Semester Project: Self-Suprervised Skeleton-based Action Recognition

Student
Ruihang Jiang
ruihang.jiang@epfl.ch

Supervisor
Mohamed Abdelfattah
mohamed.abdelfattah@epfl.ch

Abstract

In this research project, we advocate a self-supervised learning model to extract the 3D human skeleton-based representation features. The primary objective of the model is to train a network which could be utilized in various action related downstream tasks. The network has two pipeline: student and teacher network. The student line serves as the contextualized feature extraction while the teacher one is the target representation. For training and evaluation, we employ L2 loss to assess the performance of our models. We find that our model has advantages on the fully-supervised learning model and the MAMP [19] in action classification task. And the spatial information leads the accuracy comparing with the temporal transformer.

1. Introduction

Human skeleton action recognition has long been a central topic in machine intelligence. The advancements in depth sensors and pose estimation [3,4,21] techniques have significantly stimulated the development of skeleton-based 3D human action recognition, enabling applications such as human-robot interaction, video surveillance, and virtual reality. Despite the computational efficiency, background robustness, and privacy preservation advantages, existing fully-supervised skeleton-based action recognition approaches—utilizing Convolutional Neural Networks (CNNs) [2,17], Recurrent Neural Networks (RNNS) [5,11] and Graph Convolutional Networks (GCNs) [12,20]—rely heavily on annotated training sequences to achieve promising performance. However, manual annotation of data is labor-intensive and time-consuming. Furthermore, the limited amount of training data often leads to overfitting issues in supervised learning models, especially for transformers, which have weak inductive biases and high model capacity. These challenges motivated us to explore a self-supervised feature representation learning model for 3D human skeleton data, which could be applied to various action-related downstream tasks.

The main contributions of this project are:

- We propose a self-supervised learning framework based on student-teacher distillation learning. The high-level features from the teacher encoder are used as target contextualized representations to guide the student network in learning the contextual skeleton data.
- We conduct experiments on three versions of our model and compare them with two baselines: one trained using supervised learning from scratch and the other using the MAMP [19] model. Our proposed model, featuring a spatial transformer, outperforms the baseline models in 3D action classification downstream tasks.

2. Related Work

2.1. Supervised 3D Action Recognition

How to model human dynamic body actions for fully-supervised learning has been a long-standing research topic. In early works, Recurrent Neural Networks (RNNs) were favored for dynamic human action recognition due to their promising sequential processing capabilities. Notable examples include the hierarchical RNN model proposed in [16] and the 2D Spatio-Temporal Long Short-Term Memory (LSTM) model in [6, 7]. As Convolutional Neural Networks (CNNs) [1, 8] achieved success in the field of image processing, methods utilizing CNNs were also explored by treating skeleton data as pseudo-images [2, 18]. Additionally, considering the relational sequences among joints and frames, Spatial-Temporal Graph Convolutional Networks (ST-GCNs) [13] introduced Graph Neural Networks (GCNs) for skeleton learning. The skeleton topology is effectively modeled by the designed convolution kernels.

However, under limited training skeleton data, transformers, which have weak inductive biases, cannot be trained effectively. In our model, we demonstrate that pretraining can address the issue of limited data. The pretrained model has the ability to extract high-dimensional skeleton features from raw skeleton coordinates and can be adapted to various downstream tasks through fine-tuning.

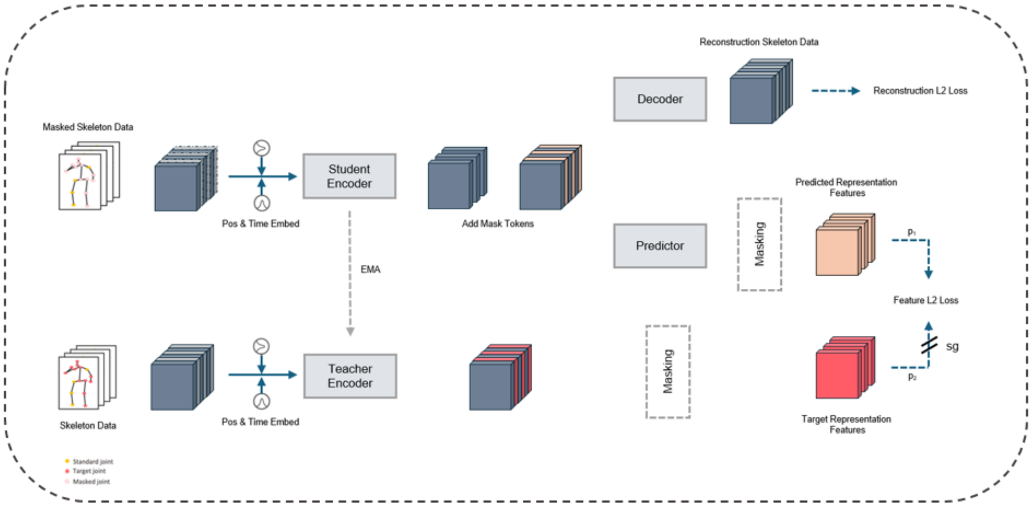

Figure 1. The overall pipeline of proposed model: there are two sections in the model. The student network serves as the contextualized skeleton representation and the teacher network is the target feature.

