Multiview Transformers for Video Recognition

Sarath Kumar Mannam Anshuman Gaharwar Abhishek Maruturi

Akashdeep Chakraborty Parth Parag Kulkarni

Contents

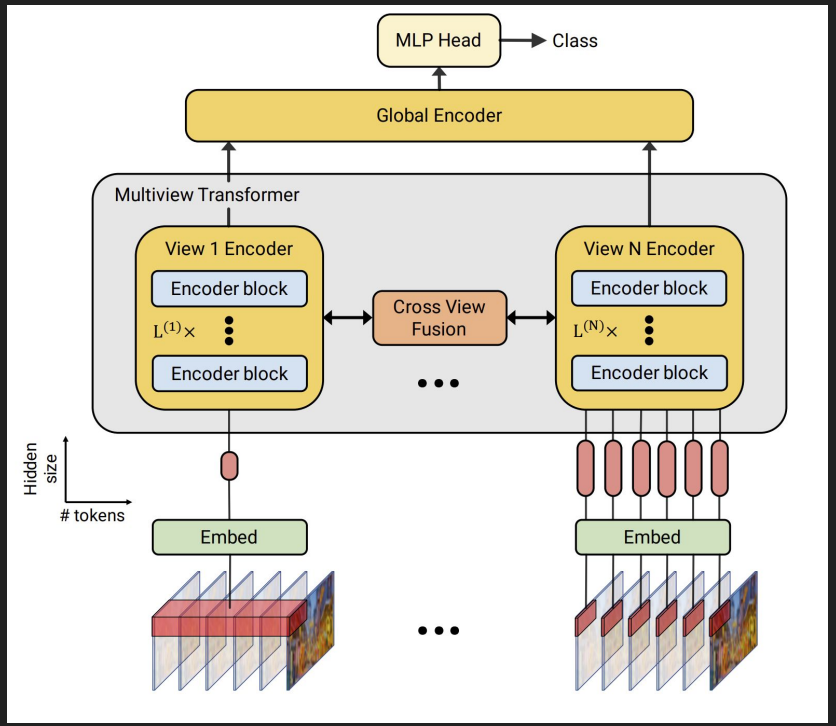
- Introduction
- Related Work
- Model Proposed
- Experiments
- Ablation Study
- Conclusion

Overview

- Multiview Transformers for Video Recognition (MTV)
- separate encoders to represent different views
- lateral connections to fuse information across views
- consistently performs better than single-view counterparts
- achieves state-of-the-art results on five standard datasets
- improve even further with large-scale pretraining

Concept

- No pyramidal structures or subsampling
- Multiple input representations, or “views”
- Tokens over multiple temporal durations



Tokens

long time interval tokens -> capture the gist of the scene

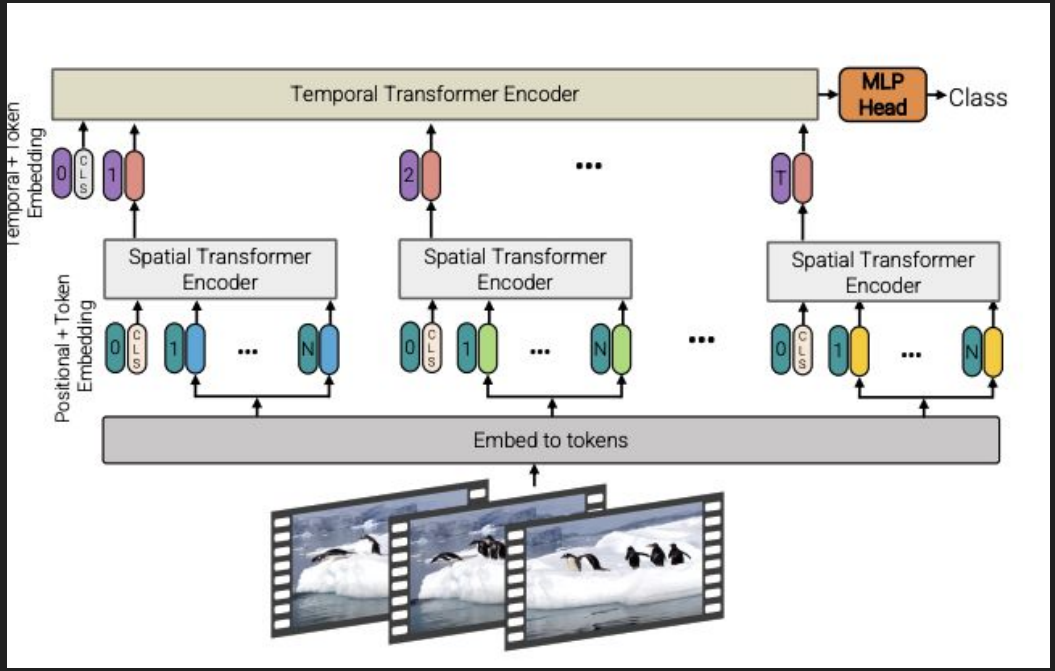
short segment tokens -> capture fine-grained details

