Action Recognition Using Vision Transformers

Saizalpreet Kaur (6915210)

https://github.com/Sayzal28/EEEM068_Action_Recognition

Abstract

This study explores the Vision Transformer, specifically Timesformer to deal with human action recognition. It demonstrates the effectiveness of the model by fine-tuning on a subset of HMDB51 with 25 classes and 50 videos in each class. The research included tuning the hyperparameters to find the best configuration, fine-tuning a timesformer pretrained on different datasets, and comparing it with other Vision Transformer for videos. The effect of different sampling techniques is also explored by using a random sampling, an equidistant and a uniform sampling. The best configuration has achieved a Top-1 score of 93.6% and a Top-5 accuracy of 99.2%. Key findings highlight the importance of using a relevant dataset pre-trained model and appropriate sampling.

1. Introduction

Human action recognition from videos is a reasonably complex but important part of Computer Vision. It is widely used in Security surveillance, healthcare and even sports. It requires the understanding of spatio-temporal patterns across different frames. Traditionally 2D Convolutional Neural networks have been used for image data but their inability to capture the temporal information between frames render them useless to work with video data. The introduction of 3D CNNs was a significant improvement because it was able to capture temporal dimensions along with the spatial convolutions. However 3D CNNs had difficulties with long range temporal dependencies which led us to Vision Transformers, adapted from the transformers from Natural Language Processing. ViTs demonstrate an equal or even better performance when compared to CNNs for image classification tasks. ViT uses an attention mechanism.

Building upon the Vit, Timesformer was introduced as a transformer architecture that deals with videos. It uses divided space-time attention, which processes spatial and temporal information separately. Timesformer improves the transformers' ability to deal with computational challenges while maintaining their ability to capture long-range dependencies.

In this project, the aim is to fine-tune the Timesformer using the HMDB_simp dataset containing 25 action classes, with 50 videos in each class. It is a subset of HMDB51. Throughout the project, experiments were done to improve the performance of the model.

2. Literature Review

Initially 2D CNNs were used to process video data by treating each frame independently, which worked fine for the spatial features but failed to efficiently capture the temporal relationship between the frames, which led to the loss of temporal context from the video. The next advancement made in the field were 3D CNNs[1]. These Neural Networks extend the 2D convolutions of a 2D CNN by adding a temporal dimension. They use 3D kernels, which cover height, width and time. 3D CNNs had limited temporal receptive fields, which limited the long-range temporal modeling.

The introduction of transformers in Natural language processing inspired the ViTs[2]. Vision Transformers reshaped the processing of image data by replacing the convolutions with self-attention mechanisms. They divide images into patches of fixed size. The patches are then flattened and are given a positional embedding to maintain the spatial information. These embeddings, along with a special token[CLS] are fed into encoder layers of the transformer where multi-headed self-attention is used to simultaneously model the relationship between all patches. The CLS token is used for image classification.

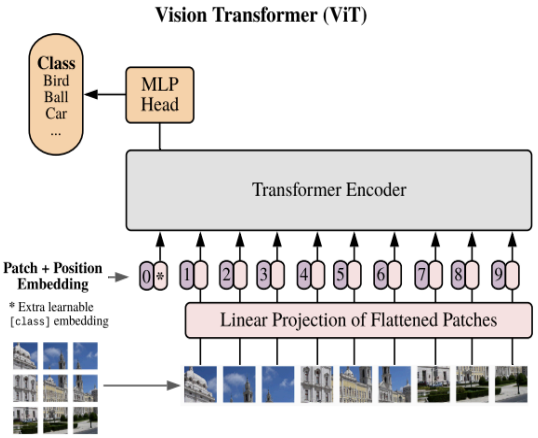

Figure 1: ViT [2]

Even though ViT's initial purpose was to deal with image data, attempts were made to use it for videos as well[3].

Timesformer[4] uses transformer architecture to deal with video data. It divides each frame of the video into a sequence of patches. The most important part of its implementation is the divided

space-time attention. It divides the standard self attention into a spatial and a temporal attention. Spatial attention is applied within each frame to model the relationships between patches in the same temporal point. Secondly, temporal attention is applied across different frames at the same spatial location to model the relationship between the patches. It was the best performing among the five attentions used in the original paper.
Each patch embedding includes both spatial and temporal positional encodings to preserve location and time information, and the model uses a classification token whose final representation is used for video-level predictions.
In the original paper[4], Timesformer is evaluated on four different datasets- Kinetics-400[5], Kinetics-600[6], Something-Something V2[7] and Diving 48. [5] and [6] contain 400 and 600 classes of human actions respectively. [7] contains clips of human-object interactions .[8] contains data about diving actions as suggested by the name.

