Visual transformer with depthwise separable convolution projections for video-based human action recognition

Yu Cao[1,2,*], *Fang* Wang[1,**], and *Qiusheng* Zheng[2,***]

[1]Department of Computer Science, Brunel University of London, Kingston Lane, Uxbridge, United Kingdom
[2]Henan Key Laboratory on Public Opinion Intelligent Analysis, Zhongyuan University of Technology, Zhengzhou, China

Abstract. Human action recognition is a task that utilizes algorithms to recognize human actions from videos. Transformer-based algorithms have attracted growing attention in recent years. However, transformer networks often suffer from slow convergence and require large amounts of training data, due to their inability to prioritize information from neighboring pixels. To address these issues, we propose a novel network architecture that combines a depthwise separable convolution layer with transformer modules. The proposed network has been evaluated on the medium-sized benchmark dataset UCF101 and the results have demonstrated that the proposed model converges quickly during training and achieves competitive performance compared with SOTA pure transformer network, while reducing approximately 7.4 million parameters.

1 Introduction

Human action recognition (HAR)[1] remains a significant challenge in computer vision and serves as a foundational task in video understanding. It has a wide range of applications, including video surveillance systems and sports action analysis. HAR algorithms have been widely explored in last decades using various backbone networks such as convolution neural networks[2], Long-Short Term Memory network[3] and Transformer networks [4]. Mainstream methods fall into two main categories: convolutional networks and transformer-based networks. Most recently, transformer-based networks have become increasingly popular for human action recognition [5]. In particular, the Vision Transformer(ViT) was the first visual transformer model applied to computer vision tasks, achieving state-of-the-art performance on large-scale benchmark datasets. Specifically, the ViT network linearly encodes image patches and computes the correlations between the patches so as to focus on the most relevant areas of the image. However, due to the self-attention mechanism, ViT lacks the inductive bias of locality, which results in slower convergence and a requirement for large amounts of training data [6].

Some research studies such as Video Swin Transformer [7] have adopted convolutional concepts to utilize a window-based transformer mechanism to compute the self-attention within windows, limiting the self-attention computing in small range of area. To show the importance of locality induct bias, research [8] analyzing the difference of cognitive field between CNN and self-attention in lower and higher layers. the research reveal that local information aggregation at lower layers are important to ViT networks, the research also demonstrates that self-attention plays a crucial role in visual transformer throughout analyzing the internal representation structure of ViT and CNN networks, which enables early aggregation of global information. Residual connection in visual transformer strongly propagate features from lower to higher layers. In the last few years, comprehensive research[9, 10] has proved the importance of locality inductive bias in computer vision task.

To improve the local spatial and temporal feature extraction, we proposed a new architecture of transformer-based network by using depthwise separable convolution layer for patch embeddings instead of linear layer, which learn the neighbor feature preferentially and without the significant increase of the number of parameters. We also build a novel self-attention mechanism that contains depthwise separable network to generate query, key and value tokens of patches, and the position embedding is provided by convolution layers. We hypothesise that the convolution layer can provide high spatial and temporal position structure in Query, Key and Value vectors due to sliding kernel in convolution layer, thus the position embedding is not required in the standard self-attention mechanism. In order to reduce the computational cost, we use the depthwise convolution layer[11] in all modules rather standard convolution layers to reduce the parameters of the model.

2 Related work

2.1 Human action recognition

Human Action Recognition is to extract appearance and temporal features from short and untrimmed videos, and

*e-mail: sdcaoyu2019@gmail.com
**e-mail: Fang.wang@brunel.ac.uk
***e-mail: zqs@zut.edu.cn

recognize human actions based on the spatial and temporal features[1, 12]. Traditional methods[13, 14] rely on handcrafted features to learn human shapes and movements. Research[15] proposed an action recognition method by using contour points of human silhouettes to represents the body shape. With the progress on deep learning methods[16], convolution networks[17, 18] and transformer networks[6] are developed to extract deep spatial features such as color ,body shape and semantic information, and temporal features of action motions.

