TokenLearner

Michael S. Ryoo[1,2]
AJ Piergiovanni[1]
Anurag Arnab[1]
Mostafa Dehghani[1]
Anelia Angelova[1]

1 Google Research
2 Stony Brook University

What is TokenLearner for?

A **module** to go inside Vision Transformers (ViT)

Faster: TokenLearner reduces the amount of computation in Transformer models.

- Cuts the computation by ½ or even more.

What is TokenLearner for?

A **module** to go inside Vision Transformers (ViT)

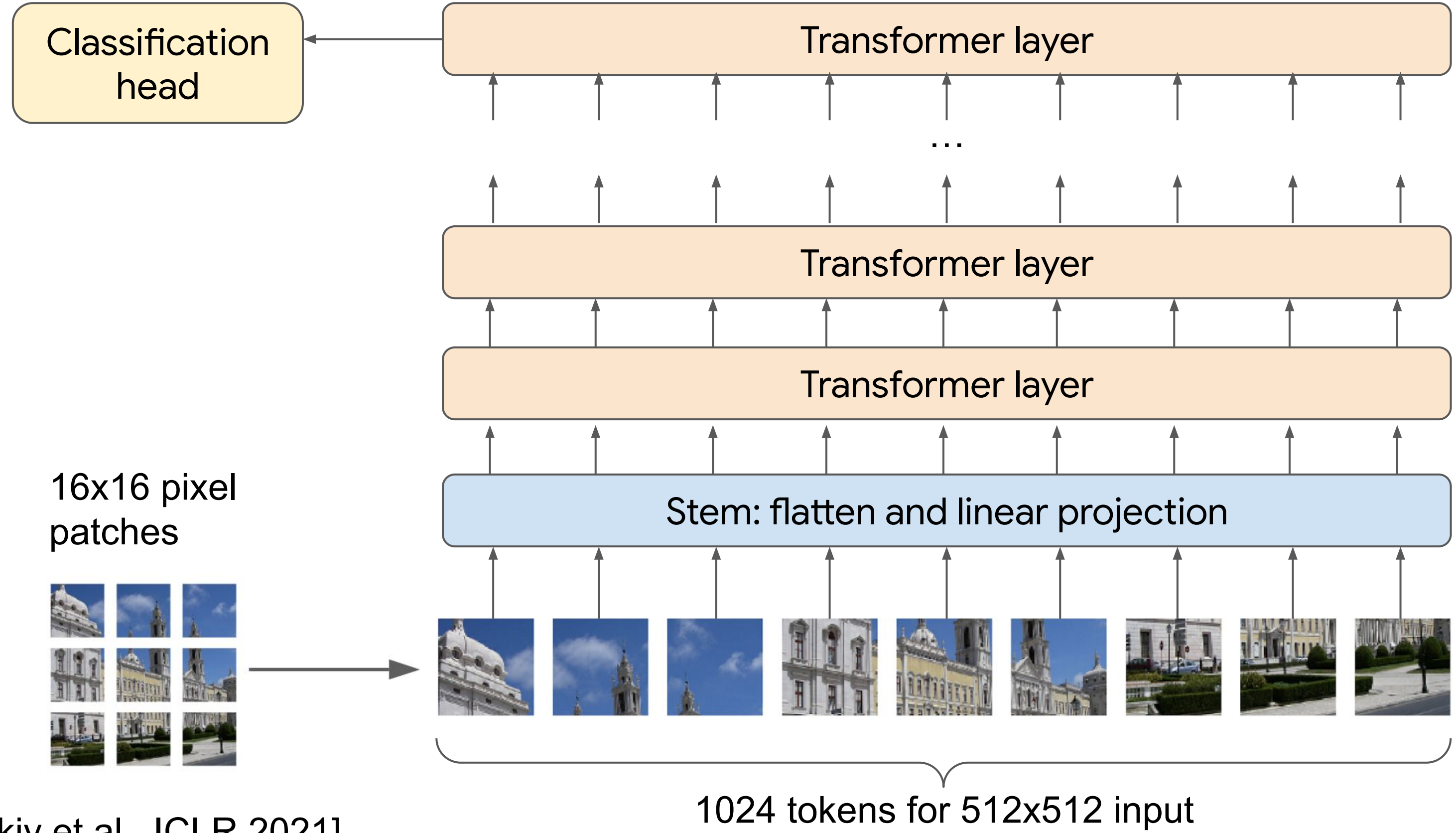
Faster: TokenLearner reduces the amount of computation in Transformer models.

