Multimodal Contrastive Learning with Hard Negative Sampling for Human Activity Recognition

Hyeongju Choi[1], Apoorva Beedu[1], Irfan Essa[1,2]
[1]Georgia Insitute of Technology, Atlanta, GA
[2]Google Research, Atlanta, GA
hchoi375,abeedu3,irfan@gatech.edu

Abstract

Human Activity Recognition (HAR) systems have been extensively studied by the vision and ubiquitous computing communities due to their practical applications in daily life, such as smart homes, surveillance, and health monitoring. Typically, this process is supervised in nature and the development of such systems requires access to large quantities of annotated data. However, the higher costs and challenges associated with obtaining good quality annotations have rendered the application of self-supervised methods an attractive option and contrastive learning comprises one such method. However, a major component of successful contrastive learning is the selection of good positive and negative samples. Although positive samples are directly obtainable, sampling good negative samples remain a challenge. As human activities can be recorded by several modalities like camera and IMU sensors, we propose a hard negative sampling method for multimodal HAR with a hard negative sampling loss for skeleton and IMU data pairs. We exploit hard negatives that have different labels from the anchor but are projected nearby in the latent space using an adjustable concentration parameter. Through extensive experiments on two benchmark datasets: UTD-MHAD and MMAct, we demonstrate the robustness of our approach for learning strong feature representation for HAR tasks, and on the limited data setting. We further show that our model outperforms all other state-of-the-art methods for UTD-MHAD dataset, and self-supervised methods for MMAct: Cross session, even when uni-modal data are used during downstream activity recognition.

1. Introduction

As large-scale datasets comprising diverse samples are increasingly helping deploy models in the real world, the need for self-supervised and unsupervised pre-training models is growing to alleviate the data annotation costs and the substantial effort needed in encoding the domain-specific knowledge. One such family of methods, Contrastive Self-Supervised Learning (SSL), has shown great effectiveness in learning strong feature representations in many domains, including computer vision [20, 42, 8, 50, 20], natural language processing (NLP) [16], and sensor domains [25, 27, 19]. Nevertheless, one of the challenges of contrastive SSL is its dependence on the sampling strategy for deriving informative positive and negative pairs, and the need to train in large batches [54, 3]. Successful sampling strategies for both positive and negative pairs have been introduced and contributed to the recent progress in contrastive learning [8, 45, 57, 50]. Most constrastive methods uniformly sample the negative pairs from the data, resulting in false negative samples that contribute to slower learning,

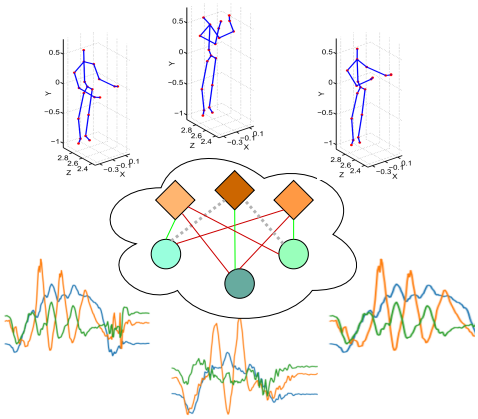

Figure 1. Illustration of negative and positive sampling methods (Best viewed in color). Uniform negative sampling would consider all $< i_n, s_k >$ samples when $n \neq k$ as negative samples, while with hard negative sampling samples that are incorrectly close to anchor(similar shades of color) is selected as negative samples while the samples very distinct(different shades) from the anchors are less preferred (marked in dotted gray).

ICCV 2023 Workshop on PerDream: PERception, Decision making and REAsoning through Multimodal foundational modeling.

or use very large batch sizes that provide enough varies negative samples [42, 20]. The effect of negative samples are more pronounced in a multimodal setting where the model now needs to learn features for two different modalities. The effectiveness of hard negative samples (i.e., instances that are difficult to distinguish from an anchor/positive instance) has not been studied extensively in multimodal Human Activity Recognition (HAR) despite its ability to guide learning to correct its mistake more quickly.

Several works have shown that models that use multimodal data learn stronger feature representations compared to uni-modal data setting [58, 1, 40, 56, 48, 22, 10]. Although HAR has made significant progress in self-supervised methods [51, 28, 19] including learning from multiple devices [24, 13], there has been comparatively less research exploring SSL methods in multimodal settings, and consequently, there has been limited exploration of sampling strategies for negative pairs in multimodal HAR systems. Therefore, this research aims to develop a novel multimodal Human Activity Recognition (HAR) system that overcomes the challenges faced in traditional contrastive SSL methods by leveraging hard negative samples that are close to the anchor and are likely to provide the most meaningful gradient information during training as illustrated in Figure 1. By doing so, the proposed approach aims to be a step towards improving the foundation models for the HAR system by addressing the limitations of traditional contrastive SSL methods.

