DOI 10.5162/iCCC2024/2.2

Minimizing the Latency of Freezing of Gait Detection on Wearable Devices

Ali Haddadi Esfahani [1], Oliver Maye[1], Max Frohberg[1], Steffen Ortmann[3], Peter Langendörfer[1,2]
[1] IHP–Leibniz-Institut für innovative Mikroelektronik, Im Technologiepark 25, 15236 Frankfurt (Oder), Germany
[2] BTU Cottbus-Senftenberg, 03046 Cottbus, Germany
[3] Carl-Thiem-Klinikum, 03048 Cottbus, Germany
haddadi, maye, frohberg, langendoerfer@ihp-microelectronics.com , s.ortmann@ctk.de

This research has been partially funded by the Federal Ministry of Education and Research of Germany under grant number 03ZZ5301E

Summary:
Freezing of Gait (FoG) is a common and severe symptom that impacts persons who have been diagnosed with Parkinson's disease. The detection of FoG is of greatest significance for precise diagnosis, preventing falls, and obtaining accurate measurements for severity of FoG episodes. These factors are critical for optimizing treatment approaches and enhancing the overall quality of life for those affected by FoG. Real-time FoG detection may be achieved by using a wearable system that can be worn by the patient. This configuration includes a sensor that is coupled with inference software on a computing device and a vibrator. The patient is alerted of a FoG incident by the vibrator that is activated on demand. The key role in FoG detection is the latency between the initial moments of imminent FoG episode and the moment at which the patient is a notified by the vibrator. By using more advanced FoG detection algorithms, the duration between incidents may be reduced, hence aiding to prevent patient falls and avoiding injuries. In this paper, we modified the model which was used in our previous work and improved the latency from 50ms to 3 ms. The dataset and input features remain unchanged from our earlier study to ensure comparability between performances of the two models. The individual models for one-second windows running on a PC of nine patients achieved a mean of 90% and 88% of sensitivity and specificities, respectively. The models that were converted and executed on Google Coral exhibited comparable performance, with a maximum variance of 1% compared to the performance attained on a personal computer.

Keywords: Machine Learning, Time series classification, AI model on Edge, Freezing of Gait, Neural network compression

Introduction

Parkinson's Disease (PD) is a persistent neurodegenerative disorder that affects a significant number of people worldwide. FoG is one of the main motor symptoms of PD patients. The presence of FoG may increase the risk of injuries resulting from falls. As most of the PD patients are old people, the long recovery time from an injury causes long pain suffering for the patient and imposes significant costs to the medical systems.

The time delay in FOG detection is a crucial key factor as faster classification leaves more time for other tasks. The process includes FOG detection, transmitting cue commands, and activating a device like a vibrator, FOG can be indicated and avoided through cues such as haptic or auditory signals within 200 ms of onset. Lower latency allows the model to apply more inference tasks in the mentioned time interval.

Recently, machine learning (ML) has shown effectiveness in FoG detection, utilizing datasets and advanced wearable technology. This study focuses on employing ML on wearable devices for rapid FoG detection.

Nine patients were used to create our own dataset and the patients cover different age and sex groups. They executed a variety of activities during test sessions. The same training and test set from our previous research work was used to train our new model in order to ensure fair comparison of performance of the new small and the older big models [1]. Our previous study showed the feasibility of FOG detection using a deep LSTM model on a wearable device. The patient-dependent models classified the real-time 3-channel acceleration data with 170000 parameters on the battery-powered device in 50 ms. In this study, the new model architecture – uses 17000 parameters which is 10 times less than in our previous research

work. It is a modified version of the previous model implemented by us [1] . The proposed model focuses more on latency reduction. Like previous research, our new model uses 1 second windows i.e. 100 data – for classification. The trained models were compressed and quantized using a symmetric quantization method and finally the compressed models were compiled to run on the CPU and TPU of a Google Coral mini board. The classification latency was reduced to 3ms. The average model accuracy on all patients is 89%. The converted models deployed on the Google Coral mini CPU produced very similar results to the PC models and have a maximum of 2% deviation compared to the performance metrics of the older models.