3. Methodology

HMDB_simp, the dataset being used has 25 classes in total, with each class having 50 videos each, totalling to 1250 videos. It includes the frames for each video. The train:val:test split is 10% validation, and 80:20 for train:val of the remaining 90, resulting in 900/225/125 clips respectively. Class representation is preserved across all splits.
Timesformer expects 8 sampled frames[4], with a resolution 224*224, sampled uniformly at the rate 1/32. Using these conditions, it is observed that not every video has enough frames to sample at least 8 using the uniform sampling rate.When this happens, a sequential augmentation pipeline is applied. The augmentation techniques include frame interpolation/ averaging between two consecutive frames using pixel averaging; Temporal reversal sequences, which append reversed frames and a brightness variation, which applies factors of 0.6, 0.8, 1.2 and 1.4 to the existing frames. They are applied in this order until the required number of frames is met.
Sampling of the frames is done after the augmentations have been applied. Different sampling techniques are used along with the uniform sampling. Random sampling randomly selects frames from videos, avoiding repetition if the video has at least 8 frames. Whereas, Equidistant sampling systematically selects frames at regular intervals across the entire video duration to ensure temporal coverage.
The Timesformer model trained on Kinetics-400 is initialised from hugging face[9] and is fine-tuned on HMDB_simp dataset. The pre-trained model is adapted fine-tuning with custom classification head mapping to the 25 classes present in the dataset. Hugging Face trainer is used for the moodle training and testing. The logging is done using tensorboard. The basic hyperparameters are found after tuning:
Learning rate of 0.0035, a batch size of 8 samples for both training and testing phases, SGD optimizer with weight decay of 0.003 and 0.9 momentum and the number of epochs being 8. Images undergo standardised preprocessing including resize to 224*224, tensor conversion and normalisation.
Primary evaluation uses top-1 accuracy which measures the exact correctness of the prediction being made. Secondary evaluation used is top-5 accuracy which assesses whether the correct class appears in the top-5 predictions. The loss function used is cross-entropy.
Model performance is evaluated during every epoch while training. Final assessment on the test set is made using the best checkpoint.

4. Experimentation and Analysis

I. Hyperparameter tuning

While experimenting, different learning rates and batch sizes were tried out to find the best performing model. The two optimizers tried out are SGD and ADAM. Batch size of 16 could not be tried because of system restrictions.

S.No	BS	Optimiser	Epoch	LR	Top-1	Top-5
1	8	SGD	8	0.0035	93.6	99.2
2	8	SGD	8	0.0045	92.0	98.4
3	8	SGD	8	0.0007	90.4	97.6
4	4	SGD	8	0.0035	93.6	98.4
6	8	ADAM	8	0.0035	17.6	47.20
7	4	ADAM	8	0.0001	89.6	96.8

Table 1: Hyperparameter Tuning

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. SGD’s momentum based updates provide more stable optimisation of the pre-trained weights.

II. Different Pretrained Timesformers

S.No	Dataset	Optimiser	
		Top-1	Top-5
1	K-400	93.6	99.2
2	K-600	93.6	98.4
3	SSV2	84.8	96.8

Table 2: Fine-tuning the model trained on different datasets

The Kinetic-400 and Kinetic-600 datasets perform far better than SSV2. This can be due to the similarity of HMDB_simp dataset with both of these. SSV2 performs the worst because of the domain mismatch with the fine-tuned dataset. The SSV2 dataset focuses on human-object interactions while HMDB_simp dataset is mainly of human actions, regardless of the objects. The relatively smaller gaps in top-5 accuracy suggest that while SSV2 pre-training affects primary predictions, it can still capture relevant features within the top predictions.