In summary, our contributions in this paper are as follows:

- We implement the hard negative sampling strategy [45] into a multimodal HAR framework that mitigates false negatives and leverages hard negative samples to boost performance on feature representation learning using IMU signals and skeleton data.
- We perform an in-depth analysis of the effect of the adjustable concentration parameter, β, for the hard negative sampling strategy for multimodal HAR.
- We show the effectiveness of multimodal foundation models by using uni-modal data during during downstream task.
- We perform extensive experiments to evaluate our proposed method against other multimodal HAR frameworks on two publicly available multimodal datasets: UTD-MHAD [6] and MMAct [32].

2. Related Work

Our work is focused on contrastive learning with hard negatives samples for multimodal HAR. In what follows, we divide and summarize the existing literature into three categories: unimodal HAR, multimodal HAR, and contrastive learning for multimodal HAR.

2.1. Unimodal HAR

Unimodal HAR systems, both vision based and sensor based, have been extensively studied by both communities [12, 55, 43, 35]. IMU-based HAR is one of the most widely used unimodal HAR approaches due to its availability on commodity platforms such as smartphones and smartwatches, and its robustness against challenges that vision-based approaches are susceptible to including occlusion, viewpoint, lighting, and background variations [49, 59, 9, 21, 21]. However, IMU signals are generally noisy, and the collection of high-quality data tends to be a tedious and time-consuming endeavor. Compared to other vision-based approaches, skeleton modality has seen a wider application in human activity recognition tasks as they directly provide body structure and pose information, is scale-invariant and robust against other challenges like variations in clothing textures and backgrounds [49, 14, 47, 46, 60]. Although the unimodal-based approaches have demonstrated their effectiveness in HAR, each modality has weaknesses that other modalities can compensate for. For instance, challenges in vision-based approaches such as occlusions can be overcome by using IMU data, and the sensitivity of IMU sensors to body-worn positions [39] can be addressed by using visual modalities such as skeleton data.

2.2. multimodal HAR

The advantages of aggregating information from various data modalities have been studied in computer vision for a wide range of tasks. For instance, [53, 4] proposed a heterogeneous architecture that utilizes RGB and depth data for object pose estimation. In [2, 1, 41, 30, 34], Visual Question Answering (VQA) utilizes RGB images and dialog to accurately answer questions about videos.

The success of multimodal approaches in these fields has helped address the shortcomings in the single modality based HAR approaches. In [44], Rani *et al*. proposed a 2D CNN-based multimodal HAR to perform human activity classification on the hand-crafted features extracted from depth images and skeleton joints. Franco *et al*. [15] proposed a skeleton and RGB-based multimodal approach where the RGB frames were used to capture the temporal evolution of actions. Recent multimodal HAR approaches use attention to effectively fuse features from different modalities to produce representations [22, 23, 29]. Although these methods have shown promising results, these methods use fully labeled data, which in itself is a non-trivial task for IMU data.

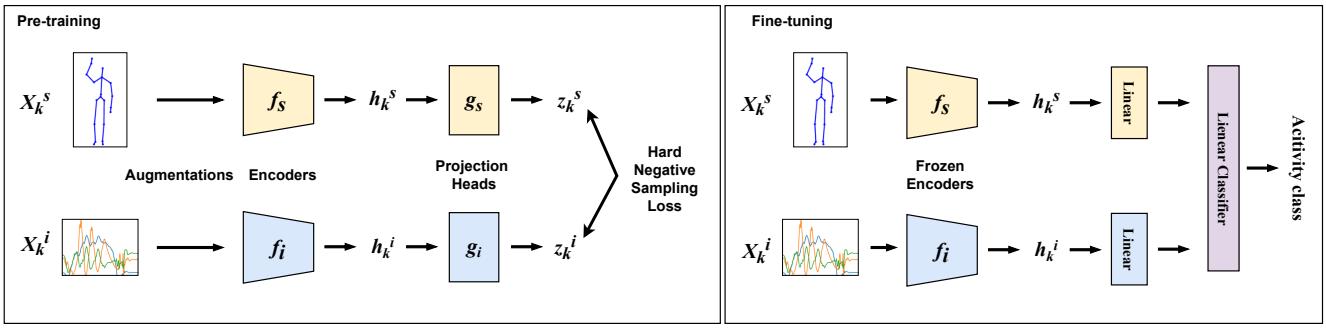

Figure 2. Pre-training (left), and fine-tuning (right) training architecture for our proposed model. During the pre-training stage, the features from the skeleton data are contrasted against IMU features. During the fine-tuning stage, these features are concatenated and trained for the action classification task keeping the encoders frozen.

