ATTENTION VISION TRANSFORMERS FOR HUMAN FALL DETECTION

Satyake Bakshi
Carleton University
Department of Systems and Computer Engineering
Ottawa
{satyakebakshi@cmail.carleton.ca}

ABSTRACT

In recent years, attention transformers have proven to be instrumental in natural language processing (NLP)-based tasks such as sentence classification and language translation. However, their application has been recently extended to large-scale object recognition tasks. In this work, Vision Transformer with attention has been investigated for the detection of human falls and ADLs (Activities of Daily Living) from time series-based signals. The Vision Transformer model has been trained and validated using the acceleration signals of waist-worn Inertial Measurement Unit (IMU) sensors obtained from the IMU Falls dataset[1]. The model is also trained and validated on the popular SiSFall dataset[2]. The model is also investigated by independently training three different cases of patch size and attention heads. It was observed that a larger patch size resulted in significant performance deterioration. Additionally, a smaller patch size took longer to train and was computationally expensive. The model performed (best case) with an Accuracy (%) of 99.9 ± 0.1 and a True Positive Rate (%) of 99.9 ± 0.1 on the SFU-IMU dataset and with an Accuracy (%) of 99.8 ± 0.25 and a true positive rate (%) of 99.87 ± 0.3 on the SISFALL dataset. Overall, the results show that Transformers are highly robust in the detection of human falls and nonfalls/ADLs, subject to the appropriate patch size.

Keywords Fall Detection · Vision Transformers · Deep Learning

1 Introduction

Falls, typically a rare event, have negative effects on the lives of individuals. Most of the methods in fall detection make use of deep learning, such as convolutional neural networks (CNNs) and long-short-term memory networks (LSTM). In deep learning, various methods are dependent on the availability of data sources. Generally, the data available in the event of falls are rare and are usually simulated in controlled laboratory settings. Due to this difficulty in collecting falls data, particularly for elderly subjects, fall detection often needs to be accomplished with a limited amount of data.

In such cases, few-shot learning architectures have proven to be robust in learning from the limited pool of data. Recently introduced Siamese architectures [3][4][5] have proven that these models can learn from a variety of data types, from time-series data to acoustic signals. If sufficient data are available, conventional state-of-the-art CNN (SOTA) architectures such as ResNet, DenseNet, and Inception have proven to be successful[6][7][8]. Synthetic data for activity recognition have also shown promising results with the use of unsupervised reconstructive algorithms, namely, variational autoencoders and generative adversarial networks. These algorithms can create synthetic reconstructions that can be used as a data augmentation strategy for further training of any deep learning-based model. To ensure that the synthetic data are representative, these models sample from the distribution within the layers of the network to ensure that the output generated is representative of the imbalanced class[9]. Apart from state of the art (SOTA) approaches, there are also classical approaches to detecting fall events using standard machine learning algorithms such as naive Bayes, Random forest, and SVMs[10]. However, it should be noted that most of the models that achieve state of the art (SOTA) performance revolve around the use of the convolution operation. It should also be considered that the convolution operation on images is computationally expensive, particularly when the kernel sizes and the

number of kernels are large and are stacked back to back, as seen in the case of architectures such as VGG16[11]. To combat this, researchers have experimented with simpler convolutional models using smaller kernel sizes. In a previous work, the author showed the use of 1 × 1 kernels in Siamese[3] to reduce the number of parameters and also to ensure robust detection of falls. However, there is still an important issue relating to the dependency and the relationship within the data points, which are not captured. To solve this problem, the use of transformers in NLP to capture the intent of sentences and the relationship between words has been a primary motivation for this paper. Transformers can effectively learn inductive biases depending on the objective task by using the concept of attention. In simple terms, attention can be viewed as a weighted average of inputs[12]. Attention models aggregate information to form context-encoded vectors. This has been shown to outperform older approaches for sentence classification, which made use of conventional RNNs/LSTMs[13]. The use of transformers to capture these dependencies in images has recently increased[14]. In one study [15] it is observed that transformers have outperformed the popularly used ResNet architecture by a significant margin when trained on a large pool of data. Transformers have also been used in the classification of human activities based on the acceleration data from smartphones[16]. Transformers have not yet been investigated for the detection of fall events based on time series data from waist-worn IMU sensors, and therefore this paper considers the use of attention-based Vision Transformers for the detection of fall events.