Related Work

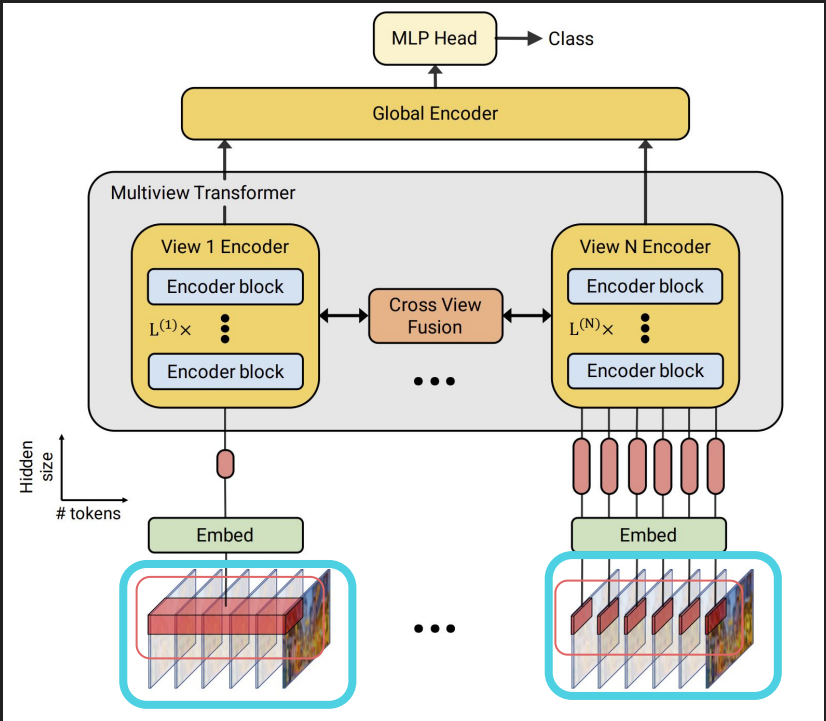
- Evolution of video understanding models
  - 3D CNNs
  - Insertion of Transformer blocks in CNNs
  - ViViT, TimeSformer
- Multiscale processing in computer vision
  - “Pyramid” (hierarchical) structures
  - SlowFast
  - spatiotemporal “tubelets”

ViViT

- $V \in R^{T x H x W x C}$
- N Patches for each Frame
- T Spatial Transformer for N Patches
- Temporal Transformer
- Classification