2.3. Contrastive Learning for HAR

To overcome the challenges with large annotation data, self supervised learning such as contrastive learning [7, 19], or transformer based methods [18] have been widely studied in many fields including HAR for their comparable, or sometimes even better performance than the supervised learning methods. The main advantage of contrastive SSL approaches is that the model can be finetuned using limited labeled data when annotations are expensive to obtain. Since IMU data are hard to annotate and require domain-specific expertise [61], contrastive learning has been explored and applied in various unsupervised settings [51, 28]. There have been a few explorations of contrastive learning in the multimodal setting similar to our own [52, 42, 36, 17, 38, 20]. Contrastive Multiview Coding (CMC) [52] maximizes mutual information between different views of the same scene, particularly between different image channels. Alternatively, in a multimodal setting, Razvan *et al*. [5] proposed a contrastive SSL framework that exploits modality-specific knowledge to mitigate the problem of false negatives. Khaertdinov *et al*. [26], on the other hand, used temporal feature alignment using a dynamic time warping (DTW) in a latent space to align features in a temporal dimension. ImageBind [17] use CLIP [42] like architecture, but extend it to siz different modalities including images, text, audio, depth, thermal and IMU data. However, the effect of hard negative samples for multimodal HAR, to the best of our knowledge, hasn't been explored extensively. Inspired by [45] and [52], we introduce a hard negative sampling method with an adjustable hardness for multimodal HAR framework. This approach allows the contrastive SSL framework to exploit hard negative samples for faster training and generalize better than using randomly sampled negatives from the batch.

3. Methodology

To improve the selection of negative samples in contrastive SSL for multimodal HAR, we implement a hard negative sampling method with an adjustable hardness using two modalities: skeleton and IMU data. An overview of the network is shown in Figure 2. We discuss our proposed work and hard negative sampling for contrastive learning in the sections below.

3.1. Contrastive Learning for multimodal HAR

Contrastive learning is a method that learns to distinguish between similar and dissimilar samples. To this end, as a pre-training stage shown in Figure 2(left) each sample $\{X_k^s, X_k^i\}$ representing input data for skeleton and IMU data respectively, undergoes modality-specific augmentations and is passed through modality specific encoders, f_s, f_i. The resulting representations are then passed through projection heads, g_s, g_i, to generate projections (z_k^s, z_k^i), which comprises the positive pair. The negative pairs are all the other inter-modal combinations of projections from different input instances, $(z_k^s, z_{n\neq k}^i)$ in the batch. Contrastive loss for the inputs can be calculated as:

$$\mathcal{L} = \sum_{k=1}^{N} (l_k^{s\to i} + l_k^{i\to s}) \tag{1}$$

$$where,\ l_k^{i\to s} = -\log \frac{\frac{exp(s(z_k^s, z_k^i))}{\tau}}{\sum_{n=1}^{N} \frac{exp(s(z_k^s, z_n^i))}{\tau}}$$

and $s(z_k^s, z_k^i)$ is a cosine similarity function between z_k^s and z_k^i, and τ is a temperature parameter.

In the fine-tuning stage shown in Figure 2(right), the frozen modal-specific encoders are used to generate representations h_k^s and h_k^i. These representations are then passed through a linear layer to map them into the same size, concatenated, and passed through a simple linear classification layer for the activity class prediction.

3.2. Hard Negative Sampling for HAR

Despite the success of contrastive learning in many fields, a challenge that remains is the selection of good negative samples as it has a significant impact on performance.

Inspired by [45], we introduce a hard negative sampling method for HAR to sample true negatives that have different labels from the anchor and are projected near the anchor instead of using all inter-modality pairs as negative samples as discussed in Section 3.1. In equations to follow, superscript $^+$ indicates positive samples, and $^-$ indicates negative samples, and $h(x)$ is the class label for the given input x, and p is the distribution. The hard negative sampling method samples negatives from the distribution defined as:

$$\begin{gathered} q_\beta^- := q_\beta(x^- | h(x) \neq h(x^-)), \\ \text{where } q_\beta(x^-) \propto \exp^{\beta f(x)^\top f(x^-)} \cdot p(x^-), \\ h : input(x) \rightarrow labels(c) \end{gathered} \tag{2}$$

q_B^- guarantees that the anchor and the negative sample correspond to different latent classes and the concentration parameter β term up-weights the negative samples that are similar to the anchor x. The hardness term β is a hyperparameter that can be adjusted to achieve a balance between improved learning from hard negatives and the potential harm from the approximate correction of false negatives.

