Action Recognition Using Vision Transformers (ViT)

Mohansree Vijayakumar[1] mv00582@surrey.ac.uk
Saizalpreet Kaur[2] zs00774@surrey.ac.uk
Ilakkiyavarsha Kannappan[3] ik00660@surrey.ac.uk
Remith Ravi[4] rr01073@surrey.ac.uk

Abstract

Recent progress in deep learning has enabled the development of advanced video analysis systems capable of capturing complex spatiotemporal patterns for effective classification tasks. Through this project, we investigate the application of TimeSformer[11]- a Transformer-based architecture, for human action recognition. The dataset used is a subset of HMDB51, containing 25 action classes. The model is trained and fine-tuned with various hyperparameter settings. The best combination of hyperparameters is used to carry out experiments, including different models, preprocessing techniques and sampling rates. Future work can include optimization efficiency using techniques like cascaded group attention from EfficientViT.

Keyword - Vision Transformer ,Data Efficient Image Transformer, Multi-layer perception,Stochastic Gradient Descent, Adaptive Moment Estimation

1. Introduction

Unlike traditional image classification, video classification presents unique challenges due to the need to capture both spatial and temporal dynamics present in the videos. These challenges have led to the development of advanced deep learning architectures that can effectively model the intricate interplay between appearance and motion. Transformer-based models have recently gained prominence in the field of video classification, offering an alternative to conventional convolutional approaches. The TimeSformer architecture exemplifies this shift by employing self-attention mechanisms across both spatial and temporal dimensions, thereby eliminating the need for convolutions and enabling the model to capture long-range dependencies within video data. This feature allows TimeSformer to efficiently process sequences of frames spatially and temporally, making it well-suited for video action recognition tasks where temporal context is critical.
The primary difference between a Vision Transformer (ViT) and a TimeSformer is how they utilise temporal information. The ViT architecture being built for image classification does not use temporal information; hence, it divides the image into fixed-size patches and applies spatial self-attention to obtain the intra-frame relationships [8]. Whereas, TimeSformer uses divided space-time attention to deal with videos. This division allows TimeSformer to represent short-term and long-range temporal relationships with improvements in processing and accuracy over action recognition benchmarks.

In this project, the TimeSformer400[11] model was adopted for video classification. This variant of TimeSformer is pretrained on the Kinetics-400 dataset, which comprises 400 action categories. The target dataset for this study, "HMDB_simp," consists of 1,250 videos distributed across 25 action categories. Following the preprocessing guidelines outlined in the original TimeSformer paper, each video was segmented into clips of size 8×224×224, with frames sampled at a rate of 1/32 to ensure temporal consistency and manageable input dimensions. The experimental phase of the project focused on systematic hyperparameter tuning to assess their impact on model convergence and accuracy. Comparative benchmarking with other transformer-based architectures provided additional insights into the strengths and limitations of the TimeSformer model. The project aims to demonstrate the efficacy of TimeSformer in action recognition on the HMDB_simp dataset for various setups.

2. Literature Review

Dosovitskiy et al. [1] introduced the Vision Transformer (ViT), a foundational model that applies transformers to image recognition by dividing images into 16x16 patches and using self-attention to capture intra-frame relationships. This work provides the theoretical basis for transformer-based architectures in vision tasks, which TimeSformer builds upon.
Bertasius et al. [2] proposed TimeSformer, a space-time attention model specifically designed for video understanding, eliminating the need for convolutions by applying self-attention across both spatial and temporal dimensions. TimeSformer's ability to capture long-range dependencies makes it well-suited for action recognition tasks where temporal context is critical. The paper's preprocessing guidelines were adopted to ensure temporal consistency.
Touvron et al. [8] proposed DeiT, a data-efficient training method for ViT using distillation, emphasizing efficient transformer training for image tasks. While focused on images, DeiT's training strategies (e.g., attention-based distillation) inform the project's hyperparameter tuning for TimeSformer, ensuring efficient convergence on the limited dataset. This helps achieve better performance with fewer computational resources.

3. Methodology

The proposed approach consists of four significant stages: video preprocessing and temporal augmentation, embedding patches, time-space attention modelling with the TimeSformer, and classification.