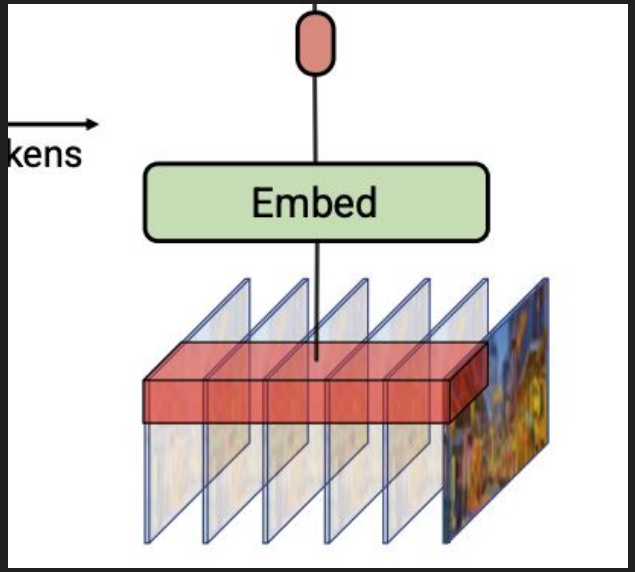


The Idea

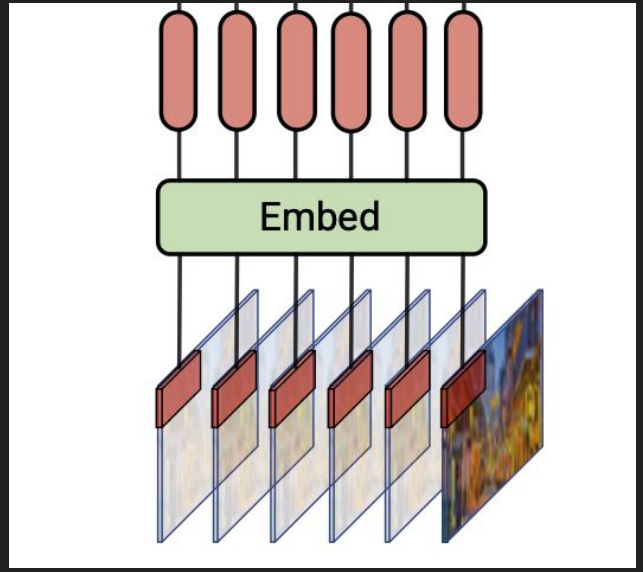


- Input Video
- Multiple Views
- Tubelets for Each View

Multiview Tokenization

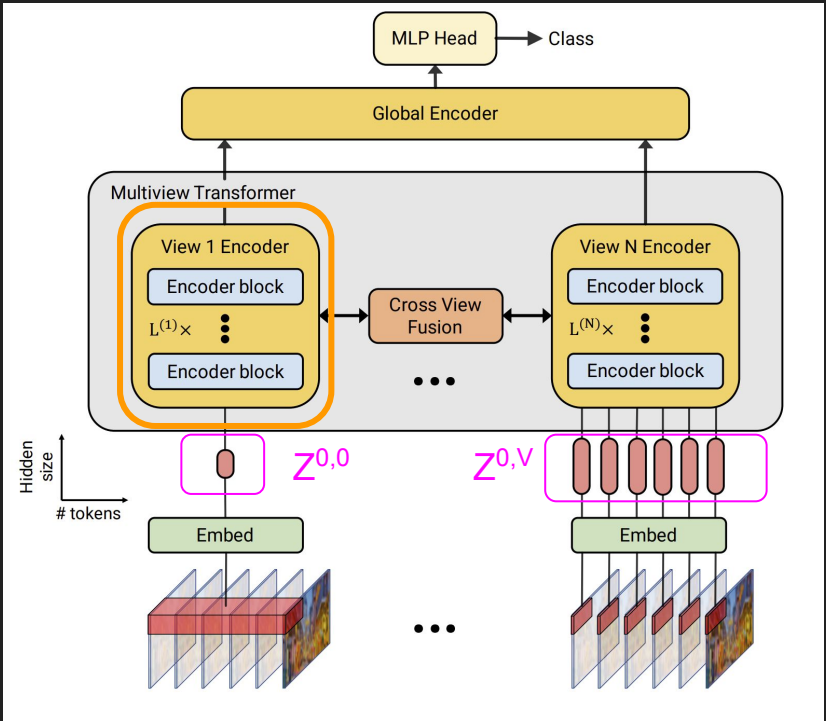


Large Tubelets - Less Tokens
Gist of Scene



Small Tubelets - More Tokens
Fine Grained Motion

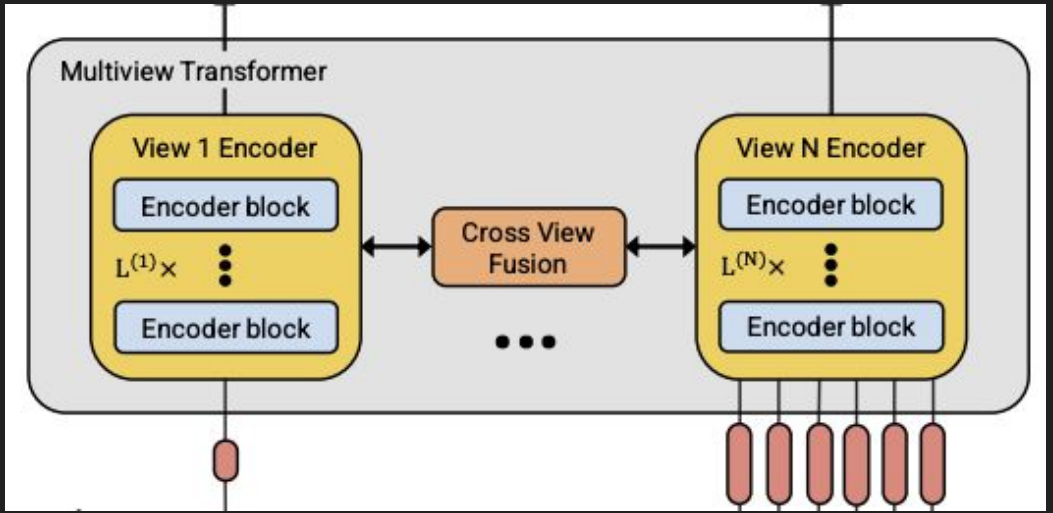
The Idea



- Personalized Encoder per view
  - Sharing Info via Fusion
- Encoder Block
  - Standard Transformer
  - L Blocks
- Varying Hidden Dim across views