2.2 3D convolution network

Convolution neural networks[2, 16, 18] has achieved great success on video action recognition. In particular, the SlowFast[16] network is a network consisting of slow and fast pathways to extract spatial and temporal feature respectively via 3D convolution modules[17]. This kind of design enhances the power of motion analysis and achieved superb performance on video-based action recognition. 3D Convolution Network[2] is a fully convolutional network with small kernel sizes to naturally extract spatiotemporal features. The Inflated 3D[18] is a network with inflated 3D convolution network, which extends 2D kernel to 3D kernel. The design of I3D kernel allows reusing pretrained 2D models weights on ImageNet dataset, to alleviate the random initialization problem. The Depthwise Separable Network[19] is designed to extract the image features with significantly drops the number of parameters, the network applies single filter per channel and output the same number of channels as input, followed a pointwise convolution to output the desired channels. However, convolutional networks are limited in their ability to capture global spatial features across entire video frames and long-range temporal dependencies within videos.

2.3 Vision transformer network

Vision Transformer[6] is a transformer network for image classification and has gained competitive performance when evaluated on large datasets such as ImageNet-22K[20]. Vision Transformers divide the original image into several non-overlapping patches, which are then linearly embedded into fixed-length tokens. These tokens are subsequently processed by the Multi-Head Self-Attention (MHSA) mechanism [21]. and positional embeddings are further added to provide position information for each token, in order to learn the global feature representation and build corelation between each patches. The Video Swin Transformer [4, 7] introduces a novel architecture to enhance local feature analysis. Its shifted window attention mechanism enables the extraction of both spatial and temporal features within local windows. ConvFormer[10] proposes a network that integrates convolution and transformer to learn local and global features for image segmentation. In [22], a combination of convolution and transformer network is designed to enhance the local feature for video understanding.

3 Visual transformer with depthwise separable convolution projections

3.1 Overall architecture

The overall framework of the proposed Visual Transformer with Depthwise Separable Convolution Projections (VT-DSCP) is shown in Fig 1. Short and trimmed video inputs are first converted into feature maps using depthwise separable convolution layer. Then, a Depthwise Separable Convolution (DSC) layer is used to generate the query, key, and value representations for computing the spatiotemporal attention regions in the video. The resolution of the feature maps are then progressively reduced by a hierarchical architecture composed of several novel transformer blocks, so as to learn deeper and stronger representations for action recognition. The downsampling in the transformer blocks is performed using a depthwise separable convolution layer as well. The final video feature map is converted to 1D feature, and classified via a fully connected network. The detailed information of each component is explained in the following sections.

3.2 Feature embedding

Given a short and trimmed video input with the shape of $V \in \mathbb{R}^{3\times D\times H\times W}$, RandomHorizontalFlip[23] and Normalization are utilized to pre-process the input video. The preprocessed data is embedded by a DSC layer to generate the feature map of video $F \in \mathbb{R}^{C\times D'\times H'\times W'}$. The DSC layer employs both spatial and temporal convolution. The spatial convolution layer uses a kernel size of $1 \times K \times K$, focusing solely on the spatial dimensions. Then,followed by a depthwise convolution layer with a kernel size of $K \times 1 \times 1$, which learns only the temporal dependencies in the videos. Finally, the spatial and temporal features are combined through addition to form the video representation. The architecture of DSC layer is illustrated in Fig 2

3.3 Depthwise separable enhanced transformer

The novel transformer enhanced by depthwise separable convolution is shown in the Fig 3, which utilizes spatial convolution and temporal convolution to generate tokens for self-attention. Compared with linear embedding, a convolutional layer scans features from top-left to bottom-right while inherently capturing positional information through its structure. Moreover, the temporal convolution layer sequentially extracts the temporal information frame by frame.

