Study on Temperature Variance for SimCLR based Activity Recognition

Pranjal Kumar*

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. In this paper, we propose a robust SimCLR model for human activity recognition with a temperature variance study. In this work, SimCLR, a contrasting learning technique is optimized via regulating the temperature for visual representations, is incorporated for improving the HAR performance in healthcare.

Keywords Contrastive learning · activity recognition

1 Introduction

Precise human activity recognition involves consideration of links between actors, objects and their surroundings, often over long time periods. One reason why video comprehension is so difficult is because it requires an understanding of the interactions between actors, objects and other contexts on the scene. Moreover, these interactions cannot always be seen from a single frame, and therefore require reasoning over long periods of time. As such, some of them only model spatial relationships between actors and objects, but not the evolution of those interactions with time. Alternative approaches model long-range time interactions[1], but do not capture and do not train spatial relations. Although certain methods model spatio-temporary interactions between objects[2, 3], further supervision is required for their explicit representations of objects. Early works in this field included modelling human-object interaction[4, 5], various objects[6], and human actions/scenes context relations[7, 8]. Moreover, human vision has also proven to be context dependent[9].
A major problem in video understanding is recognition of human action and recognition of group activities[10]. The techniques of action and activity recognition have been widely used, for example in the fields of social behavior understanding, sport video analysis and video monitoring. It is important to better understand a video scene with several people and to understand the action and collective activity of all individuals.
Recently, SimCLR was incorporated for healthcare and HAR in particular for the first time[11]. In this paper, we suggest several ways to improve the functionality and efficiency of the Human action recognition using optimizations in contrastive loss[12, 13]. Main contribution of the proposed methodology is summarized below:

- We provide a detailed study for understanding the behaviour of contrastive learning(special emphasis on temperature coefficient) in sensor data context for human activity recognition.
- We improve the SimCLR performance by regulating the temperature coefficient.

2 Related Work

2.1 Action Recognition

In earlier works, hand-crafted attributes for encoding information from motion were used[14, 15]. Advances

Pranjal Kumar
NIT Hamirpur, H.P, India-177005
Tel.: +918637511985
Fax: +91-1972-223834
E-mail: pranjal@nith.ac.in

in deep learning saw first the repurposes of video ”two stream” networks of 2D image-convolutionary neural networks (CNNs)[16, 17], and then the space-time 3D CNN[18–21]. These architectures, however, concentrate on extracting broad features, video-based features and are not suitable for studying fine grain relationships. Graph neural network (GNN), by modelling them as nodes in a directed, undirected graph, explicitly models the interaction between entities[22–24] through a neighbourhood defined in each node. Each feature maps element in each function is a node, and all nodes are fully connected. The self-attention [25] and non-local operators [26] are also considered GNNs. Such models have been outstanding in several processing tasks for the natural language and informatics, inspiring numerous follow-up methods[27–31].

2.2 Human Object Interaction

The objective of Human Object Interaction (HOI) detection is to locate humans and objects and to recognise their interactions. Previous studies[32–37] show promising results of HOI sensing by decoupling it into the detection and classification of objects. In particular, the results of human and object detection first come from an object detector pre-trained, and then a pair of combined proposals for human objects interaction classification. In recent approaches[38, 37, 36], a substitute detection problem was introduced, which would indirectly optimise the HOI detection. Firstly, the proposal of interaction was predefined on the basis of human priors. UnionDet[37], for example, defines the proposal for interactions as a union box for human and object boxes. As an interaction point, the central point from the human to the object is used by PPDM[36].

3 Methodology

3.1 Framework

SimCLR[39] 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 contrasting prediction task.

4 Results & Discussion

4.1 Dataset

MotionSense[40] 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% tri-axial accelerometer.

4.2 Experiments

- *Experiment 1: Temperature variation and loss function*
  In this section we conduct extensive studies on the temperature coefficient, in order to understand the modeling relationship of the proposed network using activity prediction precision as the assessment metric. Fig 1, 2, 3 and 4 shows loss polt results for T= 0.07, 0.1, 0.2, 1 respectively.
- *Experiment 2: Visualisation of results via t-SNE plots after full model evalutaion*
  In this experiment, we analyze the performance of the group activity recognition with different settings. Fig 5, 6, 7 and 8 below shows the output from visualization model for T= 0.07, 0.1, 0.2, 1 respectively.

5 Comparative Study with Baseline Models

In this section, we compare our best models with the most advanced methods. A linear and finally defined evaluation was conducted using the MotionSense dataset to evaluate the impact of using different transformations for SimCLR pre-training. Results are shown in Table 1.

6 Conclusion

In this work, we have adapted to HAR, one of the most relevant tasks for digital health applications, the SimCLR contrasting learning framework from visual representation learning. We have studied the effect of temperature variance on contrastive loss adhering to and thereby improving the performance of HAR.

doi:10.20944/preprints202107.0138.v1

Table 1 Results

Model	Supervised(only)	Self-Supervised	SimCLR(optimised)
Weighted F1	0.922	0.923	0.944

