

An Image is Worth 16x16 Words, What is a Video Worth?

Gilad Sharir Asaf Noy Lihi Zelnik-Manor
DAMO Academy, Alibaba Group

Abstract

Leading methods in the domain of action recognition try to distill information from both the spatial and temporal dimensions of an input video. Methods that reach State of the Art (SotA) accuracy, usually make use of 3D convolution layers as a way to abstract the temporal information from video frames. The use of such convolutions requires sampling short clips from the input video, where each clip is a collection of closely sampled frames. Since each short clip covers a small fraction of an input video, multiple clips are sampled at inference in order to cover the whole temporal length of the video. This leads to increased computational load and is impractical for real-world applications. We address the computational bottleneck by significantly reducing the number of frames required for inference. Our approach relies on a temporal transformer that applies global attention over video frames, and thus better exploits the salient information in each frame. Therefore our approach is very input efficient, and can achieve SotA results (on Kinetics dataset) with a fraction of the data (frames per video), computation and latency. Specifically on Kinetics-400, we reach 78.8 top-1 accuracy with $\times 30$ less frames per video, and $\times 40$ faster inference than the current leading method.¹

1. Introduction

The stellar growth in video content urges the need for more efficient video recognition. Increased camera coverage and constantly growing network bandwidth for video streaming are making online recognition [19, 41] essential in varied domains such as robotics, security and human-computer interaction. Additional applications like large-scale video retrieval benefit directly from faster recognition [1], as well as from efficient utilization of video frames transcoding.

In action recognition the task is to classify a video by extracting relevant information from its individual frames. The success of Convolutional Neural Networks (CNN) over images has been utilized for action recognition via 3D convolutions, extracting both spatial and temporal information

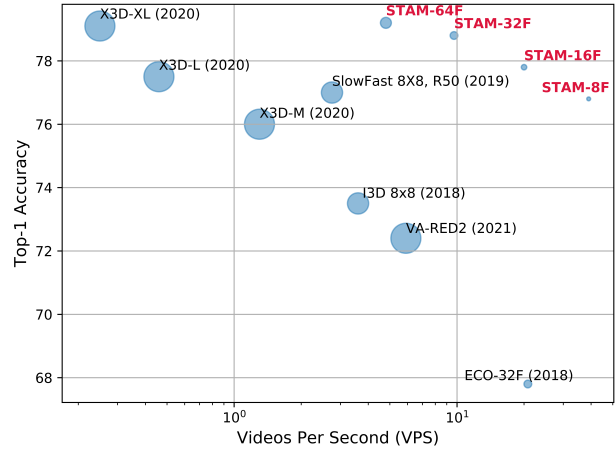


Figure 1. **Kinetics-400** top-1 Accuracy vs Runtime, measured over Nvidia V100 GPU and presented in log-scale. Markers sizes are proportional to the number of frames used per video by leading methods. Our method provides dominating trade-off for those three properties.

out of consecutive frames. Since 3D convolutions are computationally expensive, the common practice is to apply those on a predefined number of short video clips, each composed of densely sampled frames, and average the predictions over these clips. Since the clips should cover the entire video for accurate predictions, a large fraction of the video frames is used by such methods, leading to computational bottlenecks of frames processing and transcoding. Recent methods addressed the processing bottleneck from different angles: more efficient per-frame architectures [11, 34] and 3D modules [33, 4], clip sampling [38] and two-stream networks [38, 30]. While the trade-off between accuracy and efficiency is continuously improving, for many real-time applications the required runtime is lower by orders of magnitude than the ones offered by current state-of-the-art methods.

In this work, we take a different path towards efficient action recognition. We train classifiers to learn spatio-temporal representations from small numbers of uniformly sampled video frames via an end-to-end attention mechanism. Our approach is motivated by human action recog-

¹Code is available at: <https://github.com/Alibaba-MIIL/STAM>



Figure 2. **Frame attention** 16 frames uniformly sampled from a 10 second input video depicting 'beekeeping'. These frames are used as input to our model. Each frame's border displays the attention weight of that frame corresponding to the classification token (in heatmap range). We see that more attention is given to frames in which the action can clearly be identified.

nition that is shown to maintain a similar accuracy when a small numbers of frames is viewed instead of the entire video [28]. Conversely, multiple clips inference with 3D convolutions is often involved with redundant computations as consecutive video frames tend to be similar. In addition, its scope is limited by design to short actions, while real-world applications often span over larger intervals.

Inspired by recent breakthroughs in sequence modeling in the field of Natural language Processing (NLP), we propose a natural extension of Visual Transformers (ViT) [9] to videos. We view a video as analog to a text paragraph to be classified efficiently. To that end, we sample sentences (images) uniformly from it and divide those to words (patches). In NLP, the Transformer model [35] has proven superior to other sequence modeling techniques such as RNNs. The Transformer builds on a multi-head self attention layer, that learns global attention over the elements in the sequence. Similarly, our approach relies entirely on transformers, for both the spatial and the temporal dimensions.