The hard negative sampling objective $l_{HNL}(f)$ can be empirically obtained by adopting PU-learning (Positive unlabelled) and importance sampling to the standard contrastive learning objective, defined as:

$$\mathbb{E}_{\substack{x\sim p, \\ x^+\sim p_x^+ \\ x_{i=1:N}^-\sim p^N}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + \delta(x, x_{i=1:N}^-, x^+)} \right] \tag{3}$$

where $\delta(x, x_{i=1:N}^-, x^+)$ is

$$max\{\frac{1}{\tau^-}(\frac{\sum_{i=1}^N e^{(\beta+1)f(x)^T f(x_i^-)}}{\frac{1}{N}\sum_{i=1}^N e^{\beta f(x)^T f(x_i^-)}} - \tau^+ N e^{f(x)^T f(x^+)}),\ Ne^{\frac{-1}{t}}\}$$

where τ^+ is the probability of anchor class, τ^- is the probability of observing a different class, N is the number of negative samples and *t* is the temperature.

We introduce hard negative sampling loss for HAR by simply replacing the standard contrastive loss in equation 1 with the hard negative sampling loss in equation 3. Furthermore, for additional baseline comparisons, we implement debiased contrastive loss [11] adapted to CMC framework: CMC-Debiased, which claims to address the issue of sampling same-label datapoints by setting $\beta = 0$ and $\tau^+ > 0$ in equation 3. In addition, we introduce hard negative sampling in SimCLR for unimodal HAR to examine the effectiveness of leveraging hard negative samples in unimodal settings.

3.3. Encoders

For the inertial encoder, we implemented CSSHAR framework [28], a transformer-like encoder, consisting of three 1D-CNN layers with batch normalization and ReLU activation followed by positional encoding and a transformer encoder with multiple self-attention blocks to adaptively focus on the most important parts of the sensor signals [37]. For the skeleton encoder, we implemented a hierarchical co-occurrence network introduced in [33]. The network takes the skeleton keypoints as input and splits the data into two unique inputs for the network: skeleton sequence and skeleton motion, i.e., the temporal difference between two consecutive frames.

Method	Dataset	Modality	Batch Size	lr	Temperature	τ^+	β
SimCLR	UTD-MHAD	Inertial	128	0.001	0.5	-	-
SimCLR-HNL		Inertial	128	0.001	0.5	0.037	0.5
SimCLR		Skeleton	64	0.005	0.5	-	-
SimCLR-HNL		Skeleton	64	0.005	0.5	0.037	0.25
SimCLR	MMAct: Cross Subject	Inertial	64	0.001	0.4	-	-
SimCLR-HNL		Inertial	64	0.001	0.4	0.027	0.6
SimCLR		Skeleton	128	0.005	0.4	-	-
SimCLR-HNL		Skeleton	128	0.005	0.4	0.027	0.5
SimCLR	MMAct: Cross Session	Inertial	64	0.001	0.4	-	-
SimCLR-HNL		Inertial	64	0.001	0.4	0.027	0.5
SimCLR		Skeleton	128	0.005	0.5	-	-
SimCLR-HNL		Skeleton	128	0.005	0.5	0.027	0.5
CMC	UTD-MHAD	Mulitmodal	64	0.001	0.1	-	-
Ours(With HNL)				0.001	0.1	0.037	1.0
CMC	MMAct: Cross Subject	multimodal	128	0.001	0.5	-	-
Ours(With HNL)				0.001	0.5	0.027	1.5
CMC	MMAct: Cross Session	multimodal	128	0.001	0.5	-	-
Ours(With HNL)				0.001	0.5	0.027	0.5

Table 1. Hyperparameters for unimodal and multimodal pre-training

Pre-training As part of the pre-training, we apply a set of random modality-specific augmentations as suggested in [8] to enhance the quality of learned embeddings. The inertial augmentations include {jittering, scaling, permutation, channel shuffle} for UTD-MHAD and {scaling and rotation} for MMAct. For skeleton data, augmentations include {jittering, scaling, rotation, shearing, and resized crops}.

