# HaViT: Hybrid-Attention Based Vision Transformer for Video Classification

Li Li[1,2], Liansheng Zhuang[1(✉)], Shenghua Gao[3], and Shafei Wang[2]

[1] University of Science and Technology of China, Hefei 230026, China
lili1234@mail.ustc.edu.cn, lszhuang@ustc.edu.cn
[2] Peng Cheng Laboratory, Shenzhen 518000, China
[3] ShanghaiTech University, Shanghai 201210, China

**Abstract.** Video transformers have become a promising tool for video classification due to its great success in modeling long-range interactions through the self-attention operation. However, existing transformer models only exploit the patch dependencies within a video when doing self-attention, while ignoring the patch dependencies across the different videos. This paper argues that external patch prior information is beneficial to the performance of video transformer models for video classification. Motivated by this assumption, this paper proposes a novel Hybrid-attention based Vision Transformer (HaViT) model for video classification, which explicitly exploits both internal patch dependencies within a video and external patch dependencies across videos. Different from existing self-attention, the hybrid-attention is computed based on internal patch tokens and an external patch token dictionary which encodes external patch prior information across the different videos. Experiments on Kinetics-400, Kinetics-600, and Something-Something v2 show that our HaViT model achieves state-of-the-art performance in the video classification task against existing methods. Moreover, experiments show that our proposed hybrid-attention scheme can be integrated into existing video transformer models to improve the performance.

## 1 Introduction

The task of video classification is to understand the visual and audio features to assign one or more relevant tags to the video. With the rapid increase of video content, this task is critical for many applications such as video retrieval [1] and video surveillance [2]. Compared with image classification, video classification is more challenging due to the temporal dimension, which increases the overall size of the input and variations in sequence. Though many methods have been proposed to model spatial relationships for image classification, it is still an open problem to jointly model spatial and temporal features in a video.

The remarkable progress of the transformer [3] in natural language processing (NLP) has inspired researchers to investigate its adaptation to image classification. The transformer is notable for its use of multi-head self-attention to model long-range dependencies, which are often modeled by the large receptive fields


L. Wang et al. (Eds.): ACCV 2022, LNCS 13844, pp. 502–517, 2023.
https://doi.org/10.1007/978-3-031-26316-3_30

formed by deep stacks of convolutional operations. However, convolutional operations can only capture local neighborhood in images, and the deep stack strategies are inherently limited in capturing long-range dependencies by means of aggregation of shorter-range information [4]. Conversely, the self-attention operation attends to all elements in the input sequence, and thus can capture both local as well as global spatial relationships on non-overlapping image patches in images. Recently, a pure transformer-based architecture with the Vision Transformer (ViT) [5] has been proposed to replace convolutions completely, and outperformed its convolution counterparts in image classification [5]. Inspired by the fact that attention-based architecture is an intuitive choice for modelling long-range contextual relationships in video, several transformer-based models have been proposed for video classification [6–11]. Some models apply self-attention on top of convolutional layers [6], while others use self-attention as the exclusive building block in the video classification models [7,8].

The natural extension of Vision Transformers to 3-dimensional video signal is challenging. Specially, each encoder of a transformer contains heavy computations such as pair-wise self-attention. Meanwhile, a video has a longer sequential representation than an image due to the additional temporal axis. Consequently, it is not economical or easy to optimize if directly applying the joint space-time attention to flattened video sequences. To reduce the computation costs, some efforts [7,8] have factorized the spatial and temporal domains via a factorized encoder or factorized self-attention, and have achieved a good speed-accuracy trade-off. Though achieving promising results in video classification, all these transformer models only exploit internal patch dependencies across the spatial and temporal dimensions within a video, while ignoring external patch dependencies across the different videos. In fact, external patch dependencies across the different images or videos which capture the external patch prior information plays an important role in many low-level vision tasks such as image restore [12] and video super-resolution [13]. This paper argues that external patch prior information is beneficial to transformer-based models for video classification.

Motivated by the above assumptions, this paper introduces a novel Hybrid-attention based Vision Transformer (HaViT) for video classification, which explicitly exploits both internal patch dependencies within a video and external patch dependencies across videos. The main operation performed in this architecture is hybrid-attention, which is computed on a sequence of spatio-temporal tokens extracted from a video and an external token dictionary extracted from extra videos. The spatio-temporal tokens encode internal patch information within a video, while the external token dictionary encodes external patch information across the different videos. HaViT uses the hybrid-attention instead of self-attention to model both internal patch dependencies within a video and external patch dependencies across the different videos. To improve the model performance, HaViT inserts the class token later in the transformer. This choice eliminates the discrepancy on the first layer of the transformer, which ithus used to perform hybrid-attention between patches and an external token dictionary only. Extensive experiments on three public datasets (Kinetics-400 [14],

Kinetics-600 [15] and Something-Something v2 [16]) show that HaViT achieves competitive results on video classification against existing state-of-the-art models.