The paper is organized as follows: Section II elaborates on the dataset(s) considered for the work. Section III contains details of the preprocessing steps. Section IV elaborates on the proposed model. Section V shows the training and performance of the model. Section VI concludes the work with future directions.

2 Dataset

The IMU Dataset[1] has been used to train the model. The dataset contains acceleration data of 10 subjects, healthy young adults with ages between 22 to 32 years. The subjects were students at Simon Fraser University. The signals were sampled at 128 Hz. Each subject underwent 60 trials, 15 activities of daily living (ADL) tests, 24 falls, and 15 near falls. Each subject underwent 60 trials (15 Activity of Daily Livings ADLs, 24 Falls and 15 Near Falls). The experiment environment and scenario were designed to generate many activity primitives realistically. The dataset contains signals of IMU sensors obtained from the ankle, thigh, sternum, waist, and head. For this work, the accelerometer signals of the waist area have been chosen, as previous work suggests that the IMU sensors in the waist region offered the best performance [17]. The other dataset used was the SISFALL dataset[2] which contained activities related to 15 different types of falls and 15 types of nonfalls of both older people and the young. The signals in this dataset were also acquired from a waist-worn accelerometer at a 200Hz sampling rate.

3 Preprocessing

The accelerometer signals in the x, y and z directions $a_x(n)$, $a_y(n)$, $a_z(n)$ respectively were summed to obtain s(n) and the short-time Fourier transform (STFT) of s(n) was obtained as shown below.

$$S(n, \omega) = \sum_{m=-\infty}^{+\infty} s(m)w(n - m)e^{-i\omega m}, \tag{1}$$

where

$$w(n) = 0.5 - 0.5\cos\left(\frac{2\pi n}{M - 1}\right), \quad 0 \leq n \leq M - 1. \tag{2}$$

X(n, ω), is the STFT representation of a signal. w(n) is the causal von-Hann window of length 50 with 50% overlap to capture the transition between the human participants' activities within a short time interval. Also, the activity before the fall and after the fall event is important to understand when a fall truly happens. The magnitude of S(n, ω) was concatenated to generate a time-frequency map of the entire signal. This resulted in an STFT spectrogram of dimension 26 × 121 for the SISFALL dataset and a dimension of 26 × 78 for the SFU-IMU dataset.

Figure 1 shows a spectrogram of a sample fall signal and a non-fall signal. In this figure, the x-axis denotes the time in seconds, while the y-axis represents the frequency in Hz. This raw input spectrogram was used as an input to the proposed model.

4 Proposed Vision Transformer

The transformer encoder is the core backbone of the Vision Transformer, however, the input to the transformer requires some transformation before it can be passed through[14]. The transformer normally receives input in the form of a 1D

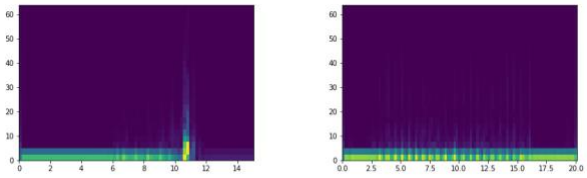

Figure 1: Spectograms showing sample Fall (left) and non-fall (right) event .

sequence of vectors. In order to take in a 2D input, the input of size H × W × N is reshaped to sequence of flattened patches $N \times P^2 \cdot N$, where (H, W, N) denote the height, width and the number of channels. (H, W) forms the resolution of the input. P is the resolution of each patch. n is the number of patches created, where n is given by $n = HW/P^2$.

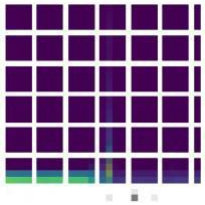

Figure 2: Patched Spectogram (for visualisation).

Figure 2 shows a visual representation of a patched spectrogram. The patches are flattened using a dense layer, resulting in an output of some dimension d. Let the output of the dense layer be $e_{flattened}$. Similarly to BERT, a learnable token embedding is included in the form of x_{class} that denotes the output. In addition to this, an embedding $p_{positional}$ is for each position in the patch sequence is learned and added to the output of the dense layer resulting in a z vector: $z = x_{class}, e_{flatten} + p_{positional}$. The z vector is obtained by using a patch encoder which includes a Dense layer and a positional embedding layer.

