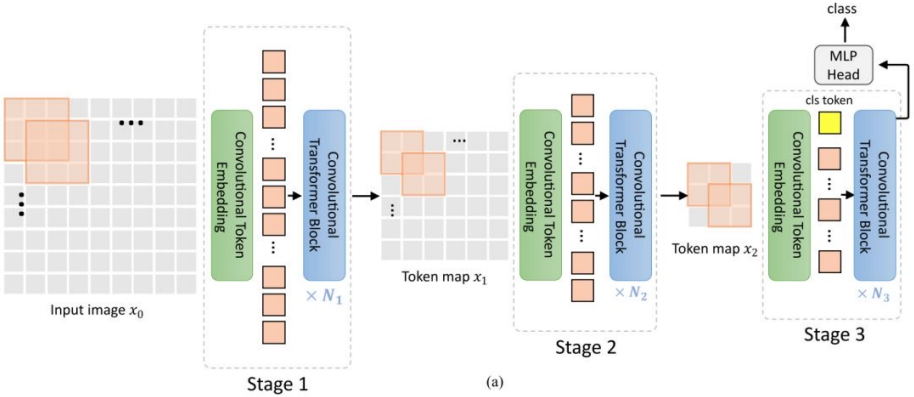
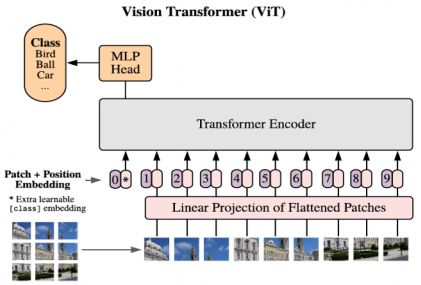
TokenLearner: What Can 8 Learned Tokens Do for Images and Videos?

Michael S. Ryoo[1,2], **AJ Piergiovanni**[1], **Anurag Arnab**[1], **Mostafa Dehghani**[1], **Anelia Angelova**[1]
[1]Google Research
[2]Stony Brook University
{mryoo,ajpiergi,aarnab,dehghani,anelia}@google.com

ViT



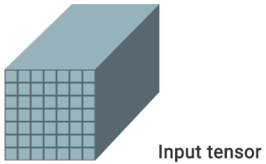


Method	GFLOPS	Accuracy
ViT S/32	3.4	77.87
ViT B/32	19.8	80.69
ViT B/16	55.6	84.73
TokenLearner S/32	1.9	76.13
TokenLearner B/16	28.7	83.65
TokenLearner S/32 (22)	3.3	79.42
TokenLearner B/32 (20)	11.5	82.74
TokenLearner B/16 (21)	47.1	85.21
16-TokenLearner B/16 (21)	47.7	85.45

