HHS Public Access

Author manuscript

Med Image Comput Comput Assist Interv. Author manuscript; available in PMC 2022 November 05.

Published in final edited form as:

Med Image Comput Comput Assist Interv. 2022 September ; 13438: 130–139.

doi:10.1007/978-3-031-16452-1_13.

GaitForeMer: Self-Supervised Pre-Training of Transformers via Human Motion Forecasting for Few-Shot Gait Impairment Severity Estimation

Mark Endo[1], **Kathleen L. Poston**[1], **Edith V. Sullivan**[1], **Li Fei-Fei**[1], **Kilian M. Pohl**[1,2], **Ehsan Adeli**[1]

[1]Stanford University, Stanford, CA 94305, USA

[2]SRI International, Menlo Park, CA 94025, USA

Abstract

Parkinson's disease (PD) is a neurological disorder that has a variety of observable motor-related symptoms such as slow movement, tremor, muscular rigidity, and impaired posture. PD is typically diagnosed by evaluating the severity of motor impairments according to scoring systems such as the Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS). Automated severity prediction using video recordings of individuals provides a promising route for non-intrusive monitoring of motor impairments. However, the limited size of PD gait data hinders model ability and clinical potential. Because of this clinical data scarcity and inspired by the recent advances in self-supervised large-scale language models like GPT-3, we use human motion forecasting as an effective self-supervised pre-training task for the estimation of motor impairment severity. We introduce **GaitForeMer**, Gait Forecasting and impairment estimation transforMer, which is first pre-trained on public datasets to forecast gait movements and then applied to clinical data to predict MDS-UPDRS gait impairment severity. Our method outperforms previous approaches that rely solely on clinical data by a large margin, achieving an F_1 score of 0.76, precision of 0.79, and recall of 0.75. Using GaitForeMer, we show how public human movement data repositories can assist clinical use cases through learning universal motion representations. The code is available at https://github.com/markendo/GaitForeMer.

Keywords

Few-shot learning; Gait analysis; Transformer

1 Introduction

Large-scale language models [1], such as the third generation Generative Pre-trained Transformer (GPT-3) [2] and Contrastive Language-Image Pre-Training (CLIP) [20], have gained great success in solving challenging problems under few or zero-shot settings. They owe their success to self-supervised pre-training on an abundant amount of raw and unlabeled data while completing a downstream task fine-tuned on small datasets. These

eadeli@stanford.edu .

methods can be of great interest in clinical applications, where data are regularly scarce and limited. For instance, one such application could be automatically estimating motor and gait impairments. This is a crucial step in the early diagnosis of Parkinson's disease (PD).

PD is a chronic, progressive brain disorder with degenerative effects on mobility and muscle control [5]. It is one of the most common neurodegenerative disorders, affecting around 0.57% of people age 45 and over [16]. Motor symptoms are typically used to diagnose PD, as non-motor symptoms such as cognitive symptoms lack specificity and are complicated to assess [27]. Most of the prior methods for automated prediction of motor-related PD signs and symptoms use wearable sensors [4,11,12], but these systems can be expensive and intrusive [14].

Recent methods have shown that video technology can be a scalable, contactless, and non-intrusive solution for quantifying gait impairment severity [15]. They use recordings of clinical gait tests from the Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [8], a widely used clinical rating scale for the severity and progression of PD. One of its components is a motor examination where participants walk 10 meters away from and toward an examiner. Specialists assess the severity of motor impairment on a score from 0 to 4. Score 0 indicates no motor impairment, 1 means slight impairment, 2 indicates mild impairment, 3 specifies moderate impairment, and 4 means severe impairment. The goal is automated prediction of this universally-accepted scale.

At the same time, most clinical studies are limited by the difficulty of large-scale data collection, since recruiting patients requires costly multi-site national or international efforts. Recent self-supervised pre-training frameworks have been used to learn useful representations that can be applied to downstream tasks [6,10,3,9,21]. Yet, their translation to clinical applications is under-explored. There is a growing number of 3D human motion capture repositories that can be used for self-supervised pre-training of motion representation learning, the same way GPT-3 [2] and CLIP [20] are trained for language-related tasks. Such pre-trained models can then be adapted for motor impairment estimation.