In summary, our main contributions are as follows:

- A new Hybrid-attention based Vision Transformer (HaViT) is proposed for video classification, which is mainly built on the hybrid-attention module. To our best knowledge, this is the first work on transformer architecture for video classification which explicitly exploits external patch dependencies across videos.
- A new hybrid-attention mechanism is introduced to model both long-range patch dependencies within a video and external patch dependencies across the different videos, which is easily integrated into existing transformer models to improve their performance.
- Extensive experiments on public datasets demonstrate that our proposed HaViT model outperforms most existing video transformer models in most cases. When combined with existing video transformer models, the hybrid-attention does improve their model performance on video classification.

## 2 Related Work

Early works on video classification use hand-crafted features to encode appearance and motion information [17,18]. With the success of AlexNet in image classification [19], deep learning increasingly dominates visual modeling for video classification. Previously for convolutional models, backbone architectures for the video were adapted from those for images simply by extending the modeling through the temporal axis. Consequently, 3D convolution neural networks (3D-CNNs) have become a de-facto standard for video classification [20–24]. Compared with their image counterparts, 3D-CNNs have significantly more parameters and thus require more computation. To alleviate this, a large body of works (such as P3D [25], R(2+1)D [26], and S3D [27]) factorize convolutions across spatial and temporal dimensions to achieve a better speed-accuracy trade-off. However, the potential of convolution based approaches is limited by the small receptive field of the convolution operator. With a self-attention mechanism, the receptive field can be broadened with fewer parameters and lower computation costs, which leads to better performance. In [4], non-local network introduces self-attention on top of CNNs. Further, CBA-QSA CNN [28] extends self-attention with compact bilinear mapping for fine-grained action classification.

With the success of Vision Transformer (ViT) in image classification [5], a shift in backbone architectures is currently underway for video classification, from Convlutional Neural Networks (CNNs) to attention-based transformers [7–11]. Attention-based transformers use self-attention blocks at each layer to understand a frame's role with respect to other frames in the video. Since performing full spatio-temporal attention is computationally prohibitive, many efforts have been devoted to reducing computation costs via factorizing temporal and spatial

domains. In TimeSformer [8], the authors propose applying spatial and temporal attention in an alternating manner reducing the complexity of calculating attention weights. Similarly, ViViT [7] explores several methods of space-time factorization. In addition, they also proposed to adapt the patch embedding process from [5] to 3D data. However, all existing works only exploit internal patch dependencies information within a video via self-attention, while ignoring external patch dependencies information across the different video. In fact, external patch dependencies have been proven to be important for many vision tasks. This paper will try to exploit the external information in transformer model for video classification.

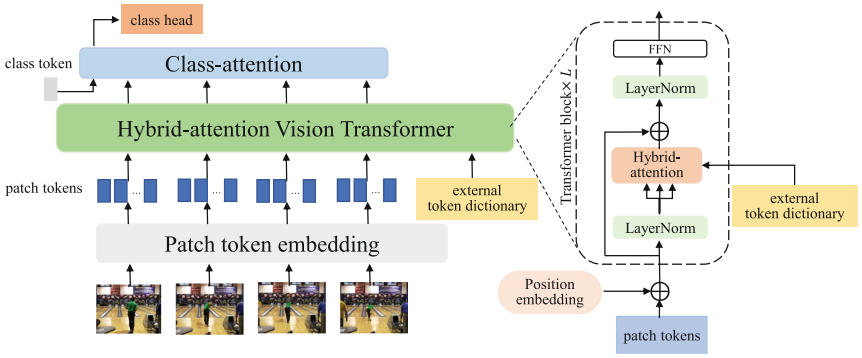



**Fig. 1.** Diagram of HaViT. The patches extracted from the frames are linearly projected into the token embedding, and position embedding is added. The patches and the class token are fed into the transformer with $L$ hybrid-attention layers.