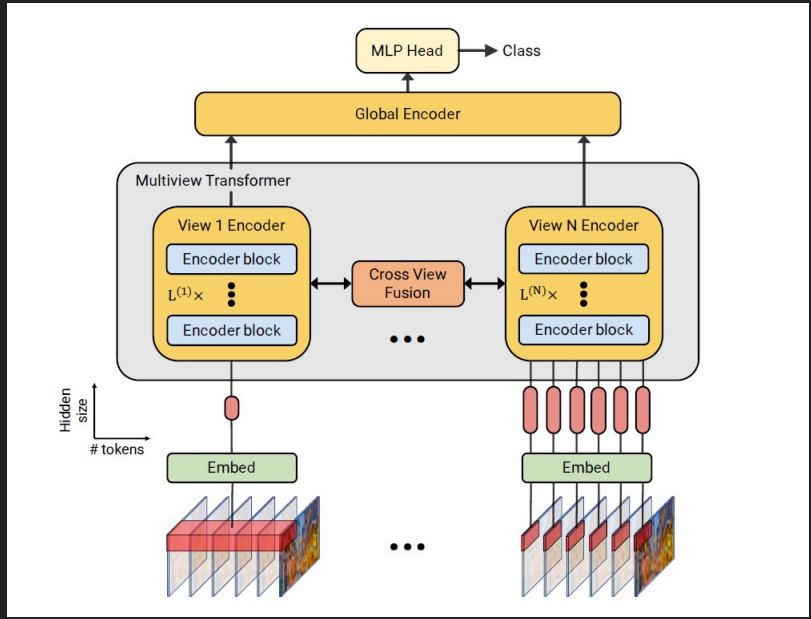
Multi-View Transformer

- Encoder Block
  - Standard Transformer
  - L Blocks
- Varying Hidden Dim across views



Multi-View Transformer

- Input : Multi-View Tokens
- Multi - view Encoder
- Fusing the information between the encoders
- Global Encoder - Information Aggregation
- MLP - Final Classification



Cross-View Fusion

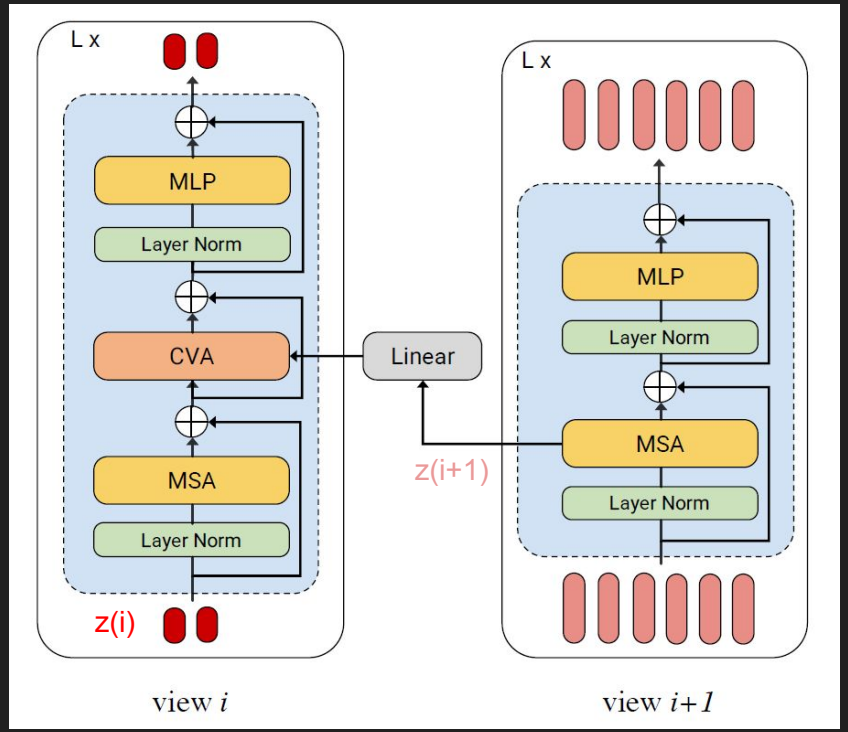
- Cross - View Attention
- Bottleneck Tokens
- MLP fusion

Cross- View Attention

- Fusing information by applying self-attention jointly on all tokens.
- Views are ordered in an ascending manner.
- Queries are z(i), the keys and values are z(i+1).

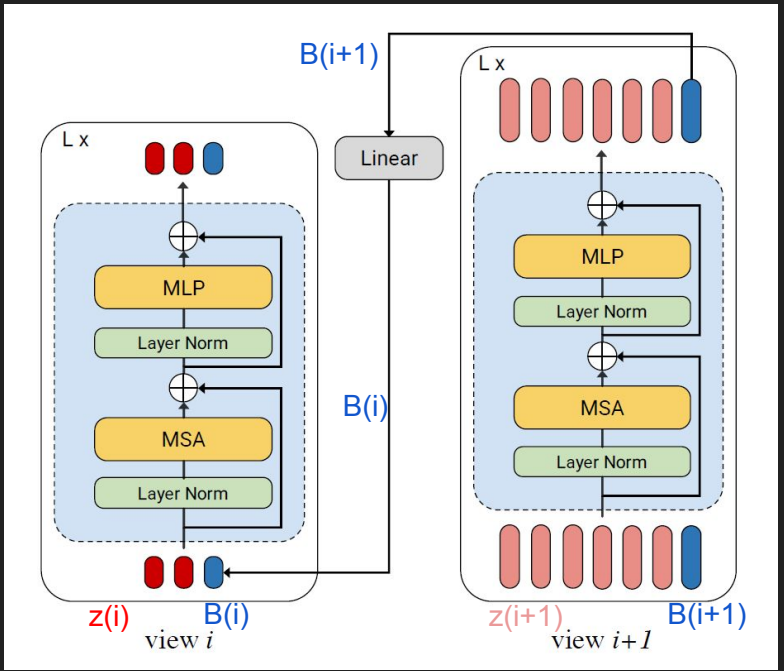
$$\mathbf{z}^{(i)} = \mathrm{CVA}(\mathbf{z}^{(i)}, \mathbf{W}^{\mathrm{proj}}\mathbf{z}^{(i+1)}),$$