We introduce an action classification model composed exclusively of self-attention layers operating in the spatial and temporal directions. We name our model **STAM** (Space Time Attention Model). The input sequence, in our case, is the sequence of image patches extracted from the individual frames and linearly projected onto a patch embedding space. First a spatial and then a temporal Transformer encoder models are applied on top of this embedding sequence to extract a video level representation or attention weighting of the frames. By leveraging this attention mechanism, we claim that the video sequence can be temporally subsampled by a larger factor than has previously been achieved, without degradation of the classification accuracy.

Figure 1 demonstrates the trade-off between the accuracy and runtime of top action recognition methods. While previous models are either accurate or efficient, models trained with our method offer a good combination of both. For example, STAM-32F achieves comparable accuracy to X3D-XL while being $\times 40$ faster.

The motivation behind the proposed method is that applying global self-attention over a sequence of input frames is the key to reduce the number of required frames, by allowing information from individual frames to be propagated globally across the entire sequence. 3D convolutions, on the

other hand, extract information locally over a small temporal (and spatial) scale, and therefore require frames to be sampled on a lower scale (i.e. dense sampling). In the NLP domain, where transformers are mostly being used, there is no issue of temporal continuity (and sampling density) since the words in a sentence are not temporally continuous as frames in the video are. Hence, our approach is a unique advantage of using Transformers on video data.

Subsampling the input video substantially reduces the computational load during training and inference, and furthermore, has the additional benefit of lowering the cost of retrieving input data. Indeed, in several applications there is a cost associated with retrieving input data from storage, or across a communications network. In such bandwidth limited applications most action recognition methods are prohibitively expensive for deployment, and methods like ours that rely on significantly less input data to operate, possess a clear advantage. To give an idea of such a scenario, suppose that there is a cost incurred for every access to a video frame located in storage. For typical methods, $30 \times 16 = 480$ frames are accessed in order to perform inference on a video. Compared to our method which requires 16 frames for the same video, the cost reduction we achieve is 30-fold, in addition to the reduction in run-time.

An additional advantage of STAM is that it is an end-to-end trainable model. This is both simpler to implement (requiring the same sampling strategy and model for train and inference) and has fewer hyperparameters. Methods employing multi-clip averaging during test-time cannot be considered end-to-end trainable, since they add an additional temporal aggregation (averaging) layer during inference. Since it is used only at inference time, this additional layer is an ad-hoc component, whose effects are not taken into account during training. (See Figure 4)

Our contributions can be summarized as follows:

- We propose a novel method for video action recognition that is entirely based on transformers for representation of spatio-temporal visual data. It is very simple, end-to-end trainable, and able to capture video information using only 3% of the data processed by leading efficient methods.
- Our method matches state-of-the-art accuracy while

being more efficient by orders of magnitude. Specifically, on Kinetics-400 benchmark it achieves 78.8 top1-accuracy with $\times 40$ faster inference time, or alternatively improving the efficient ECO [41] accuracy by +8% while being twice as fast. This makes it a leading solution for latency-sensitive video understanding.

2. Related Work

2.1. Transformers and Self Attention

Self-attention, sometimes called intra-attention is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence. Self-attention layers are the building blocks of an encoder-decoder architecture called Transformer [35].

The Transformer architecture has become the dominant model in the field of NLP, outperforming previous methods on tasks, such as language translation, and text generation ([8], [35]). Attempts to introduce Transformers to the computer vision domain [9, 32] use only attention based layers instead of the commonly used convolutional layers, and produce state-of-the-art results on image classification benchmarks such as ImageNet. Other methods ([3], [14]) combine convolutional networks with transformers for object detection. Other methods that apply self-attention in vision tasks include [36, 23, 25, 7]. These methods apply the Transformer model on the image pixel level, and therefore have to resort to approximations (either downsampling the image, or applying local attention instead of global). More similar to our approach, [15] applies a Transformer model in the domain of action recognition. However, their model is a hybrid of a 3D convolutional model and a Transformer model that acts on the CNN’s output feature vectors. Since it relies on 3D convolutions as part of the network, it has the same disadvantages of convolutional models (requiring dense sampling and a large number of frames), while our method is fully based on the Transformer model.

2.2. Action Recognition

Action recognition method typically operate by applying layers of 3D or 2D convolutions on spatio-temporal data [12, 11], [4, 33].

X3D [11] use network search and hyperparameter optimisation to find the network and sampling parameters (network depth, width, input spatial resolution, temporal resolution), while SlowFast [12] train two networks operating on different temporal resolutions. R3D [33] decomposes the 3D convolution operator into two separate convolutions operating on the temporal and spatial dimensions. Although these works improve classification efficiency by modifying the network structure, they still require densely sampled frames as input. We overcome this limitation by removing the dependence of 3D convolutions, and modeling the

temporal information via a self-attention sequence model.

Several other works apply 2D convolutional networks on individual frames [18], [10], and capture the temporal dependence by shifting feature maps in the temporal direction. In these methods information is propagated in a local neighborhood of frames, while in our case the global self-attention allows interaction across the whole spatial temporal dimensions.

