Investigating Transformers for human action forecasting

Pierrick Bournez[1], Philippe Rambaud[1,2], Raphael Fauches[2], Arpad Rimmel[1], Jean Bergounioux[3], Joanna Tomasik[1]

[1]Laboratoire Interdisciplinaire des Sciences du Numériques (LISN).
[2]SogetiLabs, Capgemini.
[2]Hôpital Raymond Poincaré, AP-HP.

Contributing authors: pierrick.bournez@student-cs.fr; philippe.rambaud@lisn.fr;

Abstract

This research is centered on the broader objective of developing pre-diagnostic tools for early detection by leveraging the analysis of spontaneous motor skills in newborns. Within the scope of this study, and without enough newborn motricity data to work on, our primary focus lies in the capacity of attention-based models as to perform forecasting human skeleton actions. We conducted a comprehensive comparison of several state-of-the-art models and demonstrated that, through the implementation of specific preprocessing steps, FEDformer[1] exhibit promising outcomes. However, subsequent visual inspection revealed discrepancies that were rather disappointing, pointing to various underlying reasons. This divergence highlights the significance of incorporating visual assessments alongside quantitative measures. The forthcoming steps involve expanding datasets capturing human movements, and refining the model architecture through contextual data integration. The inclusion of data from hospitals' maternity facilities will further augment our research.

Keywords: attention, transformer, short time series, forecasting, skeleton

1 Motivation

The study at hand exists within the broader framework of identifying neurological pathologies in newborns through the analysis of spontaneous motor skills. We aim to develop pre-diagnostic tools for early detection. The central goal of this study is to assess transformer-based models' accuracy in forecasting motor skills for individuals without pathologies (normal data). We anticipate neurological abnormalities (abnormal data) to disrupt movement patterns, heightening prediction complexity.

Presently, an inadequate number of videos capturing newborns' motor activities are available for analysis. Thus, we turned our attention to the NTU RGB+D dataset[2]. This dataset emerged as a suitable substitute for our research objectives.

Main contributions

1. We improved FEDformer performance for forecasting human action by removing the mean of time series and asking to predict both the input and output data.
2. We underscored the necessity of incorporating visual assessment within the evaluation process. This assertion was based on our findings, which revealed a potential disparity between favorable loss metrics and actual predictive performance.

In the subsequent sections, we detail our initial experiment. Section 2 presents the methodology, Section 3 covers the results, and Section 4 concludes and outlines our ongoing work.

2 Methods

We conducted various tests involving dataset modifications, preprocessing steps, and model enhancements. As we cannot present all the experiments, we focus here on FEDformer and specific preprocessing steps that yielded the best results.

Dataset We used the NTU RGB+D[2] dataset, which features motion sequences of adults spanning durations of 2 to 10 seconds. Each 3D skeletal data contains the 3D coordinates of 25 body joints at each frame.
Preprocessing Each data is normalized by removing the mean of each time series. We divided the data into input and output segments, each consisting of 16 frames (approximately 1 second) at a precise point corresponding to a change in movement.
Model FEDformer implementation was taken from TSLib library[1].
Experimental Setup All computational tasks were performed within the LabIA[2] computing environment at the University of Paris-Saclay. Using the given input data for training, the model is required to predict both the input and output sequences.
Evaluation Metrics We adopted widely-used metrics such as Mean Squared Error (MSE) and Mean Absolute Error (MAE). We also conducted visual evaluations by comparing the ground truth movement patterns with the predictions generated by our models.

[1] https://github.com/thuml/Time-Series-Library/tree/main/models
[2] https://doc.lab-ia.fr/specifications/

3 Results

Table 1 showcases the performance of FEDformer both with and without specific preprocessing steps aimed at enhancing its efficacy in human action forecasting. Visual outputs are available, along with our code, on GitHub[3].

FEDformer	Training set		Validation set
	MSE	MAE	MSE
Without preprocessing	0.012	0.06	0.02
With preprocessing	0.0007	0.003	0.014

Table 1 FEDformer's performance with and without specific preprocessing steps.

The table illustrates the enhancement in FEDformer's performance due to our preprocessing interventions. In terms of the training set, we observed a 1.7-fold reduction in MSE and a 20-fold decrease in MAE.

Nevertheless, upon visual inspection of the results, we noticed that the model encountered difficulties in accurately predicting substantial movements. Its predictions extended the termination of input movements, presenting a misalignment with actual patterns.

We may explore alternative metrics and evaluation methodologies to gain insights into the underlying reasons for the model's suboptimal learning.

4 Conclusion and discussion

During this study, after comparing various Transformers and implementing preprocessing steps for human action forecasting, we have shown significant improvements in the initial performance of FEDformer. However, it is crucial to emphasize the necessity of incorporating visual assessments into the evaluation process. Existing evaluation methods have limitations that might obscure vital insights.

In our future endeavors, we plan to explore additional datasets capturing human movements, investigate preprocessing techniques, and enhance the architecture of FEDformer by integrating contextual data. Furthermore, as data from maternity wards becomes available, we intend to align it with our overarching research theme by incorporating it into our analyses.

References

[1] Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., Jin, R.: FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In: Proc. 39th International Conference on Machine Learning (ICML 2022) (2022)

[2] Shahroudy, A., Liu, J., Ng, T.-T., Wang, G.: Ntu rgb+d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1010–1019 (2016)

[3] https://github.com/gardiens/Time-Series-Library_babygarches