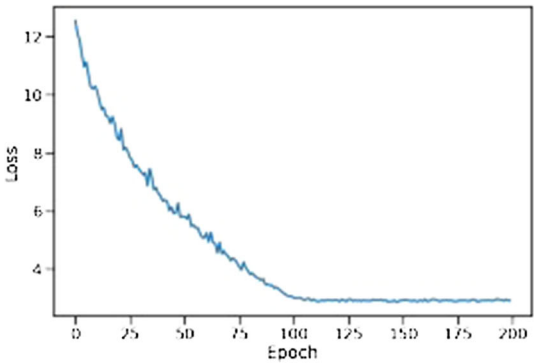
5.4 Qualitative results

If the loss value is extremely minimal, the contrastive loss function will inflict substantial penalties on closest neighbors. Semantically similar instances of data will very likely be distributed with the anchor point. Considering the depictions in T-SNE plots in Fig. 2, we follow that embedding with $\tau = 0.07$ is distributed better and evenly, although the embedding with $\tau = 1$ is more reasonable and locally clustered and globally separated.

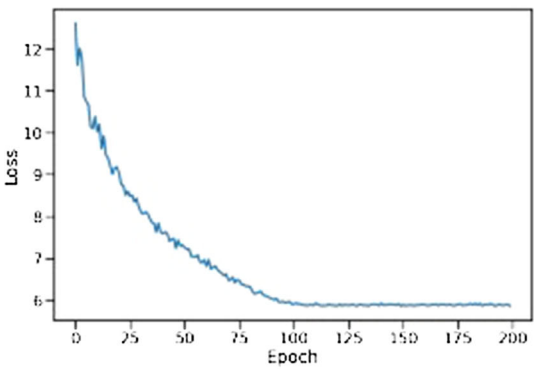
- With the τ decreasing, there is a larger gap from positive samples to other misleading negatives, i.e., more distinguishable positive and negative samples.
- Indeed, as shown in Fig. 2, small temperatures tend to increase the impact of the hard negative samples.
- Results demonstrate that the positive samples
- are more aligned with the increased temperature and that the model tends to develop more invariant features with regard to the different transformations applied to sensor data.