Another line of research is focused on reducing the computation cost of existing action recognition methods. These works introduce techniques to improve network efficiency by adaptive resolution sampling [21], importance clip sampling [17] or reducing redundant computation using linear approximations of feature maps [22]. However, these works still rely on multi-clip testing for inference, and thus suffer from the same type of inefficiency which our method proposes to solve. We tackle the problem of computation by reducing the required input frames sampled from the video.

Additional works focus on action recognition with sub-sampled data. Mauthner et al. [20] suggested a method that uses a single-frame representation for short video sequences based on appearance and on motion information. Other methods proposed encoding techniques for video representations [24, 37]. Similarly to the methods that use video clips, they processed samples separately and fused the results over time. ECO [41] sampled frames uniformly and applied a long-term spatio-temporal architecture over those. They learned per-frame representations with 2D CNNs and fed them into a 3D CNN afterwards. Our motivation and input data is similar, and the use of spatial and temporal encoders offers a significant improvement over their method.

3. Method

As shown by the recent work on Visual Transformers (ViT) [9], self-attention models provide powerful representations for images, by viewing an image as a sequence of words, where each word embedding is equivalent to a linear projection of an image patch.

In this section, We design a convolution-free model that is fully based on self-attention blocks for the spatio-temporal domain. By that, we offer an extension of ViT that captures temporal representations. While the proposed method is focused on action recognition, we note that it can be easily modified for additional video understanding tasks.

While the attention mechanism can be extended to temporal dependencies in several ways, our design is modular, making its implementation simpler and intuitive. Most importantly, our design naturally leverages the advantages of attention mechanisms compared to convolution operators when it comes to better utilization of temporal information.

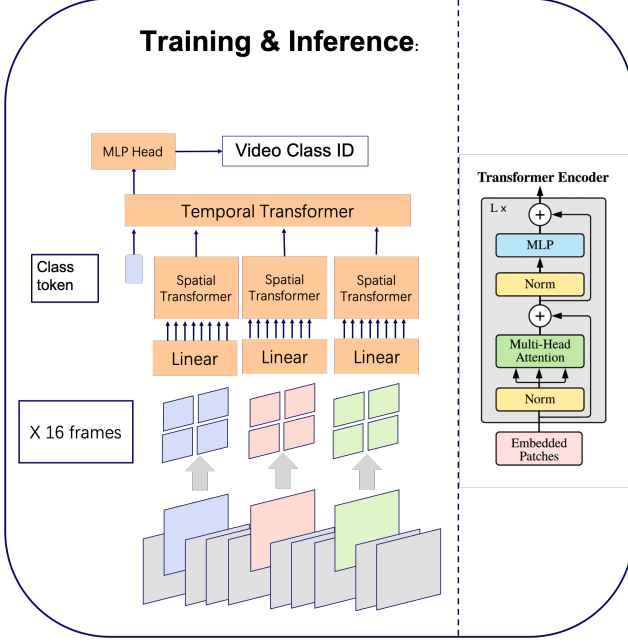


Figure 3. Our proposed Transformer Network for video

Our goal is to provide a model that can utilize sparsely subsampled temporal data for accurate predictions. Such model need to be able to capture long-term dependencies as well. While 2D convolutions filters are tailor-made for the structure of images, utilizing local connections and providing desired properties for object recognition and detection [16], the same properties might negatively affect the processing of subsampled temporal data. While a series of 3D-convolutions *can* learn long-term interactions due to increased receptive field, they are biased towards local ones. In order to verify this, we conducted an experiment: we fed leading methods that are based on 3D convolutions with the same subsampled data as in our method. The results are presented in Table 5. The performance of both methods degraded significantly, the error of X3D increased by 23% and SlowFast error by 50%.

Transformers offer advantages over their convolutional counterparts regarding modeling long-term dependencies. While a multi-head self-attention layer with sufficient number of heads is at least as expressive as any convolutional layer [6], it also has the ability of directly model long-distance interactions [26].

We propose a combined *spatial and temporal transformer* (STAM) which takes a sequence of frames sampled from the video as input, and outputs a video level classification prediction. As illustrated in Figure 3, we process the sampled frames with a spatial transformer following the method of [9], and aggregate the resulting frame embeddings with a *temporal transformer*. In this way we separate the spatial attention (on each frame) from the temporal

attention applied to the sequence of frame embedding vectors. This separation between the spatial and temporal attention components has several advantages. First, this reduces the computation by breaking down the input sequence into two shorter sequences. In the first stage each patch is compared to N other patches within a frames. The second stage compares each frame embedding vector to F other vectors, resulting in less overall computation than comparing each patch to NF other patches.

The second advantage stems from the understanding that temporal information is better exploited on a higher (more abstract) level of the network. In many works the where 2D and 3D convolutions are used in the same network, the 3D components are only used on the top layers. Using the same reasoning, we apply the temporal attention on frame embeddings rather than on individual patches, since frame level representations provide more sense of what’s going on in a video compared to individual patches.