The transformer attention equation can be shown as follows: For sequences of length t and dimension d. A query sequence Q , a key sequence K and a value sequence V . Each head computes a weighted aggregation of V with respect to Q. This is shown in the following equation: Figure 3 shows the visual representation of this concept.

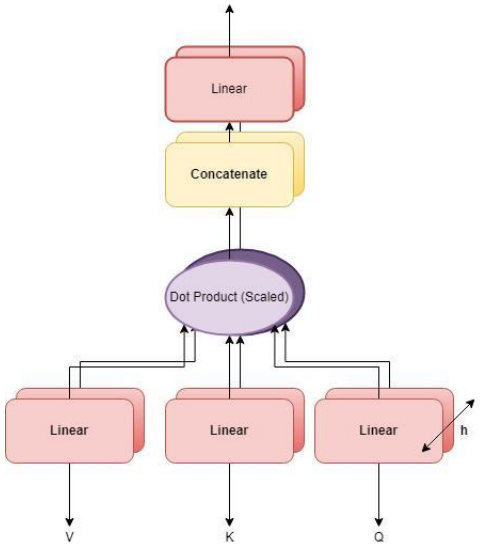

Figure 3: Multi-Head Attention Mechanism

$$h(Q,K,V)=\text{softmax}\left(\frac{Q_h(K^T)}{\sqrt{d}}\right)V_h \quad (3)$$

In the equation above , the term h represents the number of parallel attention layer also referred to as heads. The proposed transformer architecture used for this study has been shown in figure 4. The z vector is passed through this transformer encoder. The encoder uses Multi Headed Attention layer, this computes the averaged attention vector. The output of the attention block is then passed through the Multilayered Perceptron (MLP) which generates the encoded patches. These encoded patches are then flattened and passed through another MLP perceptron and then finally passed through a Dense Layer with a sof tmax activation function. The output of sof tmax produces probabilities for classification. There are three normalizations applied at the input, the output of the attention block, and before the flatten layer. There are two skip connections within the layers to avoid vanishing gradients.

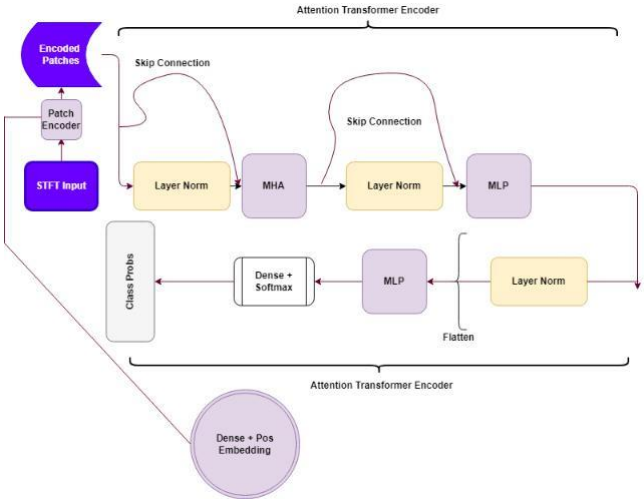

Figure 4: Proposed model structure.

5 Experimental Results

The model was trained for 100 epochs using the Adam optimizer. The learning rate was set at 0.001 and a weight decay rate of 0.0001. The batch size was set to 256. Callbacks were used to ensure that model validation loss is tracked throughout the training and the best weights were saved to ensure reproducibility. The model was trained on training-testing splits of 60:40 (60% of the dataset was attributed to training and 40 % of the dataset was attributed to testing). Due to resource constraints, the testing was divided into multiple random sampled trials. This was done to better approximate the performance of the model. The training and testing splits were sampled randomly from the dataset for 20 independent trials. In each of the trials, the training and testing splits were randomized by setting a unique seed parameter. The model is trained on each of these splits and tested. 15 % of the training data in each of the trials was reserved for validation. The model was trained and evaluated using multiple cases: patchsize = 4,8,12 and h = 4,8,12 The parameter h indicates the number of attention vectors generated. These vectors are weighted to generate the final attention vector, as discussed earlier. The number of parameters for each of these cases has been tabulated in table 1.

