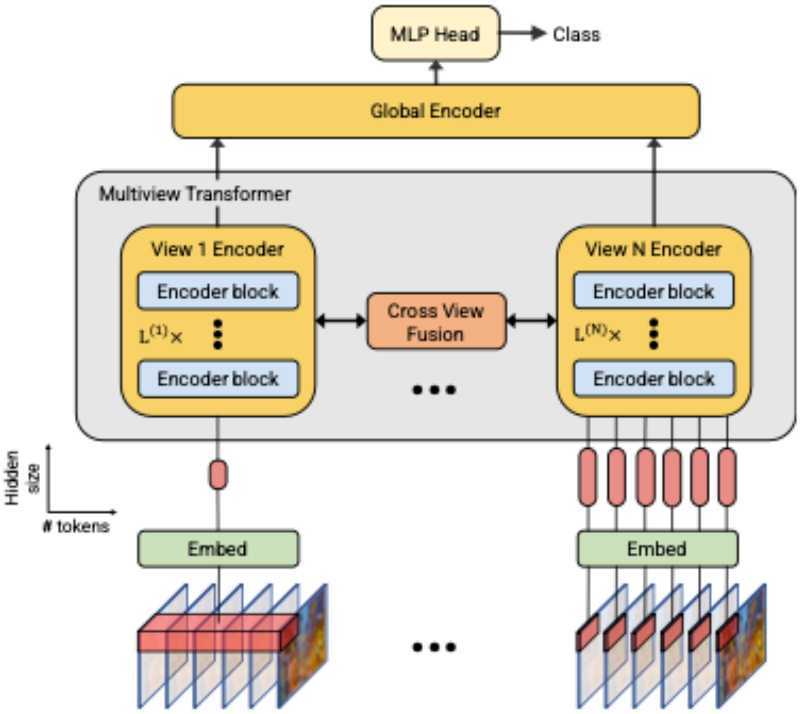
Multiview Transformers for Video Recognition

1、Motivation

在过往的视频理解方法中，时空信息常常因为pooling或下采样的操作而丢失。

作者提出了一种基于transformer的架构用于捕获多multi-resolution时间上下文的信息。从long segments中提取到的tokens中包含场景的主旨（如事件发生的背景），从short segments中提取到的tokens中包含了细粒度动作信息（如姿态信息。）



2、Apporach

2.1 multi-view tokenization

传统的transformer只提取一组token $z^0=[z_{cls},Ex_1,Ex_2,...,Ex_N]+p$，而在这篇文章中，作者提取了多组tokens：$z^{0,(1)},z^{0,(2)},...,z^{0,(V)}$，作为不同的views，其中V为views的数量，而$z^{l,(i)}$表示第i组tokens通过了l层transformer以后得到的结果。

作者使用了不同的3D卷积核和不同层数的网络来提取tokens。越小的卷积核将得到越多tokens的view

2.2 multi-view transformer

首先将不同组的tokens分别通过属于自己的一个encoder，每个encoder中间设置了一个cross view fusion模块。完成encode以后将得到的信息再通过global encoder实现特征融合

①multiview encoder

对每个view有不同的encoder，每个encoder block就是一个基本的transformer模块，不同的是其中加入了一个cross view fusion模块。

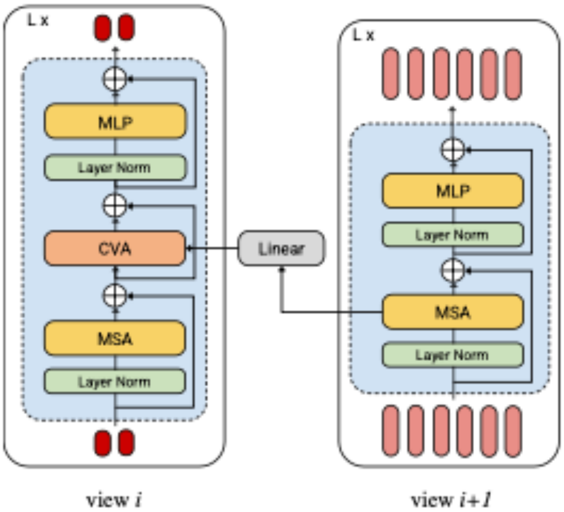
②cross-view fusion

作者提出了三种不同的fusion策略：

1. cross-view attention(CVA)：将不同tokens的views按照token数量从小到大排序。cross-view fusion将在每对i和i+1个view上进行。更新策略如下：

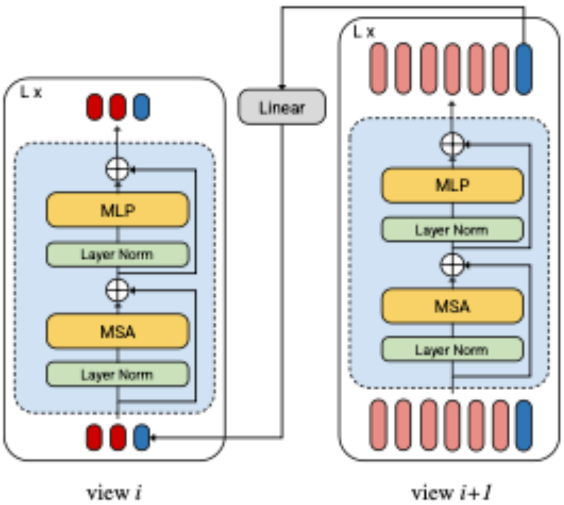
$$z^{(i)} = CVA(z^{(i)}, W^{proj} z^{(i+1)})$$