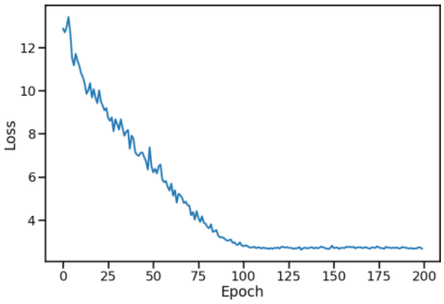

Fig. 1

T=0.0.07

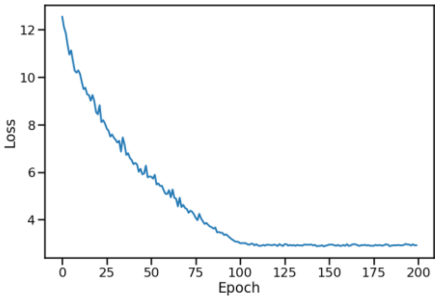

Fig. 2

T=0.1

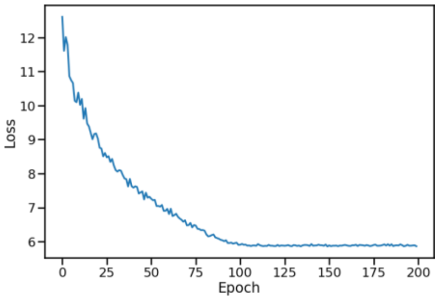

Fig. 3

T=0.20

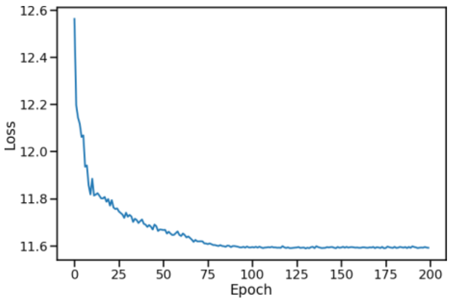

Fig. 4

T=1

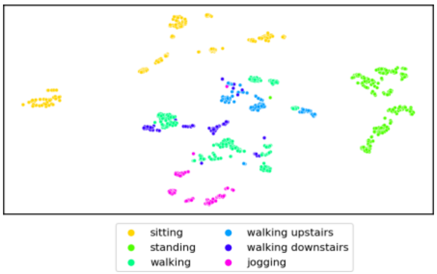

Fig. 5

T=0.07

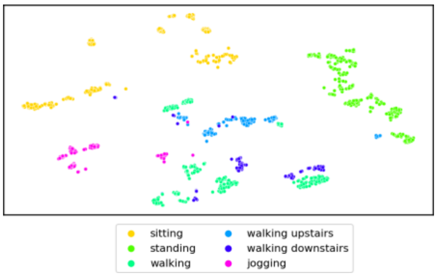

Fig. 6

T=0.1

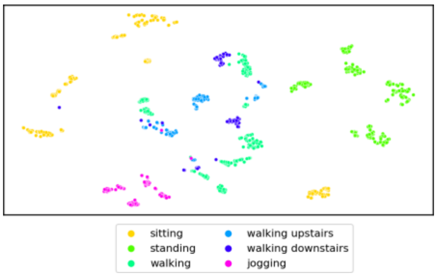

Fig. 7

T=0.2

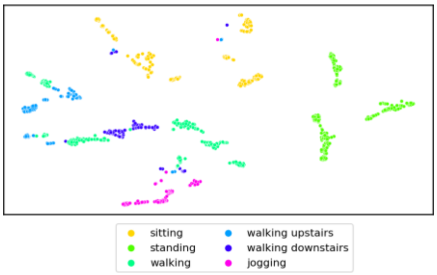

Fig. 8

T=1

References

1. Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019.
2. Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European conference on computer vision (ECCV)*, pages 399–417, 2018.
3. Fabien Baradel, Natalia Neverova, Christian Wolf, Julien Mille, and Greg Mori. Object level visual reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 105–121, 2018.
4. Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE transactions on pattern analysis and machine intelligence*, 31(10):1775–1789, 2009.
5. Bangpeng Yao and Li Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 17–24. IEEE, 2010.
6. Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge Belongie. Objects in context. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.
7. Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2929–2936. IEEE, 2009.
8. Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual action recognition with r* cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1080–1088, 2015.
9. Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in cognitive sciences*, 11(12):520–527, 2007.
10. Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9964–9974, 2019.
11. Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, and Cecilia Mascolo. Exploring contrastive learning in human activity recognition for healthcare, 2021.
12. Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2021.
13. Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss, 2021.
14. Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, 103(1):60–79, 2013.
15. Ivan Laptev. On space-time interest points. *International journal of computer vision*, 64(2-3):107–123, 2005.
16. Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.
17. Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014.
18. Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
19. R Christoph and Feichtenhofer Axel Pinz. Spatiotemporal residual networks for video action recognition. *Advances in Neural Information Processing Systems*, pages 3468–3476, 2016.
20. Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.
21. Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018.
22. Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.
23. Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.
24. Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
25. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.
26. Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.
27. Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.
28. Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. $a\hat{}2$-nets: Double attention networks. *arXiv preprint arXiv:1810.11579*, 2018.
29. Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 433–442, 2019.
30. Felix Wu, Angela Fan, Alexei Baevski, Yann N Dauphin, and Michael Auli. Pay less attention with lightweight and dynamic convolutions. *arXiv preprint arXiv:1901.10430*, 2019.
31. Li Zhang, Dan Xu, Anurag Arnab, and Philip HS Torr. Dynamic graph message passing networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3726–3735, 2020.
32. Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018.
33. Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction

detection. In *European Conference on Computer Vision*, pages 696–712. Springer, 2020.

34. Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367, 2018.
35. Yang Liu, Qingchao Chen, and Andrew Zisserman. Amplifying key cues for human-object-interaction detection. In *European Conference on Computer Vision*, pages 248–265. Springer, 2020.
36. Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–490, 2020.
37. Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *European Conference on Computer Vision*, pages 498–514. Springer, 2020.
38. Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4116–4125, 2020.
39. Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.
40. Mohammad Malekzadeh, Richard G Clegg, Andrea Cavallaro, and Hamed Haddadi. Protecting sensory data against sensitive inferences. In *Proceedings of the 1st Workshop on Privacy by Design in Distributed Systems*, pages 1–6, 2018.