VIVIT: A VIDEO VISION TRANSFORMER

Review Paper - 1

GNR 650

By
Aditya Prakash (210260004)
K.S Saketh (210260025)

1 Introduction

After the massive success of Visual Transformers aka ViT and the fact that attention-based architectures are an intuitive choice for modelling long-range contextual relationships in video, they developed several transformer-based models for video classification.Performing object detection, classification, and segmentation on videos are different from just applying the same image algorithm on every frame because this approach ignores, the temporal relations between frames, which is a very important factor for developing a strong understanding of content.One of the main differences between image and video classification is the addition of a fourth dimension, time, to the input. This paper is from Google's research lab and in it, authors explore five different model variants for performing video classification using pure transformer-based architecture on different video classification benchmark datasets.

2 Patch Embedding

In videos, we don't have just one image, in fact, each sample is a video and is made of multiple frames. The authors of ViViT propose two methods for embedding video samples for passing through a model.

2.1 Uniform Frames Sampling

In this embedding approach, we apply ViT like patch creation on each frame in the video sample. In this embedding approach, each token is a patch extracted from a frame.

2.2 Tubelet Embedding

Instead of extracting a flat patch from each frame of the video sample, the authors suggest that we extract a series of patches from a video clip, aka a tube.

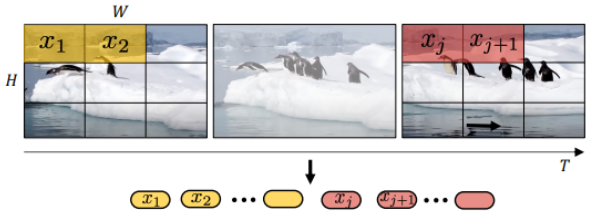

Figure 2: Uniform frame sampling: We simply sample n_t frames, and embed each 2D frame independently following ViT [18].

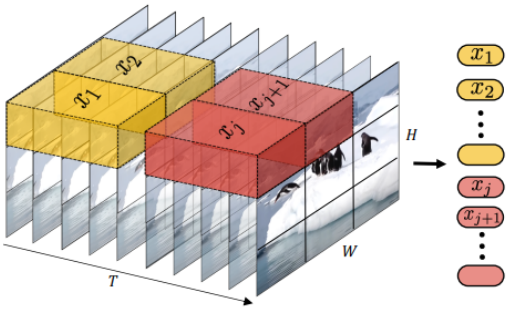

Figure 3: Tubelet embedding. We extract and linearly embed non-overlapping tubelets that span the spatio-temporal input volume.

3 Models

In this paper authors proposed 4 different variants of pure transformer-based video classification models inspired by ViT, for performing video classification.

3.1 Spatio Temporal attention

The first model is pretty straightforward, We tokenize a video sample using a Tubelet embedding-based approach and treat each Tubelet as a token(spatio-temporal tokens). Next up, each token is passed through a patch embedding layer and we add a position encoding to it and pass all tokens through a standard transformer encoder.

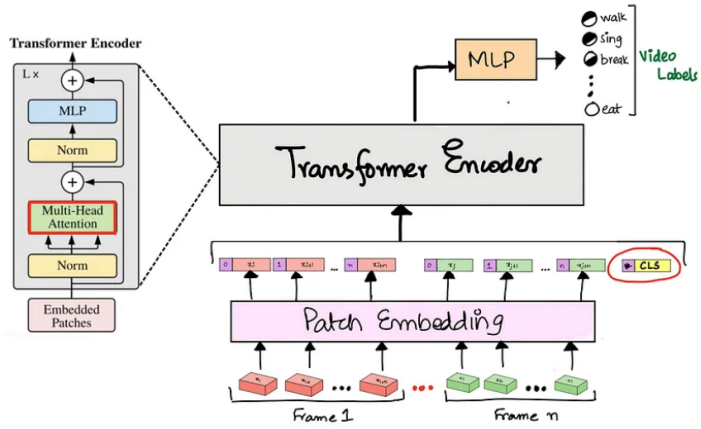

Figure 1: Model-1

3.2 Factorized Encoder

As shown in Fig.1, this model consists of two separate transformer encoders. The first, spatial encoder, only models interactions between tokens extracted from the same temporal index.Each clip is passed through the spatial transformer and we get an encoding vector for each clip. These encoding vectors, one per clip, along with a CLS token get added to respective position embedding and we pass them through a temporal transformer encoder.For temporal transformer, each token is a vector extracted from a clip, so each token is from a different temporal index in the video.