A. **Overview:** The proposed method aims to recognise human actions accurately from videos with a video-based action recognition transformer that models both the spatial

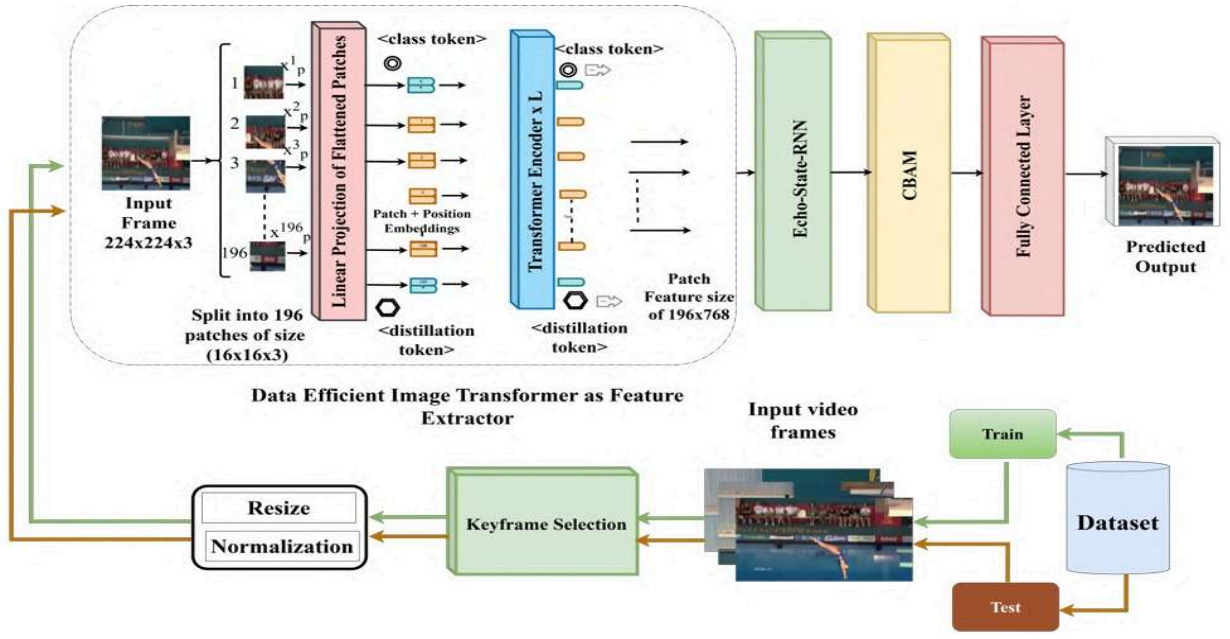


Fig 1: The overall proposed methodology for human activity recognition, containing data processing, frame selection, spatial and temporal feature extraction, and spatio-temporal attention mechanism described pictorially

and temporal features. The TimeSformer model is an extended ViT model that utilises a divided space-time attention to process the video inputs. Our pipeline is to simply preprocess the input videos into constant-size sequences of frames, apply temporal augmentations, and classify the sequence of frames using the TimeSformer network, for action recognition. An overview of the entire proposed framework is shown in Fig.1.

B. **Preprocessing and Temporal Augmentation:** The videos in the HMDB-simp dataset were transformed into sequences of 8 frames, resized to 224 × 224 pixels. Frames were sampled at a fixed frame rate of 1/32, resulting in a consistent temporal sampling structure across all actions. To meet these requirements and to increase the temporal robustness, we experimented with three different combinations of data augmentation:

Frame Padding: If after the sampling process, the number of frames remains less than 8, the whole sequence of frames is repeated and padded after the last frame till the desired number of frames is achieved.

Temporal Reversal, Frame Averaging and Flickering: For 2 or more sampled frames, frame averaging is used to average two consecutive frames to mimic motion blur. If the desired number of frames is not achieved, we use temporal reversal to reverse the frame sequence, which helps the model learn bidirectional temporal dependencies. If we get only a single frame after sampling, image augmentation- Flickering is used to vary the brightness levels until 8 frames are achieved.

Optical flow-based interpolation and Image Augmentation: Optical flow-based interpolation generates synthetic intermediate frames by estimating the pixel motion between consecutive sampled frames. For a single frame achieved after sampling, random application of colour jittering, random crop, horizontal flip, rotation and Gaussian noise is applied till we get 8 frames.