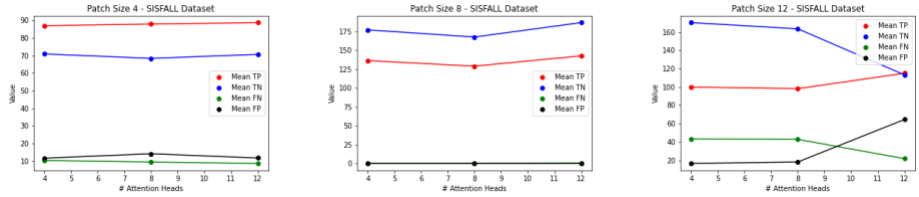

Figure 5: Case by Case Mean TP, FP, FN and TN (SISFALL)

Figures 6 and 5 show the visualization of the mean True Positive (TP), the mean True Negative (TN), the mean False Positive (FP) and the mean False Negative (FN) of the model in the SFU-IMU and SISFALL datasets.

Table 1: Parameters by patch size and #h heads.

CRITERIA (PATCH SIZE, #h)	# PARAMETERS
(4, 4)	11, 163, 842
(4, 8)	11, 694, 274
(4, 12)	12, 224, 706
(8, 4)	4, 878, 530
(8, 8)	5, 408, 962
(8, 12)	5, 939, 394
(12, 4)	3, 320, 258
(12, 8)	3, 850, 690
(12, 12)	4, 381, 122

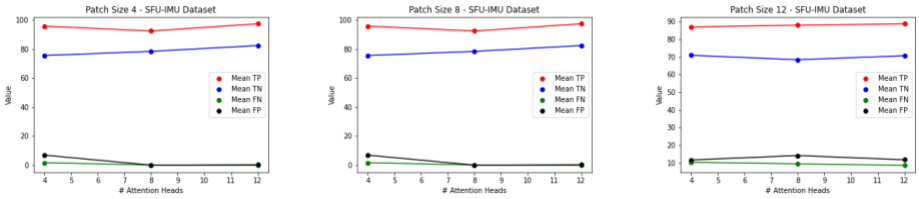

Figure 6: Case by Case Mean TP, FP, FN and TN (SFUIMU)

Table 3 and 2 show the Mean Accuracy and F score along with the variance in percentage for both the datasets. The metrics are computed accross 20 trials. It is observed that smaller patch sizes with a higher number of attention heads have resulted in better performance. A reduction in performance is observed as the patch size is increased to 12, regardless of the attention heads the performance becomes inconsistent. In the best case, the model performed with an F Score (%) of 99.9 ± 0.1 with a patch size of 8 and #h of 12 in the SFU IMU dataset and an F Score (%) of 99.87 ± 0.3 with a patch size of 8 and #h of 4 in the SISFALL dataset. The worst case, the model performed with an F Score (%) of 88.6 ± 2.2 with a patch size of 12 and #h of 8 in the SFU IMU dataset and an F Score (%) of 74.75 ± 7.7 with a patch size of 12 and #h of 12 respectively.

Smaller patch sizes with a larger number of heads were computationally expensive with longer training times compared to larger patch sizes. This was due to the fact that smaller patch sizes and increased number of attention heads resulted in more number of trainable parameters as seen in table 1. Overall, the proposed model shows good performance on both datasets, subject to the appropriate selection of patch size and number of heads.

6 Conclusion

Given the above results, Vision Attention transformers are observed to be robust in the problem of human fall detection. In this work, the binary classification instance of falls vs non-falls was considered using three different cases of patch size and the number of heads on the two datasets. The higher number of attention heads combined with a smaller patch size resulted in better performance. The future direction of this work would involve further investigation into the use of CNN in place of the MLP for the Vision Transformer. In addition, the possibility of an interclass classification of fall