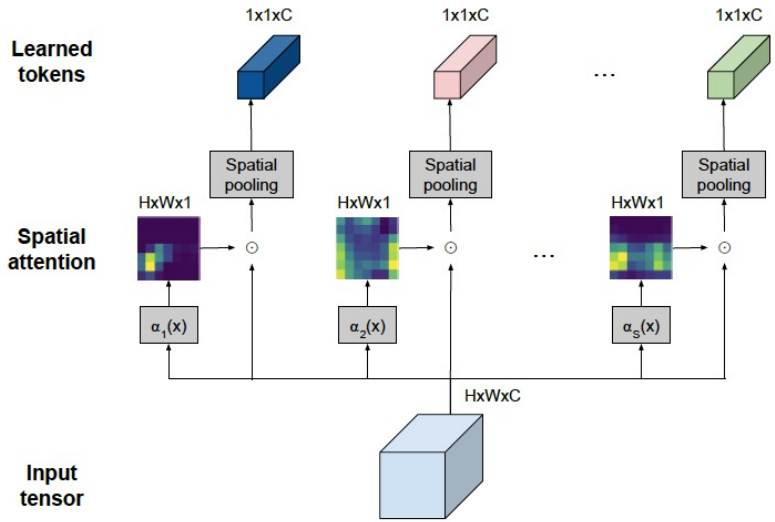
TokenLearner

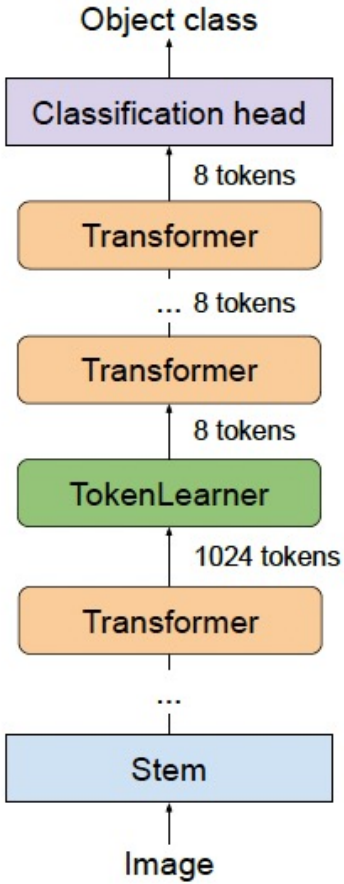
Input tensor: $X \in \mathbb{R}^{T \times H \times W \times C}$

output tensor: $Z_t = [z_i]_{i=1}^{S} \in \mathbb{R}^{S \times C}$



$$z_i = A_i(X_t) = \rho(X_t \odot A_{iw}) = \rho(X_t \odot \gamma(\alpha_i(X_t)))$$