## 3 Our Method

### 3.1 The Overall Architecture

The core idea of our proposed method is to introduce an external patch token dictionary to represent the texture feature space of all image patches. It uses the similarity of image patches in videos to calculate the internal attention and uses the similarity of image patches in current videos and the external token dictionary to calculate the external attention. In the calculation of hybrid-attention, the external token dictionary is used to expand the elements involved in the attention for the encoding of prior information. Specifically, when using the attention mechanism for feature embedding, the embedding of one image patch is not only linearly combined by the image patch features of the current video, but also by the elements of the external patch token dictionary. As shown in Fig. 1, the overall architecture includes a patch token embedding module, a video transformer module with hybrid-attention, and a class-attention layer. In the following, we elaborate on the processing flow: Firstly, $F$ frame images are sampled

from the video sequence to form a multidimensional tensor $x \in R^{H \times W \times C \times F}$ as the model input, where $H$, $W$ and $C$ denote the height, the width and the number of channels of each frame, respectively. Then, each frame in the video is divided into a fixed number of non-overlapping image patches and the image patches are reshaped into a flatten vector $x_{\mathrm{p,t}}$, with $\mathrm{p} = 1, ..., N$ denoting the spatial locations and $\mathrm{t} = 1, ..., F$ denoting the index of frames. Then, we get the image patch feature sequence $z^0_{\mathrm{p,t}}$ through the patch token embedding module. Next, we input the image patch feature sequence to the video transformer with hybrid-attention. The module uses the image patch feature sequence and the external token dictionary to calculate the feature representation of image patches through the multi-head hybrid-attention mechanism. Finally, the feature for classification is calculated using the class-attention layer, combining the feature representation of image patches with the class token.

### 3.2 Hybrid-attention Vision Transformer

The hybrid-attention module is designed to model the relationship between image patches in a video and the relationship across the different videos. It consists of the internal attention based on patch tokens within a video and the external attention based on an external token dictionary $D^{token}$. In the following, we describe the hybrid-attention module and its several variants.

**Internal Attention**. The internal attention is designed to model the internal patch dependencies in a video. Each patch representation is projected into the query, key, and value vector. The vectors are computed from $z^{(l-1)}_{\mathrm{p,t}}$ encoded by the preceding block:

$$\begin{aligned} q^{(l,a)}_{\mathrm{p,t}} &= W^{(l,a)}_Q f_{\mathrm{LN}}\left(z^{(l-1)}_{\mathrm{p,t}}\right) \in R^{d_h}, \\ k^{(l,a)}_{\mathrm{p,t}} &= W^{(l,a)}_K f_{\mathrm{LN}}\left(z^{(l-1)}_{\mathrm{p,t}}\right) \in R^{d_h}, \\ v^{(l,a)}_{\mathrm{p,t}} &= W^{(l,a)}_V f_{\mathrm{LN}}\left(z^{(l-1)}_{\mathrm{p,t}}\right) \in R^{d_h}, \\ \mathrm{p,t} \in \{\mathrm{p}', \mathrm{t}' | & \begin{matrix} \mathrm{p}' = 1, ..., N \\ \mathrm{t}' = 1, ..., F \end{matrix} \} \cup \{0, 0\} \end{aligned} \tag{1}$$

where $a = 1, 2, ..., A$ is the index over the multiple attention heads, and $d_h = d/A$ is the hidden dim of each head. $W_Q, W_K, W_V \in R^{d_h \times d}$ are learnable parameter matrices for projecting the queries, keys and values. Next, the attention weights $\alpha^{(l,a)}_{\mathrm{p,t}} \in R^{NF+1}$ are computed via the dot products of the query $q^{(l,a)}_{\mathrm{p,t}}$ with all keys:

$$\alpha_{\mathrm{p,t}}^{(l,a)} = \sigma\left(\frac{q_{\mathrm{p,t}}^{(l,a)^T}}{\sqrt{d_h}} \cdot \begin{bmatrix} k_{0,0}^{(l,a)}\{k_{\mathrm{p}',\mathrm{t}'}^{(l,a)}\}_{\substack{\mathrm{p}'=1,...,N \\ \mathrm{t}'=1,...,F}} \end{bmatrix}\right), \tag{2}$$

where $\sigma$ denotes the activation function. Above attention weights are used as coefficients in a weighted summation over value vectors to obtain the result of each attention head $s_{\mathrm{p,t}}^{(l,a)}$. Then, these outputs from each attention head are concatenated and passed through embedding matrix $W_O$ and the feed-forward network (FFN) which contains two MLP layers with GeLU activation:

$$\tilde{z}_{\mathrm{p,t}}^{(l)} = W_O \begin{bmatrix} s_{\mathrm{p,t}}^{(l,1)} \\ \vdots \\ s_{\mathrm{p,t}}^{(l,A)} \end{bmatrix} + z_{\mathrm{p,t}}^{(l-1)} \tag{3}$$