In this paper, we develop a novel method, **GaitForeMer**, that forecasts motion and gait (pretext task) while estimating impairment severity (downstream task). Our model is based on the recent advances in convolution-free, attentionbased Transformer models. GaitForeMer can take advantage of large-scale public datasets on human motion and activity recognition for learning reliable motion representations. To this end, we pre-train the motion representation learning on 3D motion data from the NTU-RGB+D dataset [22]. The learned motion representations are then fine-tuned to estimate the MDS-UPDRS gait scores. Our approach outperforms previous methods on MDS-UPDRS score prediction by a large margin. The benefits of GaitForeMer are twofold: (1) we use motion forecasting as a self-supervised pre-training task to learn useful motion features for the task of motor impairment severity prediction; (2) the joint training of motion and MDS-UPDRS score prediction helps improve the model's understanding of motion, therefore leading to enhanced differentiation of motor cues. To the best of our knowledge, we are the first to use motion prediction from natural human activity data as an effective pre-training task for the downstream task of

motor impairment severity estimation. We expect this method to be useful in decreasing the reliance on large-scale clinical datasets for the detection of factors defining gait disturbance.

2 GaitForeMer: Gait Forecasting & Impairment Estimation TransforMer

We first introduce a Transformer model that operates on sequences of 3D human body skeletons for the pre-training task of human motion forecasting, and we subsequently adapt it to the downstream task of MDS-UPDRS gait score estimation. Given a sequence of t 3D skeletons $\mathbf{x}_{1:t}$, we predict the next M skeletons $\mathbf{x}_{t+1:T}$ and the motion class y (either activity or MDS-UPDRS score). We follow the original setup of the pose Transformer model defined in [17]. Specifically, our model comprises a skeleton encoding neural network ϕ, a Transformer encoder and decoder, a pose decoding neural network ψ that reconstructs the 3D poses, and a linear classifier for MDS-UPDRS score/activity prediction. The model takes $\mathbf{x}_{1:t}$ as input, and the skeleton encoding network ϕ embeds this joint data into dimension D for each skeleton vector. Then, the Transformer encoder takes in this sequence of skeleton embeddings aggregated with positional embeddings and computes a latent representation $\mathbf{z}_{1:t}$ using L multi-head self-attention layers. The outputs of the encoder embeddings, $\mathbf{z}_{1:t}$, are fed into a single linear layer to predict class probabilities. The classification loss L_c (a multi-class cross-entropy) is used to train activity or score prediction. Note that the class prediction uses the motion representation (i.e., the latent space of the Transformer) as input and outputs the activity classes in the pre-training stage and MDS-UPDRS scores in the downstream task.

In addition, the encoder outputs $\mathbf{z}_{1:t}$ and a query sequence $\mathbf{q}_{1:M}$ are fed into the Transformer decoder. The query sequence is filled with the last element of the input sequence $\mathbf{x}_t$. The decoder uses L multi-head self- and encoder-decoder attention layers. The output of the decoder is fed into a skeleton decoding network ψ to generate the future skeleton predictions $\mathbf{x}_{t+1:T}$. The motion forecasting branch is trained using a layerwise loss calculated as:

$$L_l = \frac{1}{M \cdot N} \sum_{m=t+1}^{T} \left\| \hat{\mathbf{x}}_m^l - \mathbf{x}_m^* \right\|_1,$$

where $\hat{\mathbf{x}}_m^l$ is the predicted sequence of N-dimensional skeleton vectors at layer l of the Transformer decoder, and $\mathbf{x}_m^*$ is the ground-truth future skeleton. The motion forecasting loss L_f is then computed by averaging the layerwise loss over all decoder layers L_l. See Figure 1 for an overview of the model architecture.

2.1 Pre-training Procedure

For pre-training, we jointly train the activity and motion branches of the GaitForeMer. For the classification loss L_c, we use a standard categorical cross-entropy loss. The final loss is calculated as $L_{pre} = L_c + L_f$, where there is an equal weighting of the two different losses. We train the model for 100 epochs using an initial learning rate of 0.0001.

2.2 Fine-tuning Procedure

For our downstream task of MDS-UPDRS score prediction, we initialize the model using the learned weights from pre-training. We set the classification loss L_c as a weighted categorical cross-entropy loss since there is a significant class imbalance in the clinical data. We experiment with a variety of training procedures for fine-tuning the model. In one setup, we solely fine-tune the class prediction branch by setting $L_{fine} = L_c$. In another setup, we fine-tune both the class prediction branch and the motion prediction branch by setting $L_{fine} = L_c + L_f$. We also experiment with first fine-tuning both branches for 50 epochs then additionally solely fine-tuning the class prediction branch for 50 epochs. All fine-tuning setups are trained for 100 epochs using an initial learning rate of 0.0001.

