# Attention Vision Transformers for Human Fall Detection

**Satyake Bakshi** ( satyakebakshi@cmail.carleton.ca )
Carleton University

**Research Article**

**Keywords:**

**Posted Date:** May 3rd, 2022

**DOI:** https://doi.org/10.21203/rs.3.rs-1614908/v1

**Noname manuscript No.**
(will be inserted by the editor)

# Attention Vision Transformers for Human Fall Detection

**Satyake Bakshi**

Received: date / Accepted: date

**Abstract** In recent years, Attention transformers have proven to be instrumental in Natural Language Processing (NLP) based tasks like 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 accelerometer 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 popularly used SiSFall dataset[15]. The model is also investigated by independently training on 3 different cases of patch size and attention heads. It is observed that a larger patch size has resulted in significant performance deterioration. Additionally, smaller patch size took longer to train and was computationally expensive. The model performed (best-case) with an Accuracy (%) of $99.9 \pm 0.1$ and a True Positive Rate (%) of $99.9 \pm 0.1$ on the SFU-IMU dataset and with an Accuracy (%) of $99.8 \pm 0.25$ and a True Positive Rate (%) of $99.87 \pm 0.3$ on the SISFALL dataset. Overall, the results show that Transformers are highly robust in the detection of human falls and non-falls/ADL subject to the appropriate patch size.

**Keywords** Sensor signal processing, Transformers, Fall Detection

Satyake Bakshi
Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada
E-mail: satyakebakshi@cmail.carleton.ca

## 1 Introduction

Falls, typically a rare event, has negative effects on the life of individuals. Most of the methods in fall detection make use of deep learning such as convolutional neural networks (CNN), Long short-term memory networks (LSTM), etc. In deep learning, various methods are dependent on the availability of data sources. Generally, the data available in the case of falls are rare and is usually simulated in controlled lab settings. Due to this difficulty in data collection of falls particularly for elderly subjects, fall detection 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. Few-shot Siamese networks have been a success[8][2]. Recently introduced Siamese architectures [3][4] have proved that these models can learn from a variety of data types, from time-series data to acoustic signals. If there are sufficient data available conventional state of the art (SOTA) CNN architectures like ResNet, DenseNet, Inception have proven to be a success[16][9][11]. Synthetic data for activity recognition has also proven to be a success, with the use of unsupervised reconstructive algorithms, namely Variational Autoencoders and Generative Adversarial Networks. These algorithms can create synthetic reconstructions which can be used as a data augmentation strategy for further training any deep learning-based models. To ensure that the synthetic data is representative, these models sample from distribution within the layers of the network to ensure that the output generated is representative of the imbalanced class[18]. Apart from the SOTA approaches, there are also the classical approaches to detecting fall events by using standard machine learning algorithms like Naive Bayes,

Random-forest, SVMs[17]. However, it is to be noted, that most of the models attaining SOTA performance, revolve around the use of the Convolution operation. It is to be also considered that 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 like the VGG16[14]. To combat this, researchers have experimented with simpler convolutional models using smaller kernel sizes. In earlier work, the author has shown the use of $1 \times 1$ kernels in Siamese[3] for the detection of falls. The approach resulted in robust performance with limited data. However, there is still an important issue relating to the dependency and the relationship within the data points, which are not modeled. To solve this problem, the use of transformers in NLP to capture sentence intent and the relationship between the words has been a prime motivation for this paper. Transformers can effectively learn the 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[6]. 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[10]. The use of transformers to capture these dependencies in images has recently picked up pace [7].In one study [5] 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 smartphone acceleration data[13]. 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 in the following manner. Section 1 elaborates on the dataset considered for the work. Section 2 contains details of the proposed model and also elaborates briefly on the concept of transformers. Section 3 explains the preprocessing. Section 4 shows the training and the performance of the architecture. Section 5 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 Activity of Daily Living (ADL)s 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 from the waist area have been chosen as past work suggests that IMU sensors in the waist region offered the best performance [12]. The other dataset used was the SISFALL dataset[15] which contained activities relating to 15 different types of falls and 15 types of non-falls of both elderlies and the youth. 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

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

$X(n,\omega)$, 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 $S(n,\omega)$ was concatenated to generate a time-frequency map of the entire signal. This resulted in an STFT spectrogram of dimension $26 \times 121$ for the SISFALL dataset and a dimension of $26 \times 78$ for the SFU-IMU dataset.

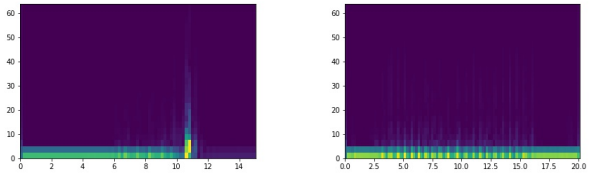

**Fig. 1** Spectograms showing sample Fall (left) and Non Fall (right) event .

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 has been 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[7]. The transformer normally receives input in the form of a 1D sequence of vectors. In order to take in a 2D input, the input of size $H \times W \times N$ is reshaped to sequence of flattened patches $N \times \left(P^2 \cdot N\right)$, 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 $n = HW/P^2$.