- Cuts the computation by ½ or even more.

Better: Simultaneously, it increases the accuracy of the models

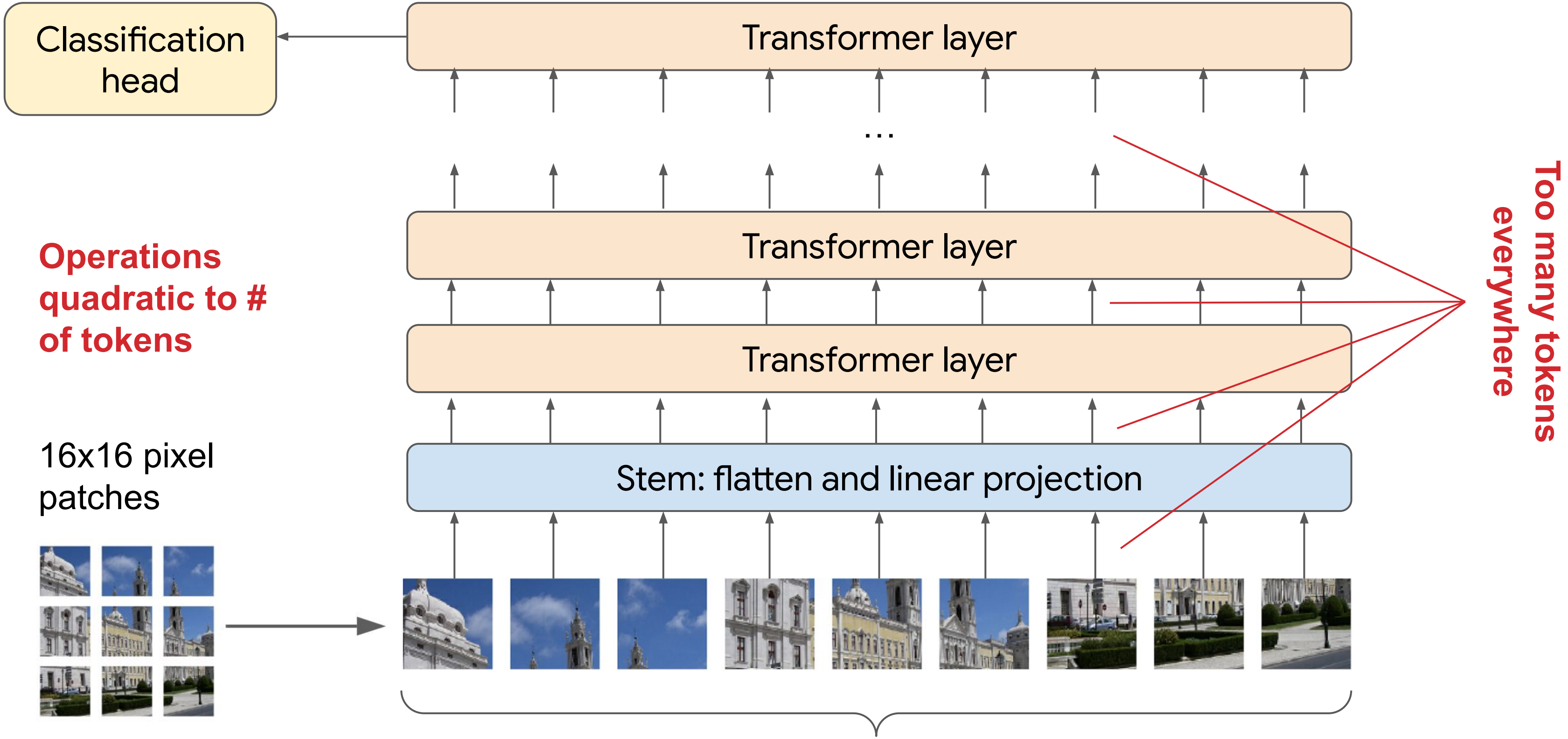
- Better than the full ViT models on image classification and video recognition
- New SOTA on Kinetics-400, Kinetics-600, Charades, and AViD.

