3D CNN with BERT and Vision Transformer for Video Recognition

Bao Thai Duong[1] and Thai Hoang Le[2,3](✉)

[1] Faculty of Information Technology, Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam
bao.dt@ou.edu.vn

[2] Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam
lhthai@fit.hcmus.edu.vn

[3] Vietnam National University, Ho Chi Minh City, Vietnam

Abstract. According to the development of the monitor system, detection and recognition are the major areas of interest within the field of computer vision. In recent years, due to their capacity to filter spatiotemporal video features, 3D CNN architectures with BERT have proven to be the best solution to this problem. Vision Transformer (ViT) has performed exceptionally well in recent benchmarks for image classification, object detection, and semantic image segmentation, among other computer vision applications. Transferring knowledge from such powerful ViT is an intriguing opportunity for developing excellent video recognition models. In this work, we discuss and evaluate the methods on HDMB-51 dataset to address the advantages and disadvantages. As a result, the study shows that two methods improve performance and accuracy of video recognition.

Keywords: Action Recognition · Video Recognition · Vision Transformers · 3D Convolution Neural Networks (3D CNNs)

1 Introduction

Action recognition is one of the most fundamental yet challenging tasks in video understanding. Human action recognition encompasses numerous computer vision research topics, such as medical supervision [17], micro video recommendation [18], autonomous driving [19], and so on.

Convolutional-based models that are optimized on the ImageNet dataset in a supervised fashion dominated this discipline over the past decade [1]. Based upon convolutional neural networks and now transformers, video recognition has achieved remarkable progress [20]. The CNN-based technique to extract meaningful characteristics from images is built around the convolution operation. Convolution operations include one-dimensional (1D), two-dimensional (2D), and three-dimensional (3D) convolution. Because of the importance of action recognition and other computer vision tasks in general, 2D and 3D convolution are naturally used more than 1D convolution for feature

P. Cong Vinh and N. Thanh Tung (Eds.): ICCASA 2023, LNICST 579, pp. 149–160, 2024.
https://doi.org/10.1007/978-3-031-58878-5_11

extraction. However, in action identification, where the goal is to acquire context and motion throughout the video, 3D convolution outperforms 2D convolution due to the capacity to collect spatiotemporal information in video at the same time.

Recently, researchers have shown an increased interest in Vision Transformer (ViT). ViT has a remarkable performance, good robustness, and smooth operation, which has received considerable critical attention in various visual recognition tasks such as image captioning, visual question answering, and multimodal understanding. CLIP, for example, can embed images and words into the same semantic space for similarity computation using 400M image-text pairs for training [5]. Furthermore, CLIP4Clip applies CLIP's image text expertise to the Video-Text Retrieval (VTR) problem, resulting in considerable performance increases across a variety of video-text retrieval datasets [6].

The objective of this study is to investigate the performance of 3D CNNs and ViT in the context of action recognition on video datasets. Both 3D CNNs and ViT have shown promising results in various computer vision tasks, but their effectiveness in action recognition remains an open question. In fact, the Human-Daily Activities with Multiple Cameras (HDMB-51) dataset is thought to be a good and famous dataset for action recognition tasks in the field of computer vision. It was designed to facilitate the study and evaluation of action identification algorithms and models. By conducting this comparison, we aim to shed light on the strengths and limitations of each method, ultimately providing insights into their suitability for action recognition tasks.

The rest of the paper is organized as follows. Section 2 discusses an overview of the related works in Sect. 2. A description of our approach is presented in Sect. 3. Section 4 details the experimentation carried out on. Finally, conclusions are given in Sect. 5.

2 Related Works

Over the past decade, convolutional networks have long been the standard architecture in video recognition. Supervised convolutional models that are optimized on the ImageNet dataset have dominated this discipline [1]. Simonyan et al. introduced a two-stream strategy in 2014 with the idea of having a CNN trained with raw RGB frames and another CNN trained with optical flow, which represents the movement vectors between two successive frames [10]. A combination of convolutional neural networks (CNNs) and long short-term memories (LSTMs) is presented by Wang et al. on untrimmed movies for poorly supervised action recognition and detection [14]. Convolutional neural networks (CNNs) have provided a high level of accuracy in image classification, so Krizhevsky et al. primarily employ this learning method among the other deep learning methods [4]. In 2018, Tran et al. discovered that disentangling spatial and temporal convolution improves the speed-accuracy tradeoff over the original 3D convolution [7]. Combining the 3DCNN and two-stream approaches is something that Wan et al. proposed by applying 3D convolutions to the spatial-stream and the VGG16 CNN to the temporal-stream. For the final prediction, they combined features from both streams and then utilized a support vector machine to improve accuracy to 70.2% [11].