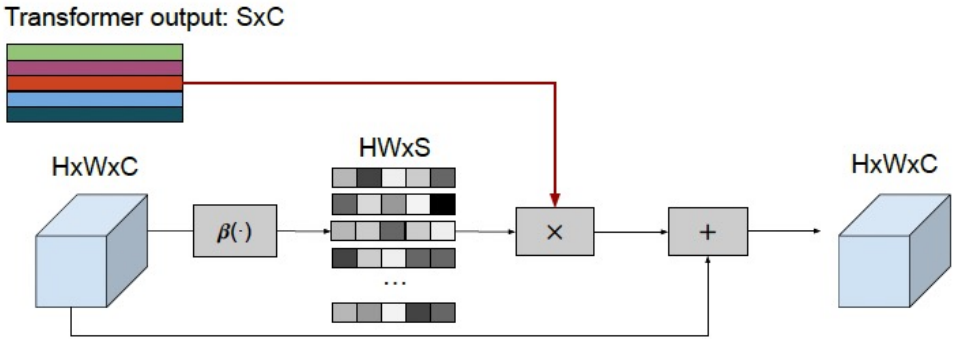


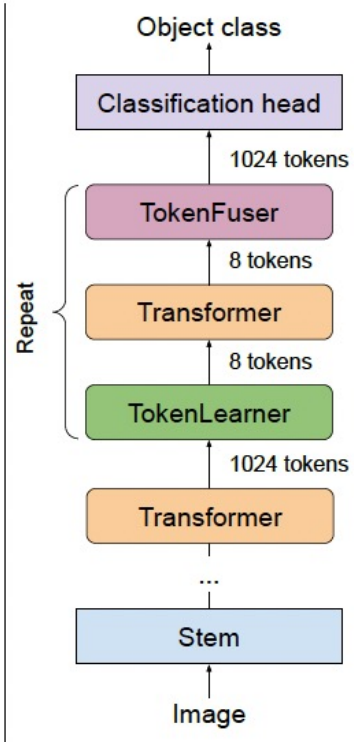
TokenFuser

Y: $\mathbb{R}^{ST\times C} \rightarrow \mathbb{R}^{ST\times C}$

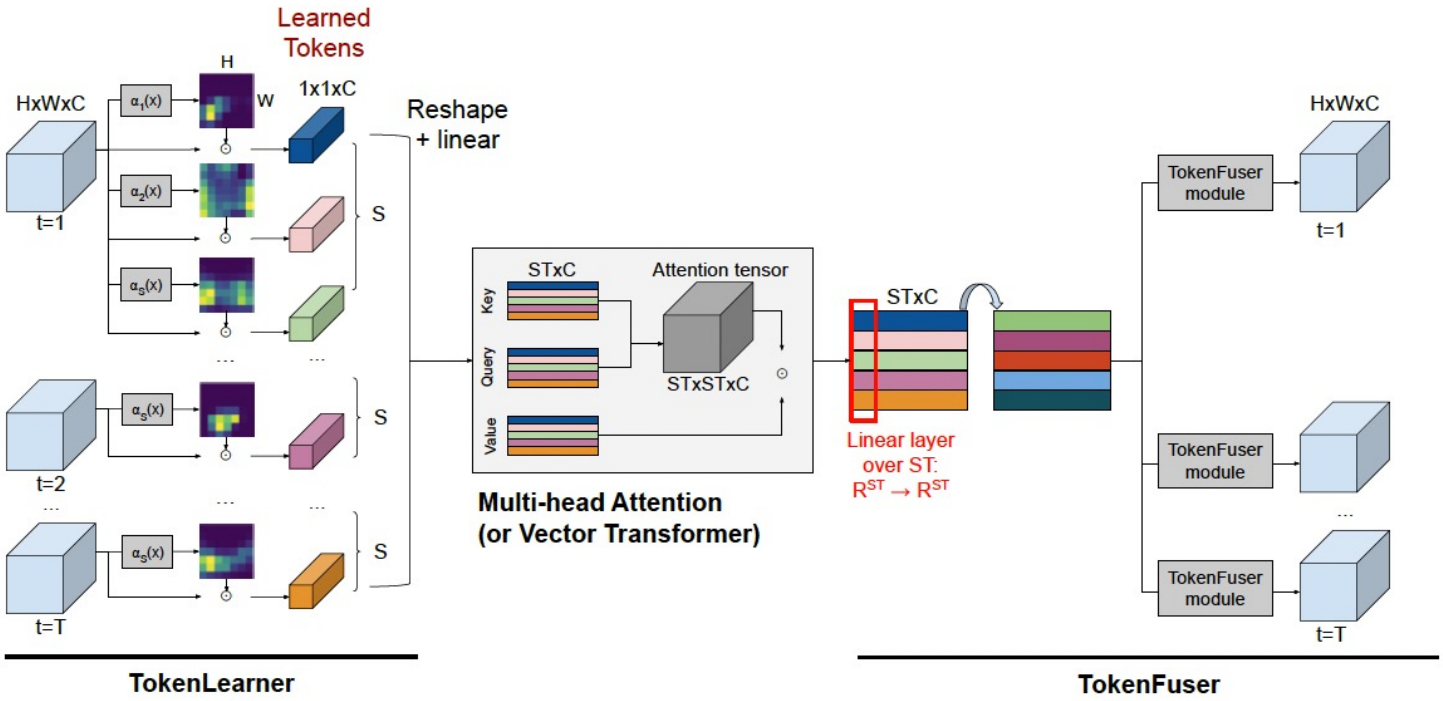
$$Y = (Y^T M)^T$$

$$X_t^{j+1} = B(Y_t, X_t^j) = B_w Y_t + X_t^j = \beta_i(X_t^j) Y_t + X_t^j$$

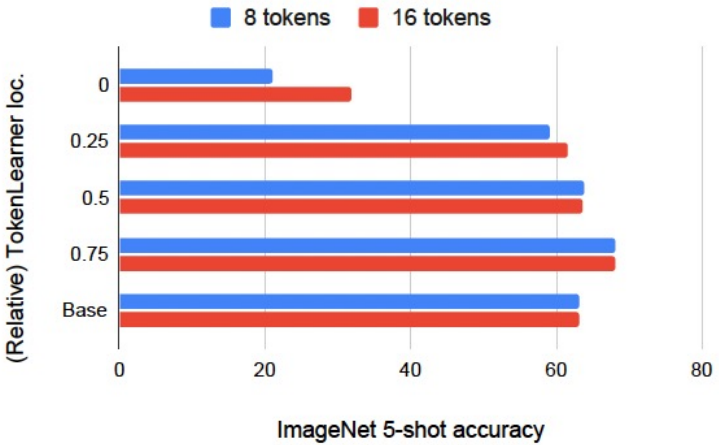


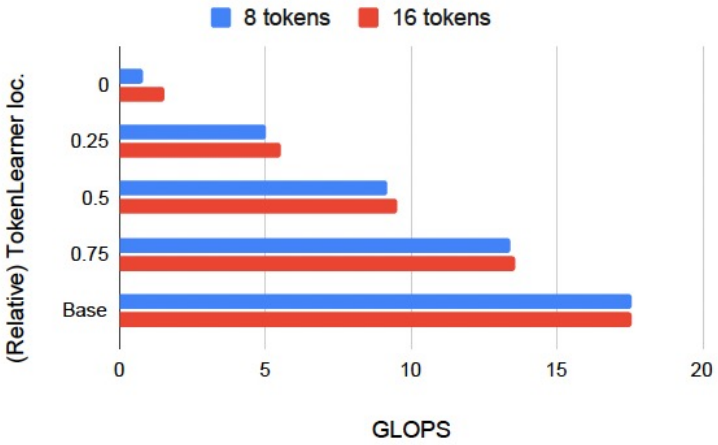


video representation learning



Where should we have TokenLearner