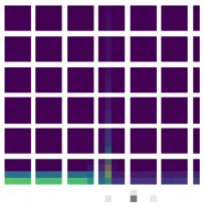

**Fig. 2** Patched Spectogram (for visualisation).

Figure 2 shows a visual representation of a patched Spectrogram. The patches are flattened by using a dense layer resulting in an output of $d$ dimension. Let the output of the dense layer be $e_{flattened}$. Similar to BERT, A learnable token embedding is included in the form of $x_{class}$ which 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: $\mathbf{z} = \mathbf{x}_{class}, e_{flatten} + p_{positional}$. 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 weighed aggregation of V with respect to Q. This is shown in the equation below: Figure 3 shows the visual representation of this concept.

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

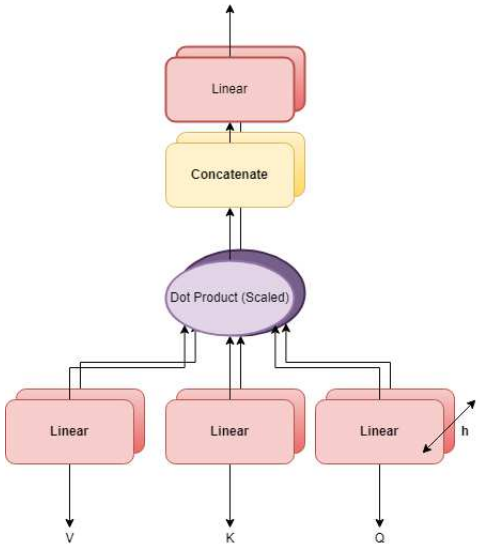


**Fig. 3** Multi-Head Attention Mechanism

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 MLP which generates the encoded patches. These encoded patches are then flattened and passed through another MLP and then finally passed through a Dense Layer with a $softmax$ activation function. The output of the $softmax$ yields 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.

## 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 at 256. Callbacks were used to ensure that the model validation loss is tracked though-out the training and to account for overfitting. 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 randomly sampled trials. This was done to better approximate the model performance. The training and testing splits were sampled randomly from the dataset for 20 independent trials. In

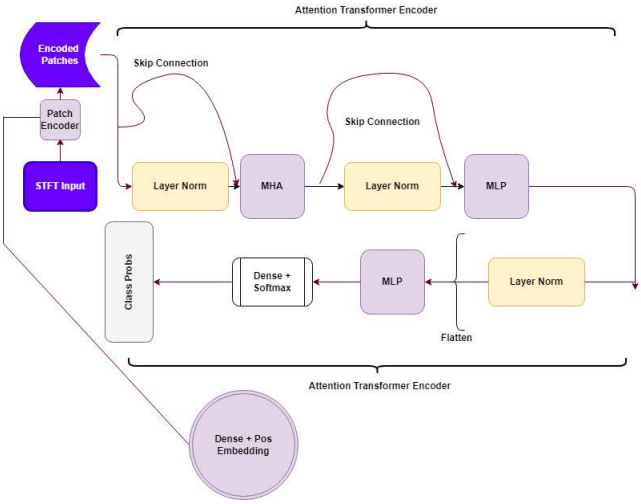


**Fig. 4** Proposed model structure.

**Table 1** Parameters by patch size and $\#h$ heads.

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

**Table 2** Accuracy and F score on SFU-IMU

| (PATCH SIZE, #h) | Accuracy (%) | F Score (%) |
|---|---|---|
| $(4, 4)$SFUIMU | $95.17 \pm 8.7$ | $96.1 \pm 5.4$ |
| $(4, 8)$SFU IMU | $99.83 \pm 0.37$ | $99.83 \pm 0.4$ |
| $(4, 12)$SFU IMU | $99.8 \pm 0.27$ | $99.82 \pm 0.2$ |
| $(8, 4)$SFUIMU | $99.86 \pm 0.3$ | $99.8 \pm 0.3$ |
| $(8, 8)$SFU IMU | $99.8 \pm 0.2$ | $99.7 \pm 0.2$ |
| $(8, 12)$SFU IMU | $99.9 \pm 0.1$ | $99.9 \pm 0.1$ |
| $(12, 4)$SFU IMU | $87.4 \pm 6.2$ | $88.6 \pm 2.2$ |
| $(12, 8)$SFU IMU | $87.3 \pm 2.4$ | $89.6 \pm 2.5$ |
| $(12, 12)$ SFU IMU | $88.6 \pm 1.7$ | $89.6 \pm 1.7$ |

each of the trials, the training and testing splits were randomized by setting a unique seed parameter. This was done to ensure that no two trials were the same. 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 $h$ parameter 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 shows the visualization of the True Positive Rate and the F score performance of the model on the