In (2), the attention coefficient is calculated by using the query vector of the image patch and the key vector of all image patches in the video. The internal attention is jointly computed by the spatial and the temporal dimension. A reduction in computation can be achieved by disentangling the spatial and the temporal dimension. For the spatial dimension, only $N+1$ query-key comparisons are made, using keys from the same frame as the query patch token exclusively:

$$\alpha_{\mathrm{p,t}}^{(l,a)space} = \sigma\left(\frac{q_{\mathrm{p,t}}^{(l,a)^T}}{\sqrt{d_h}} \cdot \left[k_{0,0}^{(l,a)}\{k_{\mathrm{p}',\mathrm{t}}^{(l,a)}\}_{\mathrm{p}'=1,...,N}\right]\right). \tag{4}$$

If we only consider the space-attention, the model becomes the ViT which encodes the feature from each frame, and the classifier vector is the global average of the features of all the frames. The baseline of time dimensional dependencies are proposed by TimeSformer [8], only making $F+1$ query-key comparisons and using the patches from the other frames in the same location as the query patch:

$$\alpha_{\mathrm{p,t}}^{(l,a)time} = \sigma\left(\frac{q_{\mathrm{p,t}}^{(l,a)^T}}{\sqrt{d_h}} \cdot \left[k_{0,0}^{(l,a)}\{k_{\mathrm{p},\mathrm{t}'}^{(l,a)}\}_{\mathrm{t}'=1,...,F}\right]\right) \tag{5}$$

The internal attention with divided space-time dimension can effectively reduce the computation costs, but it will increase the number of parameters of the model. In (4) and (5), the query vector and key vector are used twice, which are calculated with two different parameters. Compared with joint space-time attention mechanism, this method increases the number of parameters but reduces the costs of computation.

**External Attention.** The external attention is to model the external patch dependencies across the different videos. Since the number of external patches is huge, this paper proposes to use an external patch token dictionary to encode the external patch information. In this way, each patch is represented by the

combination of the element of the dictionary. Corresponding attention coefficients encode the dependencies between internal patches and external patches in different videos. The external token dictionary includes two trainable parts, the value set $\{v_1^e, v_2^e, ..., v_n^e\}$ and its corresponding key set $\{k_1^e, k_2^e, ..., k_n^e\}$. Given a patch embedding $z_{\mathrm{p,t}}^{(l-1)}$ from the layer, the attention coefficient is calculated via the dot products with the key set of the external token dictionary. Then the value set of the external token dictionary is linearly weighted to obtain the external feature $\hat{z}_{\mathrm{p,t}}^{(l)}$:

$$\begin{aligned} \alpha_i &= \sigma\left(\frac{q_{\mathrm{p,t}}^{(l)^T}}{\sqrt{d}} \cdot \{k_i^e\}_{i=1,\ldots,n}\right), \\ \hat{z} &= \sum_{i=1}^{n} \alpha_i v_i^e, \\ \hat{z}_{\mathrm{p,t}}^{(l)} &= f_{\mathrm{FFN}}\left(f_{\mathrm{LN}}\left(\hat{z}\right)\right) + \hat{z}. \end{aligned} \tag{6}$$

where $q_{\mathrm{p,t}}^{(l)}$ is the query vector of $z_{\mathrm{p,t}}^{(l-1)}$. Note that the external token dictionary in the external attention aims to learn general visual features from the dictionary. For the sake of simplicity, we only give the formula of the attention. We can also use the multi-head attention to calculate values of the external attention (Fig. 2).

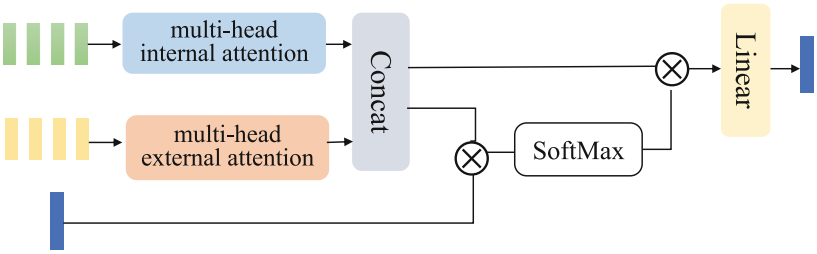



**Fig. 2.** Diagram of the combined-correlation hybrid-attention mechanism. The green block represents the image patches in the video and the yellow block represents the external patch token dictionary. (Color figure online)

**Hybrid Schemes for Hybrid-attention.** First of all, it is necessary to explain how to model the external prior information of videos using the external token dictionary. This paper assumes that the features of image patches are composed of two parts. The one is a linear combination of features of image patches in the video, utilizing the self-attention mechanism (i.e., the internal attention mechanism introduced earlier); The other is constructed from shared visual features instead of the current video. We use a learnable dictionary to construct shared visual features, which are related to all videos.