2.3 Baselines

We compare our GaitForeMer method to a similar setup without pre-training on the NTU RGB+D dataset as well as various other motion impairment severity estimation models for MDS-UPDRS score prediction. The GaitForeMer model trained from scratch (GaitForeMer-Scratch) uses the loss function $L = L_c + L_f$ and follows the same configuration as the fine-tuning setups except it is trained for an additional 100 epochs. Hybrid Ordinal Focal DDNet (OF-DDNet) [15] uses a Double-Features Double-Motion Network with a hybrid ordinal-focal objective. This previous method has shown promising results on this MDS-UPDRS dataset. Spatial-Temporal Graph Convolutional Network (ST-GCN) is a graphical approach for learning spatial and temporal characteristics that can be used for action recognition [26]. For this method, we add slow and fast motion features and pass the input through a Graph Attention Network (GAT) [23] layer first to allow for additional spatial message passing. DeepRank [18] is a ranking CNN, and the Support Vector Machine (SVM) [24] is using the raw 3D joints.

3 Datasets

In this work, we use a clinical dataset for estimation of gait impairment severity from MDS-UPDRS videos and a public 3D human gait dataset for pre-training the motion forecasting component. Both datasets are described below.

3.1 NTU RGB+D Dataset

We use NTU RGB+D [22] to pre-train our GaitForeMer model. This dataset includes 56,880 video samples with 60 action classes. We use the skeletal data containing 3D coordinates of 25 joints and class labels. We pre-train our model using the joints as input and the activity labels for supervision of the activity branch. The activity branch is the same as our MDS-UPDRS branch (see Fig. 1), except that during pre-training we train the linear layers to predict the activity class in the NTU RGB+D dataset.

3.2 MDS-UPDRS Dataset

For the downstream task of gait impairment severity estimation, we use the MDS-UPDRS dataset defined in [15]. This dataset contains video recordings of MDS-UPDRS exams from 54 participants. Following previously published protocols [19], all participants are recorded during the off-medication state. During the examinations, participants are recorded walking

towards and away from the camera twice at 30 frames per second. Each sample is scored by three boardcertified movement disorders neurologists and we use the majority vote among the raters as the ground-truth score, randomly breaking ties. Note that, in this work, we do not aim to model the uncertainty among raters. The raters score the videos on a scale from 0 to 4 based on MDS-UPDRS Section 3.10 [7]. In our work, we combine scores 3 and 4 due to the difficulty of obtaining video recordings for participants with extreme motor impairment. The data setup and protocols are the same as in [15], except in their previous work two sets of scores were counted from one of the neurologists. Using the gait recordings as input, we use Video Inference for Body Pose and Shape Estimation (VIBE) to extract 3D skeletons [13]. This joint data is then preprocessed by normalization and splitting of samples into clips of 100 frames each. We then use these clips for estimating motor impairment severity.

4 Experiments

In this section, we first evaluate how motion forecasting helps improve a system estimating the MDS-UPDRS scores. We compare our results with several baselines (Section 4.1). We then evaluate how the fine-tuning strategy contributes to better results (Section 4.2). We further experiment on how our few-shot learning paradigm can be adopted for clinical approaches using pre-training (Section 4.3). Qualitative results on motion forecasting of PD patients validate that GaitForeMer is able to learn good motion representations (Section 4.4).

4.1 Using Motion Forecasting as an Effective Pre-training Task

We investigate the efficacy of using human motion forecasting as a self-supervised pre-training task for the downstream task of motor impairment severity estimation. We evaluate each model using macro F_1 score, precision, and recall. These metrics are calculated on a per subject level with leave-one-out-cross-validation settings. We compare our GaitForeMer method to baseline methods in Table 1.

We find that our GaitForeMer method pre-trained on the NTU RGB+D dataset results in improved performance over training the model from scratch and all baselines trained on the MDS-UPDRS dataset. Our best setup achieves an F_1 score of 0.76, precision of 0.79, and recall of 0.75. In comparison, training the model from scratch results in an F_1 score of 0.60, precision of 0.64, and recall of 0.58, which is still superior to other baselines. The OF-DDNet baseline (previous state-of-the-art approach in MDS-UPDRS score prediction) has an F_1 score of 0.58, precision of 0.59, and recall of 0.58.