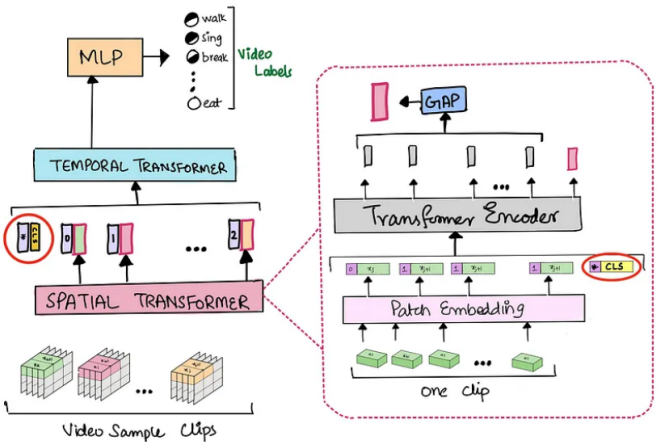

Figure 2: Model-2

3.3 Factorized Self attention

This is exactly similar to the first one, but the only change here is the transformer encoder block that is used is not the standard block used in original transformers. This new transformer block is pretty similar to the original standard transformer block, the only difference being the multi-headed self-attention. The MSA layers are factorized or broken into two parts. So unlike model 1 attention won't be computed among all the tokens.The first multi-headed self-attention layer computes attention between all the tokens extracted from the same spatial index (as in among all the tokens extracted from the same clip) and then the temporal self-attention layer computes attention between all tokens extracted from a different temporal index.

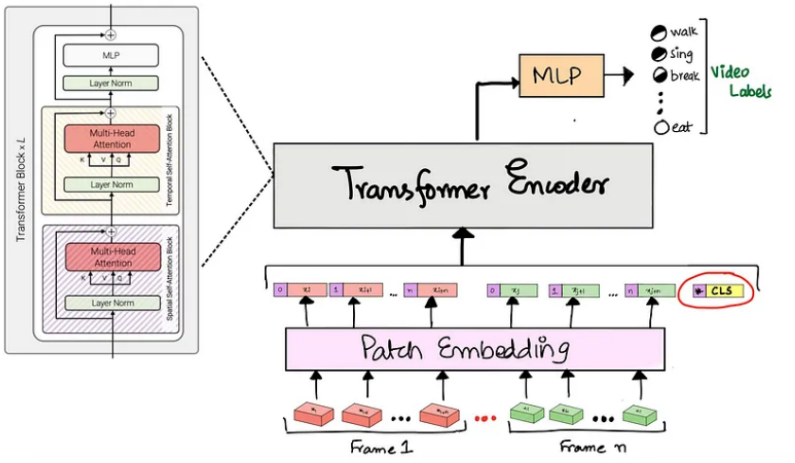

Figure 3: Model-3

3.4 Factorized dot product attention

So in model 4, the authors went on the more fine-grained level and factorized the dot product attention heads between the spatial and temporal aspects. Here half of the heads in the MSA layer were computing the dot product self-attention between tokens extracted from the same spatial index and the remaining heads were computing the dot product self-attention between the tokens extracted from different temporal indexes.his model is different from model 3 because in that model authors were trying to compute first spatial and temporal attention using all heads and two separate MSA layers, whereas, in model 4, authors used different heads in the same MSA layer to compute temporal and spatial self-attention.