The keys and values calculated from the external dictionary and internal patches can be spliced together for subsequent calculation. In internal attention, the feature information of all image patches in the video is aggregated by calculating attention. It should be noted that not only the key vectors of all image patches but also all key sets from the external token dictionary are used to calculate the attention coefficient. In this way, an expanded attention coefficient is calculated as follows:

$$\alpha_{\mathrm{p,t}}^{l,a} = \sigma\left(\frac{q_{\mathrm{p,t}}^{l,a^T}}{\sqrt{d_h}} \cdot \{k_i^{l,a}\}\right), \tag{7}$$

where $i$ is the index of the image patch feature sequence and the external token dictionary. In this scheme, the model splices the attention coefficient matrix between the internal and external attention of a query patch to form an extended correlation matrix (called the combined-correlation type). With the attention coefficients calculated by (7), the final output of hybrid-attention is obtained by using the weighted summation over corresponding value vectors.

### 3.3 Class-attention

As we all know, the class token has two functions: guiding the learning of attention weights between patches and aggregating overall information to the linear classifier for classification [29]. Recent work has shown that separating two functions is beneficial to the classification. In this paper, we will explore whether this method influences the performance of video classification. Our implementation includes two stages: the hybrid-attention stage and the class-attention stage. In the hybrid-attention attention stage, we get the space-temporal feature of patches without the class token. In the class-attention stage, we only update the class token embedding while keeping patch features frozen.

Specifically, we first calculate the query, key, and value vectors:

$$\begin{aligned} q_{0,0}^{(l,a)} &= W_Q^{(l,a)} f_{\mathrm{LN}}\left(z_{0,0}^{(l-1)}\right), \\ k_{\mathrm{p,t}}^{(l,a)} &= W_K^{(l,a)} f_{\mathrm{LN}}\left(z_{\mathrm{p,t}}^{(l-1)}\right), \\ v_{\mathrm{p,t}}^{(l,a)} &= W_V^{(l,a)} f_{\mathrm{LN}}\left(z_{\mathrm{p,t}}^{(l-1)}\right). \end{aligned} \tag{8}$$

Next, the attention weights are given by

$$\alpha_{0,0}^{(l,a)} = \sigma\left(\frac{q_{0,0}^{(l,a)^T}}{\sqrt{d_h}} \cdot \left[k_{\mathrm{p',t'}}^{(l,a)}\right]_{\substack{\mathrm{p'}=1,...,N \\ \mathrm{t'}=1,...,F}}\right) \tag{9}$$

Then, we use the (3) to calculate the $z_{0,0}^{(l)}$ as the output of class-attention block. Finally, HaViT has a hybrid-attention module (which combines the internal attention and the external attention) and several class-attention layers (at which only the values of class tokens are updated).

# 4 Experiment

## 4.1 Setup

**Datasets**. All the experiments in this paper were conducted on the following datasets. The Kinetics dataset contains short clips sampled from YouTube. Since some videos on YouTube have been deleted or privatized, the dataset versions used in this paper include about 260k clips of Kinetics-400 and 397k clips of Kinetics-600. Note that, these numbers are lower than the original dataset and thus might induce a negative performance bias when compared with previous works. The Something-Something v2 (SSv2) consists of about 220k short videos with a length of 2 to 6 s, depicting human beings performing predefined basic actions on daily objects. Since the objects and backgrounds in the videos are consistent across the different action classes, this dataset tends to require stronger temporal modeling.

**Network Architecture**. The backbone modules closely follow the ViT architecture. Most of the experiments were performed using the HaViT-B/16($L = 12$, $A = 12$, $d = 768$, $P = 16$)and HaViT-S/16($L = 12$, $A = 6$, $d = 384$, $P = 16$), where L, $A$, d, P denotes the number of transformer layers, the number of heads, the embedding dimension, and the patch size.

**Training and Inference.** Unless otherwise stated, we sample frames uniformly across the video. For the training stage, we resize the smaller dimension of each frame to a value $\in [256, 320]$ and take a random crop of size $224 \times 224$ from the same location for all frames of the same video. In the inference, we give the accuracy results for $4 \times 3$ views (4 temporal clips and 3 spatial crops). The models are implemented by python and pytorch, and were trained on a DGX-v1 server (Table 1).