4.2 Evaluating Fine-tuning Strategies

We experiment with various fine-tuning strategies in order to evaluate different approaches for training our GaitForeMer method. Our best approach of first fine-tuning both the class prediction and motion prediction branches then solely fine-tuning the class prediction branch achieves an F_1 score of 0.76, precision of 0.79, and recall of 0.75. Another approach of fine-tuning both branches achieves an F_1 score of 0.72, precision of 0.75, and recall of 0.71. We observe that solely fine-tuning the class branch results in worse performance than also training the motion branch with an F_1 score of 0.66, precision of 0.72, and recall of 0.63. The relatively poor performance could be due to the data shift between the NTU RGB+D

and MDS-UPDRS datasets that requires training of the motion forecasting branch. Results are shown in Table 2.

4.3 Few-Shot Estimation of Gait Scores

To better understand the few-shot capabilities of GaitForeMer, we experiment with limiting the training dataset size and evaluating performance compared to ST-GCN. We sample either 25%, 50%, or 75% of the data (analogous to 13, 26, or 39 videos) for training in each fold, preserving the same samples across the two methods. We maintain class balance by sampling one-fourth of the required subsamples from each class when permitted. We resample and run each method three times. The results are illustrated in Figure 2.

We find that our GaitForeMer method maintains relatively strong performance with only a fraction of the data. GaitForeMer with access to 25% of training data achieves an average F_1 score of 0.56, which is higher than ST-GCN using 100% of training data with an F_1 score of 0.52 and comparable to OF-DDNet (second-best performing method) using 100% training data with an F_1 score of 0.58. This shows the power of using motion forecasting as a self-supervised pre-training task for few-shot gait impairment severity estimation.

4.4 Motion Forecasting Visualization

In Figure 3, we visualize the predicted outputs of the GaitForeMer model jointly trained on MDS-UPDRS score prediction and motion forecasting. Although accurate pose forecasting is not necessary for the prediction of MDS-UPDRS scores, it can help demonstrate the utility of learned motion features. Qualitatively, we see that the predicted poses most closely match the ground-truth at the beginning of the output. This might be because using the last input entry $\mathbf{x}_t$ as the query sequence $\mathbf{q}_{1:M}$ helps the prediction in the short term [17]. A larger error exists for longer horizons where the outputted poses become less similar to the query sequence. In addition, the non-autoregressive approach of GaitForeMer can lead to an increased accumulation of error.

Clinically, the results illustrated in Figure 3 show normal movement behavior in class 0 (normal), while classes 1 and 2 show increased stiffness, decreased mobility, and reduced arm swing and pedal motion. Participants in class 3 are imbalanced and require assistive devices for safe walking. These results verify that the forecasting module is able to properly predict future motion that encodes motor impairments.

5 Conclusion

Herein, we presented a model, GaitForeMer, based on transformers for forecasting human motion from video data that we use to predict MDS-UPDRS gait impairment severity scores. We found that human motion forecasting serves as an effective pre-training task to learn useful motion features that can subsequently be applied to the task of motor impairment severity estimation, even in few-shot settings. The pre-trained GaitForeMer outperformed training from scratch and other methods for motor impairment severity estimation that solely use the MDS-UPDRS dataset for training. Our approach demonstrates the utility of using motion pre-training tasks in data-limited settings.

Acknowledgements

This work was supported in part by NIH grants (AA010723, NS115114, P30AG066515), the Michael J Fox Foundation for Parkinson's Research, UST (a Stanford AI Lab alliance member), and the Stanford Institute for Human-Centered Artificial Intelligence (HAI) Google Cloud credits.