Specifically, we adopt the depthwise separable convolution layer to tokenize the input feature $F \in \mathbb{R}^{C\times D'\times H'\times W'}$ by tripling the channel, and to split the feature into query, key and value tokens. Then we follow the idea of [6] that divides the feature map into non-overlapping patches with a patch size of P. Therefore, the query, key and values are $f_w = \mathbb{R}^{3\times N\times L\times C}$ respectively, where N is the number of patches ($N = \frac{D'}{P} \times \frac{H'}{P} \times \frac{W'}{P}$) and L is the size of window ($L = P\times P\times P$). We also adopt the multi-head self-attention

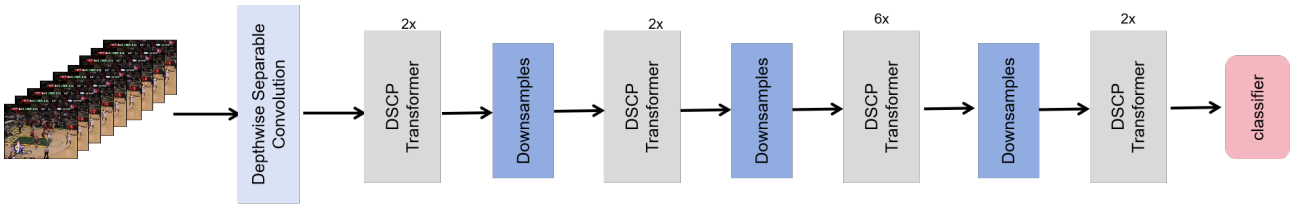

Figure 1. Framework of depthwise separable convolution involved transformer.

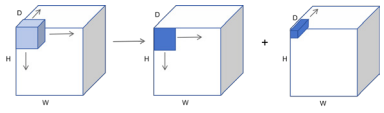

Figure 2. The figure on the left illustrates a standard 3D convolution network. The figures on the right depict a depthwise separable network, where convolution filter first learns spatial features, followed by a depthwise convolution that focuses on temporal dependency learning.

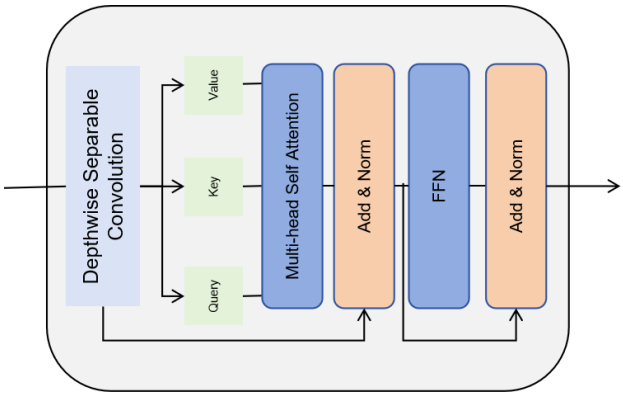

Figure 3. Paradigm of Depthwise Separable Convolution Projection enhanced Transformer.

to enrich the representation of visual features. As the convolution layer provides positional information, additional position embedding is not required in the novel DSC based module.

The multi-head attention mechanism computes attention scores between query and key tokens and captures correlations among all the patches. These scores weight the value tokens to highlight the most important patches. A Layer Normalization followed by a residual addition then normalizes the embedded features. Next, a Feed-Forward Network (FFN), again followed by Layer Normalization and a residual connection, further processes the spatiotemporal features and improves model generalization. Whereas convolution extracts spatial and temporal features within its kernel's receptive field, which prioritizes local neighbors of feature. Multi-head self-attention models associations across all patches, enhancing the network's understanding of global visual and temporal context. Our model adopts a hierarchical architecture that downsamples both spatial and temporal dimensions using a depthwise-separable convolution layer.

3.4 Classifier

The novel network uses a feed-forward classifier comprising two fully connected layers with a ReLU activation between them. A softmax function then produces the class-probability distribution. To evaluate the model performance, we conducted experiments on the public UCF101 benchmark dataset.