TokenLearner on larger models

Table 2: TokenLearner with ViT L/16 and L/14. 512x512 input images used.

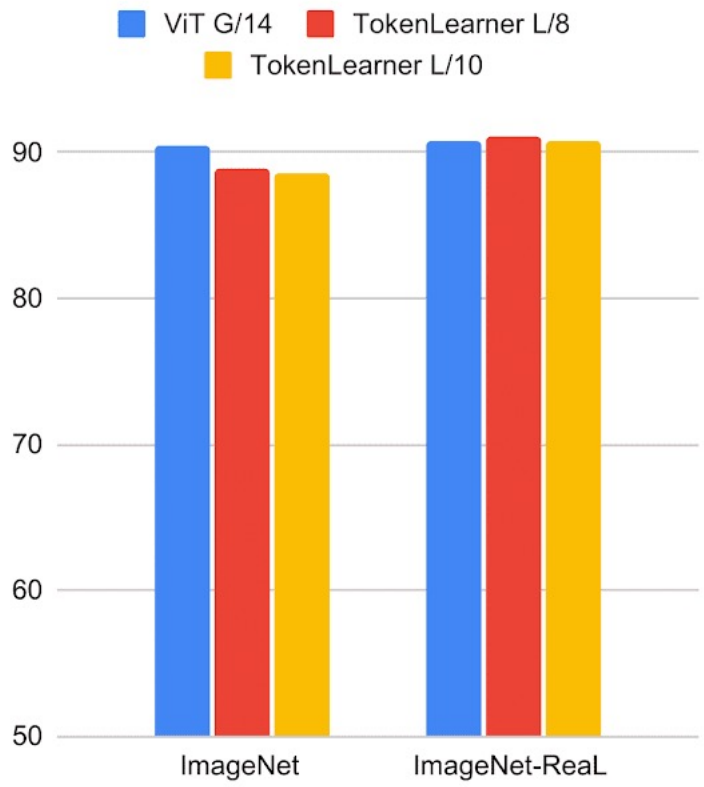
Base	# layers	TokenLearner	GFLOPS	ImageNet Top1
ViT L/16	24	-	363.1	87.35
ViT L/16	24	16-TL at 12	178.1	87.68
ViT L/16	24+11	16-TL at 12	186.8	87.47
ViT L/16	24+6	8-TL at 18	274.2	88.11
ViT L/14	24+11	16-TL at 18	361.6	88.37

Table 3: Comparison to state-of-the-art ViT models.