References

1. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, et al. : On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
2. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al. : Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901 (2020)
3. Chen T, Kornblith S, Norouzi M, Hinton G: A simple framework for contrastive learning of visual representations (2020)
4. Daneault J, Vergara-Diaz GP, Parisi F, Admati C, Alfonso C, Bertoli M, Bonizzoni E, Carvalho GF, Costante G, Fabara EE, Fixler N, Golabchi FN, Growdon J, Sapienza S, Snyder P, Shpigelman S, Sudarsky LR, Daeschler M, Bataille L, Sieberts SK, Omberg L, Moore S, Bonato P: Accelerometer data collected with a minimum set of wearable sensors from subjects with parkinson's disease. Scientific Data 8 (2021)
5. DeMaagd G, Philip A: Parkinson's disease and its management: Part 1: Diseaseentity, risk factors, pathophysiology, clinical presentation, and diagnosis. P & T : a peer-reviewed journal for formulary management 40, 504–32 (08 2015) [PubMed: 26236139]
6. Devlin J, Chang MW, Lee K, Toutanova K: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
7. Goetz C, Fahn S, Martinez-Martin P, Poewe W, Sampaio C: The mds-sponsored revision of the unified parkinson's disease rating scale. J. Mov. Disord 1, 1–33 (2008)
8. Goetz C, Tilley B, Shaftman S, Stebbins G, Fahn S, Martinez-Martin P, Poewe W, Sampaio C, Stern M, Dodel R, Dubois B, Holloway R, Jankovic J, Kulisevsky J, Lang A, Lees A, Leurgans S, Lewitt P, Nyenhuis D, Lapelle N: Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): Scale presentation and clinimetric testing results. Movement disorders : official journal of the Movement Disorder Society 23, 2129–70 (11 2008). 10.1002/mds.22340 [PubMed: 19025984]
9. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R: Masked autoencoders arescalable vision learners (2021)
10. He K, Fan H, Wu Y, Xie S, Girshick R: Momentum contrast for unsupervisedvisual representation learning (2020)
11. Hobert MA, Nussbaum S, Heger T, Berg D, Maetzler W, Heinzel S: Progressive gait deficits in parkinson's disease: A wearable-based biannual 5-year prospective study. Frontiers in Aging Neuroscience 11 (2019). 10.3389/fnagi.2019.00022, https://www.frontiersin.org/article/10.3389/fnagi.2019.00022
12. Hssayeni MD, Jimenez-Shahed J, Burack MA, Ghoraani B: Wearable sensorsfor estimation of parkinsonian tremor severity during free body movements. Sensors (Basel, Switzerland) 19 (2019)
13. Kocabas M, Athanasiou N, Black MJ: Vibe: Video inference for human bodypose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
14. Lu M, Poston K, Pfefferbaum A, Sullivan EV, Fei-Fei L, Pohl KM, Niebles JC, Adeli E: Vision-based estimation of mds-updrs gait scores for assessing parkinson's disease motor severity. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III. p. 637–647. Springer-Verlag, Berlin, Heidelberg (2020)
15. Lu M, Zhao Q, Poston KL, Sullivan EV, Pfefferbaum A, Shahid M, Katz M, Kouhsari LM, Schulman K, Milstein A, Niebles JC, Henderson VW, Fei-Fei L, Pohl KM, Adeli E: Quantifying

parkinson's disease motor severity under uncertainty using mds-updrs videos. Medical Image Analysis 73, 102179 (2021) [PubMed: 34340101]

16. Marras C, Beck J, Bower J, Roberts E, Ritz B, Ross G, Abbott R, Savica R, Van Den Eeden S, Willis A, Tanner C: Prevalence of parkinson's disease across north america. npj Parkinson's Disease 4 (12 2018). 10.1038/s41531-018-0058-0
17. Martínez-González A, Villamizar M, Odobez JM: Pose transformers (potr): Human motion prediction with non-autoregressive transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 2276–2284 (October 2021)
18. Pang L, Lan Y, Guo J, Xu J, Xu J, Cheng X: Deeprank. Proceedingsof the 2017 ACM on Conference on Information and Knowledge Management (Nov 2017). 10.1145/3132847.3132914, 10.1145/3132847.3132914
19. Poston KL, YorkWilliams S, Zhang K, Cai W, Everling D, Tayim FM, Llanes S, Menon V: Compensatory neural mechanisms in cognitively unimpaired parkinson disease. Annals of Neurology 79(3), 448–463 (2016). 10.1002/ana.24585, https://onlinelibrary.wiley.com/doi/abs/10.1002/ana.24585 [PubMed: 26696272]
20. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
21. Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M, Sutskever I: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
22. Shahroudy A, Liu J, Ng TT, Wang G: NTU RGB+D: A large scale datasetfor 3d human activity analysis. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1010–1019 (2016). 10.1109/CVPR.2016.115
23. Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y: Graphattention networks (2018)
24. Weston J, Watkins C: Support vector machines for multi-class pattern recognition (1999)
25. Wilcoxon F: Individual comparisons by ranking methods. In: Breakthroughs instatistics, pp. 196–202. Springer (1992)
26. Yan S, Xiong Y, Lin D: Spatial temporal graph convolutional networks forskeleton-based action recognition (2018)
27. Zesiewicz TA, Sullivan KL, Hauser RA: Nonmotor symptoms of Parkinson'sdisease. Expert Review of Neurotherapeutics 6(12), 1811–1822 (Dec 2006) [PubMed: 17181428]