Recently, Vision Transformers has emerged as a new trend in image recognition backbones. Transformers have also been adopted for video recognition. The Natural Language Processing scaling successes of Vaswani et al. [3]. Alexey Dosovitskiy et al.

presented a method for image recognition tasks based on transformers, which were originally devised for natural language processing (NLP) tasks [9]. Touvron et al. suggested a method for performing state-of-the-art picture classification tasks using fewer computer resources. They used knowledge distillation and provided data-efficient training approaches to boost the performance of smaller models [13]. The Swin Transformer, developed by Liu et al., employs a hierarchical architecture that grows quickly to accommodate high-resolution images. It employs shifting windows to properly capture local and global context [12]. Yuan-Hong Liao et al. investigated training Vision Transformers from scratch without using huge datasets like ImageNet for pretraining. Their strategy outperforms models pretrained on huge datasets in terms of performance [15]. Li et al. proposed the Vision Permutator, an architecture that employs permutation operations to replace self-attention layers in the traditional ViT model. It achieves competitive results with reduced computation [16].

Considering the above explorations, we propose BERT-based Temporal Modeling with 3D CNNs for action recognition tasks. By combining the advantages of 3D CNNs and BERT-based language modeling, this method efficiently captures the temporal dynamics and spatial aspects of films [21]. Additionally, several recent studies investigate ViT, which have drawn substantial attention and have displayed astounding performance in several computer vision applications [12, 13, 15, 16]. These methods have achieved the best possible results on a variety of standard datasets [22]. In Sect. 3 of the paper, we propose and describe two methods for evaluating the performance and accuracy of video recognition on the HDMB-51 dataset.

3 Methodology

3.1 Advantages BERT-Based Temporal Modeling with 3D CNNs

BERT-based Temporal Modeling with 3D CNNs provides several benefits for capturing temporal information and fusing textual and visual modalities for video comprehension. Combining the potential of BERT-based language modeling with 3D CNNs enables the capturing of temporal dependencies. This enables the model to comprehend the temporal relationships between frames or segments in a video, resulting in a more thorough analysis of the video's content. Second, the method enhances representations by combining the semantic understanding of BERT with the spatial-temporal features of 3D CNNs. The combination of these modalities improves the model's ability to capture the video's content and temporal context, resulting in representations that are richer and more informative. Moreover, BERT-based temporal modeling with 3D CNNs improves video comprehension by combining the strengths of textual and visual data. By combining these modalities, the model obtains a more comprehensive and holistic understanding of the video content, thereby enhancing its performance on tasks such as action recognition, video captioning, and video summarization. The approach also benefits from transfer learning, as the pretrained BERT model can capture high-level semantic information and the 3D CNNs can acquire specific spatial and temporal characteristics from video data. This transfer learning allows the model to apply knowledge from large-scale textual data to video tasks, despite having limited labeled video data. Overall, BERT-based Temporal

Modeling with 3D CNNs offers advantages in capturing temporal dependencies, enhancing representations, enhancing video comprehension, and facilitating transfer learning, making it a valuable technique for a variety of video-related applications.

3.2 Architecture of BERT-Based Temporal Modeling with 3D CNNs

The architecture of BERT-based Temporal Modeling with 3D CNNs is summarized in Fig. 1. It contains 3D CNN without applying temporal global average pooling. More details are provided in the original paper [8].

