ViViT: A Video Vision Transformer

School of Industrial and Management Engineering, Korea University

Eun Ji Koh

KOREA UNIVERSITY

DMQA

Contents

Research Purpose

❖ ViViT: A Video Vision Transformer (arXiv, 2021)

- Google Research에서 연구하였으며 2021년 08월 15일 기준으로 34회 인용됨

ViViT: A Video Vision Transformer

Anurag Arnab* Mostafa Dehghani* Georg Heigold Chen Sun Mario Lučić† Cordelia Schmid†
Google Research
{aarnab, dehghani, heigold, chensun, lucic, cordelias}@google.com

Abstract

We present pure-transformer based models for video classification, drawing upon the recent success of such models in image classification. Our model extracts spatio-temporal tokens from the input video, which are then encoded by a series of transformer layers. In order to handle the long sequences of tokens encountered in video, we propose several, efficient variants of our model which factorise the spatial- and temporal-dimensions of the input. Although transformer-based models are known to only be effective when large training datasets are available, we show how we can effectively regularise the model during training and leverage pretrained image models to be able to train on comparatively small datasets. We conduct thorough ablation studies, and achieve state-of-the-art results on multiple video classification benchmarks including Kinetics 400 and 600, Epic Kitchens, Something-Something v2 and Moments in Time, outperforming prior methods based on deep 3D convolutional networks. To facilitate further research, we will release code and models.

only very recently with the Vision Transformer (ViT) [15], that a pure-transformer based architecture has outperformed its convolutional counterparts in image classification. Dosovitskiy *et al*. [15] closely followed the original transformer architecture of [65], and noticed that its main benefits were observed at large scale – as transformers lack some of the inductive biases of convolutions (such as translational equivariance), they seem to require more data [15] or stronger regularisation [61].

Inspired by ViT, and the fact that attention-based architectures are an intuitive choice for modelling long-range contextual relationships in video, we develop several transformer-based models for video classification. Currently, the most performant models are based on deep 3D convolutional architectures [6, 17, 18] which were a natural extension of image classification CNNs [24, 57]. Recently, these models were augmented by incorporating self-attention into their later layers to better capture long-range dependencies [72, 20, 76].

As shown in Fig. 1, we propose pure-transformer mod-

DMQA

Research Purpose

❖ Motivation & Background

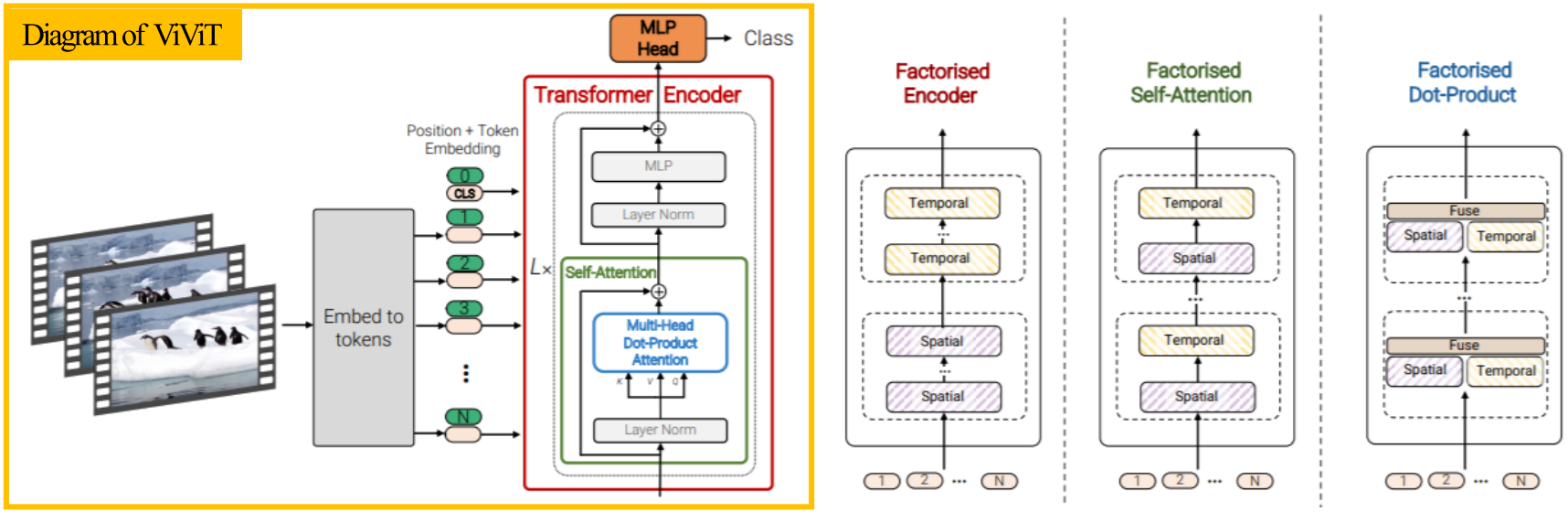
- Transformer는 sequence to sequence modeling에서 중요한 architecture
  - ➢ 비디오의 long-range contextual relationship modeling에 유리할 것으로 판단