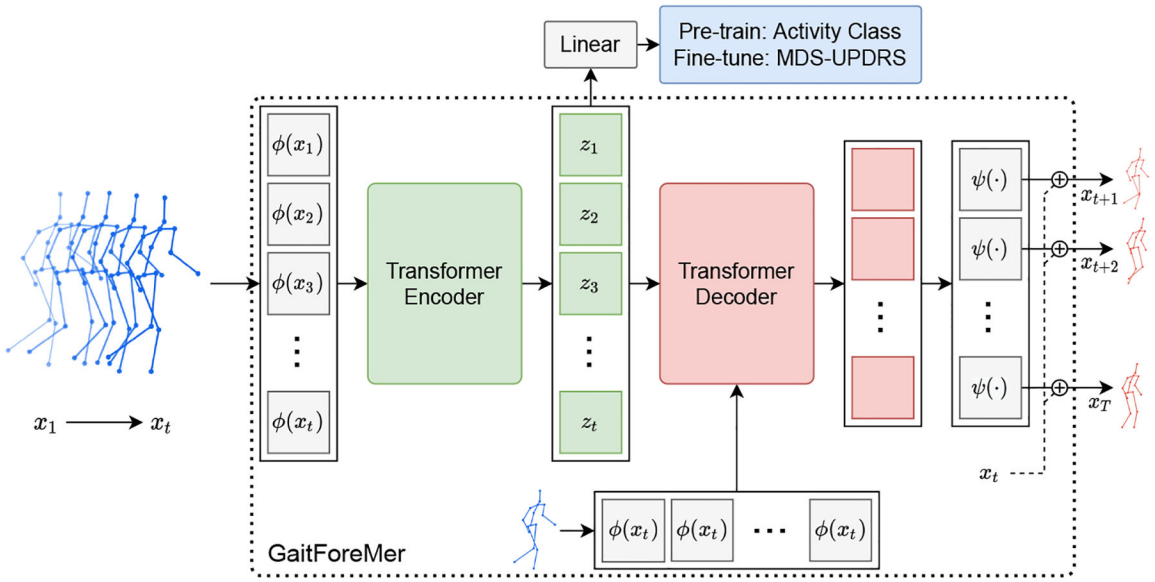

Fig. 1.

The proposed GaitForeMer framework for motor impairment severity estimation. A Transformer model based on body skeletons is first pre-trained on a large dataset jointly for human motion prediction and activity prediction. The model is subsequently fine-tuned on the task of MDS-UPDRS gait score prediction using extracted 3D skeleton sequences from clinical study participants using VIBE (Video Inference for Body pose and shape Estimation) [13].

