YILDIZ TECHNICAL UNIVERSITY
FACULTY OF ELECTRICAL AND ELECTRONICS
DEPARTMENT OF COMPUTER ENGINEERING

ACTIVITY RETRIEVAL FROM VIDEOS

Engin Memiş, Elif Sena Yılmaz, Mine Elif Karslıgil
engin.memis@std.yildiz.edu.tr, sena.yilmaz4@std.yildiz.edu.tr, elif@yildiz.edu.tr

Özet

Aktivite çıkarımı, günümüzde video verilerinin hızla artışıyla beraber önemi artan yapay zekâ çözümlerinden biridir. Bu çalışmada, aktivite çıkarımı için yaygın olarak tercih edilen Kinetics400 veri seti kullanılarak 3 Boyutlu Evrişimsel Sinir Ağları (3DCNN) ve Video Vision Transformers (ViViT) karşılaştırılmıştır. Çalışmanın motivasyonu, video verisinin hızlı artışı nedeniyle bu videoların otomatik olarak etiketlenmesinin çeşitli sektörlere fayda sağlayacağı düşüncesidir. Bu iki model, veri setinin bir altkümesi ile denenmiş ve karşılaştırılmıştır. Çalışmanın sonucu, bu modellerde elde edilen başarıların çok farklı olmadığını, ancak kullanılacağı alana göre tercih yapılabileceğini göstermektedir.

Anahtar Kelimeler: aktivite çıkarımı, 3 Boyutlu Evrişimsel Sinir Ağları, Video Vision Transformers.

Abstract

Activity recognition is an artificial intelligence solution whose importance has increased with the rapid growth of video data in recent times. In this study, 3D Convolutional Neural Networks (3DCNN) and Video Vision Transformers (ViViT) were compared using the Kinetics400 dataset, which is widely used for activity recognition. The motivation of the study is the belief that automatically labeling these videos due to the rapid increase in video data will benefit various sectors. These two were tested by with a subset of this dataset and compared. The result of the study indicates that the successes achieved with these two different models are not significantly different, but the choice can be made depending on the area of application.

Keywords: activity recognition, 3D Convolutional Neural Networks, Video Vision Transformers.

I. Introduction

In recent years, the rise of video content on digital platforms has highlighted the importance of systems that can extract information from videos. Advances in artificial intelligence and deep learning have enabled the development of activity recognition systems, which analyze videos to identify activities. These systems are useful in various sectors, including healthcare for patient monitoring, security for detecting suspicious behavior, and sports for performance evaluation. Video processing involves preprocessing, feature extraction, and analysis, utilizing various machine learning and deep learning models. Traditionally, three-dimensional convolutional neural networks (3DCNN) have been widely used and proven effective. Recently, transformers, known for their success in natural language processing, have been adapted for image processing and are competing with traditional CNNs. This study aims to compare 3DCNN and transformer approaches to determine which is more advantageous.

II. System Design

To train video data using models, videos must first undergo pre-processing. Given that videos can vary in length and size, frames are selected according to the model architecture. These frames are then normalized, and pixel values are scaled to the range of zero to one. This ensures that videos are suitable for training. Due to limited resources, the models were tested on smaller datasets derived from Kinetics400, specifically using 20 classes with 400 videos each.

Traditional CNN models, initially developed for image processing, were adapted for video data but faced limitations due to the sequential nature of videos. While 2DCNNs were successful with static images, they fell short with sequential video data. In contrast, 3DCNN models, designed specifically for video classification, have been more successful [1]. The 3DCNN model used in this project, which includes three convolution layers, processes pre-processed RGB video data through 3x3x3 filters, followed by pooling layers, progressively extracting detailed features. The final layers are fully connected neural networks, reducing the data to the number of classes for classification.