**Table 1.** Training hyperparamters for experiments.

| Config | Value |
|---|---|
| Optimizer | AdamW [30] |
| Momentum | $\beta_1, \beta_2 = 0.9, 0.999$ |
| Batch size | 64 |
| Learning rate | $2.5e^{-4}$ |
| Weight decay | 0.05 |
| Learning rate schedule | Cosine with linear warmup |
| Linear warmup epochs | 3 |
| Epochs | 30 |
| Dropout | 0.1 |

**Table 2.** Comparison with state-of-the-art methods on the Kinetics-400 dataset. T × frames are used in our experiments.

| Method | Top-1 | Top-5 | Views | TFLOPs |
|---|---|---|---|---|
| blVNet [31] | 73.5 | 91.2 | 3 × 3 | 0.84 |
| STM [32] | 73.7 | 91.6 | 10 × 3 | 2.01 |
| TEA [33] | 76.1 | 92.5 | 10 × 3 | 2.10 |
| CorrNet-101 [35] | 79.2 | – | 10 × 3 | 6.70 |
| ip-CSB-152 [23] | 79.3 | 93.8 | 10 × 3 | 3.27 |
| LGD-R101 [36] | 79.4 | 94.4 | 10 × 3 | – |
| SlowFast [24] | 79.8 | 93.9 | 10 × 3 | 7.02 |
| X3D-XXL [22] | 80.4 | 94.6 | 10 × 3 | 5.82 |
| TimeSformer-L [8] | 80.7 | 94.7 | 1 × 3 | 7.14 |
| ViViT-L/16 × 2 [7] | 80.6 | 94.7 | 4 × 3 | 17.35 |
| Swin-B [9] | 80.6 | 94.6 | 4 × 3 | 3.38 |
| **Our model**(8×) | 80.6 | 94.3 | 4 × 3 | 6.96 |
| **Our model**(16×) | **81.7** | **95.2** | 4 × 3 | 13.12 |

### 4.2 Comparison with State-of-the-Art

In this subsection, we compare our HaViT model with state-of-the-art models on three mentioned datasets. The results are shown in the Table 2-Table 4. Unless otherwise stated, we report the results on all the datasets using the 4 × 3 views.

Table 2 gives a comparison with the state-of-the-art on Kinetics-400, including convolution based networks and transformer-based networks. Compared with transformer-based model TimeSformer-L, the performance of our proposed structure is largely improved, and the classification accuracy is improved by 1.1%. Compared with the most advanced convolution network X3d, our model also improves the classification accuracy by 1.3% and uses fewer temporal views.

**Table 3.** Comparison with state-of-the-art on the Kinetics-600.

| Method | Top-1 | Top-5 | TFLOPs |
|---|---|---|---|
| AttentionNAS [6] | 79.8 | 94.4 | 1.03 |
| LGD-R101 [36] | 81.5 | 95.6 | – |
| SlowFast [24] | 81.8 | 95.1 | 7.02 |
| X3D-XL [22] | 81.9 | 95.5 | 1.05 |
| TimeSformer [8] | 82.4 | 95.3 | 5.11 |
| ViViT-L/16 × 2 [7] | 82.5 | – | 17.35 |
| Swin-B [9] | 84.0 | **96.5** | 3.38 |
| Our model(16×) | **84.5** | 96.1 | 13.12 |

Table 3 shows the comparison between our model and the state-of-the-art on Kinetics-600. The classification accuracy is much higher than the previous convolution network based method (+2.6%) and transformer-based method (+0.5%). Compared with the Kinetics-400, the size of the Kinetics-600 dataset is 0.6 times larger, which also shows that the performance of the transformer model will be improved when the dataset is large.

**Table 4.** Comparison with state-of-the-art on the SSv2.

| Method | Top-1 | Top-5 | TFLOPs |
|---|---|---|---|
| TRN [37] | 48.8 | 77.6 | – |
| SlowFast [24] | 61.7 | – | 7.02 |
| TSM [34] | 63.4 | 88.5 | 0.95 |
| STM [32] | 64.2 | 89.7 | 2.01 |
| TEA [33] | 65.1 | – | 2.10 |
| blVNet [31] | 65.2 | 90.3 | 0.84 |
| TimeSformer-L [8] | 62.5 | – | 7.14 |
| ViViT-L/16 × 2 [7] | 65.4 | 89.8 | 17.35 |
| **Our model**(16×) | **67.3** | **90.5** | 13.12 |

Table 4 compares our model with the state-of-the-art on the Something-Something v2 dataset. In terms of classification accuracy, our model is 2.1% higher than the previous convolution network blvnet, However, it is 1.9% higher than the previous transformer model ViViT-L/16.