Related work

An extensive investigations into FoG have been conducted in order to improve the accuracy of its detection. These studies have used a range of technologies including IMU sensors, cameras, and more recently, WiFi signals [2][3][4][5]. However, a key observation from recent researches is that the models for detecting FOG have been limited to high-performance computing devices, like personal computers and high-end servers. There has been a lack of focus on evaluating real-time capabilities of FOG detection systems or exploring the prospects of wearable technology that can perform real-time AI processing and gather data on-the-go. Furthermore, the field of FOG detection research has seen a variety of datasets, location of sensor placements and methods being utilized. This diversity makes it challenging to directly compare the effectiveness of these models with each other, or to set a standard benchmark for the research conducted to date.

Guo et al. used a proxy measurement model, integrating acceleration and pseudo-EEG data, to detect FoG [6]. Like our research he used LSTM followed by SVM models to classify FoG and nonFoG in a patient dependent manner. The chosen window is 2 seconds and the sliding window moves 0.25 s forward. The average sensitivity over 8 patients in the dataset is 88%±10.

Kun Hu et al. designed a polynomial transformer for FoG detection [3]. It incorporates pose and appearance feature sequences to formulate detailed FoG patterns. The HP-Transformer uses a higher-order self-attention mechanism based on linear, bilinear, and trilinear transformers. The window length is 1-second and the classification latency 120 ms on a high end GPU in PC.

Also FPGA architectures were used for implementing machine learning models for FOG detection. Mikos et al. implemented a neural network on an FPGA, creating a custom feature selection process [7]. Their model achieved 95.6% sensitivity and 90.2% specificity with a 4.5s window length, inferring in 20ms, and operated for 9 hours on an 800mAh battery. Langer et al. trained and tested a Temporal Convolutional Neural Network using the Daphnet dataset, later implementing it on a Xilinx FPGA with VitisAI [8]. Their model labeled data in under a millisecond, enabling a cueing device trigger in less than 250 ms, and achieved 78% sensitivity and 90% specificity.

Dataset, Methodology, and Result

Kliniken Schmieder Allensbach received ethical approval and patient consent for a study to analyze Freezing of Gait (FoG) in Parkinson's disease. Nine patients underwent parkour sessions to challenge their walking, while data was collected via an ankle-mounted sensor. The study's novelty includes real-time FOG detection, patient engagement, and a custom-designed board developed by IHP microelectroncis that records the acceleration data at 100Hz, see Fig. 1.

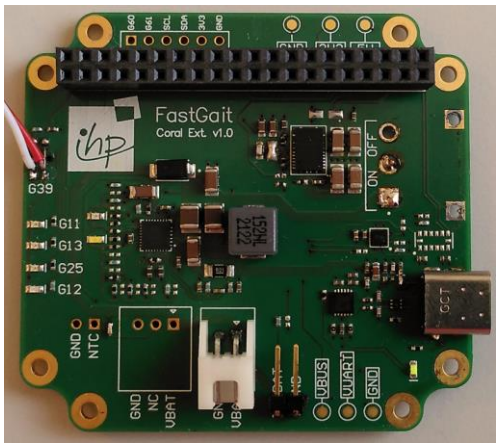

Fig. 1: Custom-design data acquisition board that records movement of patients by using an acceleration sensor.

Recorded data from the acceleration sensor, along with a video analysis by clinical experts, was used for post analysis. Video recordings helped to identify FOG instances, which were labelled in MATLAB for classification.

The data, predominantly lossless thanks to IHP's robust firmware, was post-processed in MATLAB and prepared for supervised machine learning using TensorFlow 2.9 and its Lite version for both building the model running on a PC and the compressed model for embedded compatible version [9]. Training and testing involved a sliding window analysis with 100

samples per second from three acceleration channels, ensuring each window contained data from one label type, only. The dataset, with FOG data was increased to balance the training set. Upon tripling the size of the FoG windows, the training set's FoG data increased from 15.3% to 45.9%, thereby significantly mitigating the imbalance between FOG and non-FOG data. 10% of whole dataset were used for testing and 90% for training and validation, maintaining an even distribution of labels across the sets.

Model architecture and conversion

The primary benefit of LSTM modeling is its ability to capture nonlinear patterns in time series data [10]. This model architecture provides a significant ability to identify complex patterns in freezing episodes as stored by accelerometer signals.

Reduced latency combined with high model performance are key indicators for real-time FoG detection. A model that maintains low latency without a substantial decrease in performance enables a greater number of inferences within the given time.

In our prior research, the model utilized a multi-stack LSTM layer followed by a single classification layer for FoG detection [1]. In the current study, we have shifted our focus towards enhancing the classification aspect of the model while reducing the temporal-dependency layer and minimizing LSTM layer parameters. The revised model comprises a single LSTM layer with 80 hidden units, accompanied by four fully connected layers. The total number of parameters has been significantly reduced to 17,000, which is a tenfold decrease from the previous model's 170,000 parameters.