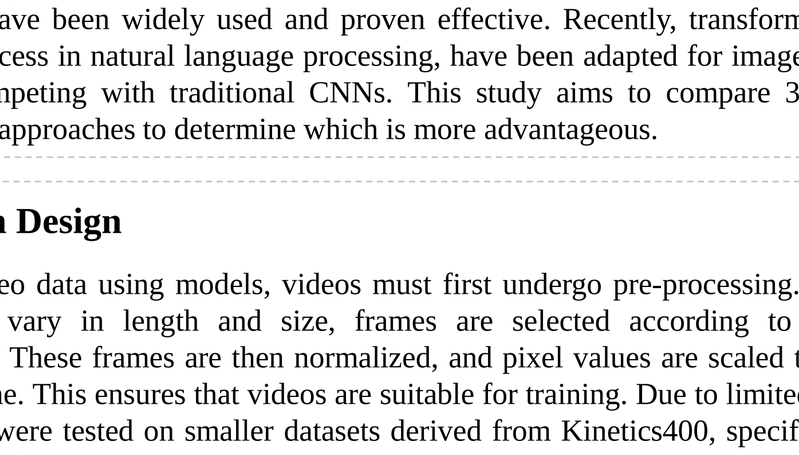

Figure 1. 3D CNN Structure [2]

Alternatively, the ViViT model, a transformer-based approach with an ‘attention’ mechanism, divides videos into smaller structures for detailed feature extraction. By utilizing Attention mechanisms, ViViT decodes relationships within videos, culminating in a flatten layer and a softmax layer to classify the video [4].

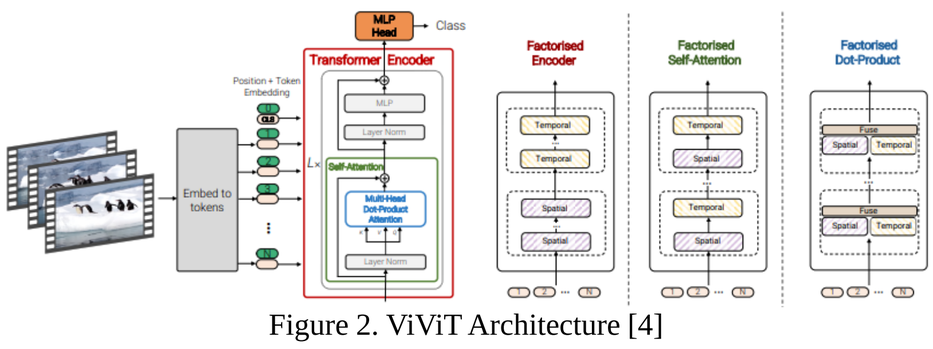

Figure 2. ViViT Architecture [4]

III. Experimental Results

The analysis revealed that the 3DCNN model is more resource-intensive than the ViViT model, requiring more trainable parameters and GPU memory, leading to longer training times. The 3DCNN model achieved an accuracy of 59.89%, compared to the ViViT model's 54.98%. Despite starting with a lower success rate, the ViViT model showed a steady increase in performance over 10 epochs. In contrast, the 3DCNN model reached near-peak performance by the 4th epoch and fluctuated within a narrow success range thereafter. The ViViT model's efficiency, suggests it could surpass the 3DCNN model with further training. Therefore, the ViViT model presents a viable option depending on resource constraints and specific application needs, despite the 3DCNN model's higher overall accuracy in this study.

Table 1. Results of Experiments

	ViViT	3DCNN
Accuracy	54.98%	59.89%
Trainable Parameters	52,585,748	822,373,652
GPU Usage	4 GB	21.8 GB
Training Time	309 min	569 min