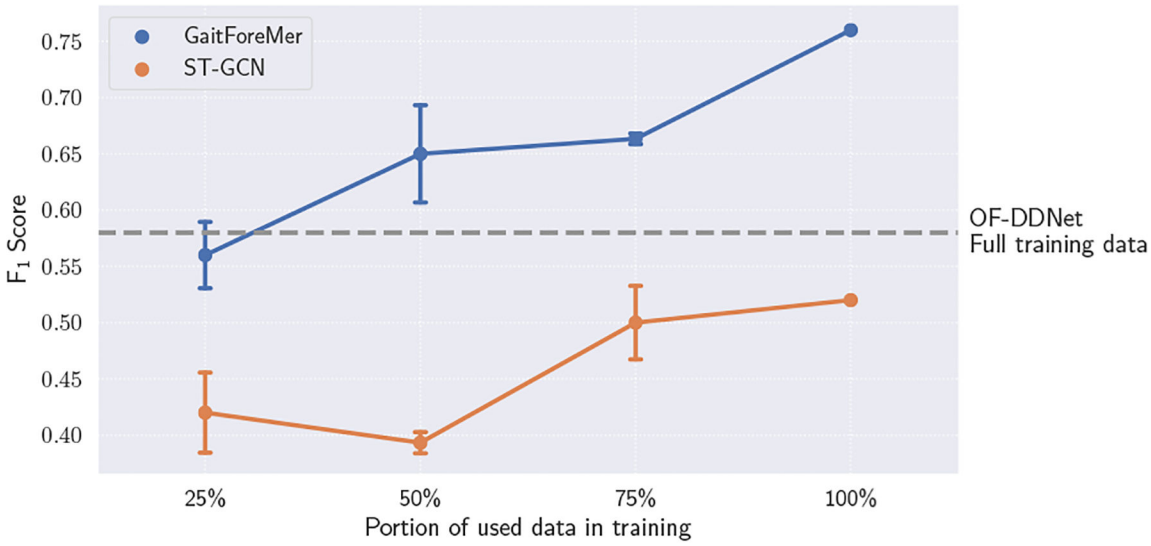

Fig. 2.

Few-shot performance of GaitForeMer compared to ST-GCN with different portions of data used in training. Error bars represent standard deviation across 3 runs. Our GaitForeMer method using only 25% of training data maintains comparable performance to the second-best performing method with full training data (OF-DDNet).

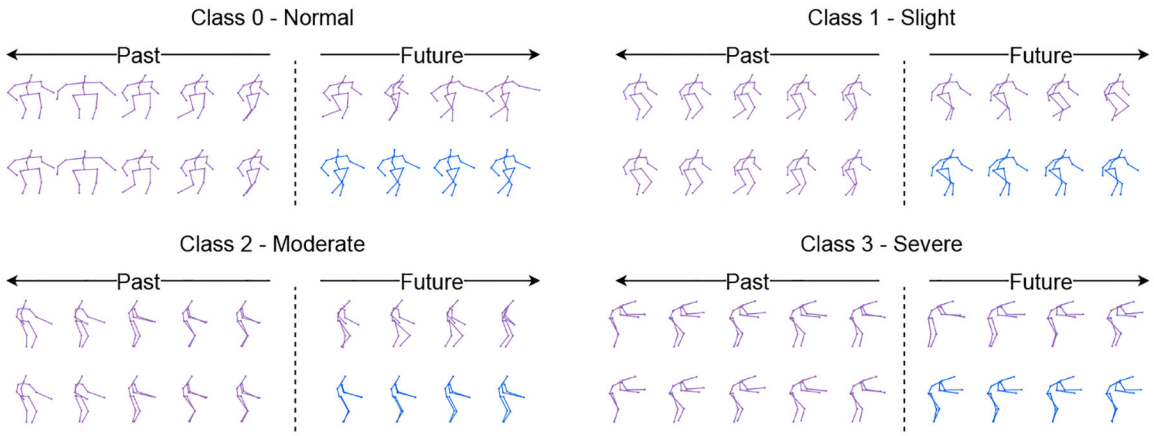

Fig. 3.

Visualization of human motion forecasting for different levels of motor impairment severity. The purple skeletons are ground-truth data, and the blue ones are predictions from GaitForeMer with fine-tuning both branches.

Table 1.

Comparison with baseline methods. Performance is evaluated using macro F_1 score, precision, and recall. We find that pre-training results in significantly improved performance over training from scratch and the other methods.

Method	F_1	Pre	Rec
GaitForeMer (Ours)	**0.76**	**0.79**	**0.75**
GaitForeMer-Scratch (Ours)	0.60	0.64	0.58
OF-DDNet$^{\natural *}$ [15]	0.58	0.59	0.58
ST-GCN [26]*	0.52	0.55	0.52
DeepRank$^{\natural *}$ [18]	0.56	0.53	0.58
SVM$^{\natural *}$ [24]	0.44	0.49	0.40

$^{\natural}$refers to results directly cited from [15].

*indicates statistical difference at ($p < 0.05$) compared with our method, measured by the Wilcoxon signed rank test [25]. Note that this is a 4-class classification problem and hence 0.25 recall implies a random classifier. Best results are in bold. See text for details about compared methods.

Table 2.

Comparison of different training/fine-tuning strategies of our method (ablation study on fine-tuning strategy). Performance is evaluated using macro F_1 score, precision, and recall. We find that first fine-tuning both branches (forecasting and score prediction) then additionally fine-tuning the score prediction branch yields best results.

Pre-trained	Fine-tune strategy	F_1	Pre	Rec
Yes	Both branches then class branch	**0.76**	**0.79**	**0.75**
Yes	Both branches	0.72	0.75	0.71
Yes	Class branch	0.66	0.72	0.63
No		0.60	0.64	0.58