C. **TimeSformer Architecture:** Our framework primarily uses TimeSformer [1], a transformer-based model built for video understanding. Unlike standard CNNs or RNNs, TimeSformer learns space-time representations through an attention mechanism that operates in both space and time.

Patch Embedding: Each input video has *T* frames (for example, T=8), where each frame has size H×W. Each frame has non-overlapping square patches (for example, 16×16) which are flattened and linearly projected into a fixed-dimensional token embedding. This results in a sequence of patch tokens for each frame.

Positional Encoding: To encode spatial and temporal ordering, spatial-positional encodings are included to identify the patch location relative to all patches in each frame, and temporal positional encodings will identify the patch index in a sequence of frames. These embeddings help the model in distinguishing patches across time and space.

Divided Space-Time Attention: The attention-time implementation in TimeSformer is the divided attention method, where attention is calculated into two separate modules, one at a time.

Spatial Attention: This attention is used independently at each frame, imitating the interactions of the patches in the same frame.

$$Attention_t^{spatial} = \text{Softmax}\left(\frac{Q_t K_t^T}{\sqrt{d_k}}\right) V_t$$

where $Q_t K_t V_t \in R^{N \times D}$ represent the queries, keys, and values of the patches in frame t.

Temporal Attention: Temporal attention is conducted across the same spatial patch location found in different frames, as the model can learn motion and temporal progression.

$$Attention_n^{temporal} = \text{Softmax}\left(\frac{Q_n K_n^T}{\sqrt{d_k}}\right) V_n$$

where $Q_n K_n V_n \in R^{T \times D}$ correspond to the token embeddings at the same patch index across all frames.

This separation promotes memory efficiency and scalability without compromising modelling capacity. As with a Conventional Transformer, the attention blocks are followed by LayerNorm, residual connections, and feed-forward multi-layer perceptrons.

Classification Head: The token sequence has a learnable classification token [CLS] as a prefix, which acts as a robust aggregation of the global representation of the entire video input. After passing through the TimeSformer encoder blocks, the last embedding of the [CLS] token represents spatiotemporal features necessary for action classification.

This representation is passed through a Multi-Layer Perceptron (MLP) head, consisting of two fully connected layers with a non-linear activation in between, to produce the final logits for classification:

$$FFN(x) = GELU\left(x W_1 + b_1\right) W_2 + b_2$$

The output logits are passed through a softmax layer for classification :

$$\hat{y} = Softmax\left(FFN\left(x_{[CLS]}\right)\right)$$

The model is trained using cross-entropy loss:

$$l_{CE} = \sum_{i=1}^{C} y_i \log\left(\hat{y}_i\right)$$

Where C is the number of classes, $\hat{y}_i$ *is the true label*, *and* $\hat{y}_i$ *is the predicted probability*

D. **Baseline Models:** To show the effectiveness of TimeSformer we implement Vision Transformer (ViT) and Data-Efficient Image Transformer (DeiT) as baseline models that only model static frame-level information without explicit temporal modeling capability, which limits capabilities for motion-dependent tasks. The importance of spatiotemporal attention in TimeSformer is observed throughout the comparative study.

4. Experimentation and analysis

In this section, we describe the experimental framework we employed to analyse the performance of our proposed TimeSformer-based framework. All experiments were conducted with the HMDB-simp dataset.. We examine the effects of various data augmentations, model configurations, hyperparameter tuning, and frame sampling. We also include TimeSformer400 in this set of results, along with two other baseline transformer-based approaches: the Vision Transformer (ViT), and the Data-Efficient Image Transformer (DeiT), to help illustrate the benefit of applying spatiotemporal attention to spatiotemporal-based action recognition.

4.1 **Hyperparameter tuning** was done using the combination of Temporal Reversal, Frame Averaging and Flickering augmentations, using the TimeSformer-400. An 80-10-10 split was used for training, validation and testing. The loss function used throughout the experimentation was Cross-Entropy Loss. The best results are obtained for a batch size of 8, SGD optimiser, and a learning rate of 0.001, evaluated for 10 epochs. The momentum value of 0.9 for SGD gives the best results. TimeSformer seems to benefit from SGD's tendency to explore flatter minima, as it might help the model capture more robust and generalizable temporal patterns[8].