### 4.3 Ablation Studies

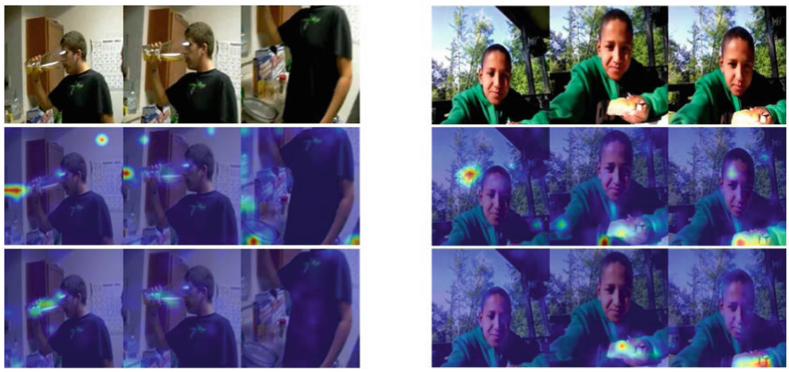
This subsection studies the impact of different components on the HaViT performance. For all experiments in this subsection, we use a lightweight model HaViT-S/16 with the model dim of 384 and adopt the Kinetics-400 dataset.

**Table 5.** Effect of different hybrid schemes for hybrid-attention.

| Hybrid scheme | Top-1 | Top-5 |
|---|---|---|
| Simplified | 77.1 | 92.5 |
| Multi-view | 77.5 | 92.7 |
| Combined-correlation | **78.2** | **93.1** |

**Hybrid Schemes for Hybrid-attention.** First, we consider the effect of different hybrid schemes for hybrid-attention on the final performance. Three

schemes of hybrid-attention are discussed, including simplified scheme, multi-view scheme, and combined-correlation scheme. The baseline is the simplified scheme, which directly adds the internal attention result and the external attention results. As shown in Table 5, compared to the simplified scheme, the multi-view scheme is more flexible and achieves better performance. For combined-correlation scheme, each head of attention is influenced by internal attention and external attention weights, and each query patch's feature is decided by the attention which influences it most. There's a trade-off between the internal attention and the external attention in the combined-correlation scheme. So, it's not surprising that the combined-correlation scheme has the best performance. Our model also adopts the combined-correlation scheme. Here, we give the visualization results of HaViT model. It can be seen from the Fig. 3 that compared with the existing divided space-time self-attention scheme, the proposed hybrid-attention scheme can better reflect the attention mechanism to the related objects in the video.



(a) drinking (b) eating hot dog

**Fig. 3.** Here are the results of attention visualization of the two models. "drinking" and "eating hot dog" represent two categories of data respectively; three of them represent the visualization results of the original video image, divided space-time self-attention and hybrid-attention respectively

**Effect of Attention Realization.** Previously, we introduced different ways to achieve attention, and here we will give the experimental results of different ways to achieve attention. The models using joint space-time self-attention and divided space-time self-attention are pre-trained on the image classification dataset Imagenet. From the Table 6, it is not difficult to find that the parameters of the model using the divided self-attention mechanism are more than those of the implementation of joint space-time self-attention. There are three modeling methods of joint space-time self-attention: the original attention method and

two linear computational complexity methods (linear activation and cosine re-weighting) [38]. The accuracy of the proposed linear activation method is lower than that of the original joint space-time method (−2.7%), but the inference speed is about 5 times faster. After using cosine re-weighting technology, the performance of the model is improved (+2.2%), but the classification accuracy is still not as good as the original joint space-time attention method. Cosine re-weighting technology attaches a larger weight of the attention coefficient to the query value vector, so that the query image patches pay more attention to the surrounding image patches, so the classification effect of this linear computational complexity method is better. Although the space-time joint method of linear computational complexity introduced above is fast, there are also unstable problems in the training process. Compared with the joint space-time self-attention method, the accuracy of divided space-time self-attention method is improved by about 1%, and the calculation speed of the model is about 3 times faster. Therefore, divided space-time attention is used in internal attention modeling.

**Table 6.** Effect of internal attention realization.

| Internal attention | Top-1(%) | Parameters(M) |
|---|---|---|
| Joint space-time | 77.3 | 26.4 |
| Linear activation | 74.6 | 26.4 |
| Cosine re-weighting | 76.8 | 26.4 |
| Divided space-time | **78.1** | 34.5 |

### 4.4 With Different Vision Transformers

To verify that the hybrid-attention scheme proposed in this chapter can be combined with different transformer models, different vision transformers are used for experiments. Firstly, different vision transformer models are extended to 3-dimensional space to get the corresponding video transformer model. Then different models are used to experiment on the Kinetics-400 dataset and test the classification accuracy. Finally, the models obtained by combining different video transformers with hybrid-attention schemes are trained and tested. The classification accuracy results are shown in Table 7. It is not difficult to find that in the three different transformer models, the hybrid-attention scheme obtained by using the external token dictionary can achieve better results. The proposed hybrid-attention scheme can be combined with other vision transformer models and improve the performance of its models.