- Transformer의 특징
  - ➢ Convolution layer를 사용하지 않고 multi-headed self-attention을 기초로 함
  - ➢ Long-range dependencies model에서 효율적으로 작동
  - ➢ Input sequence에 특정 부분이 아닌 all element에 대해 attend 가능
  - ➢ ViT의 경우 inductive bias가 부족하기 때문에 large dataset이 필요

Research Purpose

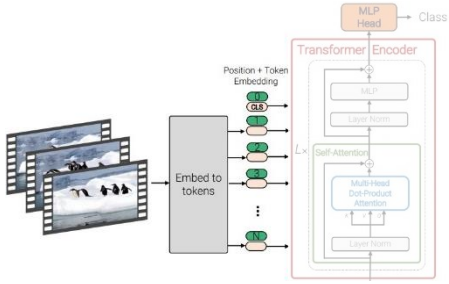
❖ ViViT: A Video Vision Transformer (arXiv, 2021)

- 비디오 classification을 위한 transformer-based model 제안 (ViViT)
  - ➢ Input 비디오에서 추출한 sequence of spatio-temporal token 연산을 위한 self-attention 사용
  - ➢ Spatial-temporal dimensions에 따른 model factorizing 방법 제시
  - ➢ Smaller dataset에서의 효율적 학습을 위한 model regularise 방식 및 pre-train image model 사용 방식 제시



DMQA

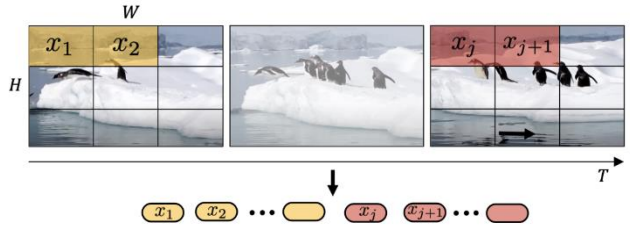
ViViT: A Video Vision Transformer



❖ Embedding video clips

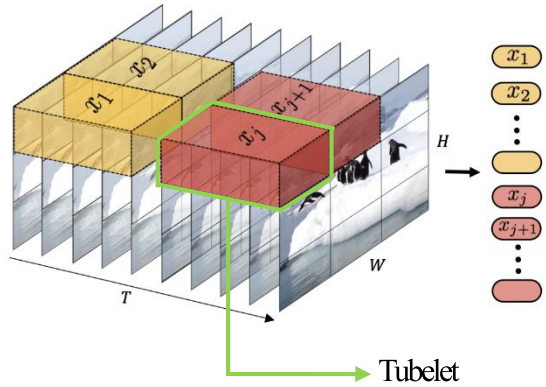
- Video를 a sequence of token에 mapping 하는 방법 2가지 고려

1) Uniform frame sampling



- Input 비디오 clip의 n_t frames을 균일하게 나누어 token 생성
- 각 token을 독립적으로 embedding한 후, concatenate
- ViT의 embedding 방식과 동일
- 다른 frame에서 나온 token 간의 temporal information이 fuse 됨