$$\mathrm{CVA}(\mathbf{x}, \mathbf{y}) = \mathrm{Softmax}\left(\frac{\mathbf{W}^{Q}\mathbf{x}\mathbf{W}^{K}\mathbf{y}^{\top}}{\sqrt{d_k}}\right)\mathbf{W}^{V}\mathbf{y}.$$



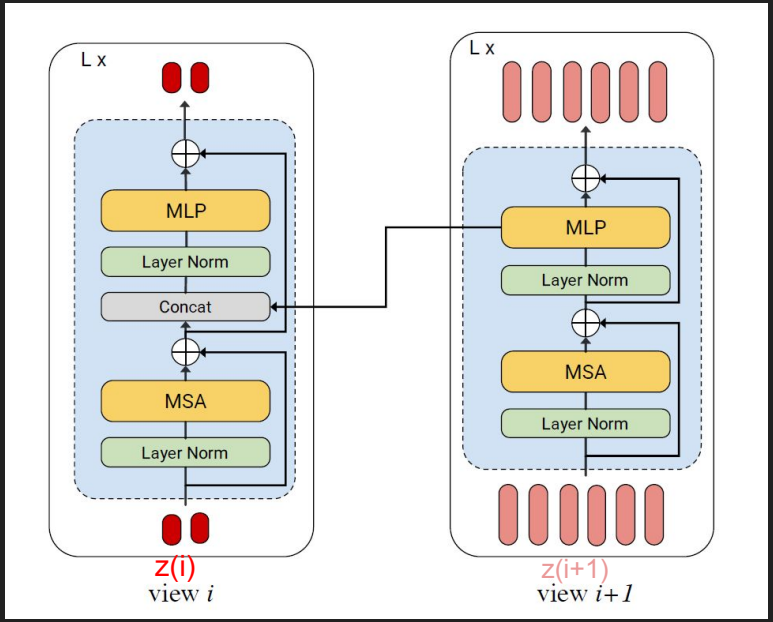
Bottleneck Tokens

- Concatenation(z(i+1),B(i+1))
- Self-Attention, MSA(z(i+1),B(i+1))
- B(i) = Project [z(i+1)] → d(i)
- Concretely, transferring information from one view to other.



MLP Fusion

- Fusing information right before the MLP layer within each encoder.
- Tokens z(i+1) and depth of d(i+1) are fed into encoder for view i.
- Project [z(i+1)] → d(i)
- Concatenation (z(i+1), z(i))

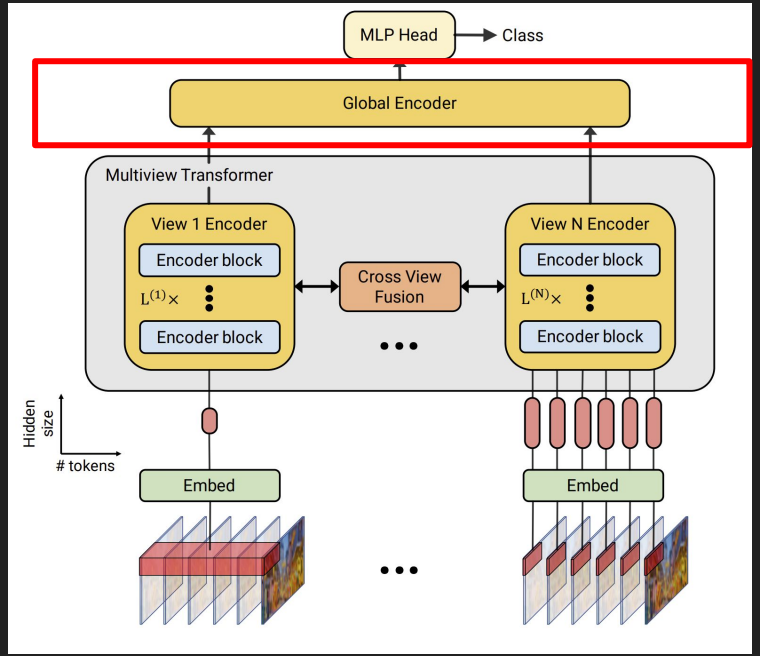


Fusion Locations

- Not necessary to perform cross-view fusion at each layer of cross-view encoder.
- Fusion operation has a “global receptive field”
- Fusion Location is a design choice which is analyzed further in ablation study

Global Encoder

- Aggregate tokens from each view with the final global encoder.
- Extraction of classification token from each view and process with another transformer encoder which aggregates information from all the views.
- The resulting classification token is mapped to one of the C classification outputs. C is the number of classes.