4 Experiments

To evaluate the effectiveness of the new transformer network VT-DSCP, experiments have been conducted on UCF101[24], which is a medium-size video dataset. We adopt data augmentation methods including Normalization, Scale (224×224) in the spatial dimension. We also use the RandomCenterCrop method in temporal dimension to randomly obtain 32 frames per clip, thus the input size is $[3 \times 32 \times 224 \times 224]$. We use the AdamW[25] as the optimizer with a learning rate of 1e-4. 1e-5 is set as weight decay, and CosineAnnealingLR in PyTorch as the learn rate scheduler. The training epoch is 20. The kernel size of spatial and temporal convolution are $(1, 3, 3)$ and $(3, 1, 1)$, respectively. We split the total dataset into 75% for training and 25% for test. The experiments ran on 2 GPUs by using the DistributedDataParallel method.

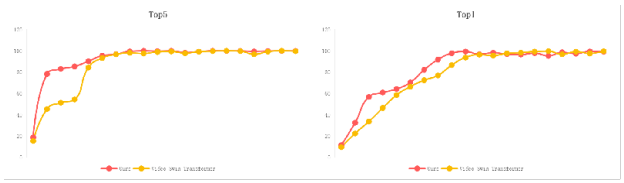

Figure 4. Training comparison of VT-DSCP and Video Swin Transformer on UCF-101.

4.1 Results

As shown in Fig 4, in the training stage, VT-DSCP converged much faster than Video Swin transformer. Compared with other CNN-based approaches, our network achieved higher accuracy on the UCF101 dataset, by taking advantages of the attention mechanism to capture overall spatial content on the spatial dimension. However, the accuracy of VT-DSCP is lower than the video swin transformer network which has no convolution layers. The reason may be the small size of our model. The total number

Table 1. Difference between VT-DSCP and Swin transformer baseline.

model	projection	position encoding	parameter
VT-DSCP	Depthwise separable CNN	no	20.8M
Video Swin Transformer-Tiny	linear	yes	28.2M

Table 2. Performance Comparison of Various methods on UCF101.

Model	UCF101(Accuracy)
C3D	82.3
TSN	84.5
I3D	88.8
TSN	91.7
Video Swin Transformer	93.1
Our	92.3

of parameters of VT-DSCP is 20.8M, 7.4M parameters are reduced compared to Video Swin transformer with similar architectures, the architecture difference is shown in Table 1. The performance comparison between various models is shown in Table 2. Compared with other CNN based approaches, our network has fewer parameters and better performance, which inherits the advantage of the convolution network and transformer network.

4.2 Ablation study

In order to validate the effectiveness of depthwise separable embedding for visual transformer, the comparison is shown in the Table 3. Linear embedding was tested and achieved a result 86.5%, with slower convergence during the training stage. Moreover, we replaced linear embedding with standard CNN layer to generate the query, key and values. The result achieved 91.9% on the UCF101 dataset with faster convergence. However, the parameters used by this model are much more than our model.

Table 3. Performance comparison of various embedding methods on the UCF101 dataset.

Embedding	Accuracy
Linear	86.5
Standard CNN	91.9
Depthwise separable CNN	92.3

5 Conclusion

In this work, we propose a new transformer network for video based action recognition with depthwise separable convolution layers. In particular, depthwise separable convolution layers with small sized kernels are used to project tokens for the self-attention mechanism. The experiments have demonstrated that the position encoding can be safely removed from the transformer network. In addition, the convolution layer introduces linear invariances in visual representations while preserving strong location information. The proposed VT-DSCP model has achieved competitive performance with fewer parameters, when compared with other networks. In the future, we will evaluate our model's performance on various HAR benchmark datasets to assess its generalization ability. Additionally, we will compare the model with state-of-the-art transformer networks to demonstrate the effectiveness.

This work is supported by the Zhongyuan University of Technology-Brunel University London (ZUT-BUL) Joint Doctoral Training Programme. This work is funded by the ZUT/BRUNEL scholarship.