Vision Transformer (ViT)



[Dosovitskiy et al., ICLR 2021]

Vision Transformer (ViT) - Limitation



1024 tokens for 512x512 input

[Dosovitskiy et al., ICLR 2021]

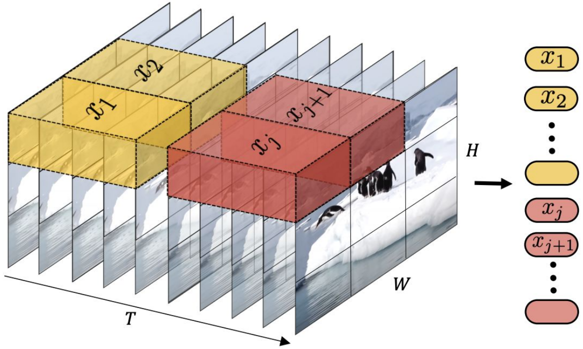
Questions

Do we really need that many tokens and process them all at every layer?

Can we not ‘learn’ to adaptively obtain much fewer tokens instead, and focus on processing them?

Motivation - TokenLearner

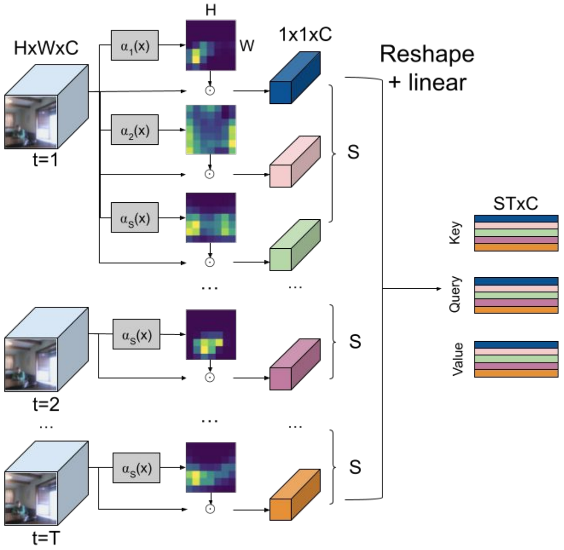
Instead of always using hand-designed tokenization, we learn to adaptively tokenize.



Previous tokenization for images/videos:
spatio-temporal cropping (ViViT)

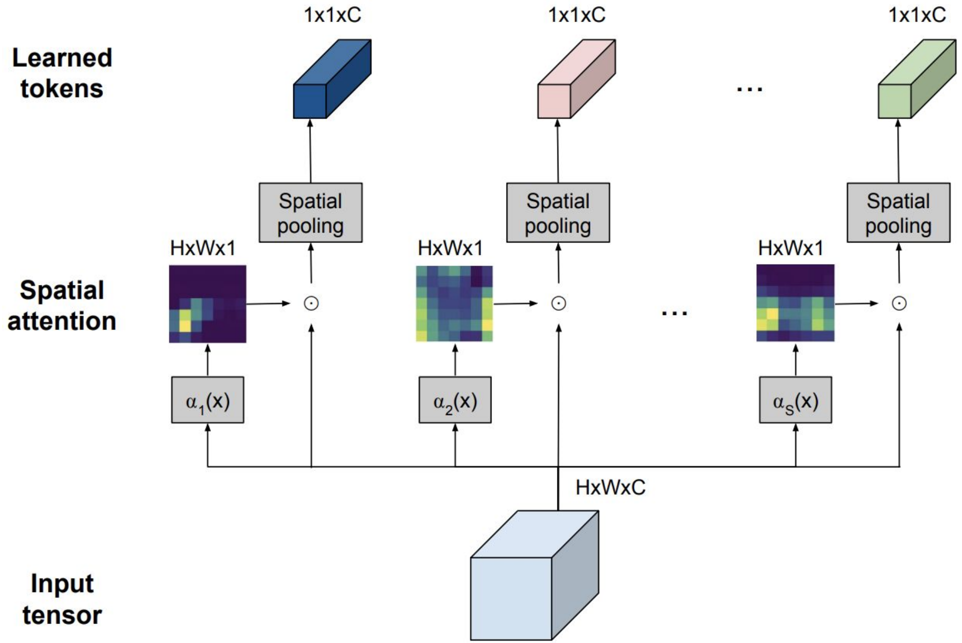
500 * 64 tokens

vs.