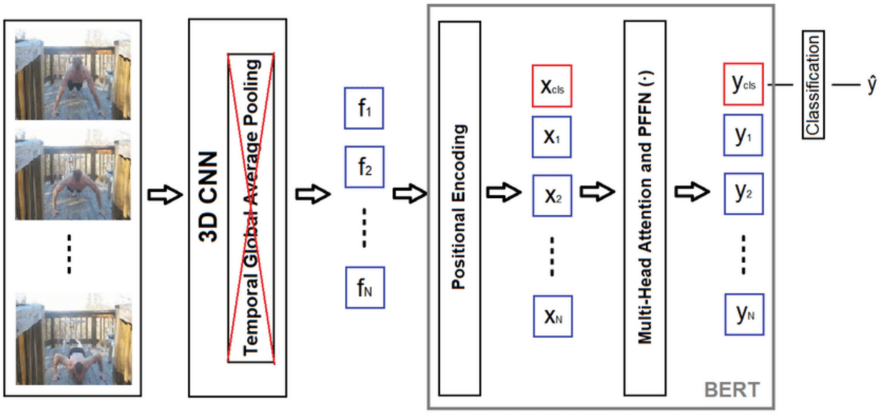

Fig. 1. The architecture of BERT-based Temporal Modeling with 3D CNNs [8].

3.3 Advantages Vision Transformer Model

Vision Transformer (ViT) has emerged as a formidable architecture in computer vision, offering several advantages over conventional convolutional neural networks (CNNs). The attention mechanism's capacity to capture global contextual information and long-range dependencies is a significant advantage. By considering the relationships between all image regions, the ViT is able to comprehend complex visual patterns and facilitate a holistic understanding of the visual content. In addition, ViT are scalable to large image resolutions, allowing for the efficient processing of high-resolution images without imposing substantial computational burdens. In addition, they offer versatility in handling variable image sizes, allowing for greater generalization and enhanced performance on datasets with varying aspect ratios. ViT's hierarchical representation learning, which enables the extraction of both local and global features by paying attention to various levels of abstraction, is an additional advantageous feature. ViT excels in transfer learning as well, as pretrained models can be fine-tuned for specific tasks, resulting in impressive performance across a broad range of computer vision applications. ViT provides interpretability and explain ability through their attention mechanism, allowing for

insights into the model's decision-making process and fostering confidence in their predictions. Together, these benefits make ViT a compelling option for a variety of computer vision tasks and contribute to their rising prominence in the scientific community.

3.4 Architecture of Vision Transformer

Dosovitskiy et al. proposed the vision transformer (ViT), the first pure transform-architecture for image processing. It can achieve comparable outcomes to contemporary convolutional neural networks [9]. Figure 2 depicts the structure of ViT. There is a summary of the model: 1) divide an image into segments of fixed size. 2) embed each of them linearly 3) add position embeddings 4) provide the resultant vector sequence to a standard Transformer encoder. More information can be found in the original paper [9] (Fig. 2).

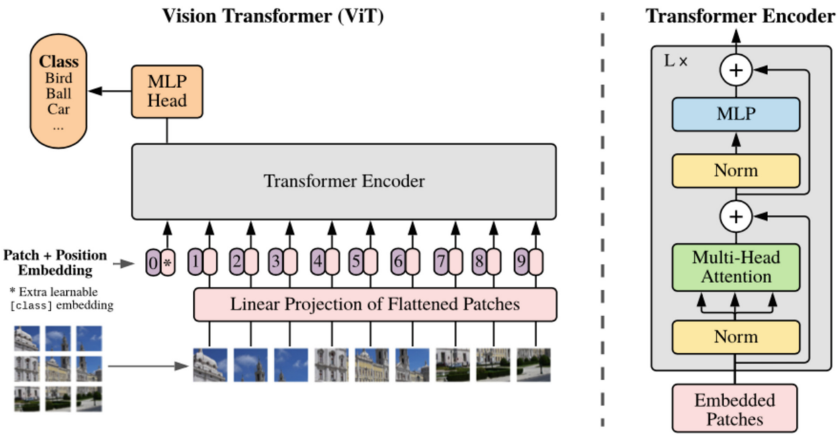

Fig. 2. Architecture of Vision Transformer [9]

3.5 Our Contribution

HDMB-51 dataset, which stands for "Human-Daily Activities with Multiple Cameras", is often considered suitable for action recognition tasks. Firstly, The HDMB-51 dataset comprises videos of fifty-one human daily activities captured from multiple camera viewpoints. It consists of many video clips, providing a substantial amount of training data. HDMB-51 focuses on fine-grained action recognition, requiring models to distinguish between subtle differences in human activities. The dataset is suitable to evaluate two methods as they provide the original videos. The HMDB-51 dataset is well-suited for evaluating 3D CNNs in action recognition tasks due to its video-based nature, temporal dynamics, diversity, and established role as a benchmark in the field. By using this dataset, researchers can effectively assess the capabilities of 3D CNNs in understanding complex actions in videos and drive advancements in the field of action recognition.