Input embeddings. The input to the spatio-temporal transformer is $X \in \mathbb{R}^{H \times W \times 3 \times F}$ consists of F RGB frames of size $H \times W$ sampled from the original video. Each frame in this input block is first divided into non-overlapping patches. For a frame of size $H \times W$, we have $N = HW/P^2$ patches of size $P \times P$.

These patches are flattened into vectors and linearly projected into an embedding vector:

$$\mathbf{z}_{(p,t)}^{(0)} = E\mathbf{x}_{(p,t)} + \mathbf{e}_{(p,t)}^{pos} \quad (1)$$

where input vector $\mathbf{x}_{(p,t)} \in \mathbb{R}^{3P^2}$, and embedding vector $\mathbf{z}_{(p,t)} \in \mathbb{R}^D$ are related by a learnable positional embedding vector $\mathbf{e}_{(p,t)}^{pos}$, and matrix E . The indices p , and t are the patch and frame index, respectively with $p = 1, \dots, N$, and $t = 1, \dots, F$. In order to use the Transformer model for classification, a learnable classification token is added in the first position in the embedding sequence $\mathbf{z}_{(0,0)}^{(0)} \in \mathbb{R}^D$. As will be shown, this classification token will be used to encode the information from each frame and propagate it temporally across the sequence of frames. For this reason we include a separate classification token for each frame in the sequence $\mathbf{z}_{(0,t)}^{(0)}$.

Multi-head Self-Attention block (MSA). STAM consists of L MSA blocks. At each block $\ell \in \{1, \dots, L\}$, and head $a \in \{1, \dots, \mathcal{A}\}$, each patch representation is transformed into query, key, and value vectors. The representation produced by the previous block $\mathbf{z}_{(p,t)}^{\ell-1}$ is used as input.

$$\mathbf{q}_{(p,t)}^{(\ell,a)} = W_Q^{(\ell,a)} \text{LN} \left(\mathbf{z}_{(p,t)}^{(\ell-1)} \right) \in \mathbb{R}^{D_h} \quad (2)$$

$$\mathbf{k}_{(p,t)}^{(\ell,a)} = W_K^{(\ell,a)} \text{LN} \left(\mathbf{z}_{(p,t)}^{(\ell-1)} \right) \in \mathbb{R}^{D_h} \quad (3)$$

$$\mathbf{v}_{(p,t)}^{(\ell,a)} = W_V^{(\ell,a)} \text{LN} \left(\mathbf{z}_{(p,t)}^{(\ell-1)} \right) \in \mathbb{R}^{D_h} \quad (4)$$

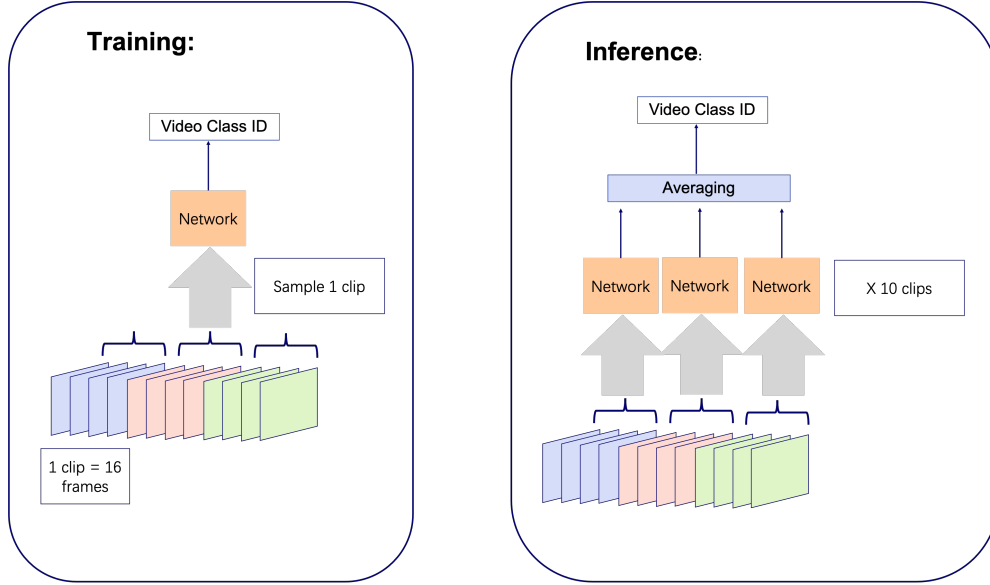


Figure 4. **Training and inference commonly used by other methods.** In the common approach multiple clips are sampled from each video, i.e., dense sampling of frames. Furthermore, training and inference have a different structure.

Where LN represents a LayerNorm [2]. The dimension of each attention head is given by $D_h = D/A$.