2.2. Self-supervised 3D Action Recognition

Self-supervised learning aims to extract domain priors from unlabeled data. In 3D human action recognition, pretext tasks have been utilized to facilitate the application of downstream tasks. LongT GAN [10] and P&C [9] aim to learn action features through autoencoder-based sequence reconstruction. Colorization [14] treats point clouds as representations of skeleton sequences, learning the representation of each joint based on its spatial and temporal orders. SkeletonMAE [reference] introduces the concept of Masked Autoencoders (MAE) [15] into 3D action representation learning based on transformers. The MAMP [19] framework, designed for explicit contextual motion modeling, introduces a new masking strategy according to motion frequencies.

In this study, we propose that utilizing high-dimensional contextualized representations as prediction targets and instructing context learning in the student network can lead to more effective 3D skeleton feature learning. This approach offers advantages in directly predicting raw joint coordinates or human motions in downstream classification tasks.

3. Method

In this section, we first provide a brief introduction to the model structure, followed by a detailed exploration of each component. The overall framework of our model is shown in Fig. 1.

3.1. Student Network

For the student network, it takes skeleton data $S \in R^{T_s \times V \times C_s}$ as input. The input data is randomly cropped from the raw skeleton sequence and resized to a fixed frame length T_s. V represents the number of joints and $C_s$denotes the channel of the coordinate features. The skeleton data then undergoes random masking, which randomly drops part of the joints and time frames. Similar to most transformers, the input joints are linearly mapped into a joint embedding $E \in R^{T_e \times V \times C_e}$. The remaining features are processed by either the spatial or temporal transformer. The encoder learns the representation features from the masked skeleton data. After adding the mask tokens to compensate for the dropped spatial and temporal data due to the masking strategy, the features pass through the predictor and decoder separately. For the predictor path, the output feature $R^{T_s \times V \times C_f}$ is considered as the learned high-dimensional predicted representation features. For the decoder path, the structure follows an encoder-decoder network, and the output is the reconstructed skeleton data, which ensures that the student encoder learns useful representation features.

3.2. Teacher Network

For the teacher network, it takes the same input $S \in R^{T_s \times V \times C_s}$ but uses the whole skeleton data without masking. The output from the spatial or temporal encoder $R^{T_s \times V \times C_f}$ serves as the target representation features, which will guide the learning of the student encoder.

3.3. Target Prediction

The outputs from the student predictor and the teacher encoder are utilized to monitor the model pre-training stage. We adopt L2 loss as the learning objective, calculating the loss only for the masked positions:

$$L_{st} = \frac{1}{M} \sum_{i \in M} ||Y_i - \hat{Y}_i||_2^2 \tag{1}$$

Model		Top-1 Accuracy	Top-5 Accuracy	Runing Time
Supervise Learning	Vanilla Transformer	58.8	89.3	14.5 min / Epoch
MAMP [19]		80.8	**95.0**	1.45 min / Epoch
Spatial Encoder		54.6	81.1	5 min / Epoch
Temporal Encoder		59.3	84.7	3.6 min / Epoch
Spatial Encoder	Patches & Masking Token	68.2	90.0	**1** min / Epoch
Temporal Encoder	Patches & Masking Token	65.8	88.8	**0.85** min / Epoch
Spatial Encoder	Patches & Masking Token + Decoder	**82.3**	**95.0**	1.58 min / Epoch

Table 1. Performance comparison with finetune in NTU 60 dataset. The accuracy is calculate with models in similar #parameters

Y_i and $\hat{Y}_i$ are the output from teacher encoder and student predictor. M denotes the set of masked positions.

$$L_{rec} = \sum_{i \in V} ||X_i - \hat{X}_i||_2^2 \tag{2}$$

X_i and $\hat{X}_i$ are the original skeleton data and the output from the student decoder. V denotes all joints in the dataset. For the final loss, we add two parts together and calculate the loss L

3.4. Teacher Parameterization

The student network parameters θ are updated through backpropagation on the loss gradients. The teacher parameters $\triangle$ are updated through exponentially moving average (EMA) based on student ones:

$$\triangle \leftarrow \alpha\triangle + (1 - \alpha)\theta \tag{3}$$

α is a hyperparameter used to control the update steps of teacher encoder.

4. Experiment

In this project, we implement three distinct versions of our model along with two baselines. One baseline is fully-supervised learning using a vanilla transformer, and the second is the MAMP model. We select human action classification as the downstream task and monitor both top-1 and top-5 accuracy.