8 * 64 tokens

TokenLearner module



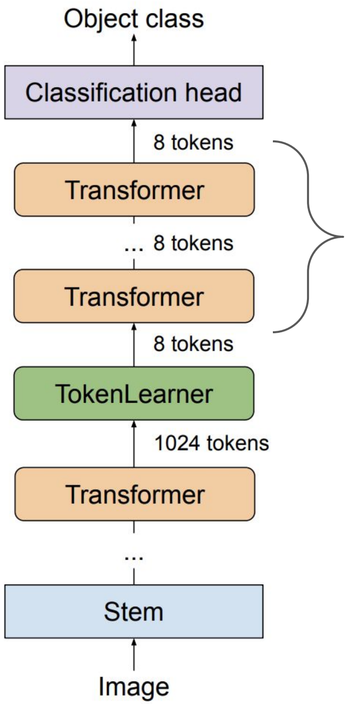
TokenLearner has a form of spatial attention mechanism

Given an image-like tensor, it

- Weights each pixel differently (i.e., focuses on a subset of pixels)
- Summarizes them as a token.
- Could be applied to intermediate tensors

Small number of tokens! 8 or 16

TokenLearner for ViT



Extremely lightweight computation

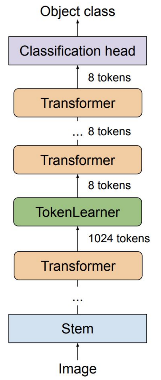
TokenLearner module inserted in the middle of Transformer architecture

- Backbone: ViT - L/16, B/16, ...

Dataset:

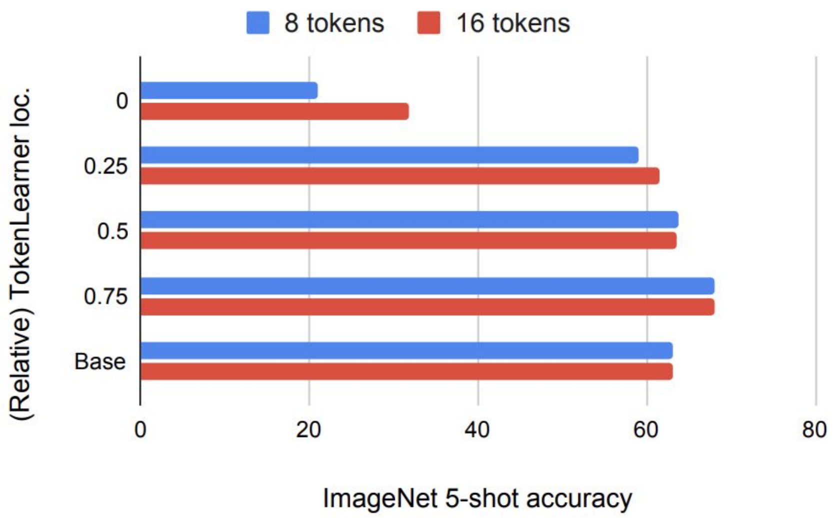
- JFT for the pretraining
- ImageNet for the fine-tuning and evaluation

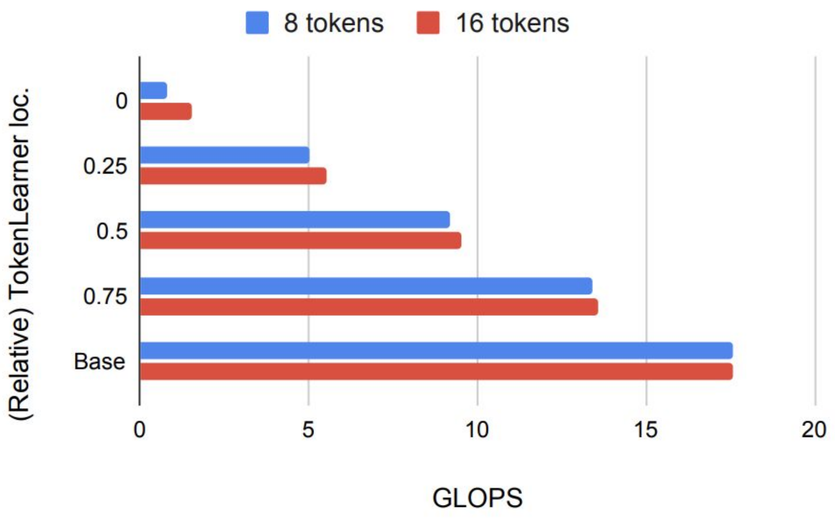
Where do we put TokenLearner?