The attention weights are computed by a dot product comparison between queries and keys. The self-attention weights $\alpha_{(p,t)}^{(\ell,a)} \in \mathbb{R}^{N \times F}$ for patch (p, t) are given by:

$$\alpha_{(p,t)}^{(\ell,a)} = \text{SM} \left(\frac{\mathbf{q}_{(p,t)}^{(\ell,a)^\top}}{\sqrt{D_h}} \cdot \left[\mathbf{k}_{(0,t)}^{(\ell,a)} \left\{ \mathbf{k}_{(p',t')}^{(\ell,a)} \right\}_{\substack{p'=1,\dots,N \\ t'=1,\dots,F}} \right] \right) \quad (5)$$

where $\text{SM}()$ denotes the softmax activation function, and $\mathbf{k}_{(0,t)}^{(\ell,a)}$ is the key value associated with the class token of frame t .

Spatial attention Applying attention over all the patches of the sequence is computationally expensive, therefore, an alternative configuration is required in order to make the spatio-temporal attention computationally tractable. A reduction in computation can be achieved by disentangling the spatial and temporal dimensions. For the spatial attention, we apply attention between patches of the same frame:

$$\alpha_{(p,t)}^{(\ell,a)\text{space}} = \text{SM} \left(\frac{\mathbf{q}_{(p,t)}^{(\ell,a)^\top}}{\sqrt{D_h}} \cdot \left[\mathbf{k}_{(0,t)}^{(\ell,a)}, \left\{ \mathbf{k}_{(p',t)}^{(\ell,a)} \right\}_{p'=1,\dots,N} \right] \right). \quad (6)$$

The attention vector entries are used as coefficients in a weighted sum over the values for each attention head:

$$\mathbf{s}_{(p,t)}^{(\ell,a)} = \alpha_{(p,t),(0,t)}^{(\ell,a)} \mathbf{v}_{(0,t)}^{(\ell,a)} + \sum_{p'=1}^N \alpha_{(p,t),(p',t)}^{(\ell,a)} \mathbf{v}_{(p',t)}^{(\ell,a)}. \quad (7)$$

These outputs from attention heads are concatenated and passed through a 2 Multi-Layer Perceptron (MLP) layers with GeLU [13] activations:

$$\mathbf{z}'_{(p,t)}^{(\ell)} = W_O \begin{bmatrix} \mathbf{s}_{(p,t)}^{(\ell,1)} \\ \vdots \\ \mathbf{s}_{(p,t)}^{(\ell,A)} \end{bmatrix} + \mathbf{z}_{(p,t)}^{(\ell-1)} \quad (8)$$

$$\mathbf{z}_{(p,t)}^{(\ell)} = \text{MLP} \left(\text{LN} \left(\mathbf{z}'_{(p,t)}^{(\ell)} \right) \right) + \mathbf{z}_{(p,t)}^{(\ell-1)}. \quad (9)$$

The MSA and MLP layers are operating as residual operators thanks to added skip-connections

After passing through the spatial Transformer layers, the class embedding from each frame is used to produce an embedding vector \mathbf{f}_t . This frame embedding will be fed into the temporal attention.

$$\mathbf{f}_t = \text{LN} \left(\mathbf{z}_{(0,t)}^{(L_{\text{space}})} \right)_{t=1,\dots,F} \in \mathbb{R}^D. \quad (10)$$

where L_{space} is the number of layers of the spatial Transformer.

Temporal attention. The spatial attention provides a powerful representation for each individual frame by applying attention between patches in the same image. However, in order to capture the temporal information across the frame sequence, a temporal attention mechanism is required. The effect of temporal modeling can be seen in table 2. The spatial attention backbone provides a good representation of the videos, however the additional temporal attention provides a significant improvement over it. In our

model, the temporal attention layers are applied on the representations produced by the spatial attention layers.

For the temporal blocks of our model, we use the frame embedding vectors from eqn. 10, stacked into a matrix $X_{time} \in \mathbb{R}^{F \times D}$ as the input sequence. As before, we add a trainable classification token at index $t = 0$. This input sequence is projected into query/key/value vectors $\mathbf{q}_t^{(\ell,a)}$, $\mathbf{k}_t^{(\ell,a)}$, $\mathbf{v}_t^{(\ell,a)}$. The temporal attention is then computed only over the frame index.

$$\alpha_t^{(\ell,a)time} = \text{SM} \left(\frac{\mathbf{q}_t^{(\ell,a)\top}}{\sqrt{D_h}} \cdot \left[\mathbf{k}_0^{(\ell,a)} \left\{ \mathbf{k}_{t'}^{(\ell,a)} \right\}_{t'=1,\dots,F} \right] \right). \quad (11)$$

Next, we apply the same equations of the attention block eqn. (7) through (9), with a single axis describing the frame indices instead of the double (p, t) index which was used in those equations. The embedding vector for a video sequence is given by applying the layer norm on the classification embedding from the top layer:

$$\mathbf{y} = \text{LN} \left(\mathbf{z}_{(0)}^{(L_{time})} \right) \in \mathbb{R}^D. \quad (12)$$

where L_{time} is the number of temporal attention layers. An additional single layer MLP is applied as the classifier, outputting a vector of dimension equal to the number of classes.