**Table 7.** Effect of hybrid schemes combined with different vision transformers

| Transformer model | External token dictionary | Top-1(%) |
|---|---|---|
| ViT-S [5] | without | 77.6 |
| | with | **78.2**↑ |
| CvT [39] | without | 76.6 |
| | with | **77.4**↑ |
| Swin-T [9] | without | 78.4 |
| | with | **78.9**↑ |

## 5 Conclusion

In this paper, we propose a new hybrid-attention based vision transformer model for video classification, which explicitly exploits external patch dependencies across videos. Instead of using self-attention, it uses hybrid-attention to model both long-range patch dependencies within a video as well as external patch dependencies across videos. Compared to existing vision transformers, our model achieves competitive or better performance on public datasets including Kinetics-400/600 and SSv2. Experiments also show that hybrid-attention can be integrated into existing transformer models and improve their performance.

**Acknowledgements.** This work was supported in part to Dr. Liansheng Zhuang by NSFC under contract No. U20B2070 and No. 61976199.

## References

1. Dong, Y., Li, J.: Video retrieval based on deep convolutional neural network. In: Proceedings of the 3rd International Conference on Multimedia Systems and Signal Processing, pp. 12–16 (2018)
2. Muhammad, K., Khan, S., Elhoseny, M., Ahmed, S.H., Baik, S.W.: Efficient fire detection for uncertain surveillance environment. IEEE Trans. Industr. Inf. **15**, 3113–3122 (2019)
3. Vaswani, A., et al.: Attention is all you need (2017)
4. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
5. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale (2020)
6. Wang, X., et al.: AttentionNAS: spatiotemporal attention cell search for video classification. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12353, pp. 449–465. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58598-3_27
7. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: a video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6836–6846 (2021)

8. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML, p. 4 (2021)
9. Liu, Z., et al.: Video swin transformer. arXiv preprint arXiv:2106.13230 (2021)
10. Zhang, Y., et al.: VidTr: video transformer without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13577–13587 (2021)
11. Neimark, D., Bar, O., Zohar, M., Asselmann, D.: Video transformer network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3163–3172 (2021)
12. Chen, F., Zhang, L., Yu, H.: External patch prior guided internal clustering for image denoising. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2015)
13. Ma, Z., Liao, R., Tao, X., Xu, L., Jia, J., Wu, E.: Handling motion blur in multi-frame super-resolution. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5224–5232 (2015)
14. Kay, W., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
15. Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., Zisserman, A.: A short note about kinetics-600 (2018)
16. Goyal, R., et al.: The "something something" video database for learning and evaluating visual common sense. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5842–5850 (2017)
17. Laptev, I.: On space-time interest points. Int. J. Comput. Vision **64**, 107–123 (2005)
18. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. Int. J. Comput. Vision **103**, 60–79 (2013)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
20. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497 (2015)
21. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
22. Feichtenhofer, C.: X3D: expanding architectures for efficient video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
23. Tran, D., Wang, H., Torresani, L., Feiszl, M.: Video classification with channel-separated convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
24. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
25. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3D residual networks. In: Proceedings of the IEEE Conference on Computer Vision (ICCV) (2017)
26. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)

27. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: speed-accuracy trade-offs in video classification. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11219, pp. 318–335. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01267-0_19
28. Hao, Y., Zhang, H., Ngo, C.W., Liu, Q., Hu, X.: Compact bilinear augmented query structured attention for sport highlights classification. In: Proceedings of the 28th ACM International Conference on Multimedia (2020)
29. Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H.: Going deeper with image transformers. arXiv preprint arXiv:2103.17239 (2021)
30. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)
31. Fan, Q., Chen, C.F.R., Kuehne, H., Pistoia, M., Cox, D.: More is less: learning efficient video representations by big-little network and depthwise temporal aggregation. In: Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, Curran Associates, Inc. (2019)
32. Jiang, B., Wang, M., Gan, W., Wu, W., Yan, J.: STM: spatiotemporal and motion encoding for action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
33. Li, Y., Ji, B., Shi, X., Zhang, J., Kang, B., Wang, L.: Tea: temporal excitation and aggregation for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 909–918 (2020)
34. Lin, J., Gan, C., Han, S.: TSM: temporal shift module for efficient video understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
35. Wang, H., Tran, D., Torresani, L., Feiszli, M.: Video modeling with correlation networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
36. Qiu, Z., Yao, T., Ngo, C.W., Tian, X., Mei, T.: Learning spatio-temporal representation with local and global diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
37. Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal relational reasoning in videos. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11205, pp. 831–846. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01246-5_49
38. Qin, Z., et al.: cosformer: Rethinking softmax in attention. In: International Conference on Learning Representations (2021)
39. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: CVT: introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22–31 (2021)