2) Tubelet embedding



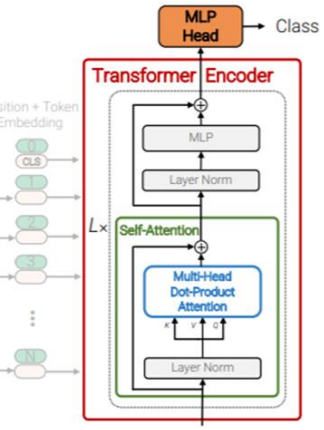
- Input Volume으로 부터 non-overlapping하게 spatio-temporal tubes를 추출한 후, d 차원으로 linear projection
- ViT의 embedding 방식을 3D로 확장한 것
- Tokenisation 과정에서 spatio-temporal information이 fuse 됨
- 작은 tubelet 사용할수록 computation complexity 증가

DMQA

ViViT: A Video Vision Transformer

❖ Transformer Models for Video

- Transformer architecture에서 video의 spatial, temporal dimension을 factorising 하기 위한 변형 제시
- **Model 1) Spatio-temporal attention**



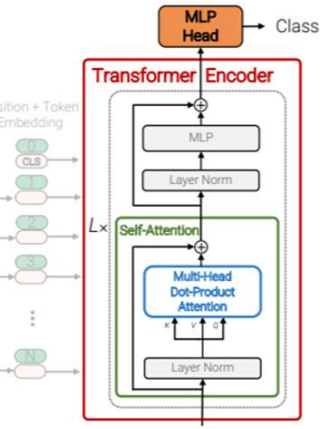
- 모든 spatio-temporal tokens을 transformer encoder에 입력
  - ➢ 모든 token에 대한 self-attention
- Multi-headed self-attention (MSA)는 token 개수에 따른 complexity 증가

DMQA

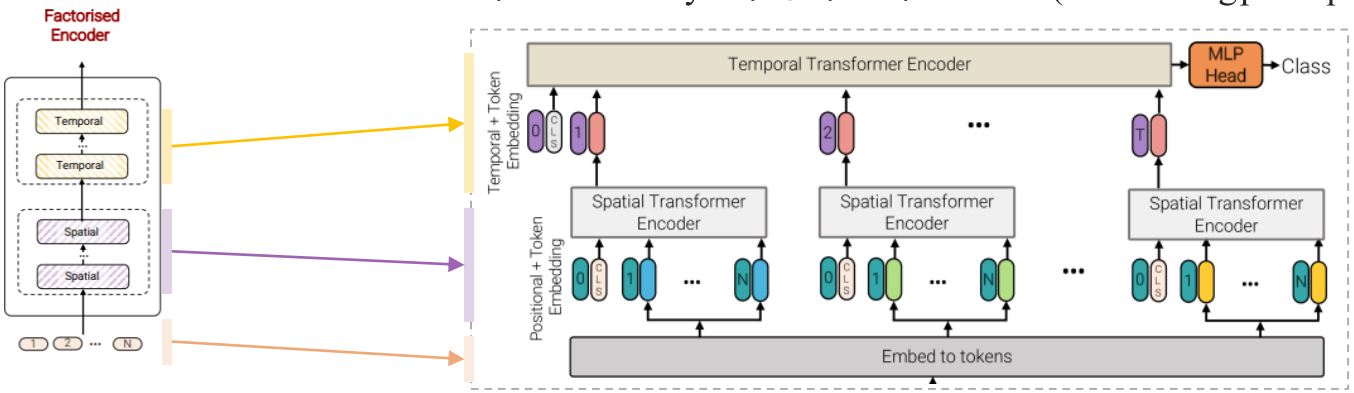
ViViT: A Video Vision Transformer

❖ Transformer Models for Video

- Transformer architecture에서 video의 spatial, temporal dimension을 factorising 하기 위한 변형 제시
- **Model 2) Factorised encoder**



- Spatial encode와 temporal encoder가 각각 존재
- **Spatial encoder**
  - ➢ 동일한 index에서 나온 token 간의 interaction이 이루어지며, CLS token도 함께 encoded됨
- **Temporal encoder**
  - ➢ spatial encoder를 통해 얻은 representation을 concat하여 temporal encoder에 넣음
  - ➢ 다른 temporal indices 간의 interaction이 이루어짐
- Model 1보다 transformer layer가 많지만 적은 FLOPs (fewer floating point operations)이 요구됨