4.1. Datasets

We utilize the datasets NTU RGB+D 60. We only use $\frac{1}{4}$ of the dataset. There are 14,220 skeleton sequences across 60 action categories performed by 40 subjects. We follow the X-sub split strategy: sequences from 20 subjects are used for training and the rest are used for testing.

4.2. Model Settings

We set the number of layers L_e in both the student and teacher encoders to 5. The student predictor and student decoder each have 3 layers. The embedding dimension is set to 256, and the number of heads in the multi-head self-attention module is set to 8. For joint embedding, the length of each segment is set to 4.

4.3. Evaluation and Comparison

We employed the L2 loss to assess the performance of all models. The evaluation loss and metric results are presented in Table 1.

We have implemented three versions of the model. The first is the basic spatial and temporal model. In the second model, we incorporate the joint embedding and masking strategy. In the third model, we add a decoder to the student network.

From the results, we find that our second model shows improved classification accuracy compared to fully-supervised learning. Additionally, our final model demonstrates a 2-3% improvement in accuracy compared to the MAMP model. Furthermore, we observe that with the patch strategy, the spatial transformer outperforms the temporal transformer. This is likely because the joint embedding reduces the number of frames, leading to a loss of some temporal sequence information.

5. Conclusion

In this project, we developed a self-supervised model to learn the representation features of 3D human skeleton data. Our model achieved the best performance in the action classification downstream task and demonstrated advantages in training time. However, the results still need to be validated on other datasets and other action-related tasks, such as skeleton position prediction or human mesh recovery. Additionally, the network structure could potentially be improved through hyper-parameter tuning. Furthermore, the Exponential Moving Average (EMA) strategy could be modified to enhance the learning capability of the encoder during the fine-tuning stage.

References

[1] Ilya Sutskever Alex Krizhevsky and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks.

In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2012. 1

[2] Di Xie Chao Li, Qiaoyong Zhong and Shiliang Pu. Skeleton-based action recognition with convolutional neural networks. *2017 IEEE international conference on multimedia expo workshops (ICMEW)*, 2017. 1

[3] Yu-Wing Tai Hao-Shu Fang, Shuqin Xie and Cewu Lu. Rmpe: Regional multi-person pose estimation. *In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017. 1

[4] Bingbing Ni-Jiancheng Yang Xiaokang Yang Jingwei Xu, Zhenbo Yu and Wenjun Zhang. Deep kinematics analysis for monocular 3d human pose estimation. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[5] Dong Xu Jun Liu, Amir Shahroudy and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. *In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11- 14, 2016, Proceedings*, 2016. 1

[6] Dong Xu Jun Liu, Amir Shahroudy and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. *In Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1

[7] Dong Xu-Alex C Kot Jun Liu, Amir Shahroudy and Gang-Wang. Skeleton-based action recognition using spatio-temporal lstm network with trust gates. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 2017. 1

[8] Shaoqing Ren Kaiming He, Xiangyu Zhang and Jian Sun. Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016. 1

[9] Xiulong Liu Kun Su and Eli Shlizerman. Predict cluster: Unsupervised skeleton based action recognition. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[10] Risheng Liu Liangqu Long Jianhua Dai Nenggan Zheng, JunWen and Zhefeng Gong. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. *In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 2

[11] Junliang Xing Wenjun Zeng Jianru Xue Pengfei Zhang, Cuiling Lan and Nanning Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. *In Proceedings of the IEEE international conference on computer vision*, 2017. 1

[12] Yuanjun Xiong Sijie Yan and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *AAAI*, 2018. 1

[13] Yuanjun Xiong Sijie Yan and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 1

[14] Shijian Lu Meng Hwa Er Siyuan Yang, Jun Liu and Alex C Kot. Skeleton cloud colorization for unsupervised 3d action representation learning. *In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[15] Ce zheng Shiqian Wu Chen Chen Wenhan Wu, Yilei Hua and Aidong Lu. Skeletonmae: Spatial-temporal masked autoencoders for self-supervised skeleton action recognition. *arXiv preprint arXiv:2209.02399*, 2022. 2

[16] Wei Wang Yong Du and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[17] Yun Fu Yong Du and Liang Wang. Skeleton based action recognition with convolutional neural network. *2015 3rd IAPR Asian conference on pattern recognition (ACPR)*, 2015. 1

[18] Yun Fu Yong Du and Liang Wang. Skeleton based action recognition with convolutional neural network. *In Proceedings of the Asian Conference on Pattern Recognition (ACPR)*, 2015. 1

[19] Wengang Zhou Yao Fang Wanli Ouyang Yunyao Mao, Jiajun Deng and Houqiang Li. Masked motion predictors are strong 3d action representation learners. *In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 3

[20] Chunfeng Yuan Bing Li Ying Deng Yuxin Chen, Ziqi Zhang and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. *In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1

[21] T. Simon S. Wei Z. Cao, G. Hidalgo and Y. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 1