Method	# params.	ImageNet	ImageNet ReaL
BiT-L	928M	87.54	90.54
ViT-H/14	654M	88.55	90.72
ViT-G/14	1843M	**90.45**	90.81
TokenLearner L/10 (24+11)	**460M**	88.5	90.75
TokenLearner L/8 (24+11)	**460M**	88.87	**91.05**

Table 4: TokenLearner inserted earlier within ViT L/16. 384x384 input images used.

Base	# layers	TokenLearner	GFLOPS	ImageNet Top1
ViT B/16	12	-	55.63	84.73
ViT L/16	24	16-TL at 2	20.91	83.89
ViT L/16	24	16-TL at 3	28.66	85.40
ViT L/16	24	16-TL at 6	51.92	86.44


ViT G/14
TokenLearner L/8
TokenLearner L/10
90
80
70
60
50
ImageNet
ImageNet-ReaL

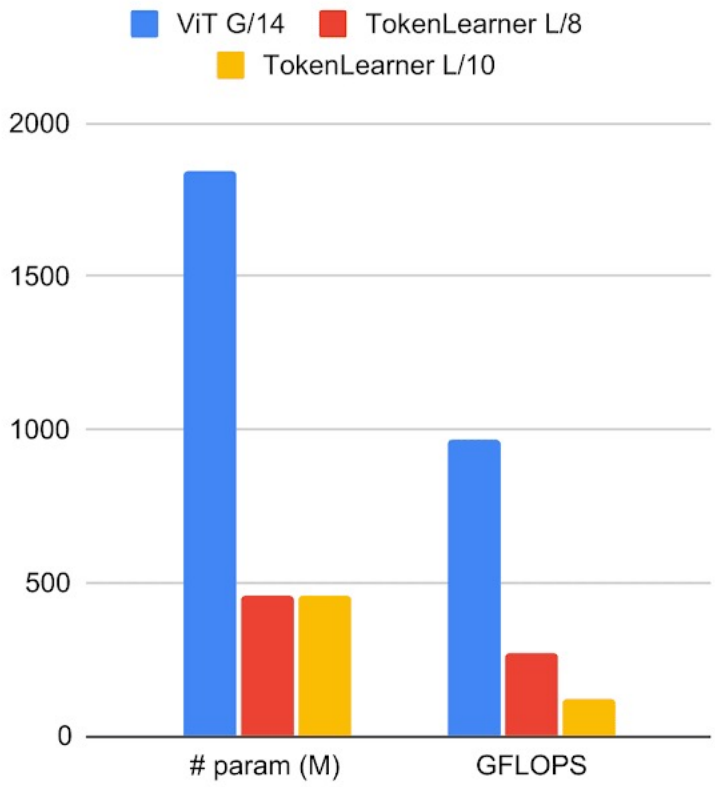

ViT G/14
TokenLearner L/8
TokenLearner L/10
2000
1500
1000
500
0
param (M)
GFLOPS

Table 6: TokenLearner compared against pooling-based token reduction.

Details	ImageNet	GFLOPS
Base ViT L/16	87.35	363.1
2x2 pool at 9 and 18	85.63	144.3
2x2 pool at 12 and 18	86.41	187.2
4x4 pool at 12	83.93	184.4
16-TL at 12	87.68	184.6

$$\mathbf{X}_{out} = \mathbf{X}_{in} + \text{SOFTMAX}_L\left((\mathbf{X}_{in}\mathbf{W}_Q)(\mathbf{T}\mathbf{W}_K)^T\right)\mathbf{T}$$

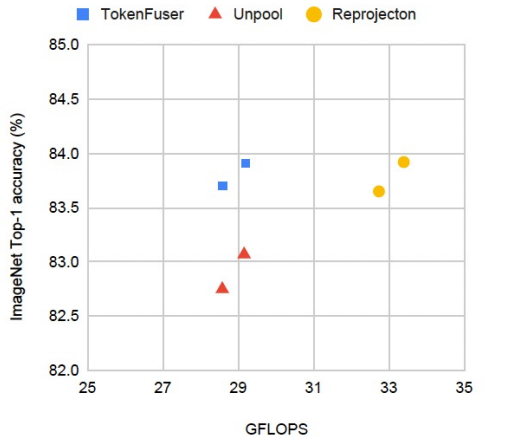


Figure 7: Ablations with TokenFuser alternatives.