Table 2: Accuracy and F score on SFU-IMU

(PATCH SIZE, #h)	Accuracy (%)	F Score (%)
(4,4) SFU IMU	95.17 ± 8.7	96.1 ± 5.4
(4,8) SFU IMU	99.83± 0.37	99.83 ± 0.4
(4, 12) SFU IMU	99.8 ± 0.27	99.82 ± 0.2
(8,4) SFU IMU	99.86 ± 0.3	99.8 ± 0.3
(8,8) SFU IMU	99.9± 0.2	99.9 ± 0.2
(8, 12) SFU IMU	99.9± 0.1	99.9 ± 0.1
(12, 4) SFU IMU	87.4± 6.2	88.6 ± 2.2
(12, 8)SFU IMU	86.8 ± 2	88.6 ± 2
(12,12) SFU IMU	88.6± 1.7	89.6 ±1.7

Table 3: Accuracy and F score on SISFALL

(PATCH SIZE, #h)	Accuracy (%)	F Score (%)
(4, 4) SISFALL	97.5 ± 2.05	97.1 ± 2.3
(4, 8) SISFALL	99.5± 0.6	99.5 ± 0.7
(4, 12) SISFALL	99.5± 0.6	99.4 ± 0.6
(8, 4) SISFALL	99.89± 0.25	99.87 ± 0.3
(8, 8) SISFALL	99.8± 0.2	99.7 ± 0.2
(8, 12) SISFALL	99.7± 0.2	99.6 ± 0.3
(12, 4) SISFALL	81.9± 1.4	76.9 ± 2.1
(12, 8) SISFALL	81.41 ± 1.4	76.5 ± 4.5
(12, 12) SISFALL	73.7 ± 12.3	74.75 ± 7.7

events could also be investigated. However, this work should serve as a pathway for the wide adoption of transformer models in the detection and monitoring of falls.

References

[1] Omar Aziz, Magnus Musngi, Edward J. Park, Greg Mori, and Stephen Neil Robinovitch. A comparison of accuracy of fall detection algorithms (threshold-based vs. machine learning) using waist-mounted tri-axial accelerometer signals from a comprehensive set of falls and non-fall trials. Medical & Biological Engineering & Computing, 55:45–55, 2016.

[2] A. Sucerquia, J. López, and J. Vargas-Bonilla. Sisfall: A fall and movement dataset. Sensors, 17:12, January 2017.

[3] Satyake Bakshi and Sreeraman Rajan. Few-shot fall detection using shallow siamese network. 2021 IEEE International Symposium on Medical Measurements and Applications (MeMeA), pages 1–5, 2021.

[4] S. Jeba Berlin and Mala John. Vision based human fall detection with siamese convolutional neural networks. Journal of Ambient Intelligence and Humanized Computing, pages 1–12, 2021.

[5] Diego Droghini, Stefano Squartini, Emanuele Principi, Leonardo Gabrielli, and Francesco Piazza. Audio metric learning by using siamese autoencoders for one-shot human fall detection. IEEE Transactions on Emerging Topics in Computational Intelligence, 5:108–118, 2021.

[6] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In AAAI, 2017.

[7] Mohamed Hammad, Pawel Plawiak, Kuanquan Wang, and U. Rajendra Acharya. Resnet-attention model for human authentication using ecg signals. Expert Systems, 38, 2021.

[8] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2261–2269, 2017.

[9] Dazhuo Wang, Jianfei Yang, Wei Cui, Lihua Xie, and Sumei Sun. Multimodal csi-based human activity recognition using gans. IEEE Internet of Things Journal, 8:17345–17355, 2021.

[10] Mrs. I. Varalakshmi, Ms. A. Mahalakshmi, and Ms P. Sriharini. Performance analysis of various machine learning algorithm for fall detection-a survey. 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), pages 1–5, 2020.

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2015.

[12] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. Multi-head attention: Collaborate instead of concatenate. ArXiv, abs/2006.16362, 2020.

[13] Thomas Hollis, Antoine Viscardi, and Seung Eun Yi. A comparison of lstms and attention mechanisms for forecasting financial time series. ArXiv, abs/1812.07699, 2018.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv, abs/2010.11929, 2021.

[15] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pretraining or strong data augmentations. ArXiv, abs/2106.01548, 2021.

[16] Yoli Shavit and Itzik Klein. Boosting inertial-based human activity recognition with transformers. IEEE Access, 9:53540–53547, 2021.

[17] Periklis Ntanasis, Evangelia Pippa, Ahmet Turan Özdemir, Billur Barshan, and Vasileios Megalooikonomou. Investigation of sensor placement for accurate fall detection. In MobiHealth, 2016.