Experimental Setup: Training

- Training settings same as ViViT
- Synchronous SGD
- Momentum = 0.9
- Cosine Learning Rate + Linear Warm up
- Resolution: 224 x 224

Experimental Setup: Other

- Initialization (View Encoders):
  - ViT model pre-trained on large-scale datasets
  - For dimension mismatch: Use CVA's (Cross View Attention) Projection

- Initialization (Global Encoder):
  - Random Initialization

- Datasets:
  - Kinetics-XX (400, 600, 700 classes)
  - Moments in Time
  - Epic-Kitchens-100
  - Something-Something V2

Ablation Study

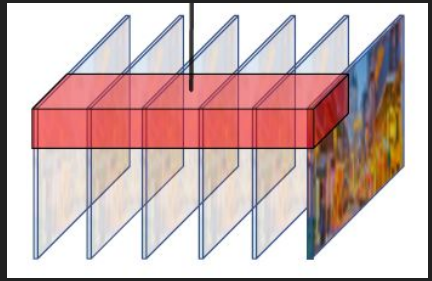
Notations

B/2+S/4+Ti/8

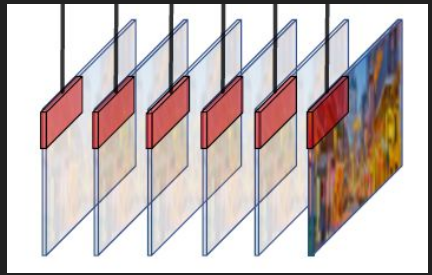
(X/a+X/b+X/c)

- X: Type of ViT encoder
  - (Base [B], Small [S], Tiny[Ti])
- a,b,c: Size of tubelet
- 16 x 16 x Tubelet Size

Tubelet Size: 6



Tubelet Size: 1



Effects of different model-view assignments

Larger model for small tubelet size

Model variants	GFLOPs	MParams	Top-1
B/8+Ti/2	81	161	77.3
B/2+Ti/8	337	221	81.3
B/8+S/4+Ti/2	202	250	78.5
B/2+S/4+Ti/8	384	310	81.8
B/4+S/8+Ti/16	195	314	81.1

Small Tubelet Size: Granular details of the scene

Large Tubelet Size: Gist of the scene

Effects of the same model applied to different views

Using complex encoder for large tubelet size is OverKill

Model variants	GFLOPs	MParams	Top-1
B/4+S/8+Ti/16	195	314	81.1
B/4+B/8+B/16	324	759	81.1
B/2+Ti/8	337	221	81.3
B/2+B/8	448	465	81.5
B/2+S/4+Ti/8	384	310	81.8
B/2+B/4+B/8	637	751	81.7

Comparison of different cross-view fusion methods

Sum the probabilities coming from each view encoder

Concatenate final embeddings from each view encoder

Model variants	Method	GFLOPs	MParams	Top-1
B/4		145	173	78.3
S/8	N/A	20	60	74.1
Ti/16		3	13	67.6
B/4+S/8+Ti/16	Ensemble	168	246	77.7
	Late fusion	187	306	80.6
	MLP	202	323	80.6
	Bottleneck	188	306	81.0
	CVA	195	314	**81.1**

Effects of increasing number of views

Saturation for number of views > 3

Model variants	GFLOPs	Top-1
B/4	145	78.3
B/4+Ti/16	168	80.8 (+2.5)
B/4+S/8+Ti/16	195	81.1 (+2.8)
B/4 (14)	168	78.1 (-0.2)
B/4 (17)	203	78.4 (+0.1)