Fine-tuning In the finetuning stage, the pre-trained encoders are frozen for both unimodal and multimodal settings. For unimodal fine-tuning, features from the frozen encoder are flattened and passed through a linear classifier that maps the features into the number of activity classes. For multimodal fine-tuning, features from inertial and skeleton encoders are flattened and passed through a modality-specific layer to map into a feature vector of size 256 followed by batch normalization and ReLU, then the two feature vectors are concatenated and passed through a linear classifier.

4. Experiments and Results

In order to evaluate the effectiveness of the proposed approach, we conducted a set of experiments and ablations. In what follows, we discuss the setup of our training protocol and discuss the results and effects of various hyperparameters used in our training.

Approach	Modality	UTD-MHAD Accuracy	MMAct: Cross Subject F-1	MMAct: Cross Subject Accuracy	MMAct: Cross Session F-1	MMAct: Cross Session Accuracy
Supervised	Inertial	74.65	62.16	64.24	80.18	79.17
SimCLR	Inertial	66.74	52.73	55.29	70.59	73.34
SimCLR-HNL	Inertial	71.55	54.59	57.71	71.75	74.03
Supervised	Skeleton	93.10	79.76	80.65	84.36	84.18
SimCLR	Skeleton	95.50	74.08	75.19	79.04	81.22
SimCLR-HNL	Skeleton	95.97	75.15	76.32	80.06	81.90
Supervised	multimodal	94.96	81.37	83.23	89.04	92.04
CMC-TFA* Supervised [26]	multimodal	-	**84.05**	-	-	-
CMC	multimodal	95.35	80.78	83.29	88.91	91.84
CMC-CMKM [5]	multimodal	94.96	80.72	82.88	-	-
CMC-Debiased [11]	multimodal	95.12	80.80	83.03	87.93	91.17
CMC-TFA* [26]	multimodal	-	83.36	-	-	-
Ours	multimodal	**96.20**	81.64	**83.61**	**89.16**	**92.06**

Table 2. Activity classification results using multimodality during pre-training and fine-tuning. The best results are in bold and the 2nd best results are underlined. * indicates that the results are reported as is from the paper, and were not reproduced.

4.1. Setup

We evaluate the performance of our proposed approaches on two benchmark multimodal datasets: UTD-MHAD [6] and MMAct [32]. For both these datasets, we use the IMU and skeleton data for training. Consistent with the UTD-MHAD protocol, we employed odd-numbered subjects for training and even-numbered subjects for testing purposes across all models. For the cross-subject evaluation in MMAct, we employed samples from the first 16 subjects for training and the remaining subjects for testing. For cross-session, we selected samples from the top 80% of sessions, arranged in ascending order based on session ID, for each subject. We report the test accuracies for UTD-MHAD, and test accuracies and F1 scores for MMAct dataset.

For both, unimodal and multimodal training, the model is trained for 150 epochs using Adam optimizer [31] and a scheduler that reduces the learning rate by half when a metric has stopped improving for more than 20 epochs. In the fine-tuning stage, both unimodal and multimodal settings are trained for 100 epochs with the Adam optimizer. In addition, for fair comparison with baselines, we implemented and retrained SimCLR [8], a unimodal SSL framework, for unimodal encoders and also adapted CMC [52] for IMU and Skeleton modalities. We performed 3 runs for every method, and report the average scores for all metrics. More details on the hyperparameters for the training can be found in Table 1.

4.2. Feature Representation Learning for HAR

In this experiment, we compared our approach to single-modality frameworks, and other state-of-the-art multi-modality frameworks. Our primary focus in this paper is to explore the effectiveness of CMC with hard negatives for multimodal HAR, but we also investigated the effectiveness of leveraging hard negative samples in unimodal HAR by implementing the hard negative sampling loss in the SimCLR framework (i.e., SimCLR-HNL). Compared to SimCLR, SimCLR-HNL achieved better results in all unimodal settings for both datasets, with a performance boost ranging from 0.5% to 4.8%. For the multimodal setting, we compare the results of the supervised model using the same encoders as described in Section 3.3, our own implementation of CMC-CMKM [5] using the same hyperparameters as specified in the paper, and our proposed model. As shown in Table 2, our proposed method outperformed or performed comparatively to all multimodal approaches, including supervised, in every test setting with a performance boost ranging from 0.3% to 0.9% compared to CMC. Overall, we also see that the multimodal HAR models performed better than the unimodal HAR models. While CMC-TFA in the supervised and self-supervised setting outperform our method, it is worth noting that their model is a stronger feature extractor as can be seen by comparing the supervised methods. We believe that using hard negative sampling on top of CMC-TFA would improve the accuracy even further.