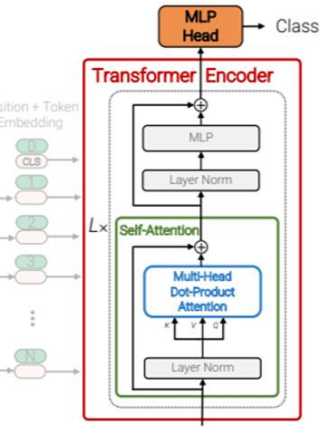


DMQA

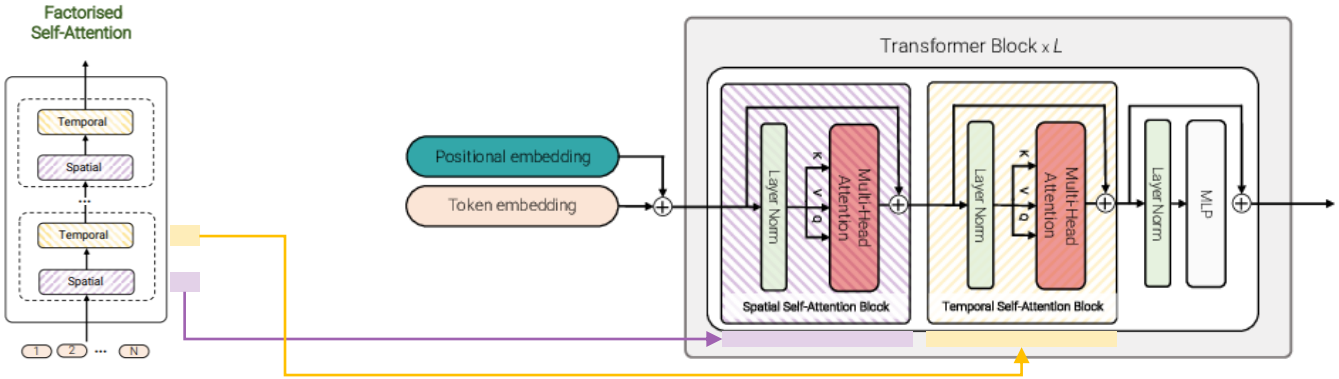
ViViT: A Video Vision Transformer

❖ Transformer Models for Video

- Transformer architecture에서 video의 spatial, temporal dimension을 factorising 하기 위한 변형 제시
- **Model 3) Factorised self-attention**



- 첫 번째 block은 동일한 temporal index에서 추출한 token을 spatially 하게 계산하고,
  두 번째 block은 동일한 spatial index에서 추출한 token을 temporally하게 self-attention
  - Block의 순서에 따른 성능 차이는 없으며 classification token은 사용하지 않음
- Spatial self-attention, temporal self-attention 계산의 효율을 위해 token z를 reshape
  - $reshape\ z \in R^{1\times n_t\cdot n_h\cdot n_w\cdot d} \rightarrow R^{n_t\times n_h\cdot n_w\cdot d}$ (Spatial)
    $R^{n_h\cdot n_w\times n_t\cdot d}$ (Temporal)
- Model 1과 transformer layer 수는 동일하지만, computation complexity는 Model 2와 유사

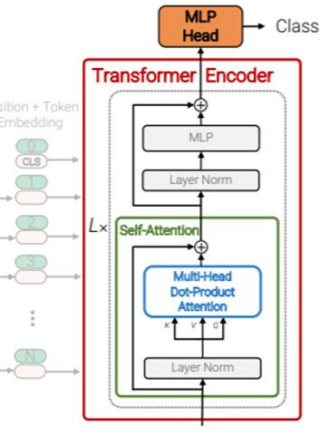


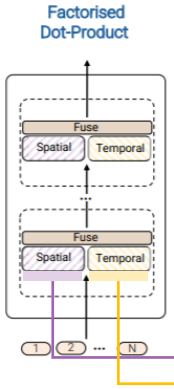
DMQA

ViViT: A Video Vision Transformer

❖ Transformer Models for Video

- Transformer architecture에서 video의 spatial, temporal dimension을 factorising 하기 위한 변형 제시
- **Model 4) Factorised dot-product attention**