Effects of applying CVA at different layers

- Stages: Early, mid & Late

Single fusion: Mid > Late > Early

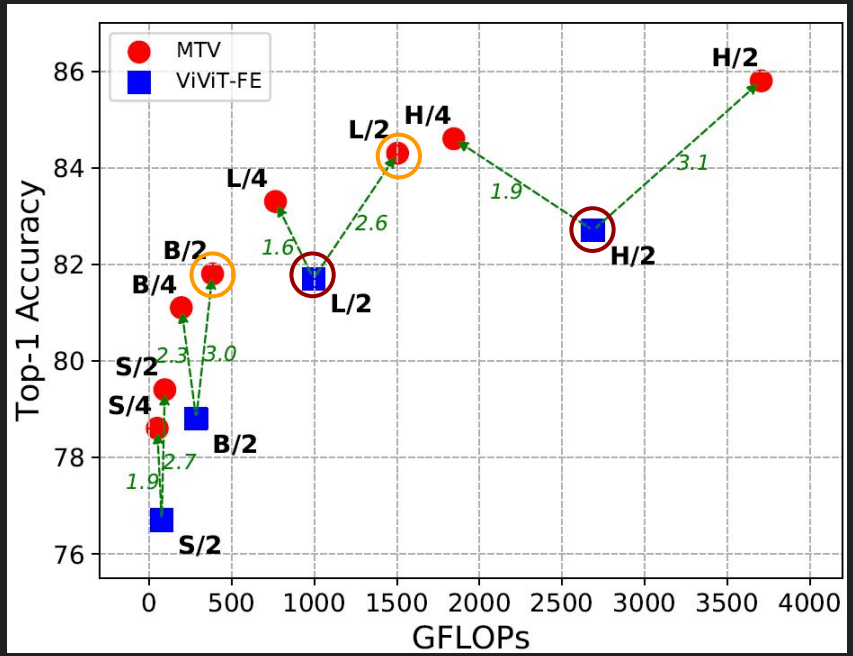
Mid + Late works best

Fusion layers		GFLOPs	MParams	Top-1
0	**Early**			80.96
5	**Mid**	195	314	81.08
11	**Late**			81.00
0, 1				80.91
5, 6		203	323	80.96
10, 11				80.81
5, 11				**81.14**
0, 5, 11		210	331	80.95

Comparison to SlowFast multi-resolution method

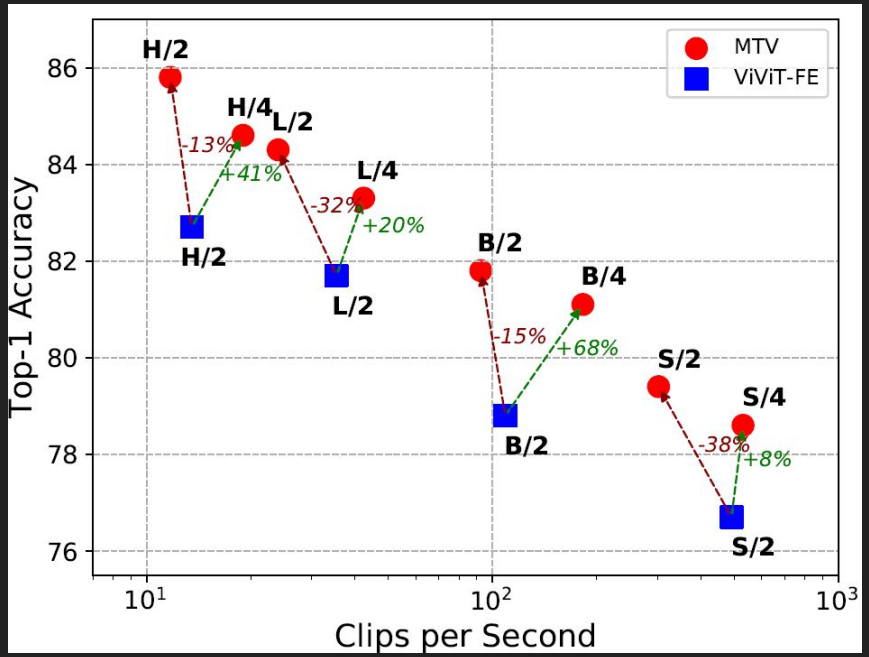
Model variants	GFLOPs	MParams	Top-1
SlowFast (transformer backbone)			
Slow-only (B)	79	87	78.0
Fast-only (Ti)	63	6	74.6
Slowfast (B+Ti)	202	105	79.7
B/4+Ti/16 (ours)	168	224	**80.8**

Accuracy/computation trade-offs



Accuracy[%] - GFLOPs comparison between MTV and ViViT-FE

Accuracy/computation trade-offs



Accuracy[%] - Throughput comparison between MTV and ViViT-FE

Comparison to SOTA - kinetics-400

Method	Top 1	Top 5	Views	TFLOPs
TEA [40]	76.1	92.5	10 × 3	2.10
TSM-ResNeXt-101 [41]	76.3	–	–	–
I3D NL [75]	77.7	93.3	10 × 3	10.77
VidTR-L [84]	79.1	93.9	10 × 3	10.53
LGD-3D R101 [52]	79.4	94.4	–	–
SlowFast R101-NL [23]	79.8	93.9	10 × 3	7.02
X3D-XXL [22]	80.4	94.6	10 × 3	5.82
OmniSource [20]	80.5	94.4	–	–
TimeSformer-L [6]	80.7	94.7	1 × 3	7.14
MFormer-HR [51]	81.1	95.2	10 × 3	28.76
MViT-B [21]	81.2	95.1	3 × 3	4.10
MoViNet-A6 [35]	81.5	95.3	1 × 1	0.39
ViViT-L FE [3]	81.7	93.8	1 × 3	11.94
MTV-B	81.8	95.0	4 × 3	4.79
MTV-B (320p)	82.4	95.2	4 × 3	11.16