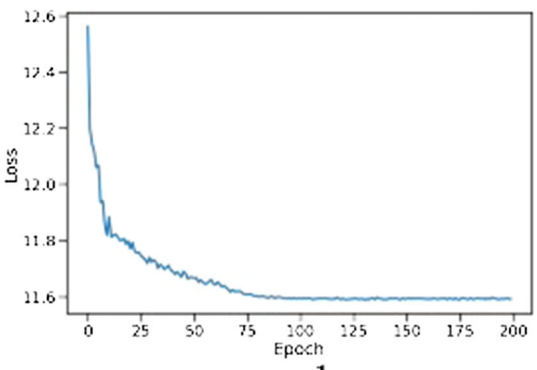
τ=0.07



τ=0.1

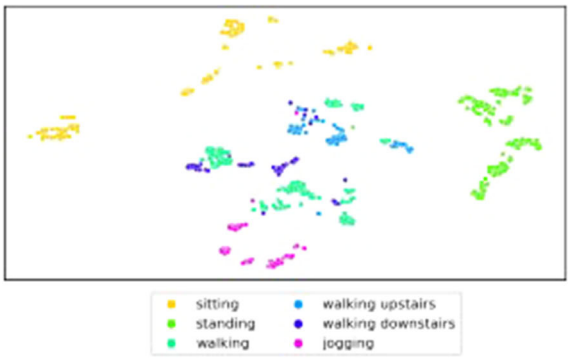


τ=0.2

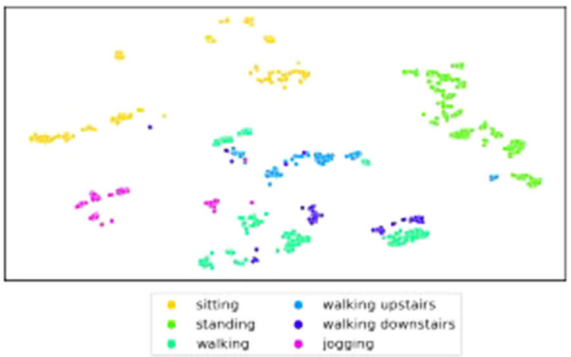


τ=1

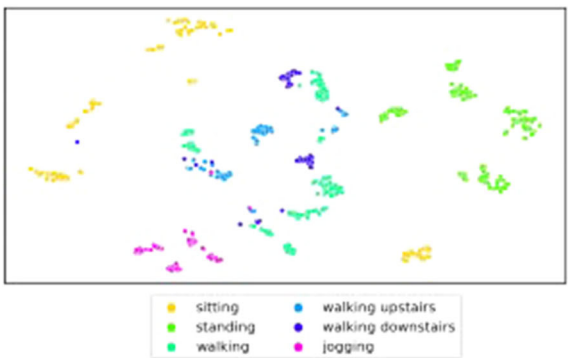
Fig. 1 Loss Variation



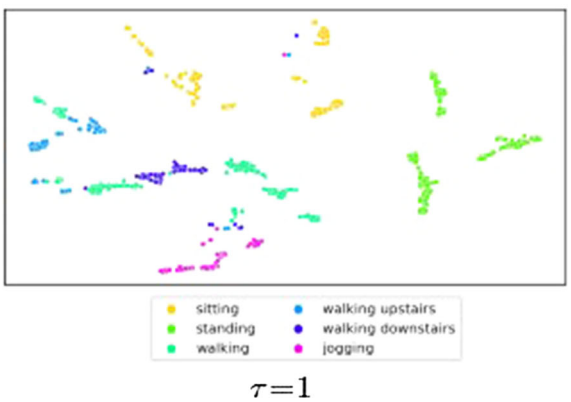
τ=0.07



τ=0.1



τ=0.2



τ=1

Fig. 2 T-SNE [36] Plot

Table 1 Comparison with baseline models

Model	Supervised (only)	Self-supervised	SimCLR (optimized)
Weighted F1	0.922	0.923	**0.955**

Table 2 Quantitative comparison on the MotionSense Validation Dataset

Temperature (τ)	F1 Macro	F1 Micro	F1 Weighted
0.07	0.858	0.870	0.874
0.1	0.900	0.921	0.922
0.2	**0.935**	**0.954**	**0.955**
0.3	**0.931**	**0.951**	**0.951**
0.7	0.894	0.914	0.916
1	0.879	0.905	0.907

5.5 Comparative study with baseline models

In this section, we compare our best model with the most advanced methods. A linear and finally defined evaluation was conducted using the MotionSense dataset to evaluate the impact of using different temperature (τ) variances for SimCLR pre-training. Results are shown in Table 1. F1 scores are taken directly from work already carried out by Tang et al. for supervised and self-supervised models.

6 Conclusion

In this work, we have studied one of the most important tasks in digital health applications i.e., human activity recognition (HAR) and how SimCLR (contrastive learning framework for visual representation learning) can be adapted efficiently to mitigate the difficulty in incorporation of data-oriented methods for digital health due to the limitations of HAR's labelled datasets. We have examined the effect of temperature (τ) changes on contrastive loss in connection with sensor data to improve the feature quality and performance for downstream tasks. SimCLR showed promising results in our evaluation, outperforming both fully supervised and semi-supervised methods. A significant effect on performance can be seen when τ is regularized. This issue will be explored further with the inclusion of additional evaluation dataset and transformations Table 2.