Approach	Modality	UTD-MHAD Accuracy	MMAct: Cross Subject F-1	MMAct: Cross Subject Accuracy	MMAct: Cross Session F-1	MMAct: Cross Session Accuracy
Supervised	Inertial	74.65	**62.16**	**64.24**	**80.18**	**79.17**
SimCLR	Inertial	66.74	52.73	55.29	70.59	73.34
SimCLR-HNL	Inertial	71.55	54.59	57.71	71.75	74.03
CMC	Multi(Pre)	69.77	50.46	52.62	71.53	67.79
CMC-CMKM [5]	Multi(Pre)	74.42	50.54	53.67	69.14	67.77
CMC-Debiased [11]	Multi(Pre)	70.70	49.79	52.08	68.68	71.95
Ours	Multi(Pre)	**75.73**	54.18	55.74	71.85	74.23

Table 3. Activity classification results using multimodality during pre-training, and using IMU data only during fine-tuning. The best results are in bold and the 2nd best results are underlined.

In Table 3, we compare the performances between SSL using uni-modal data, and multimodal data, and our method of using hard negative sampling with multimodal data. However, we fine-tune the model with only IMU data, to show the model's effectiveness in real-world wearable applications where having skeleton or video data is unfeasible. We see that our method, which was trained with hard negative sampling and multimodal data outperforms all other methods including supervised methods for UTD-MHAD dataset, while surpassing all other self-supervised method for MMAct: Cross session and comparably with SimCLR with hard negative sampling for MMAct: Cross Subject. This shows that multimodal pre-traning with hard negative sampling is an effective traning strategy for foundational models for human activity recognition.

4.3. Effect of β in CMC

For the proposed approach of hard negative sampling, the concentration parameter β determines the hardness of the negative samples and is treated as a hyperparameter. Therefore, it's crucial to tune β adequately to balance improved learning from hard negatives and the potential nega-

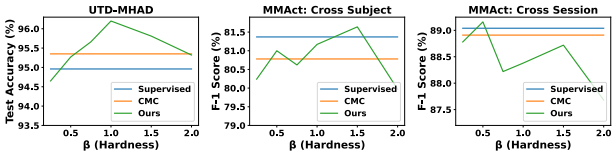

Figure 3. The effect of concentration parameter β ranging from 0.25 to 2.0.

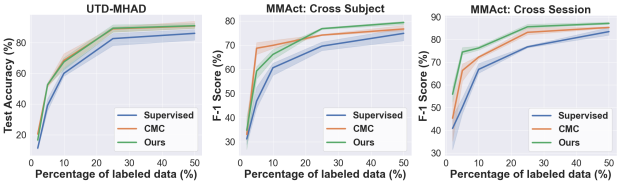

Figure 4. Performance of different multimodal models with 95% CI for semi-supervised learning setting on UTD-MHAD and cross-subject and cross-session setting for MMAct.

tive impact of correcting false negatives. We adjust β while keeping all other hyperparameters the same, using the settings specified in Table 1 and report in Figure 3. The results indicate that different datasets favor different values of β, emphasizing the importance of hyperparameter tuning to achieve optimal results even for the same dataset with different protocols.

4.4. Fine-tuning on limited annotated data

We test our framework on the targeted setting where limited annotations are available for fine-tuning and a large unannotated dataset during pre-training. For each dataset, we use 100% of unannotated samples during pre-training and limit the annotated data to 2%, 5%, 10%, 25%, and 50% for both datasets during fine-tuning. The results are reported in Figure 4. Across all experiments, both CMC and our framework consistently outperform the supervised model, with their performance increasing as the percentage of annotated samples decreases. The greatest performance boost was observed when using 5% of the available annotated samples, with a maximum boost of up to 24%. When comparing CMC and our framework, we observe no significant difference in performance for a relatively smaller UTD-MHAD dataset. However, for the MMAct with a cross-session protocol, our method outperforms CMC for all percentages of limited annotated labels, with a performance boost up to 10%. For MMAct: cross-subject protocol, our method outperforms CMC when more than 10% of annotated labels were available, with a performance boost of up to 2.5%. Based on our findings, we conclude that our method is more effective than CMC in semi-supervised learning with limited labels.