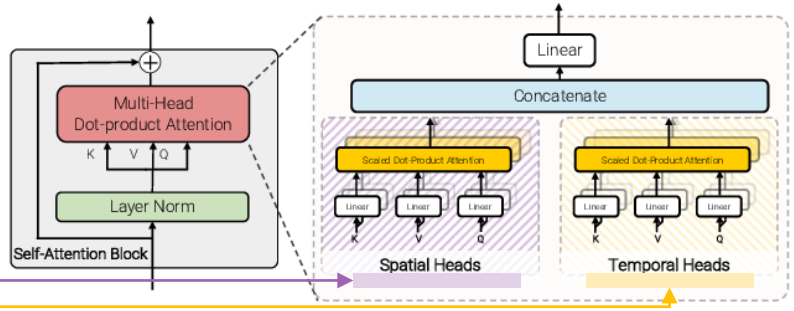


- Multi-head를 spatial head와 temporal head로 구성
  - 각 head 내에서 토큰 별로 attention weights를 계산 $Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V$
  - Attention operation의 핵심 idea

    - 동일한 spatial-temporal index를 가진 tokens에만 집중하도록 key, value를 query에 대해 수정

    $$K_s, V_s \in R^{n_h \cdot n_w \times d} \text{ (Spatial)}, \quad K_t, V_t \in R^{n_t \times d} \text{ (Temporal)}$$

    - query에 대해서만 attention neighborhood를 변경하기 때문에 각 head의 output dimension 동일
    - 각 head의 output은 concatenate
- Computational complexity는 Model 2, Model 3과 유사, parameter 수는 Model 1과 유사



ViViT: A Video Vision Transformer

❖ Initialisation by leveraging pretrained models

- Pretrained image model로부터 large-scale video classification model을 initialize하기 위한 전략 제시
- **Positional embedding**
  - ➢ Video model의 input token은 image model에 비해 n_t배 만큼 더 존재하므로 temporally하게 반복하여 positional embedding 함 (same spatial index have the same embedding)
- **Embedding weights E** (tubelet embedding tokenization method 사용하는 경우)
  - ➢ 2D filter에서 3D conv filter로 initializing하기 위해 temporal dimension에 따라 filter를 replicating 하고 평균을 취함(Inflate) $E = \frac{1}{t}[E_{image}, \ldots, E_{image}, \ldots, E_{image}]$
  - ➢ 추가적으로 zero를 사용하여 initializing하면 ($E = [0, \ldots, E_{image}, \ldots, 0]$) "Uniform frame sampling method"와 동일하게 작동하며, temporal information도 학습 가능
- **Transformer weights for Model 3**
  - ➢ Model 3의 block은 multi-head 부분에서 pretrained ViT와 차이가 있음
  - ➢ Spatial MSA는 pretrained된 module로부터 initialize하고, temporal MSA는 weights를 모두 0으로 함 (temporal MSA 시작 시에 residual connection 역할)

Experiments

❖ Experimental Setup

- Backbone architecture : ViT & BERT
- Tubelet의 height와 width는 같음
- Optimizer는 SGD와 momentum 사용
- Learning Rate: cosine learning rate schedule
- ImageNet-21K 또는 더 큰 JFT dataset을 학습한 ViT image pretrained model 사용

Experiments

❖ Ablation study

- **Input encoding**
  - ➢ Model 1과 ViViT-B (Kinetics 400 dataset)에 Uniform frame sampling과 Tubelet embedding을 적용하여 비교
  - ➢ Central frame을 사용하여 initializing한 tubelet embedding method에서 가장 성능이 좋음

Table 1: Comparison of input encoding methods using ViViT-B and spatio-temporal attention on Kinetics. Further details in text.