Methods with web-scale pretraining				
VATT-L [2] (HowTo100M)	82.1	95.5	4 × 3	29.80
ip-CSN-152 [70] (IG)	82.5	95.3	10 × 3	3.27
R3D-RS (WTS) [19]	83.5	–	10 × 3	9.21
OmniSource [20] (IG)	83.6	96.0	–	–
ViViT-H [3] (JFT)	84.9	95.8	4 × 3	47.77
TokenLearner-L/10 [55] (JFT)	85.4	96.3	4 × 3	48.91
Florence [80] (FLD-900M)	86.5	97.3	4 × 3	–
CoVeR (JFT-3B) [82]	87.2	–	1 × 3	–
MTV-L (JFT)	84.3	96.3	4 × 3	18.05
MTV-H (JFT)	85.8	96.6	4 × 3	44.47
MTV-H (WTS)	**89.1**	**98.2**	4 × 3	44.47

Comparison to SOTA - Kinetics 600

(b) Kinetics 600

Method	Top 1	Top 5
SlowFast R101-NL [23]	81.8	95.1
X3D-XL [22]	81.9	95.5
TimeSformer-L [6]	82.2	95.6
MFormer-HR [51]	82.7	96.1
ViViT-L FE [3]	82.9	94.6
MViT-B [21]	83.8	96.3
MoViNet-A6 [35]	**84.8**	**96.5**
MTV-B	83.6	96.1
MTV-B (320p)	84.0	96.2
R3D-RS (WTS) [19]	84.3	–
ViViT-H [3] (JFT)	85.8	96.5
TokenLearner-L/10 [55] (JFT)	86.3	97.0
Florence [80] (FLD-900M)	87.8	97.8
CoVeR (JFT-3B) [82]	87.9	–
MTV-L (JFT)	85.4	96.7
MTV-H (JFT)	**86.5**	**97.3**
MTV-H (WTS)	**89.6**	**98.3**

Comparison to SOTA - Kinetics 700

(d) Kinetics 700

	Top 1	Top 5
VidTR-L [84]	70.2	–
SlowFast R101 [23]	71.0	89.6
MoViNet-A6 [35]	72.3	–
MTV-L	**74.0**	**91.3**
CoVeR (JFT-3B) [82]	79.8	–
MTV-H (JFT)	**78.0**	**93.3**
MTV-H (WTS)	**82.2**	**95.7**

Comparison to SOTA - SSv2

(c) Something-Something v2

Method	Top 1	Top 5
SlowFast R50 [23, 78]	61.7	–
TimeSformer-HR [6]	62.5	–
VidTR [84]	63.0	–
ViViT-L FE [3]	65.9	89.9
MViT [21]	67.7	90.9
MFormer-L [51]	68.1	91.2
MTV-B	67.6	90.1
MTV-B (320p)	68.5	90.4

Comparison to SOTA - Epic Kitchen 100

(e) Epic Kitchens 100 Top 1 accuracy

Method	Action	Verb	Noun
SlowFast [23]	38.5	65.6	50.0
ViViT-L FE [3]	44.0	66.4	56.8
MFormer-HR [51]	44.5	67.0	58.5
MoViNet-A6 [35]	47.7	**72.2**	57.3
MTV-B	46.7	67.8	**60.5**
MTV-B (320p)	**48.6**	68.0	**63.1**

Comparison to SOTA

(f) Moments in Time

	Top 1	Top 5
AssembleNet-101 [56]	34.3	62.7
ViViT-L FE [3]	38.5	64.1
MoViNet-A6 [35]	40.2	–
MTV-L	**41.7**	**69.7**
VATT-L (HT100M) [2]	41.1	67.7
MTV-H (JFT)	**44.0**	**70.2**
MTV-H (WTS)	**45.6**	**74.7**

Conclusion

- Processing multiple “views” of the input video in parallel.
- The model performs better than increasing the depth of single-view architecture.
- They have achieved state-of-the-art results on five popular video classification datasets.

Limitations

- Room for improvement in other datasets except Kinetics

- Models pre trained on large image- or video datasets for initialization.

Future Work

- Reducing the dependency on large datasets.

- Looking into recent transformer based encoders such as MViT and Swin

Innovations

- Introduction of the concept of Multi-view input representation stating the reason that multiple spatio-temporal resolutions provide better reasoning.
- Granular and large scale understanding through the temporal patches only but of different sizes

Criticism

- It is computationally heavy owing to the usage of different encoders for different views.

References

Yan, S., Xiong, X., Arnab, A., Lu, Z., Zhang, M., Sun, C., & Schmid, C. (2022). Multiview Transformers for Video Recognition. arXiv preprint arXiv:2201.04288.

Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 6202-6211).

Kondratyuk, D., Yuan, L., Li, Y., Zhang, L., Tan, M., Brown, M., & Gong, B. (2021). Movinets: Mobile video networks for efficient video recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 16020-16030).

Akbari, H., Yuan, L., Qian, R., Chuang, W. H., Chang, S. F., Cui, Y., & Gong, B. (2021). Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. Advances in Neural Information Processing Systems, 34.