5. Conclusion

In this paper, we explore the use of contrastive learning with hard negative sampling for multimodal HAR using inertial and skeleton data. Our goal is to mitigate the problem of false negative samples and leverage hard negative samples in multimodal HAR, which we achieve by implementing a hard negative sampling loss derived from the hardness-biased objective. Through a series of experiments, our results demonstrated that our model outperforms both supervised multimodal HAR frameworks and CMC-based multimodal HAR frameworks in various experimental settings, including limited annotated setting by learning stronger representations. We also explore the effect of the concentration term β and emphasize the importance of proper tuning for optimizing the model's performance.

References

[1] Huda Alamri, Anthony Bilic, Michael Hu, Apoorva Beedu, and Irfan Essa. End-to-end multimodal representation learning for video dialog. *arXiv preprint arXiv:2210.14512*, 2022.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[3] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023.

[4] Apoorva Beedu, Huda Alamri, and Irfan Essa. Video based object 6d pose estimation using transformers. *arXiv preprint arXiv:2210.13540*, 2022.

[5] Razvan Brinzea, Bulat Khaertdinov, and Stylianos Asteriadis. Contrastive learning with cross-modal knowledge mining for multimodal human activity recognition. *arXiv preprint arXiv:2205.10071*, 2022.

[6] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International conference on image processing (ICIP)*, pages 168–172. IEEE, 2015.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Everest Hinton. A simple framework for contrastive learning of visual representations. 2020.

[9] Yuqing Chen and Yang Xue. A deep learning approach to human activity recognition based on single accelerometer. In *2015 IEEE international conference on systems, man, and cybernetics*, pages 1488–1492. IEEE, 2015.

[10] Hyeongju Choi, Apoorva Beedu, Harish Haresamudram, and Irfan Essa. Multi-stage based feature fusion of multi-

modal data for human activity recognition. *arXiv preprint arXiv:2211.04331*, 2022.

[11] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020.

[12] Federico Cruciani, Anastasios Vafeiadis, Chris Nugent, Ian Cleland, Paul McCullagh, Konstantinos Votis, Dimitrios Giakoumis, Dimitrios Tzovaras, Liming Chen, and Raouf Hamzaoui. Feature learning for human activity recognition using convolutional neural networks. *CCF Transactions on Pervasive Computing and Interaction*, 2(1):18–32, 2020.

[13] Shohreh Deldari, Hao Xue, Aaqib Saeed, Daniel V Smith, and Flora D Salim. Cocoa: Cross modality contrastive learning for sensor data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(3):1–28, 2022.

[14] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015.

[15] Annalisa Franco, Antonio Magnani, and Dario Maio. A multimodal approach for human activity recognition based on skeleton and rgb data. *Pattern Recognition Letters*, 131:293–299, 2020.

[16] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings, 2022.

[17] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.

[18] Harish Haresamudram, Apoorva Beedu, Varun Agrawal, Patrick L Grady, Irfan Essa, Judy Hoffman, and Thomas Plötz. Masked reconstruction based self-supervision for human activity recognition. In *Proceedings of the 2020 ACM International Symposium on Wearable Computers*, pages 45–49, 2020.

[19] Harish Haresamudram, Irfan Essa, and Thomas Plötz. Contrastive predictive coding for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(2):1–26, 2021.

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[21] Masaya Inoue, Sozo Inoue, and Takeshi Nishida. Deep recurrent neural network for mobile human activity recognition with high throughput. *Artificial Life and Robotics*, 23(2):173–185, 2018.

[22] Md Mofijul Islam and Tariq Iqbal. Hamlet: A hierarchical multimodal attention-based human activity recognition algorithm. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10285–10292. IEEE, 2020.

[23] Md Mofijul Islam and Tariq Iqbal. Mumu: Cooperative multitask learning-based guided multimodal fusion,". AAAI, 2022.

[24] Yash Jain, Chi Ian Tang, Chulhong Min, Fahim Kawsar, and Akhil Mathur. Collossl: Collaborative self-supervised learning for human activity recognition, 2022.

[25] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1), 2021.

[26] Bulat Khaertdinov and Stylianos Asteriadis. Temporal feature alignment in contrastive self-supervised learning for human activity recognition. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2022.

[27] Bulat Khaertdinov, Esam Ghaleb, and Stylianos Asteriadis. Contrastive self-supervised learning for sensor-based human activity recognition. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8, 2021.

[28] Bulat Khaertdinov, Esam Ghaleb, and Stylianos Asteriadis. Contrastive self-supervised learning for sensor-based human activity recognition. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021.