The added cost of the temporal encoder layers over the spatial layers is negligible since they operate on an input sequence of length F , which is an order of magnitude smaller than the number of patches N (in our experiments usually $F = 16$, while $N = 196$). In this architecture, the total complexity is $O(FN^2 + F^2)$. If the attention operation were applied over all spatio-temporal patches, the complexity would be in the order of $O((FN)^2)$, which is prohibitively large.

An alternative possible configuration for the spatio-temporal attention would be to apply the temporal attention (between patches at the same spatial position, but from different frames), after the spatial attention within each block. We found that this variation produces slightly worse results, and higher computational cost.

Using an analogy to NLP, we consider each frame to be a sentence (in which patches play the role of word tokens), and each video is a paragraph of sentences. Following this analogy, it makes sense to apply a transformer separately on the sentences, extracting a representation vector per sentence, and then an additional Transformer on these vectors to predict a class (e.g. sentiment) from the entire paragraph.

4. Experiments

Implementation Details STAM consists of two parts: the spatial attention, and the temporal attention. In our experiments we closely follow the ViT_B model proposed by [35]

as the spatial Transformer. This model is the lighter version of the ViT family of models and contains 12 MSA layers, each with 12 self-attention heads. We use the imagenet-21K pretraining provided by [35] (unless specified otherwise). For the *Temporal Transformer* we use an even smaller version of the Transformer with 6-layers and 8-head self-attention ($L_{space} = 12$, $L_{time} = 6$). The temporal layers are trained from scratch, and initialized randomly.

We sample frames uniformly across the video. For training we resize the smaller dimension of each frame to a value $\in [256, 320]$, and take a random crop of size 224×224 from the same location for all frames of the same video. We also apply random flip augmentation, Cutout with factor 0.5, and auto-augment with Imagenet policy on all frames.

For inference we resize each frame so that the smaller dimension is 256, and take a crop of size 224×224 from the center. We use the same uniform frame sampling for training and inference.

Training For Kinetics-400 we train our model on 8V100 GPUs for 30 epochs with learning rate warm-up, and a cosine learning rate schedule. Compared to X3D and SlowFast, both trained with 128 GPUs for 256 epochs, our training is much faster and requires less epochs (~ 30).

Kinetics: We compare our method to others on Kinetics-400 dataset [40]. Table 1 shows our method achieves 77.8% top-1 accuracy using 16 sampled frames per video (at 270 GFLOPS). Compared to X3D_L, which achieves similar top-1 accuracy (77.5%) using 30 clips for inference (at 744 GFLOPS), this is an 0.3 improvement in top-1 accuracy, using only 36% of the computation required by X3D_L.

The reduction in run-time is even more significant. Compared to X3D_L we observe a reduction in inference time from 2.27 to 0.05 hrs for the entire validation set. We also calculate the video per second runtime by performing inference on a single batch of clips. We find that our method (with 16 frames) is able to outperform X3D_L by a factor of 43. This substantial improvement in runtime is partly due to the fewer input frames required by our method, and in part due to the improved efficiency of the ViT_B model compared to the more complex X3D architecture. Runtime is a more tangible metric for efficiency, and therefore we focus on it.

4.1. Ablation Experiments

This section provides ablation studies on Kinetics-400 comparing accuracy and computational complexity.

First, we compare models with only spatial attention to a model that has temporal attention as well. In table 2 we compare these two types of models and verify the positive effect of the temporal attention. We compare two variants of

Model	Top-1 Accuracy [%]	Flops \times views [G]	Param [M]	Runtime [hrs]	Runtime [VPS]
Oct-I3D + NL [5]	75.7	28.9×30	33.6	—	—
SlowFast 8×8 , R50	77.0	65.7×30	34.4	2.75	1.4
X3D-M	76.0	6.2×30	3.8	1.47	1.3
X3D-L	77.5	24.8×30	6.1	2.06	0.46
X3D-XL	79.1	48.4×30	11.0	—	—
TSM R50	74.7	65×10	24.3	—	—
Nonlocal R101	77.7	359×30	54.3	—	—
STAM (16 Frames)	77.8	270×1	96	0.05	20.0
STAM (64 Frames)	79.2	1040×1	96	0.21	4.8

Table 1. **Model comparison on Kinetics400.** Time measurements were done on Nvidia V100 GPUs with mixed precision. The Runtime [hrs] measures inference on Kinetics-400 validation set (using 8 GPUs), while the videos per second (VPS) measurement was done on a single GPU. Results of various methods are as reported in the relevant publications. The proposed STAM is an order of magnitude faster while providing SOTA accuracy.

Backbone+Temporal	Top-1 Accuracy [%]	Flops [G]
ViT+Temporal-Transformer	77.8	270
TResNet-M+Temporal-Transformer	75.7	93
ViT+Mean	75.1	265
TResNet-M+Mean	71.9	88

Table 2. **Temporal Transformer vs. Mean.** Comparing the Temporal Transformer representation vs. mean of frame embeddings.