References

1. Slim, S., Atia, A., Elfattah, M., Mostafa, M.S.M.: Survey on human activity recognition based on acceleration data. Intl. J. Adv. Comput. Sci. Appl **10**, 84–98 (2019)
2. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. arXiv preprint arXiv:2006.09882, (2020)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton G.: A simple framework for contrastive learning of visual representations, 2020
4. Grill, J.B., Strub, F., Altch´e, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doesch, C., Avila Pires, B., Guo, Z.D., Azar, M.G. et al. Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733, 2020.
5. Hadsell, R., Chopra, S., LeCun Y.: Dimensionality reduction by learning an invariant mapping. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), volume 2, pages 1735–1742. IEEE, (2006)
6. Li, Y., Hu, P., Liu, Z., Peng, D., Zhou, J.T., Peng, X.: Contrastive clustering. In 2021 AAAI Conference on artificial intelligence (AAAI), (2021)
7. van den Oord, A., Li, Y., Vinyals O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018.
8. Schroff, F., Kalenichenko, D., Facenet J.P.: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, (2015)
9. Gutmann, M., Hyv¨arinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Proceedings of the thirteenth international conference on artificial intelligence and statistics, pages 297–304. JMLR Workshop and Conference Proceedings, (2010)
10. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning, (2021)
11. Wang, F., Liu, H.: Understanding the behaviour of contrastive loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2495–2504, (2021)
12. Tang, C.I., Perez-Pozuelo, I., Spathis, D., Mascolo, C.: Exploring contrastive learning in human activity recognition for healthcare, (2021)
13. Wang, J., Chen, Y., Hao, S., Peng, X., Lisha, Hu.: Deep learning for sensor-based activity recognition: a survey. Pattern Recogn. Lett. **119**, 3–11 (2019)
14. Ramamurthy, S.R., Roy, N.: Recent trends in machine learning for human activity recognition—a survey. Wiley Interdisciplinary Rev: Data Mining Knowled Discovery **8**(4), e1254 (2018)
15. Alrazzak, U., Alhalabi, B.: A survey on human activity recognition using accelerometer sensor. In 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), pages 152–159. IEEE, 2019
16. Li, X., He, Y., Jing, X.: A survey of deep learning-based human activity recognition in radar. Remote Sens **11**(9), 1068 (2019)
17. Aguileta, A.A., Brena, R.F., Mayora, O., Molino-Minero-Re, E., Trejo, L.A.: Multi-sensor fusion for activity recognition—a survey. Sensors **19**(17), 3808 (2019)
18. Lima, W.S., Souto, E., El-Khatib, K., Jalali, R., Gama, J.: Human activity recognition using inertial sensors in a smartphone: an overview. Sensors **19**(14), 3213 (2019)
19. Jobanputra, C., Bavishi, J., Doshi, N.: Human activity recognition: a survey. Procedia Computer Science **155**, 698–703 (2019)
20. Hussain, Z., Sheng, Q.Z., Zhang, W.E.: A review and categorization of techniques on device-free human activity recognition. J Network Comp Appl **167**, 102738 (2020)
21. Yousefi, B., Loo, C.K.: Biologically-inspired computational neural mechanism for human action/activity recognition: A review. Electronics **8**(10), 1169 (2019)
22. Mekruksavanich, S., Jitpattanakul, A.: Exercise activity recognition with surface electromyography sensor using machine learning

approach. In 2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON), pages 75–78. IEEE, (2020)
23. Tripathi, R.K., Jalal, A.S., Agrawal, S.C.: Suspicious human activity recognition: a review. Artif Intell Rev **50**(2), 283–339 (2018)
24. Damaˇsevi ˇcius, R., Maskeliu ¯nas, R., Ven ˇckauskas, A., Wo'zniak, M.: Smartphone user identity verification using gait characteristics. Symmetry **8**(10), 100 (2016)
25. Rault, T., Bouabdallah, A., Challal, Y., Marin, F.: A survey of energy-efficient context recognition systems using wearable sensors for healthcare applications. Pervasive Mobile Comput **37**, 23–44 (2017)
26. Dong, B., Biswas, S.: Wearable networked sensing for human mobility and activity analytics: A systems study. In 2012 Fourth International Conference on Communication Systems and Networks (COMSNETS 2012), pages 1–6. IEEE, (2012)
27. Curone, D., Bertolotti, G.M., Cristiani, A., Secco, E.L., Magenes, G.: A real-time and self-calibrating algorithm based on triaxial accelerometer signals for the detection of human posture and activity. IEEE trans Infor Tech Biomed **14**(4), 1098–1105 (2010)
28. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2014)
29. Wu, Z., Xiong, Y., Yu, S.X., Lin., D.:Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3733–3742, (2018)
30. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670, (2018)
31. Ye, M., Zhang, X., Yuen, P.C., Chang, S.F.: Unsupervised embedding learning via invariant and spreading instance feature. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 6210–6219, (2019)
32. Bachman, P., Hjelm, R.D., Buchwalter W.: Learning representations by maximizing mutual information across views. *arXiv preprint* arXiv:1906.00910, 2019.
33. Henaff, O.: Data-efficient image recognition with contrastive predictive coding. In International Conference on Machine Learning, pages 4182–4192. PMLR, (2020)
34. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23– 28, 2020, Proceedings, Part XI 16, pages 776–794. Springer, (2020)
35. Saeed, A., Ozcelebi, T., Lukkien, J.: Multi-task self-supervised learning for human activity detection. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **3**(2), 1–30 (2019)
36. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J Mach Learn Res **9**(2605), 2579–2605 (2008)
37. Malekzadeh, Md., Clegg, R.G., Cavallaro, A., Haddad, H.: Protecting sensory data against sensitive inferences. In Proceedings of the 1st Workshop on Privacy by Design in Distributed Systems, pages 1–6, 2018