[29] Zanobya N Khan and Jamil Ahmad. Attention induced multi-head convolutional neural network for human activity recognition. *Applied Soft Computing*, 110:107671, 2021.

[30] Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. Mmbert: Multimodal bert pretraining for improved medical vqa, 2021.

[31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[32] Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. Mmact: A large-scale dataset for cross modal human action understanding. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8657–8666, 2019.

[33] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *arXiv preprint arXiv:1804.06055*, 2018.

[34] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language, 2019.

[35] Jun Liu, Amir Shahroudy, Dong Xu, Alex C Kot, and Gang Wang. Skeleton-based action recognition using spatio-temporal lstm network with trust gates. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):3007–3021, 2017.

[36] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022.

[37] Saif Mahmud, M Tonmoy, Kishor Kumar Bhaumik, AK Mahbubur Rahman, M Ashraful Amin, Mohammad Shoyaib, Muhammad Asif Hossain Khan, and Amin Ahsan Ali. Human activity recognition from wearable sensor data using self-attention. *arXiv preprint arXiv:2003.09018*, 2020.

[38] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Alireza Dirafzoon, Aparajita Saraf, Amy Bearman, and Babak Damavandi. Imu2clip: Multimodal contrastive learning for imu motion sensors from egocentric videos and text. *arXiv preprint arXiv:2210.14395*, 2022.

[39] Subhas Chandra Mukhopadhyay. Wearable sensors for human activity monitoring: A review. *IEEE sensors journal*, 15(3):1321–1330, 2014.

[40] Ryumei Nakada, Halil Ibrahim Gulluk, Zhun Deng, Wenlong Ji, James Zou, and Linjun Zhang. Understanding multimodal contrastive learning and incorporating unpaired data, 2023.

[41] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition, 2022.

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[43] Hossein Rahmani and Ajmal Mian. 3d action recognition from novel viewpoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1506–1515, 2016.

[44] S Sandhya Rani, G Apparao Naidu, and V Usha Shree. Kinematic joint descriptor and depth motion descriptor with convolutional neural networks for human action recognition. *Materials Today: Proceedings*, 37:3164–3173, 2021.

[45] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

[46] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7912–7921, 2019.

[47] Tae Soo Kim and Austin Reiter. Interpretable 3d human action analysis with temporal convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 20–28, 2017.

[48] Sandya Subramanian, Emery N Brown, and Riccardo Barbieri. Multimodal vs unimodal estimation of sympathetic-driven arousal states. In *2020 Computing in Cardiology*, pages 1–4. IEEE, 2020.

[49] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *arXiv preprint arXiv:2012.11866*, 2020.

[50] Afrina Tabassum, Muntasir Wahed, Hoda Eldardiry, and Ismini Lourentzou. Hard negative sampling strategies for contrastive representation learning. *arXiv preprint arXiv:2206.01197*, 2022.

[51] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, and Cecilia Mascolo. Exploring contrastive learning in human activity recognition for healthcare. *arXiv preprint arXiv:2011.11542*, 2020.

[52] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020.

[53] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. 2019.

[54] Lilian Weng. Contrastive representation learning. *lilianweng.github.io*, May 2021.

[55] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018.

[56] Santosh Kumar Yadav, Kamlesh Tiwari, Hari Mohan Pandey, and Shaik Ali Akbar. A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions. *Knowledge-Based Systems*, 223:106970, 2021.

[57] Zhen Yang, Tinglin Huang, Ming Ding, Yuxiao Dong, Rex Ying, Yukuo Cen, Yangliao Geng, and Jie Tang. Batchsampler: Sampling mini-batches for contrastive learning in vision, language, and graphs. *arXiv preprint arXiv:2306.03355*, 2023.

[58] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning, 2021.

[59] Ming Zeng, Le T Nguyen, Bo Yu, Ole J Mengshoel, Jiang Zhu, Pang Wu, and Joy Zhang. Convolutional neural networks for human activity recognition using mobile sensors. In *6th international conference on mobile computing, applications and services*, pages 197–205. IEEE, 2014.

[60] Xikun Zhang, Chang Xu, Xinmei Tian, and Dacheng Tao. Graph edge convolutional neural networks for skeleton-based action recognition. *IEEE transactions on neural networks and learning systems*, 31(8):3047–3060, 2019.

[61] Liyue Zhao, Gita Sukthankar, and Rahul Sukthankar. Incremental relabeling for active learning with noisy crowdsourced annotations. pages 728–733, 10 2011.