III. Different Sampling Techniques

S.No	Sampling	Top-1	Top-5
1	Uniform(1/32)	93.6	99.2
2	Equidistant	92.8	99.2
3	Random	91.2	100

Table 3: Various Sampling Techniques

Uniform sampling performs the best of the three. It could be attributed to the fact that it maintains the temporal structure of the data.
While equidistant sampling ensures complete temporal coverage by dividing the video duration into equal segments, it may introduce temporal irregularities when video lengths vary significantly. This approach can lead to inconsistent frame intervals that disrupt the model's ability to learn stable temporal patterns.
The lack of temporal structure in random sampling means the model receives fragmented motion information, making it difficult to understand action progression and dynamics, which could be the reason behind the decrease in top-1
The convergence of top-5 accuracies suggests that all sampling methods capture sufficient discriminative information to distinguish between broad action categories

IV. Comparison with other models

Fine-tuned Timesformer is compared with another ViT dealing with videos- ViViT(Video Vision Transformer). The model with scaled dot product attention from hugging face is used for this task[11]. SGD with momentum is used as the optimiser. The model used is pre-trained on K-400 as well. Both models perform quite well.

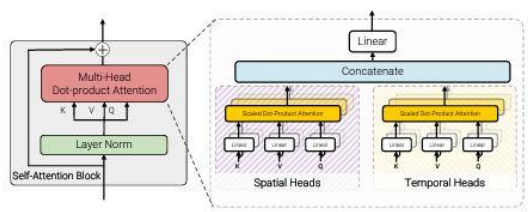

Fig. 2: ViViT with dot product attention [3]

S.No	Model	Top-1	Top-5
1	Timesformer	93.6	99.2
2	ViViT	88.8	99.2

Table 4: Comparison of two ViTs

V. Application of Sampling Augmentation

During experimentation, the impact of using Augmentation before and after sampling the frames is explored.
On augmenting the raw data, and then sampling we get a better top-1 score as compared to sampling first and then augmenting. The possible reason behind that is that augmenting before sampling helps us create more diverse frames. The other approach however, limits the augmentation to just a few frames (<8), which gives us a much smaller pool of frames to work with. Interpolation in that case gives us a very unrealistic motion.

S.No	Augmentation	Top-1	Top-5
1	Before Sampling	93.6	99.2
2	After Sampling	85.83	97.64

Table 5: Effect of augmentation order on performance of the model

The confusion matrix showing which actions were wrongly predicted during training:

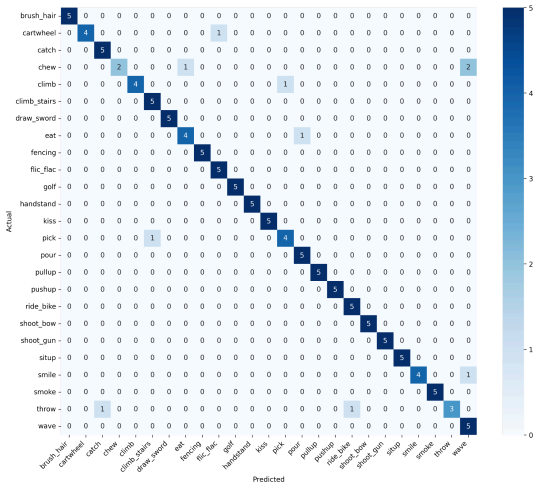

Fig. 3: Confusion Matrix for the baseline config

5. Interface Implementation

A web interface is developed using Streamlit which uses the weights of the pre-trained model and lets the user upload any video to classify. The limitation is that the classification will usually be correct only if the video belongs to one of the 25 classes . It shows the sampled frames and the prediction score along with the class prediction. It uses pre-trained weights from K400 if somehow the fine-tuned weights fail to load. The interface is shown in Appendix, Fig. 4 .The model works well and confidently on unseen data.

6. Conclusion and Future Work

Through this study, the effectiveness of Timesformer for Human Action Recognition could successfully be demonstrated on a subset of the HMDB51 dataset, achieving a 93.6% Top-1 accuracy and 99.2% Top-5 accuracy.

Uniform sampling at the rate of 1/32 turned out to be the best sampling technique, closely followed by equidistant sampling.

Timesformer models pre-trained on similar datasets (K-400,K-600) gave great accuracy, suggesting that the selection of a correct pre-training dataset is crucial for good results.

A web interface is implemented using Streamlit, showcasing the practical applicability of the fine-tuned model.