The trained models are then converted using Tensorflow Lite through *Post-training quantization.* The models are converted for two types of processors to be compared in terms of performance and latency. All trained model parameters are converted from Float32 to float16 and Int8 which is the format supported by Google Coral CPU and the corresponding Tensor Processing Unit (TPU) of Google Coral board, respectively.

Results and discussion

The reduction in parameters of our new model has led to a notable decrease in inference time on Coral's CPU, down to 3ms from 50ms of our previous implementation. The shortened classification delay allows for inference to occur between individual data transmissions from the sensor, which is optimal for a 100Hz sampling rate with a 10 ms interval between data arrivals. This enhancement in processing speed, with a latency of only 3 ms, enables the analysis of the most recent data produced by the sensor, i.e. assessing the patient's latest movement with immediacy. This low latency helps the FOG detection system to work in real-time manner and check very recently produced data.

Due to the unbalanced number of FoG and Non-FoG labels in our dataset, the performance of the models are represented as sensitivity for FOG detection rate and specificity for Non-FoG detection rate.

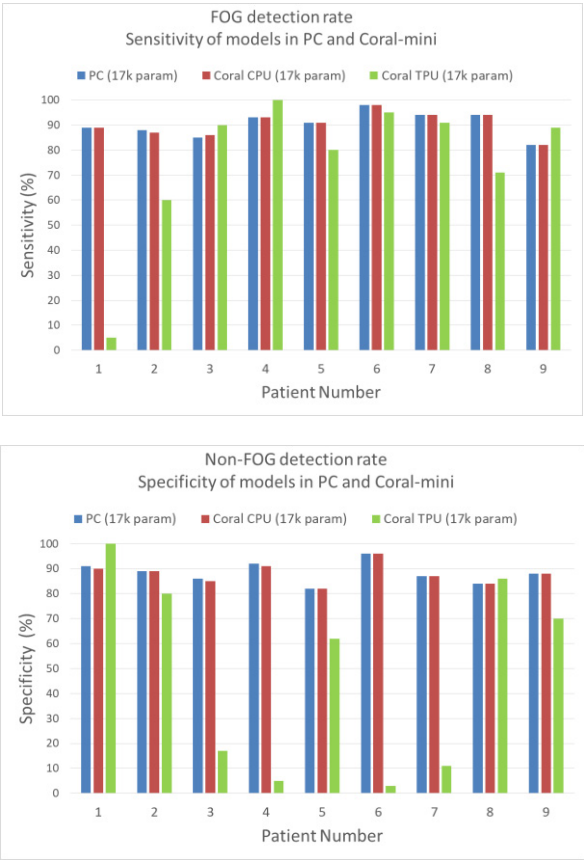

Fig. 2: Performance of the models trained on PC and the ones converted for Google Coral's CPU and TPU.

The sensitivity and specificity of the PC model closely align with the converted model for Coral's CPU, exhibiting a marginal difference of up to 2%, as shown in Fig. 2. This similarity may be attributed to the effective retention of information in the converted model, which uses float16, from the original model's float32 parameters. The reduction in bit size of parameters appears to have a minimal impact on the model's performance in terms of both sensitivity and specificity. Furthermore, all models demonstrated a sensitivity and specificity exceeding 80% across all patient data while 5 out of 9 nine patients had over 90%.

The converted models suitable for Coral's TPU, classification could be executed in as little as 2

milliseconds. However, there is a notable divergence in model performance, characterized by a tendency to exclusively identify either FoG or Non-FoG conditions across patients, see Fig. 2. Despite the presence of an imbalanced dataset, which typically biases the model towards classifying the more prevalent label (Non-FoG in this case), these models demonstrated a higher detection rate in terms of sensitivity than specificity.

Current observations reveal that the inconsistency in detection performance between TPU and CPU models remains unexplained. This difference, noted during the conversion of models to TPU-compatible format, resulted in achieving lower than 80% in sensitivity or specificity measures. The transformation from floating point to integer format reduce precision, but this alone does not conclusively explain the lower detection rate in TPU models. Therefore, at this stage, the underlying reasons for these differences remain unclear.