Moreover, ViT have shown promise in capturing fine-grained details, and their attention mechanisms allow them to address specific regions or frames in the video. ViT can effectively learn and generalize from this dataset, capturing both spatial and temporal information to recognize human activities with high accuracy. So, the HMDB-51 dataset is indeed suitable for evaluating 3D CNNs and ViT in action recognition tasks. Figure 3 shows the action categories in HDMB-51.

To demonstrate the effectiveness of the ViT and 3D CNNs, we propose two methods to evaluate based on the HDMB-51 dataset.

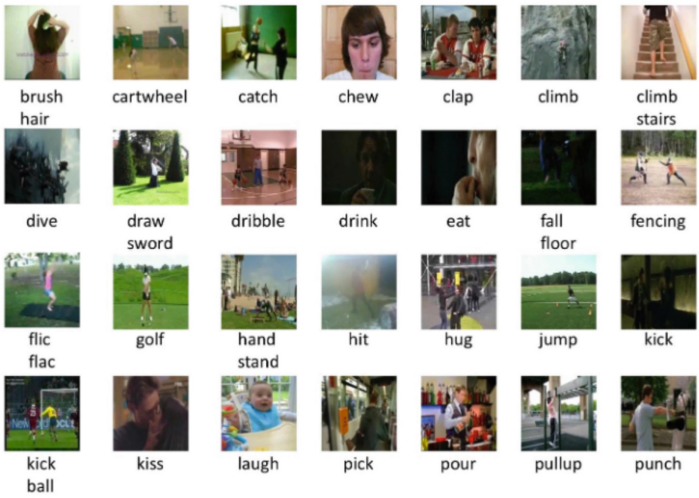

Fig. 3. Action categories in HDMB-51 [2]

We apply the general steps to apply ViT in video recognition tasks to demonstrate effective of ViT:

1. Data Preprocessing: Obtain the video dataset required for the task of recognition; Preprocessing the video entails removing frames from the video.
2. Temporal Tokenization: Extract or generate video clips with temporal context, ensuring they contain consecutive frames for each video.
3. Model Architecture: Select an appropriate ViT architecture for video recognition.
4. Model Pretrained: Initialize the ViT model with appropriate weights.
5. Model Evaluation: After training, evaluate the performance of the ViT on a validation or test set using relevant evaluation metrics.

The general steps to use ViT in video recognition are shown in Fig. 4.

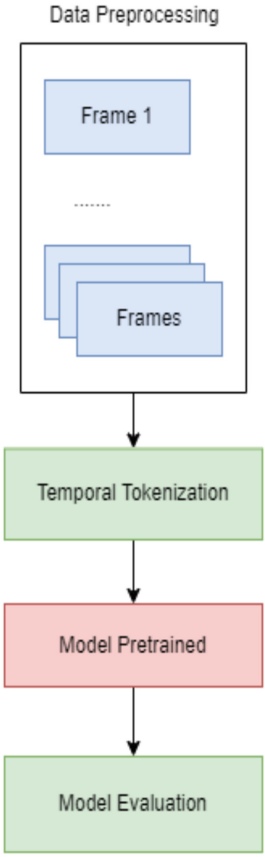

Fig. 4. General step to apply ViT in video recognition.

4 Experiments

In this part, we talk about what happened when the experiment was done two different ways. First, we talk about the information, and then we get into the details of how it will be used. Then, we do studies to find out more about the two methods and compare them.

4.1 Dataset

In this section, we perform experiments on various settings, including zero-shot followed by ablation investigations of the proposed method. HMDB-51 is a compilation of realistic videos from various sources, including films and online recordings. The dataset contains approximately seven thousand video snippets organized into 51 action class categories [2]. HDMB-51 defines three data segments that are used to calculate the results. We report the mean accuracy of the three splits as the final accuracy (Fig. 5).

Fig. 5. Action example of dataset [2].

4.2 Dataset Preparation

The classification accuracy must be evaluated so that future classification outcomes may be predicted and compared. The operation flow is presented in Fig. 6.