The attention map for one sample is given in the Appendix, which shows the localised action–relevant regions. Also, it can be observed from Fig. 6,7 from Appendix that the fine-tuned Timesformer can usually correctly predict the class. But, it can accurately predict the action from the video in top-5.

The future work inspired from this study could be fine-tuning on a larger dataset so the model can be generalised better, improving the computational efficiency by using distillation or quantisation.

References

[1] Tran, Du, et al. "Learning spatiotemporal features with 3d convolutional networks." *Proceedings of the IEEE international conference on computer vision*. 2015.

[2] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

[3] Arnab, Anurag, et al. "Vivit: A video vision transformer." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

[4] Bertasius, Gedas, Heng Wang, and Lorenzo Torresani. "Is space-time attention all you need for video understanding?." *Icml*. Vol. 2. No. 3. 2021.

[5] Kay, Will, et al. "The kinetics human action video dataset." *arXiv preprint arXiv:1705.06950* (2017).

[6] Carreira, Joao, et al. "A short note about kinetics-600." *arXiv preprint arXiv:1808.01340* (2018).

[7] Goyal, Raghav, et al. "The" something something" video database for learning and evaluating visual common sense." *Proceedings of the IEEE international conference on computer vision*. 2017.

[8] Li, Yingwei, Yi Li, and Nuno Vasconcelos. "Resound: Towards action recognition without representation bias." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

[9] https://huggingface.co/facebook/timesformer-base-finetuned-k400

[10] D. Tran, H. Wang, L. Torresani, J. Ray, Y. Lecun, and M. Paluri, “A Closer Look at Spatiotemporal Convolutions for Action Recognition.”