Conclusion

This research demonstrates the feasibility of real-time detection of FoG using a wearable device. The modification of our model introduced in our previous helped to reduce the classification latency more than 10 times and attain 3ms for inference time on a Google Coral CPU. The one layer LSTM model along deep classification layers used our dataset to make patient dependent models. Through the implementation of suitable quantization and pruning methods, the models were optimized for deployment on Google Coral CPU and TPUs, enabling them to conduct real-time inferences.

To the best of our understanding, this study represents the first instance of employing a Google Coral CPU and TPU to run a ML model, along with a custom design extension board for gathering and assessing accelerometer data from patient movements.

Our method effectively differentiated FOG from non-FOG data using one-second windows from three-axis acceleration sensor. The patient-specific models attained a sensitivity exceeding 80% for all patients, with patients reaching over 90% detection rates. The models on the Google Coral board performed comparably to PC-based models in sensitivity and specificity, with a mere +/-2% difference. With a classification time of 3ms on Coral CPU, the model's rapid inference enables timely cueing. The models adapted for Coral's TPU showed limited effectiveness in data classification, often favoring one label over the other.

Many challenges are still to be addressed in FOG detection, such as building large enough datasets allowing a more accurate detection via machine-learning techniques [12].

References

[1] Haddadi Esfahani A, Maye O, Frohberg M, Speh M, Jöbges M, Langendörfer P. Machine Learning based Real Time Detection of Freezing of Gait of Parkinson Patients Running on a Body Worn Device. 2023 IEEE/ACM Conf. Connect. Heal. Appl. Syst. Eng. Technol., 2023, p. 181–2. https://doi.org/10.1145/3580252.3589423.

[2] Habib Z, Mughal MA, Khan MA, Shabaz M. WiFOG: Integrating deep learning and hybrid feature selection for accurate freezing of gait detection. Alexandria Eng J 2024;86:481–93. https://doi.org/10.1016/j.aej.2023.11.075.

[3] Sun R, Hu K, Martens KAE, Hagenbuchner M, Tsoi AC, Bennamoun M, et al. Higher Order Polynomial Transformer for Fine-Grained Freezing of Gait Detection. IEEE Trans Neural Networks Learn Syst 2023;PP:1–14. https://doi.org/10.1109/TNNLS.2023.3264647.

[4] Yang P-K, Filtjens B, Ginis P, Goris M, Nieuwboer A, Gilat M, et al. Freezing of gait assessment with inertial measurement units and deep learning: effect of tasks, medication states, and stops. MedRxiv 2023:2023.05.05.23289387.

[5] Esfahani AH, Dyka Z, Ortmann S, Langendorfer P. Impact of Data Preparation in Freezing of Gait Detection Using Feature-Less Recurrent Neural Network. IEEE Access 2021;9:138120–31. https://doi.org/10.1109/ACCESS.2021.3117543.

[6] Guo Y, Huang D, Zhang W, Wang L, Li Y, Olmo G, et al. High-accuracy wearable detection of freezing of gait in Parkinson's disease based on pseudo-multimodal features. Comput Biol Med 2022;146:105629. https://doi.org/10.1016/j.compbiomed.2022.105629.

[7] Mikos V, Heng CH, Tay A, Yen SC, Chia NSY, Koh KML, et al. A Wearable, Patient-Adaptive Freezing of Gait Detection System for Biofeedback Cueing in Parkinson's Disease. IEEE Trans Biomed Circuits Syst 2019;13:503–15. https://doi.org/10.1109/TBCAS.2019.2914253.

[8] Langer P, Haddadi Esfahani A, Dyka Z, Langendörfer P. FPGA-Based Realtime Detection of Freezing of Gait of Parkinson Patients, 2022, p. 101–11. https://doi.org/10.1007/978-3-030-95593-9_9.

[9] TensorFlow. TensorFlow Lite n.d. https://www.tensorflow.org/lite (accessed April 17, 2022).

[10] Abbasimehr H, Shabani M, Yousefi M. An optimized model using LSTM network for demand forecasting. Comput Ind Eng 2020;143:106435. https://doi.org/10.1016/j.cie.2020.106435.

[11] Buisseret F, Dierick F, Van der Perre L. Wearable Sensors Applied in Movement Analysis. Sensors 2022;22:10–3. https://doi.org/10.3390/s22218239.

[12] Pardoel S, Kofman J, Nantel J, Lemaire ED. Wearable-sensor-based detection and prediction of freezing of gait in parkinson's disease: A review. Sensors (Switzerland) 2019;19:1–37. https://doi.org/10.3390/s19235141.