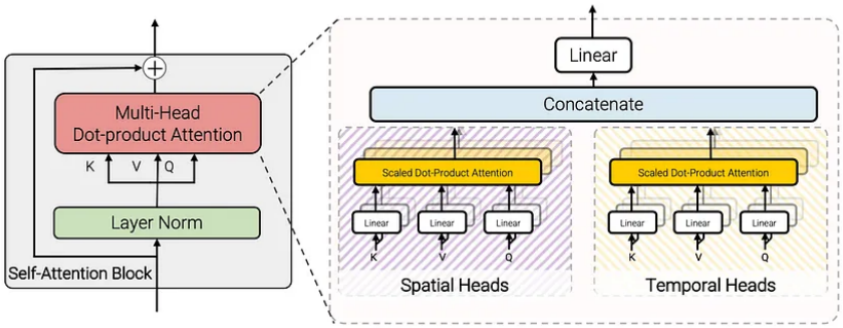

Figure 4: Model-4

4 Results

Table 6: Comparisons to state-of-the-art across multiple datasets. For "views", $x \times y$ denotes x temporal crops and y spatial crops. We report the TFLOPs to process all spatio-temporal views. "FE" denotes our Factorised Encoder model.

(a) Kinetics 400

Method	Top 1	Top 5	Views	TFLOPs
blVNet [19]	73.5	91.2	–	–
STM [33]	73.7	91.6	–	–
TEA [42]	76.1	92.5	10 × 3	2.10
TSM-ResNeXt-101 [43]	76.3	–	–	–
I3D NL [75]	77.7	93.3	10 × 3	10.77
CorrNet-101 [70]	79.2	–	10 × 3	6.72
ip-CSN-152 [66]	79.2	93.8	10 × 3	3.27
LGD-3D R101 [51]	79.4	94.4	–	–
SlowFast R101-NL [21]	79.8	93.9	10 × 3	7.02
X3D-XXL [20]	80.4	94.6	10 × 3	5.82
TimeSformer-L [4]	80.7	**94.7**	1 × 3	7.14
ViViT-L/16x2 FE	80.6	92.7	1 × 1	3.98
ViViT-L/16x2 FE	**81.7**	93.8	1 × 3	11.94
Methods with large-scale pretraining				
ip-CSN-152 [66] (IG [44])	82.5	95.3	10 × 3	3.27
ViViT-L/16x2 FE (JFT)	83.5	94.3	1 × 3	11.94
ViViT-H/14x2 (JFT)	**84.9**	**95.8**	4 × 3	47.77

(b) Kinetics 600

Method	Top 1	Top 5
AttentionNAS [76]	79.8	94.4
LGD-3D R101 [51]	81.5	95.6
SlowFast R101-NL [21]	81.8	95.1
X3D-XL [20]	81.9	95.5
TimeSformer-L [4]	82.2	**95.6**
ViViT-L/16x2 FE	**82.9**	94.6
ViViT-L/16x2 FE (JFT)	84.3	94.9
ViViT-H/14x2 (JFT)	**85.8**	**96.5**

(c) Moments in Time

	Top 1	Top 5
TSN [72]	25.3	50.1
TRN [86]	28.3	53.4
I3D [8]	29.5	56.1
blVNet [19]	31.4	59.3
AssembleNet-101 [54]	34.3	62.7
ViViT-L/16x2 FE	**38.5**	**64.1**

(d) Epic Kitchens 100 Top 1 accuracy

Method	Action	Verb	Noun
TSN [72]	33.2	60.2	46.0
TRN [86]	35.3	65.9	45.4
TBN [36]	36.7	66.0	47.2
TSM [43]	38.3	**67.9**	49.0
SlowFast [21]	38.5	65.6	50.0
ViViT-L/16x2 FE	**44.0**	66.4	**56.8**

(e) Something-Something v2

Method	Top 1	Top 5
TRN [86]	48.8	77.6
SlowFast [20, 80]	61.7	–
TimeSformer-HR [4]	62.5	–
TSM [43]	63.4	88.5
STM [33]	64.2	89.8
TEA [42]	65.1	–
blVNet [19]	65.2	**90.3**
ViVIT-L/16x2 FE	**65.9**	89.9

Figure 5: ViViT best performing models on 5 benchmark video classification datasets

It can be seen second model (Factorized encoder), the model performed better than all other model variants on all 5 datasets including Kinetics-400, 600, Epci Kitches, and Something Something v2 datasets.