4.3 Implementation

The paper runs experimentally on Google Colab platform with graphics processor 16 GB GPU P100, 12 GB RAM. Parameters are initialized in the same with the BERT-based Temporal Modeling with 3D CNNs 's original paper [8] and ViT 's original paper [9]. To save time and computational resources required for training, we use pre-trained models available, and it is compatible for action recognition [8, 9].

4.4 Result

We are able to produce the prediction score by combining the classification and selection scores. Table 1 and Fig. 7 present the five most similar labels for brush hair action.

4.5 Comparison with 3D CNN with BERT and ViT

From the values on Table 2, the result shows that the two methods show their robustness by performing efficiently in terms of action recognition. It can be shown that the ViT (ViT-B/16) produced particularly good results by achieving 84,7% accuracy. ViT-B/16 outperformed ResNeXt101 BERT slightly, indicating its potential as a strong contender in the domain of action recognition.

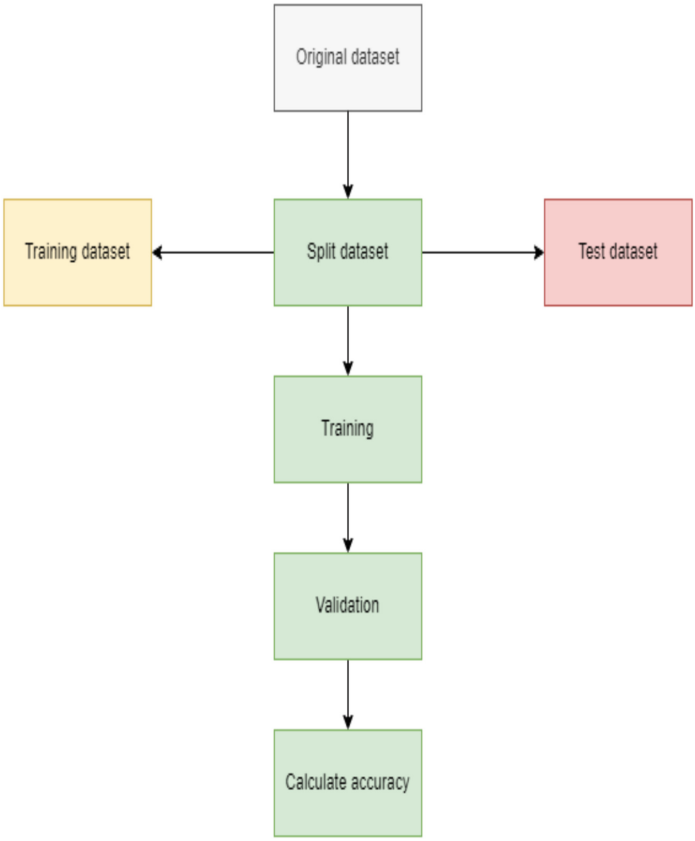

Fig. 6. The dataset's structure and operation flow

Table 1. Evaluation by using ViT

brush hair	77,2%
clap	18,3%
cartwheel	3,2%
dribble	0,6%
catch	0,1%

4.6 Zero-Shot Experiments

Zero-shot learning is a machine learning paradigm in which a model is trained to recognize and generalize to classes that were not seen during the training phase. We use ResNeXt101 BERT and ViT-B/16 to perform cross-dataset zero-shot evaluation in video dataset. We present a comprehensive comparison in Table 3.

Fig. 7. The result prediction on video by using ViT.

Table 2. Comparison with 3D CNN with BERT and ViT

Method	HMDB-51
ResNeXt101 BERT	83.5%
ViT-B/16	84.7%

Table 3. Zero-shot performance on HDMB-51

Method	HMDB-51
ResNeXt101 BERT	40.8%
ViT-B/16	44.6%

4.7 Discussion

Our study shows that ViT is a powerful video recognition architecture. ViT scale and adapt to diverse video resolutions better than 3D CNNs. Despite both architectures being pretrained on the Kinetics-400 dataset [23], we observed that the ViT outperformed the 3D CNN in terms of performance on the task of action recognition. While 3D CNNs are still effective, ViT offer a promising alternative, and their performance on the HDMB-51 dataset shows their versatility and potential in video understanding.