S.No	BS	Optimiser	Epoch	LR	Top-1	Top-5
1	4	ADAM	10	0.00001	83.2	93.6
2	4	ADAM	10	0.0001	86.4	96.8
3	4	ADAM	10	0.001	12.0	42.4
4	4	ADAM	10	0.0001	84.8	96.8
5	4	ADAM	10	0.0003	63.2	87.2
6	8	ADAM	10	0.0001	85.6	97.6
7	4	SGD	10	0.001	87.2	98.4
8	4	SGD	10	0.0001	84.0	96.0
9	8	SGD	10	0.001	88.0	98.4
10	8	SGD	10	0.003	80.8	94.4
11	8	SGD	20	0.001	88.0	96.8

Table 1: Hyperparameter Tuning

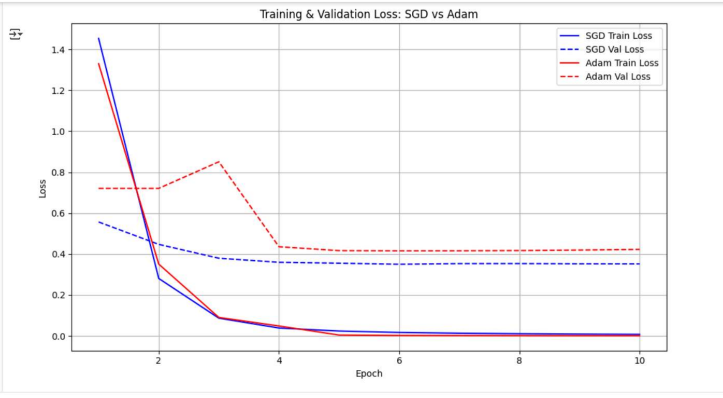


Fig 2: Loss Curve for the best configuration of SGD and ADAM

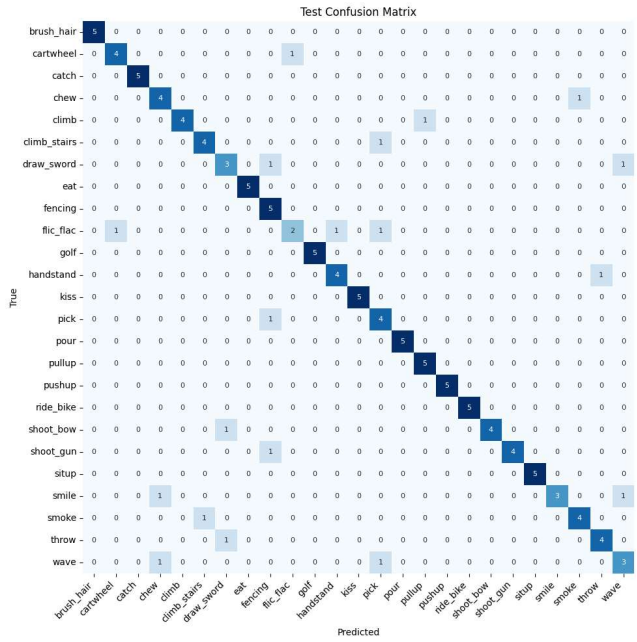


Fig 3: Confusion Matrix for the optimal configuration for the test set of 125 samples

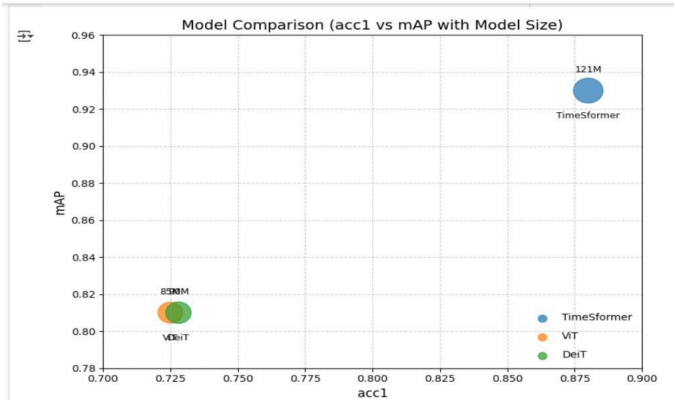
4.2 **Augmentation Techniques**: The data was pre-processed using the three augmentation techniques mentioned above. The best hyperparameters obtained in 4.1 were used for this experiment. The comparison of the results achieved by varying augmentation is stated:

S.no.	Augmentation	Top-1	Top-5
1	Padding	80.0	94.4
2	Temporal reverse+frame avg+ flicker	88.0	98.4
3	Optical flow interpolation + Image Augmentation	87.2	94.4

Table 2: Augmentation Technique Variation

Padding preserves original frames but lacks additional motion or context. The second technique is exposing the data to a broader range of temporal patterns during training, resulting in the mentioned accuracy. Optical flow interpolation enhances temporal consistency by interpolating frames, but the introduction of noise could be a cause of the reduced top-1 score.

4.3**: Model Benchmarks:** We examined the three transformer models, ViT, DeiT, and TimeSformer, for action recognition, where ViT and DeiT examine independent frames and do not have temporal modeling to leverage the motion dynamics of the action. On the other hand, TimeSformer uses a divided space-time attention block to model spatial features and temporal dependencies explicitly across the frames. TimeSformer achieved the best performance across all three models, demonstrating that it is more effective in capturing spatiotemporal patterns associated with video-based action recognition. The comparative performance results are summarised in Figure 2.



.Fig 4: Model Comparison

4.4 **Sampling Frame Rate Variation**: Besides the sampling rate of 32 provided to us, we experimented with two other values and got these outputs:

Sr. no.	Frame Rate	Top - 1	Top - 5
1	32	88.0	98.4
2	16	88.8	98.4
3	8	87.2	98.4

Table 3: Various Sampling rates

TimeSformer's space-time attention mechanism excels at modelling long-range dependencies. At the rate of 8, the Top-1 accuracy decreases, likely because the denser sampling focuses on a shorter temporal window, reducing the model's ability to capture longer-term dependencies critical for some actions. For 32, the dataset did not have enough frames for many videos, which led to the usage of augmentation, which might have led tothe disruption of TimeSformer's temporal modelling.

On randomly choosing the correct and incorrect predictions, the top 5 predicted classes for the action are shown below:

Sample 1 (Index: 48):
Actual Class: situp
Predicted Class: situp
Top-5 Predictions:

S:NO	Class	Prediction (%)
1	situp	99.29
2	pullup	0.14
3	brush_hair:	0.12
4	catch	0.09
5	pushup	0.07

Sample 1 (Index: 18):
Actual Class: handstand
Predicted Class: flic_flac

S:No	Class	Prediction(%)
1	flic_flac	49.66
2	handstand:	48.74
3	cartwheel:	0.90
4	pullup:	0.28
5	catch:	017

5. Conclusion and Future Work

Future work will focus on extending the TimeSformer model to larger and more diverse video datasets, such as UCF101, to better evaluate its generalisation capabilities [4]. To improve real-time applicability, we aim to explore lightweight transformer architectures like Video Swin Transformer [3] and MobileViT for efficient deployment. Additionally, integrating multimodal data such as audio or depth could enhance model robustness in complex environments. We also plan to investigate adaptive frame-rate sampling for better temporal modelling. We intend to incorporate attention-based interpretability methods, enabling deeper insights into how the model focuses on spatial and temporal cues during prediction. Model compression techniques such as pruning and knowledge distillation may further reduce computational cost [8]. These improvements aim to make the model more interpretable, scalable, and suitable for real-world action recognition tasks. This project demonstrated the efficacy of TimeSformer400 for action recognition on the Hmdb_simp dataset, achieving a peak Top-1 accuracy of 88.0% at a frame rate of 32. The findings underscore TimeSformer's robustness in capturing spatio-temporal dynamics and highlight the potential of tailored augmentation strategies to enhance model accuracy on small, realistic video
[2] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021.
.

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2021.

[2] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021.

[3] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Video Swin Transformer," *arXiv preprint arXiv:2106.13230*, 2021

[4] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "ViViT: A video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 6836–6846.

[5] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2014, vol. 27.

[6] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 4489–4497.

[7] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 6299–6308.

[8] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 10347–10357.

[9] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. arXiv preprint arXiv:1910.04867, 2019b.

[10] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2013, pp. 3551–3558