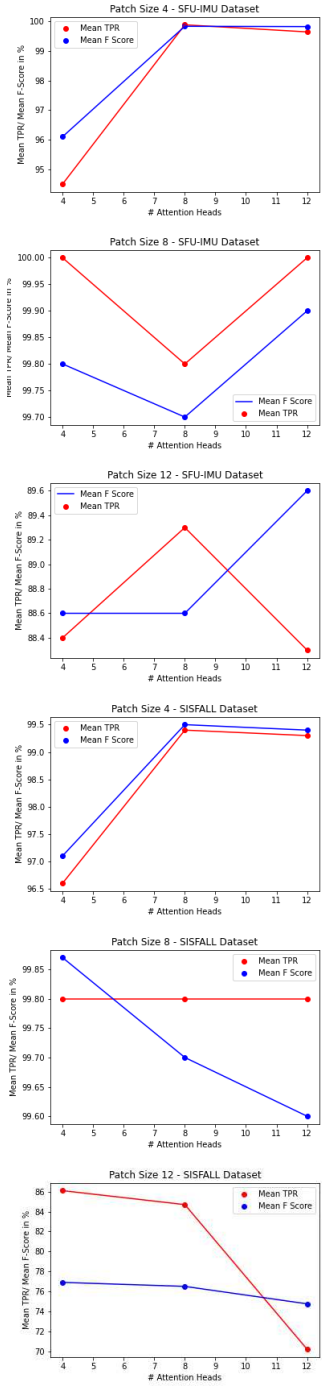


**Fig. 5** Plot of F Score and TPR by cases.

IMU signals from the SFU-IMU dataset and the SiSFall dataset using the conditions as specified.

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 across 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

**Table 3** Accuracy and F score on SISFALL

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

is increased beyond 8, regardless of the attention heads the performance becomes inconsistent. In the best case the model performed with an F Score of $99.9\% \pm 0.1$ with a patch size of 8 and $\#h$ of 12 on the SFU IMU dataset and an F Score of $99.87\% \pm 0.3$ with a patch size of 8 and $\#h$ 4 on the SISFALL dataset.

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 resulted in more number of trainable parameters as seen in table 1. Overall, the proposed model shows good performance on both of the datasets, subject to the appropriate 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 3 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. Also, the possibility of inter-class classification of fall events could also be investigated. Regardless, this work should serve as a pathway for the wide adoption of transformer models in the detection and monitoring of falls.

## References


1. Aziz, O., Musngi, M., Park, E.J., Mori, G., Robinovitch, S.N.: 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. Bakshi, S., Rajan, S.: Fall event detection system using inception-densenet inspired sparse siamese network. IEEE Sensors Letters **5**, 1–4 (2021)
3. Bakshi, S., Rajan, S.: Few-shot fall detection using shallow siamese network. 2021 IEEE International Symposium on Medical Measurements and Applications (MeMeA) pp. 1–5 (2021)
4. Berlin, S.J., John, M.: Vision based human fall detection with siamese convolutional neural networks. Journal of Ambient Intelligence and Humanized Computing pp. 1–12 (2021)
5. Chen, X., Hsieh, C.J., Gong, B.: When vision transformers outperform resnets without pretraining or strong data augmentations. ArXiv **abs/2106.01548** (2021)
6. Cordonnier, J.B., Loukas, A., Jaggi, M.: Multi-head attention: Collaborate instead of concatenate. ArXiv **abs/2006.16362** (2020)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv **abs/2010.11929** (2021)
8. Droghini, D., Squartini, S., Principi, E., Gabrielli, L., Piazza, F.: 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)
9. Hammad, M., Plawiak, P., Wang, K., Acharya, U.R.: Resnet-attention model for human authentication using ecg signals. Expert Systems **38** (2021)
10. Hollis, T., Viscardi, A., Yi, S.E.: A comparison of lstms and attention mechanisms for forecasting financial time series. ArXiv **abs/1812.07699** (2018)
11. Huang, G., Liu, Z., Weinberger, K.Q.: Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2261–2269 (2017)
12. Ntanasis, P., Pippa, E., Özdemir, A.T., Barshan, B., Megalooikonomou, V.: Investigation of sensor placement for accurate fall detection. In: MobiHealth (2016)
13. Shavit, Y., Klein, I.: Boosting inertial-based human activity recognition with transformers. IEEE Access **9**, 53540–53547 (2021)
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2015)
15. Sucerquia, A., López, J., Vargas-Bonilla, J.: Sisfall: A fall and movement dataset. Sensors **17**, 12 (2017)
16. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI (2017)
17. Varalakshmi, M.I., Mahalakshmi, M.A., Sriharini, M.P.: Performance analysis of various machine learning algorithm for fall detection-a survey. 2020 International Conference on System, Computation, Automation and Networking (ICSCAN) pp. 1–5 (2020)
18. Wang, D., Yang, J., Cui, W., Xie, L., Sun, S.: Multimodal csi-based human activity recognition using gans. IEEE Internet of Things Journal **8**, 17345–17355 (2021)