Table 5: Models with TokenLearner, with and without TokenFuser. The model without TokenFuser is described in Figure 4 (a). The model with TokenFuser uses the architecture of Figure 4 (b).

Base	# layers	TokenLearner	TokenFuser	ImageNet Top1	ImageNet ReaL	GFLOPS
B/16	12	8-TL at 6	N	83.2	88.1	28.3
B/16	12	8-TL at 6	Y	83.7	88.4	28.5
B/16	12	16-TL at 6	N	83.2	88.0	28.7
B/16	12	16-TL at 6	Y	83.9	88.7	29.1
L/16	24	16-TL at 12	N	87.6	90.4	184.6
L/16	24	16-TL at 12	Y	87.6	90.5	187.1
L/16	24	8-TL at 18	N	87.9	90.8	273.2
L/16	24	8-TL at 18	Y	88.2	90.9	273.8
L/10	24+11	16-TL at 18	N	88.5	90.7	849.0
L/10	24+11	16-TL at 18	Y	88.5	90.9	856.9

TokenLearner for Video

Table 7: Comparison of ViViT models with and without TokenLearner on Kinetics-400. GLOPS are per view. The difference in the number of parameters between the TokenLearner models (which are from Tables 2 and 5) comes from the different number of layers used after the TokenLearner module.

Method	Top-1 accuracy	Top-5 accuracy	# params.	GFLOPS
ViViT-L/16 [2]	82.8	95.5	308M	1446
ViViT-L/16 320 [2]	83.5	95.5	308M	3992
ViViT-H/14 [2]	84.8	95.8	654M	3981
ViViT-L/16 (our run)	83.4	95.6	308M	1446
TokenLearner 16at12 + L/16	83.5	95.6	308M	766
TokenLearner 8at18 + L/16	84.5	96.1	383M	1105
TokenLearner 16at18+ L/14	84.7	96.1	447M	1621
TokenLearner 16at18+ L/10	85.4	96.3	450M	4076

Table 8: ViViT + TokenLearner on Kinetics-400, compared to the state-of-the-art models. Different approaches rely on different pre-training datasets, such as ImageNet-21K (for TimeSformer and Swin) and JFT (for ViViT and TokenLearner). The multiplication in GFLOPS correponds to the number of views used for the inference, such as 4x3 = 12.

Method	Top-1 accuracy	total GFLOPS
R(2+1)D [40]	73.9	304 × 115
SlowFast 16x8, R101+NL [14]	79.8	234 × 30
TimeSformer-L [3]	80.7	2380 × 3
ViViT-L/16 [2]	82.8	1446 × 12
Swin-L [25]	83.1	604 × 12
Swin-L (384) [25]	84.6	2107 × 12
Swin-L (384) [25]	84.9	2107 × 50
TokenLearner 16at12 (L/16)	82.1	766 × 6
TokenLearner 8at18 (L/16)	83.2	1105 × 6
TokenLearner 16at12 (L/16)	83.5	766 × 12
TokenLearner 8at18 (L/16)	84.5	1105 × 12
TokenLearner 16at18 (L/14)	84.7	1621 × 12
TokenLearner 16at18 (L/10)	**85.4**	4076 × 12

	TokenLearner	Previous SOTA
Kinetics-400	85.4	84.9
Kinetics-600	86.3	86.1
Charades	66.3	63.2
AViD	53.8	50.9