$$CAV(x, y) = Softmax(\frac{W^Q x W^K y^T}{\sqrt{d_k}}) W^V y$$



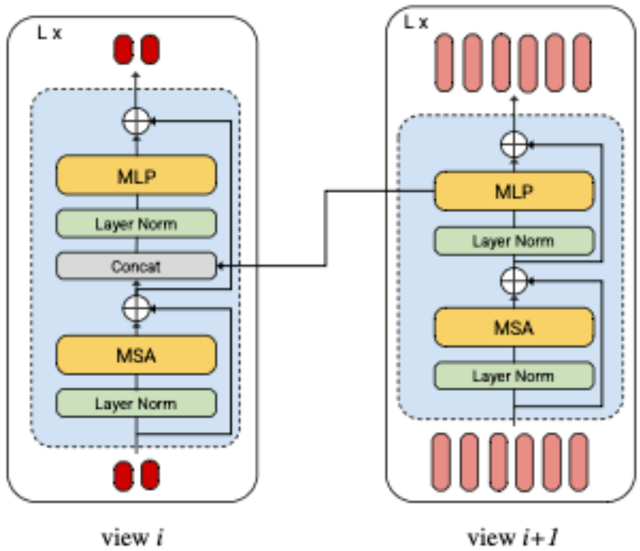
(a) An example of CVA for fusion.

2. Bottleneck tokens：这种方法需要在第i个view中引入B^i个bottleneck token，这里的B^i远小于对应view中tokens的数量。完成i+1个view的encode以后，将bottleneck token做projection以后concat到第i个view输入token的最后。这个方法要注意的是需要从包含最多tokens的view开始向前进行。



(b) An example of bottleneck tokens for fusion.

3. MLP fusion：这种比较简单，直接看图就可以理解了。



(c) An example of MLP fusion.

3、Experiment

本文的实验setting跟vivit基本一致。所以需要去读那篇文章。

模型命名规则：

For example, B/2+S/4+Ti/8 denotes a three-view model, where a “Base”, “Small”, and “Tiny” encoders are used to processes tokens from the views with tubelets of sizes 16×16×2,16×16×4,and16×16×8,respectively

3.1 Ablation study

(a) Effects of different model-view assignments.

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

(b) Effects of the same model applied to different views.

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

(c) Comparison of different cross-view fusion methods.

Model variants	Method	GFLOPs	MParams	Top-1
B/4	N/A	145	173	78.3
S/8		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**

(d) 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**

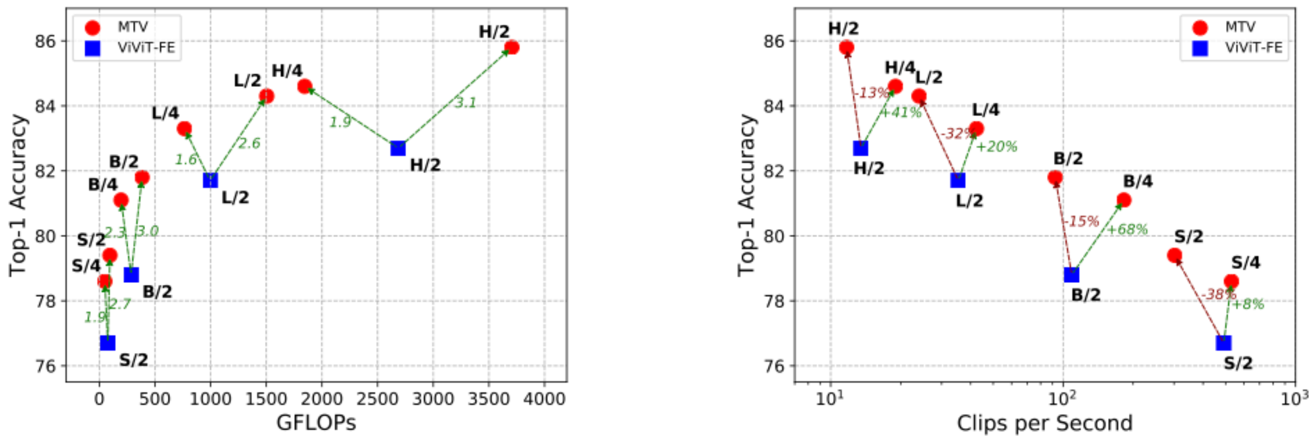
(e) Effects of increasing number of views.

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)

(f) Effects of applying CVA at different layers.

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

3.2 Comparison to the SOTA



(a) Accuracy[%] - GFLOPs comparison between MTV and ViViT-FE.

(b) Accuracy[%] - Throughput comparison between MTV and ViViT-FE.

Figure 3. Accuracy/computation trade-off between ViViT-FE [3] (blue) and our MTV (red). Figure 3a shows that MTV is consistently better and requires less FLOPs than ViViT-FE to achieve higher accuracy across different model scales (shown by the dotted green arrows pointing upper-left). With additional FLOPs, MTV shows larger accuracy gains (shown by the dotted green arrows pointing upper-right). Similarly, Fig. 3b shows that MTV can have higher throughput than ViVIT-FE, whilst still improving its accuracy, across all model scales. All speed comparisons are measured with the same hardware (Cloud TPU-v4), whilst the accuracy is computed from 4×3 view testing.

(a) 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

(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**

(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

(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**

(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**

(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)	**45.4**	**70.7**