Model (num. of frames)	Top-1 Accuracy [%]	Flops [G]
TResNet-M+Temporal (16)	75.7	93
TResNet-L+Temporal (8)	75.9	77
ResNet50+Temporal (16)	72.5	70
ViT-B+Temporal (16)	77.8	270

Table 3. **Using different backbones with the Temporal Transformer.** TResNet and ViT models use imagenet-21K pretraining, while ResNet50 is used with imagenet-1K pretraining.

Number of frames	Top-1 Accuracy [%]	Flops [G]
4	74.1	67
8	76.8	135
16	77.8	270
32	78.8	540
64	79.2	1080

Table 4. **Number of frames used for prediction** Comparing different number of frames sampled uniformly from the video as input, using the ViT+Temporal model.

Method	Top-1 Accuracy [%]
X3D-L	72.25 (77.5)
SlowFast-8x8-R50	65.5 (77.0)

Table 5. **Evaluating X3D and SlowFast with one clip (16 frames).** Applying the same sampling strategy that we use in Video Transformer on other methods. (In parentheses - accuracy at 30 views per video).

our method. The first is our full method using ViT_B backbone as the spatial attention model with a temporal transformer, and the second uses the same backbone but replace the temporal transformer with an average of the frame embeddings. The table shows that the temporal transformer provides a gain of 2.7 over the naive approach. Additionally, we repeat this experiment using a different backbone and see the consistency of the result across different backbones. We use a TRResNet-M [27] backbone, and again observe that the temporal transformer significantly improves accuracy.

In table 3, we show the effect of replacing the ViT model with a CNN model. We compare 3 different CNN variants (TRResNet-M, TRResNet-L [27], ResNet50) trained together with the temporal transformer. This experiment shows that the highest accuracy is achieved using the full spatial, and temporal transformer model. Although replacing the spatial transformer with a CNN model is possible, and achieves reasonable accuracy, it is less powerful than the combination of the spatial and temporal attention, which applies attention over all the patches of the frame sequence.

Number of frames. In table 4 we compare different sequence lengths as input to STAM. We sample 8, 16, 32, and 64 frames and compare the results. The clear trend is an increase in accuracy along with the increase in sequence length. For an increase of 16 additional frames (from 16 frames to 32 frames) we see an improvement of 1% to the accuracy. Switching the number of input frames from 32 to 64, results in a smaller gain of only 0.4. Using a larger number of frames doesn’t improve the accuracy.

The use of 4 frames to classify a video of length 10 seconds suggests the use of our method for longer range actions. For instance, if we use the same sampling rate for a 1 minute video, we would require 24 frames for inference.

Runtime comparison. In table 5 we evaluate the accuracy of two leading methods, with a reduced number of input frames. We use a single 16 frame input clip sampled uniformly, and so use these methods with the a similar running time to our method. The table clearly shows that by reducing the number of input frames, these methods suffer a large degradation in accuracy. This show that methods that rely on 3D convolutions require frames to be sampled at a higher rate than the one we use, and cannot be made to improve their runtime by sampling sparse input frames.

Figure 1 plots different action recognition methods on the accuracy vs. runtime (VPS) scale. We compare methods that are both designed for efficiency (ECO [41]), and methods that apply heavier models for increased accuracy (X3D [11], SlowFast [12]). We see that our method is on par with the accuracy of the slower methods and improves over their runtime by a significant margin. STAM’s runtime

is comparable to that of ECO, at 20 VPS, yet significantly outperforms that method in terms of accuracy by 8%. Since ECO employs a similar sampling strategy to ours (sampling individual frames across the video), we conclude that the temporal transformer is better at capturing the temporal information from separate frames.

5. Conclusion

In this work we have presented a method for efficient video action recognition that is entirely based on transformers. Inspired by NLP, we model a video as a paragraph and uniformly select frames that are modeled as sentences. This modeling allows us to utilize transformers to capture complex spatio-temporal dependencies between distinct frames, leading to accurate predictions based on a small fraction of the video data. The accuracy of our models’ predictions is comparable to state-of-the-art methods while being faster by orders of magnitude, making it a good candidate for latency-sensitive applications like real-time, such as recognition and video retrieval. In addition, our method is simple, end-to-end and generic, thus, it can be used for further video understanding tasks.