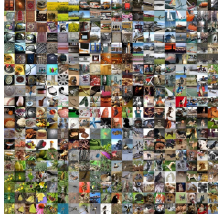
ImageNet 5-shot transfer accuracy (with ViT B/16)

- Interestingly, TokenLearner performs better, while being faster. Adaptiveness!

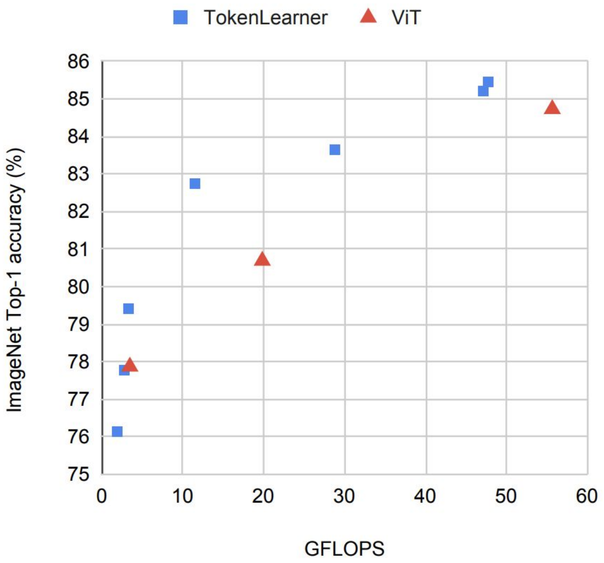


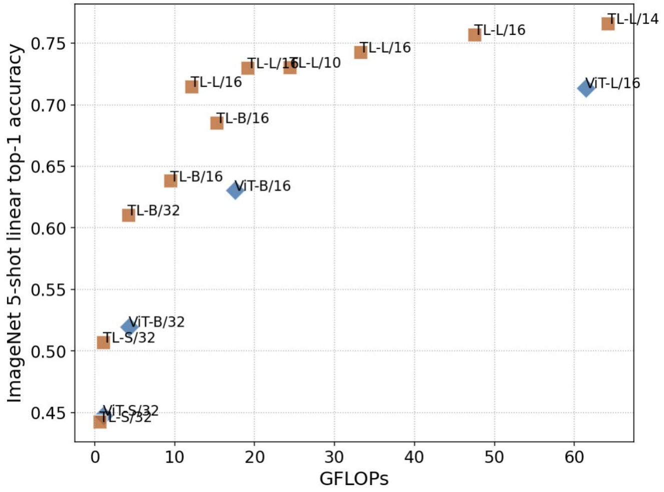


ImageNet



Pre-train with JFT, and fine-tune (or 5-shot learn) with ImageNet1K (with S/32, B/32, B/16, L/16, ...)





Scaling up - larger models

TokenLearner added to ViT L/16 (512x512 input)

- Actual images per second on TPU: ~1400 (L/16) vs. ~2500 (TokenLearner + L/16).

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 + Fuser	24+11	16-TL at 12	191.3	87.91
ViT L/14	24+11	16-TL at 18	361.6	88.37

*16 Tokens used in TokenLearner.

Scaling up - larger models

TokenLearner added to ViT L/16 (512x512 input)

- Actual images per second on TPU: ~1400 (L/16) vs. ~2500 (TokenLearner + L/16).

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 + Fuser	24+11	16-TL at 12	191.3	87.91
ViT L/14	24+11	16-TL at 18	361.6	88.37

*16 Tokens used in TokenLearner.

Scaling up - larger models

TokenLearner added to ViT L/16 (512x512 input)

- Actual images per second on TPU: ~1400 (L/16) vs. ~2500 (TokenLearner + L/16).

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 + Fuser	24+11	16-TL at 12	191.3	87.91
ViT L/14	24+11	16-TL at 18	361.6	88.37

*16 Tokens used in TokenLearner.

Scaling up - heavier models

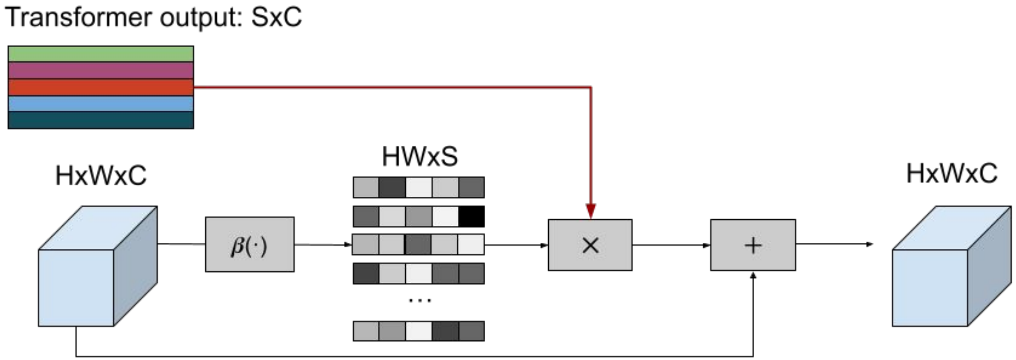
ImageNet comparison against the SOTA Transformer models