5 Conclusion

In this paper, we propose to use 3D CNNs and ViT which aim to verify the effectiveness of the methods. 3D CNNs were traditionally the dominating architecture for computer vision applications like image classification, object recognition, and segmentation. ViT produced impressive results, even outperforming 3D CNNs in some circumstances, particularly when trained on large-scale datasets. Moreover, ViT can process films of various sizes without increasing processing overhead by tokenizing video frames into patches. So, ViT can handle films of varying resolutions and aspect ratios, making them useful for video recognition jobs. Besides, ViT has some challenges with computational complexity and spatial information loss. In the future, we intend to extend our method beyond categorization to other video tasks. In the future, we will develop a system to detect the features of the actions.

References

1. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., FeiFei, L.: Imagenet: a large-scale hierarchical image database. In: CVPR, pp. 248–255 (2009)
2. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: a Large Video Database for Human Motion Recognition. In: ICCV (2011)
3. Vaswani, A., et al.: Attention is all you need. In: NeurIPS, pp. 5998–6008 (2017)
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017). https://doi.org/10.1145/3065386
5. Radford, A., et al.: Learning transferable visual models from natural language supervision. Image 2, T2 (2021)
6. Luo, H.S., et al.: CLIP4Clip: An empirical study of clip for end-to-end video clip retrieval and captioning. Neurocomputing **508**, 293–304 (2022). https://doi.org/10.1016/j.neucom.2022.07.028
7. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: CVPR (2018)
8. Esat Kalfaoglu, M., Sinan Kalkan, A., Alatan, A.: Late temporal modeling in 3D CNN architectures with bert for action recognition. In: Bartoli, A., Fusiello, A. (eds.) Computer Vision – ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V, pp. 731–747. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-68238-5_48
9. Dosovitskiy, A., et al.: An image is worth 16×16 words: Transformers for image recognition at scale. In: Proceedings of the 9th International Conference on Learning Representations (2021)
10. Simonyan, K., Zisserman, A.: Two-Stream Convolutional Networks for Action Recognition in Videos. Adv. Neural. Inf. Process. Syst. **27**, 1–9 (2014)
11. Wan, Y., Yu, Z., Wang, Y., Li, X.: Action recognition based on two-stream convolutional networks with long-short-term spatiotemporal features. IEEE Access **8**, 85284–85293 (2020). https://doi.org/10.1109/access.2020.2993227
12. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012–10022 (2021)
13. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jegou, H.: Training data-efficient image transformers & distillation through attention. In: Proceedings of the 38th International Conference on Machine Learning, PMLR vol. 139, pp. 10347–10357 (2021)
14. Wang, L., Xiong, Y., Lin, D., Van Gool, L.: Untrimmednets for weakly supervised action recognition and detection. In: CVPR (2017)
15. Yuan, L., et al.: Tokens-to-token ViT: Training vision transformers from scratch on imagenet (2021). arXiv:2101.11986. Retrieved from https://arxiv.org/abs/2101.11986
16. Hou, Q., Jiang, Z., Yuan, L., Cheng, M.M., Yan, S., Feng, J.: Vision Permutator: A Permutable MLP-Like Architecture for Visual Recognition. IEEE Trans. Pattern Anal. Mach. Intell. **45**(1), 1328–1334 (2023). https://doi.org/10.1109/TPAMI.2022.3145427
17. Hershberger, W.A.: Chapter 1 The Synergy of Voluntary and Involuntary Action. In: Volitional Action - Conation and Control, vol. 62, pp. 3–20. Elsevier (1989). https://doi.org/10.1016/S0166-4115(08)61905-6
18. Zhu, Y., et al.: A comprehensive study of deep video action recognition (2020). arXiv preprint arXiv:2012.06567
19. Herath, S., Harandi, M., Porikli, F.: Going deeper into action recognition: A survey. Image Vis. Comput. **60**, 4–21 (2017). https://doi.org/10.1016/j.imavis.2017.01.010

20. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR, pp. 1725–1732 (2014)
21. Devlin, J.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018). arXiv preprint arXiv:1810.04805
22. Tummala, S., Kadry, S., Bukhari, S.A.C., Rauf, H.T.: Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling. Curr. Oncol. **29**(10), 7498–7511 (2022). https://doi.org/10.3390/curroncol29100590
23. Kay, W., et al.: The kinetics human action video dataset (2017). arXiv preprint arXiv:1705.06950