[11] https://huggingface.co/docs/transformers/en/model_doc/vivit

APPENDIX

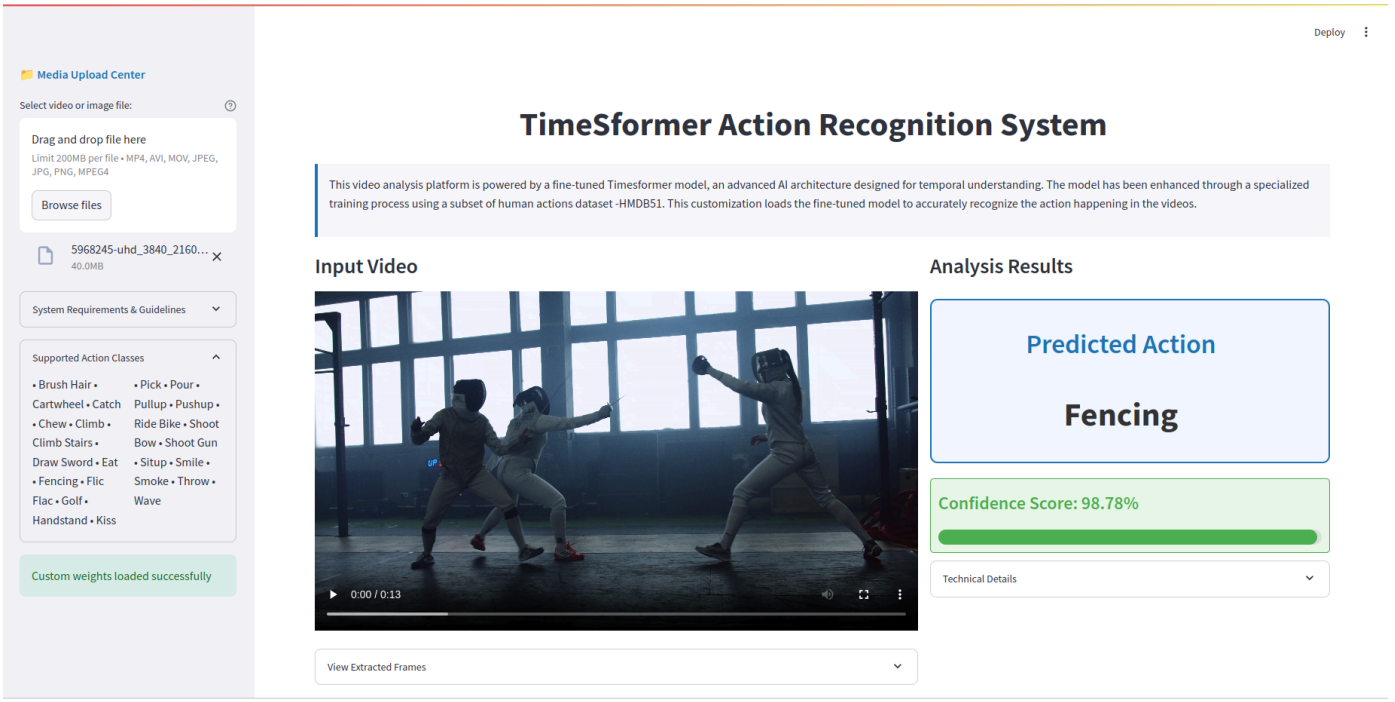

Fig. 4: Web-interface using Streamlit

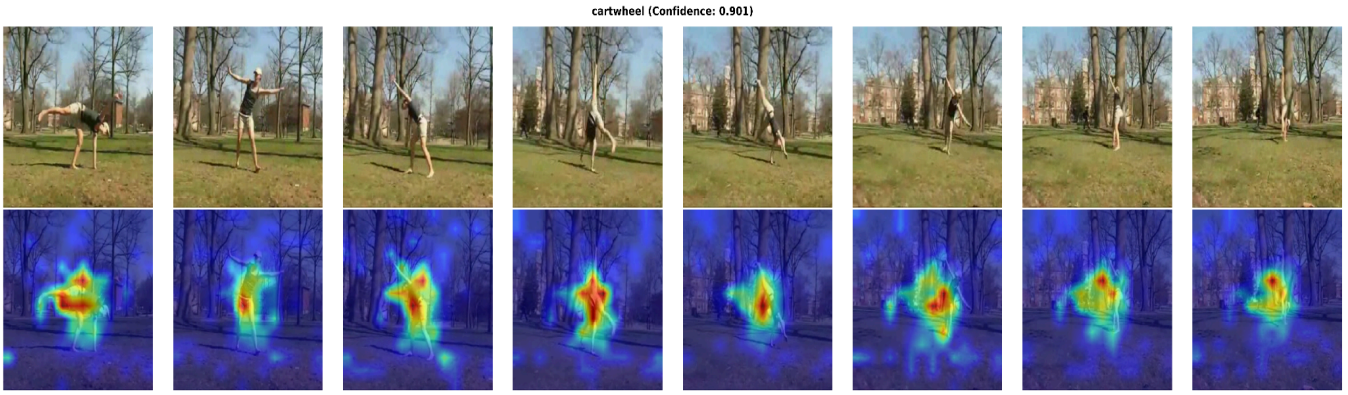

Fig. 5: Attention map for action: cartwheel

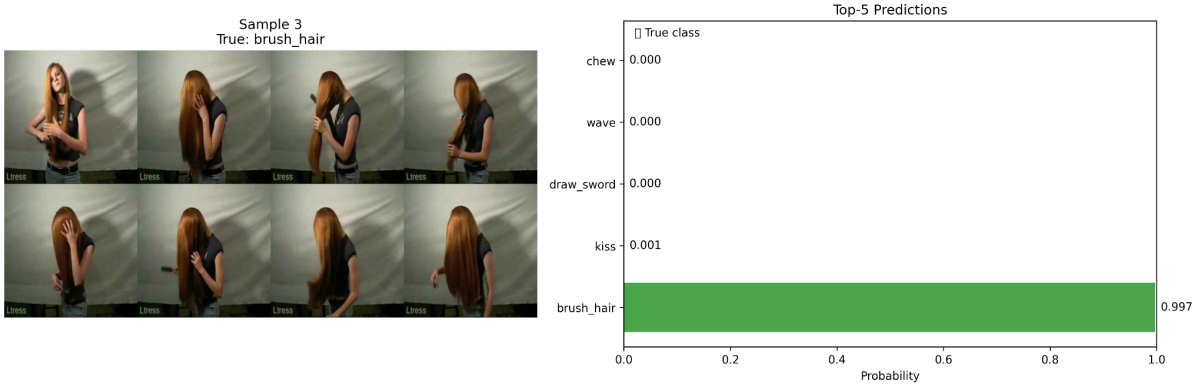

Fig. 6: Correctly predicted video with top-5 predictions

Fig. 7: Wrongly predicted video with top-5 predictions