- L/8 is with the same model size but with 4x larger number of initial tokens.

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**

*16 Tokens used in TokenLearner.

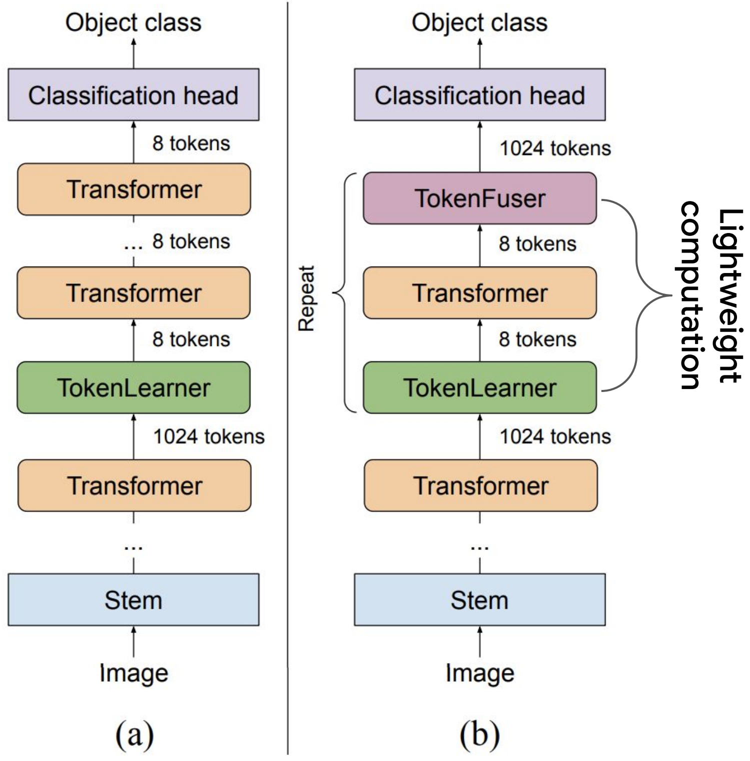
TokenFuser module



TokenFuser recombines tokens to recover the original input shape.