References

- [1] Andre Araujo and Bernd Girod. Large-scale video retrieval using image queries. *IEEE transactions on circuits and systems for video technology*, 28(6):1406–1420, 2017.
- [2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.
- [4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
- [5] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yanis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3435–3444, 2019.
- [6] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. *arXiv preprint arXiv:1911.03584*, 2019.
- [7] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. In *International Conference on Learning Representations*, 2020.
- [8] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.
- [10] Quanfu Fan, Chun-Fu (Richard) Chen, Hilde Kuehne, Marco Pistoia, and David Cox. More is less: Learning efficient video representations by big-little network and depthwise temporal aggregation. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- [11] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020.
- [12] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019.
- [13] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv: Learning*, 2016.
- [14] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [15] M Esat Kalfaoglu, Sinan Kalkan, and A Aydin Alatan. Late temporal modeling in 3d cnn architectures with bert for action recognition. In *European Conference on Computer Vision*, pages 731–747. Springer, 2020.
- [16] Eric Kauderer-Abrams. Quantifying translation-invariance in convolutional neural networks. *arXiv preprint arXiv:1801.01450*, 2017.
- [17] Bruno Korbar, Du Tran, and Lorenzo Torresani. Scsampler: Sampling salient clips from video for efficient action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6232–6242, 2019.
- [18] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019.
- [19] Jun Liu, Amir Shahroudy, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ssnet: scale selection network for online 3d action prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8349–8358, 2018.
- [20] Thomas Mauthner, Peter M Roth, and Horst Bischof. *Action recognition from a small number of frames*. Citeseer, 2009.
- [21] Yue Meng, Chung-Ching Lin, Rameswar Panda, Prasanna Sattigeri, Leonid Karlinsky, Aude Oliva, Kate Saenko, and Rogerio Feris. Ar-net: Adaptive frame resolution for efficient action recognition. In *European Conference on Computer Vision*, pages 86–104. Springer, 2020.
- [22] Bowen Pan, Rameswar Panda, Camilo Fosco, Chung-Ching Lin, Alex Andonian, Yue Meng, Kate Saenko, Aude Oliva, and Rogerio Feris. Va-red²: Video adaptive redundancy reduction. *arXiv preprint arXiv:2102.07887*, 2021.
- [23] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4055–4064, 2018.
- [24] Zhaofan Qiu, Ting Yao, and Tao Mei. Deep quantization: Encoding convolutional activations with deep generative model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6759–6768, 2017.
- [25] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- [26] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. *arXiv preprint arXiv:1906.05909*, 2019.
- [27] Tal Ridnik, Hussam Lawen, Asaf Noy, Emanuel Ben Baruch, Gilad Sharir, and Itamar Friedman. Tresnet: High performance gpu-dedicated architecture. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1400–1409, January 2021.
- [28] Konrad Schindler and Luc Van Gool. Action snippets: How many frames does human action recognition require? In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.
- [29] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016.
- [30] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014.
- [31] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- [32] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.
- [33] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.
- [34] Amin Ullah, Khan Muhammad, Weiping Ding, Vasile Palade, Ijaz Ul Haq, and Sung Wook Baik. Efficient activity recognition using lightweight cnn and ds-gru network for surveillance applications. *Applied Soft Computing*, 103:107102, 2021.
- [35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.
- [36] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

- [37] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015.
- [38] Yin-Dong Zheng, Zhaoyang Liu, Tong Lu, and Limin Wang. Dynamic sampling networks for efficient action recognition in videos. *IEEE Transactions on Image Processing*, 29:7970–7983, 2020.
- [39] Linchao Zhu, Du Tran, Laura Sevilla-Lara, Yi Yang, Matt Feiszli, and Heng Wang. Faster recurrent networks for efficient video classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13098–13105, 2020.
- [40] Andrew Zisserman, Joao Carreira, Karen Simonyan, Will Kay, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, and Mustafa Suleyman. The kinetics human action video dataset. 2017.
- [41] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. Eco: Efficient convolutional network for online video understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 695–712, 2018.

Appendix

A. Additional Datasets

We evaluate our method on 2 additional datasets: UCF101, Charades.

UCF101 UCF101 [31] is an older, yet still popular action recognition dataset. Our results on this dataset are presented in table 6. The results suggest that our model achieves a good trade-off between runtime and accuracy.

Model	Top-1 Accuracy [%]	Runtime [VPS]
ECO [41]	93.6	20.8
R(2+1)D-34	96.8	N/A
I3D	95.6	N/A
S3D	96.8	N/A
FASTER32 [39]	96.9	2.8
STAM-32	97.0	10

Table 6. **Results on UCF-101 dataset.** Results of various methods are as reported in the relevant publications. We compare methods that use only RGB frames as input (without Optical Flow), and are pretrained on Kinetics-400 or Imagenet

Charades Charades [29] is a dataset with longer range interactions, and multiple labels per video. Table 7 presents our results on this dataset. Although our model doesn’t reach SotA accuracy, it shows promise as an efficient model, requiring less input frames from each video, and less FLOPS.

Model	Top-1 Accuracy [%]	Flops \times views [G]
Nonlocal	37.5	544×30
STRG, +NL	39.7	630×30
Timeception	41.1	N/A
LFB, +NL	42.5	529×30
SlowFast, +NL	42.5	234×30
X3D-XL	43.4	48×30
STAM-64	39.7	1040×1

Table 7. **Results on Charades dataset.** Results of various methods are as reported in the relevant publications. Including methods that are pretrained on Kinetics-400 or Imagenet.