	Top-1 accuracy
Uniform frame sampling	78.5
Tubelet embedding	
Random initialisation [22]	73.2
Filter inflation [6]	77.6
Central frame	79.2

Experiments

❖ Ablation study

- **Model variant**
  - ➢ ViViT-B를 사용하여 model architecture에 따른 Top-1 accuracy (Kinetics 400 dataset, Epic Kitchens dataset)비교
  - ➢ Epic Kitchens dataset은 label이 noun과 verb로 나뉘어 있어서 verb에 대한 정확도 측정

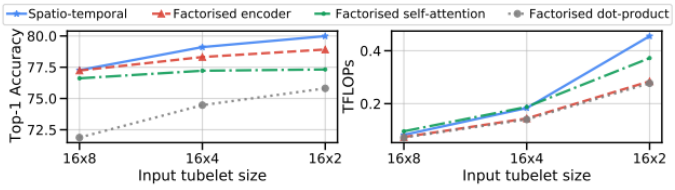
Table 2: Comparison of model architectures using ViViT-B as the backbone, and tubelet size of 16×2. We report Top-1 accuracy on Kinetics 400 (K400) and action accuracy on Epic Kitchens (EK). Runtime is during inference on a TPU-v3.

	K400	EK	FLOPs ($\times 10^9$)	Params ($\times 10^6$)	Runtime (ms)
Model 1: Spatio-temporal	80.0	43.1	455.2	88.9	58.9
Model 2: Fact. encoder	78.8	43.7	284.4	100.7	17.4
Model 3: Fact. self-attention	77.4	39.1	372.3	117.3	31.7
Model 4: Fact. dot product	76.3	39.5	277.1	88.9	22.9
Model 2: Ave. pool baseline	75.8	38.8	283.9	86.7	17.3

Experiments

❖ Ablation study

- **Varying the number of tokens and input frames**
  - ➢ Tubelet 크기가 감소함에 따라 성능이 향상되지만 동시에 계산량도 증가
  - ➢ Spatial resolution 증가에 따라 성능 향상
  - ➢ Token 수의 증가 없이도 긴 비디오에서 좋은 성능을 얻을 수 있음



(a) Accuracy (b) Compute

Figure 8: The effect of varying the number of temporal tokens on (a) accuracy and (b) computation on Kinetics 400, for different variants of our model with a ViViT-B backbone.

Table 5: The effect of spatial resolution on the performance of ViViT-L/16x2 and spatio-temporal attention on Kinetics 400.

Crop size	224	288	320
Accuracy	80.3	80.7	81.0
GFLOPs	1446	2919	3992
Runtime	58.9	147.6	238.8

Varying the number of tokens

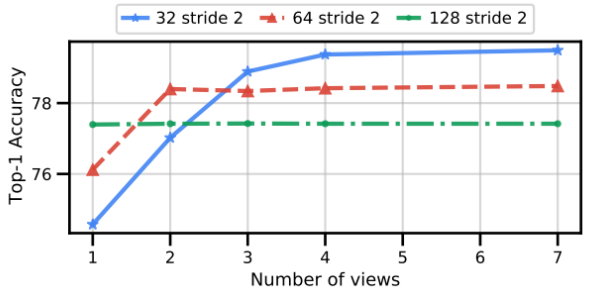


Figure 9: The effect of varying the number of frames input to the network when keeping the number of tokens constant by adjusting the tubelet length t. We use ViViT-B, and spatio-temporal attention on Kinetics 400. A Kinetics video contains 250 frames (10 seconds sampled at 25 fps) and the accuracy for each model saturates once the number of equidistant temporal views is sufficient to "see" the whole video clip.

Varying the number of input frames

Experiments

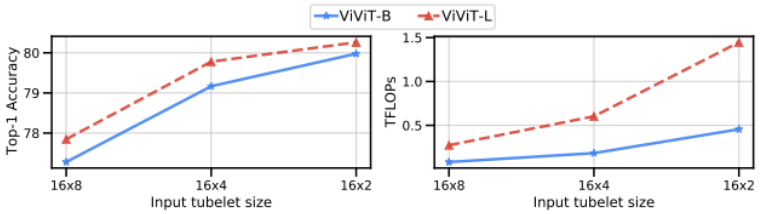
❖ Ablation study

- **Model regularisation and varying the backbone**
  - ➢ Regularisation을 추가함에 따라 성능 향상
  - ➢ Backbone capacity가 증가함에 따라 성능 향상

Table 4: The effect of progressively adding regularisation (each row includes all methods above it) on Top-1 action accuracy on Epic Kitchens. We use a Factorised encoder model with tubelet size 16×2.