References

1 R. Poppe, A Survey on Vision-Based Human Action Recognition, Image and Vision Computing, **28**, 976–990 (2010). https://doi.org/10.1016/j.imavis.2009.11.014

2 D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, Learning Spatiotemporal Features With 3D Convolutional Networks, in Proceedings of the IEEE International Conference on Computer Vision (ICCV), (2015). https://doi.org/10.1109/ICCV.2015.510

3 M. Majd and R. Safabakhsh, Correlational Convolutional LSTM for Human Action Recognition, Neurocomputing, **396**, 224–229 (2020). https://doi.org/10.1016/j.neucom.2018.10.095

4 Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, Video Swin Transformer, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3202–3211 (2022). https://doi.org/10.1109/CVPR52688.2022.00320

5 S. Khan, M. Naseer, M. Hayat, S. Zamir, F. Khan, and M. Shah, Transformers in Vision: A Survey, ACM Computing Surveys (CSUR), **54**, 1–41 (2022). https://doi.org/10.1145/3505244

6 A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, and J. Uszkoreit, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv preprint arXiv:2010.11929 (2020).

7 Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows, in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012–10022 (2021). https://doi.org/10.1109/ICCV48922.2021.00986

8 M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, Do Vision Transformers See Like Convolutional Neural Networks?, Advances in Neural Information Processing Systems, **34**, 12116–12128 (2021).

9 D. Cui, C. Xin, L. Wu, and X. Wang, ConvTransformer Attention Network for Temporal Action Detec-

tion, Knowledge-Based Systems, **300**, 112264 (2024). https://doi.org/10.1016/j.knosys.2024.112264

10 P. Gu, Y. Zhang, C. Wang, and D. Chen, ConvFormer: Combining CNN and Transformer for Medical Image Segmentation, arXiv preprint arXiv:2211.08564 (2022).

11 F. Chollet, Xception: Deep Learning with Depthwise Separable Convolutions, arXiv preprint arXiv:1610.02357 (2017).

12 H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, and D.-S. Chen, A Comprehensive Survey of Vision-Based Human Action Recognition Methods, Sensors, **19**, 1005 (2019). https://doi.org/10.3390/s19051005

13 H. Shabani, D. Clausi, and J. Zelek, Towards a Robust Spatio-Temporal Interest Point Detection for Human Action Recognition, in Proceedings of the International Conference on Image Processing (ICIP), pp. 237–243 (2009).

14 A. F. Bobick and J. W. Davis, The Recognition of Human Movement Using Temporal Templates, IEEE Transactions on Pattern Analysis and Machine Intelligence, **23**, 257–267 (2001). https://doi.org/10.1109/34.910878

15 A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, Silhouette-Based Human Action Recognition Using Sequences of Key Poses, Pattern Recognition Letters, **34**, 1799–1807 (2013). https://doi.org/10.1016/j.patrec.2013.01.021

16 C. Feichtenhofer, H. Fan, J. Malik, and K. He, SlowFast Networks for Video Recognition, in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6202–6211 (2019). https://doi.org/10.1109/ICCV.2019.00630

17 K. He, X. Zhang, S. Ren, and J. Sun, Deep Residual Learning for Image Recognition, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

18 J. Carreira and A. Zisserman, Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6299–6308 (2017). https://doi.org/10.1109/CVPR.2017.502

19 M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, MobileNetV2: Inverted Residuals and Linear Bottlenecks, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4510–4520 (2018). https://doi.org/10.1109/CVPR.2018.00474

20 A. Krizhevsky, I. Sutskever, and G. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, Communications of the ACM, **60**, 84–90 (2017). https://doi.org/10.1145/3065386

21 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, Attention is All You Need, Advances in Neural Information Processing Systems, **30** (2017).

22 C. Zhang, ConvFormer: Tracking by Fusing Convolution and Transformer Features, IEEE Access, **11**, 74855–74864 (2023). https://doi.org/10.1109/ACCESS.2023.3293592

23 T. Kumar, R. Brennan, A. Mileo, and M. Bendechache, Image Data Augmentation Approaches: A Comprehensive Survey and Future Directions, IEEE Access, (2024).

24 K. Soomro, A. RoshanZamir, and M. Shah, UCF101: A Dataset of 101 Human Actions Classes From Videos in the Wild, arXiv preprint arXiv:1212.0402 (2012).

25 I. Loshchilov and F. Hutter, Decoupled Weight Decay Regularization, arXiv preprint arXiv:1711.05101 (2017).