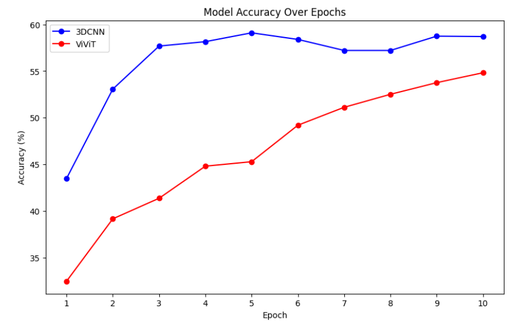

Figure 3. Accuracy Comparison

Table 2. F1-Scores For Each Class

	3DCNN	ViViT		3DCNN	ViViT
bench pressing	0.46	0.34	**presenting weather forecast**	0.89	0.86
bowling	0.48	0.38	**pull ups**	0.45	0.39
crawling baby	0.5	0.33	**riding mechanical bull**	0.64	0.56
cutting watermelon	0.57	0.56	**riding or walking with horse**	0.54	0.53
diving cliff	0.56	0.61	**shearing sheep**	0.55	0.32
filling eyebrows	0.68	0.55	**sled dog racing**	0.62	0.64
passing American football	0.77	0.76	**snowkiting**	0.75	0.71
picking fruit	0.75	0.73	**trapezing**	0.41	0.4
playing squash or racquetball	0.66	0.76	**tying tie**	0.51	0.51
playing tennis	0.59	0.64	**weaving basket**	0.5	0.4

The F1 score analysis shows that 3DCNN and ViViT models excel in different activities. The ViViT model performs better in activities such as "playing squash or racquetball" (0.76), "playing tennis" (0.64), and "sled dog racing" (0.64), while the 3DCNN model excels in activities like "filling eyebrows" (0.68) and "shearing sheep" (0.55). Although the 3DCNN model generally yields better results, the choice between these models should depend on the specific activities targeted. This highlights the importance of considering specific use cases when selecting a model for action extraction from videos.

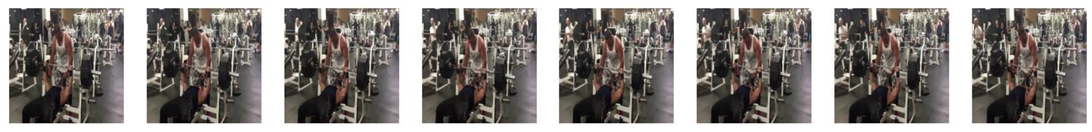

Figure 4. Bench Pressing Sample Video Frames

Table 3. Model Predictions for Bench Pressing

3DCNN		ViViT	
bench pressing	99.77%	**pull ups**	64.79%
pull ups	0.13%	**bench pressing**	29.64%
weaving basket	0.05%	**trapezing**	2.28%
shearing sheep	0.02%	**weaving basket**	1.77%
riding mechanical bull	0.01%	**shearing sheep**	0.85%

Figure 5. Sled Dog Racing Sample Video Frames

Table 4. Model Predictions for Sled Dog Racing

3DCNN		ViViT	
playing squash or racquetball	87.04%	**sled dog racing**	82.5%
trapezing	6.38%	**riding or walking with horse**	4.27%
riding mechanical bull	1.74%	**trapezing**	3.9%
playing tennis	0.99%	**snowkiting**	3.19%
passing American football	0.77%	**crawling baby**	1.85%

Conclusion

In this study, 3DCNN and ViViT models were analyzed for action extraction from videos. The 3DCNN model achieved a success rate of 59.89% after 10 epochs, while the ViViT model achieved 54.98%. However, the ViViT model demonstrated greater efficiency in terms of time and memory usage during training. It is suggested that with more extensive training, the ViViT model could potentially surpass the 3DCNN model, as it showed continuous improvement, unlike the fluctuating performance of the 3DCNN model. In conclusion, while the 3DCNN model was generally more successful, the ViViT model also showed superior performance in certain classes, suggesting that the choice between these models should consider resource availability and specific application needs.

References

[1] Helwan, A. (2023). Video classification using CNN and Transformer. Retrieved 3 December 2023, from https://abdulkaderhelwan.medium.com/video-classification-using-cnn-and-transformer-798b84c345bd
[2] Wang, C. (2023). A Review on 3D Convolutional Neural Network. 2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA), 1204–1208. doi:10.1109/ICPECA56706.2023.10075760
[3] SternDS. (2023). Video transformer(vit): A Deep Learning Model for video processing. Retrieved 3 December 2023, from https://medium.com/@nadav6stern/video-transformer-vit-a-deep-learning-model-for-video-processing-442268c8c3b4
[4] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). ViViT: A Video Vision Transformer. arXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/2103.15691