	Top-1 accuracy
Random crop, flip, colour jitter	38.4
+ Kinetics 400 initialisation	39.6
+ Stochastic depth [28]	40.2
+ Random augment [10]	41.1
+ Label smoothing [58]	43.1
+ Mixup [79]	43.7

Model regularisation



(a) Accuracy (b) Compute

Figure 7: The effect of the backbone architecture on (a) accuracy and (b) computation on Kinetics 400, for the spatio-temporal attention model (Model 1).

Varying the backbone

Experiments

❖ Comparison to state-of-the-art

- 소규모 dataset (Epic Kitchens, SSv2)에서는 Factorised encoder model (Model 2)
  대규모 dataset (Kinetics)에서는 unfactorised spatio-temporal attention model (Model 1) 사용

Table 6: Comparisons to state-of-the-art across multiple datasets. For "views", $x \times y$ denotes x temporal crops and y spatial crops. "320" denotes models trained and tested with a spatial resolution of 320 instead of 224.

(a) Kinetics 400

Method	Top 1	Top 5	Views
blVNet [16]	73.5	91.2	–
STM [30]	73.7	91.6	–
TEA [39]	76.1	92.5	10 × 3
TSM-ResNeXt-101 [40]	76.3	–	–
I3D NL [72]	77.7	93.3	10 × 3
CorrNet-101 [67]	79.2	–	10 × 3
ip-CSN-152 [63]	79.2	93.8	10 × 3
LGD-3D R101 [48]	79.4	94.4	–
SlowFast R101-NL [18]	79.8	93.9	10 × 3
X3D-XXL [17]	80.4	94.6	10 × 3
TimeSformer-L [2]	80.7	94.7	1 × 3
ViViT-L/16x2	80.6	94.7	4 × 3
ViViT-L/16x2 320	**81.3**	**94.7**	4 × 3
Methods with large-scale pretraining			
ip-CSN-152 [63] (IG [41])	82.5	95.3	10 × 3
ViViT-L/16x2 (JFT)	82.8	95.5	4 × 3
ViViT-L/16x2 320 (JFT)	83.5	95.5	4 × 3
ViViT-H/16x2 (JFT)	**84.8**	**95.8**	4 × 3

(b) Kinetics 600

Method	Top 1	Top 5	Views
AttentionNAS [73]	79.8	94.4	–
LGD-3D R101 [48]	81.5	95.6	–
SlowFast R101-NL [18]	81.8	95.1	10 × 3
X3D-XL [17]	81.9	95.5	10 × 3
TimeSformer-HR [2]	82.4	**96.0**	–
ViViT-L/16x2	82.5	95.6	4 × 3
ViViT-L/16x2 320	**83.0**	95.7	4 × 3
ViViT-L/16x2 (JFT)	84.3	96.2	4 × 3
ViViT-H/16x2 (JFT)	**85.8**	**96.5**	4 × 3

(c) Moments in Time

	Top 1	Top 5
TSN [69]	25.3	50.1
TRN [83]	28.3	53.4
I3D [6]	29.5	56.1
blVNet [16]	31.4	59.3
AssembleNet-101 [51]	34.3	62.7
ViViT-L/16x2	**38.0**	**64.9**

(d) Epic Kitchens 100 Top 1 accuracy

Method	Action	Verb	Noun
TSN [69]	33.2	60.2	46.0
TRN [83]	35.3	65.9	45.4
TBN [33]	36.7	66.0	47.2
TSM [40]	38.3	**67.9**	49.0
SlowFast [18]	38.5	65.6	50.0
ViViT-L/16x2 Fact. encoder	**44.0**	66.4	**56.8**

(e) Something-Something v2

Method	Top 1	Top 5
TRN [83]	48.8	77.6
SlowFast [17, 77]	61.7	–
TimeSformer-HR [2]	62.5	–
TSM [40]	63.4	88.5
STM [30]	64.2	89.8
TEA [39]	65.1	–
blVNet [16]	65.2	**90.3**
ViViT-L/16x2 Fact. encoder	**65.4**	89.8

Conclusion

❖ Conclusion

- This paper presented four pure-transformer models for video classification
- The models achieved state-of-the art results across five popular datasets
- Researchers showed how to effectively regularize such high-capacity models for training on smaller datasets

Reference

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). Vivit: A video vision transformer. arXiv preprint arXiv:2103.15691.


Thank You