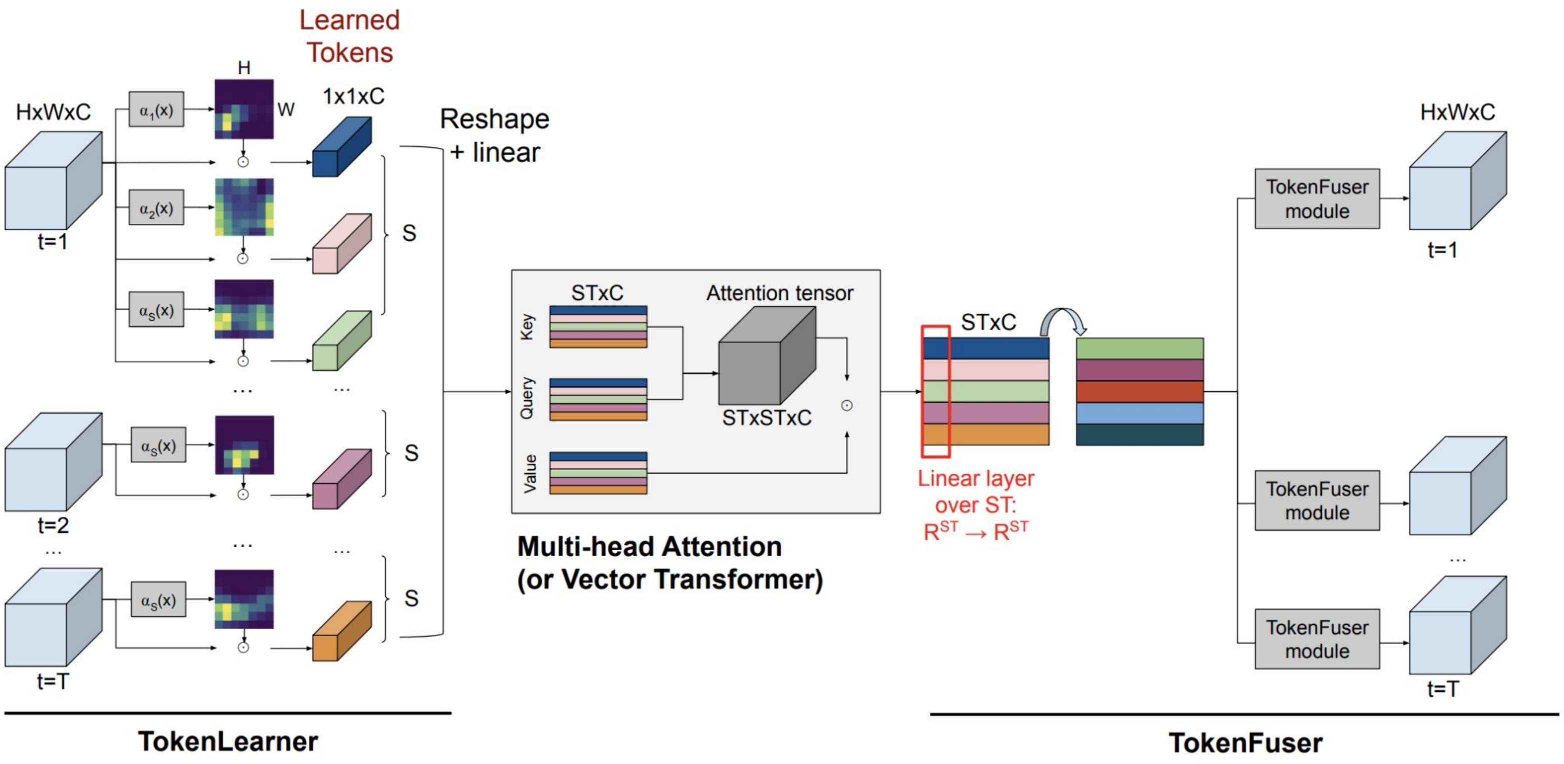
It learns to generate fusion weights per pixel location, conditioned on the input tensor.

ViT architecture with TokenFuser



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

Video architecture with TokenFuser



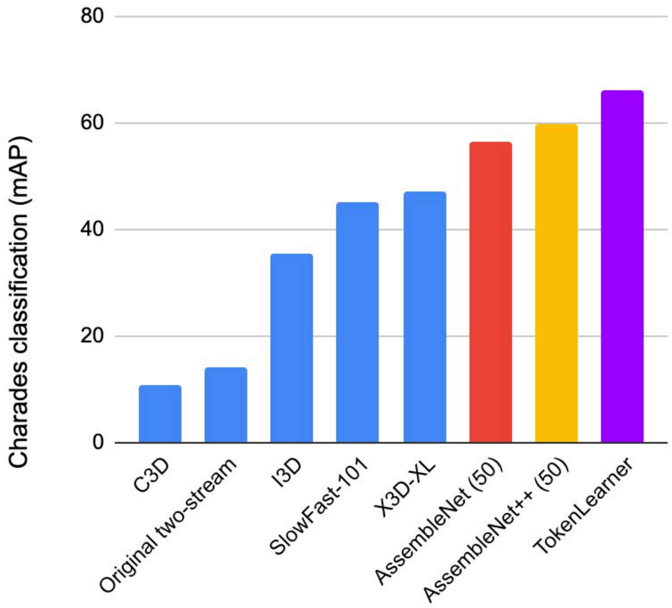
Video datasets

Kinetics400, Charades, and AViD

- Compared against prior works, including ViViT
- 85.4% on Kinetics400 is the new SOTA.

Kinetics400 results

Method	Accuracy	GFLOPS
ViViT-L/16	82.8	1446
ViViT-L/16 320	83.5	3992
ViViT-H/14	84.8	3981
ViViT-L/16 (our run)	83.4	1446
TokenLearner 16at12 + L/16	83.5	766
TokenLearner 8at18 + L/16	84.5	1105
TokenLearner 16at18+ L/14	84.7	1621
TokenLearner 16at18+ L/10	85.4	4076



Video datasets

Kinetics400, Charades, and AViD

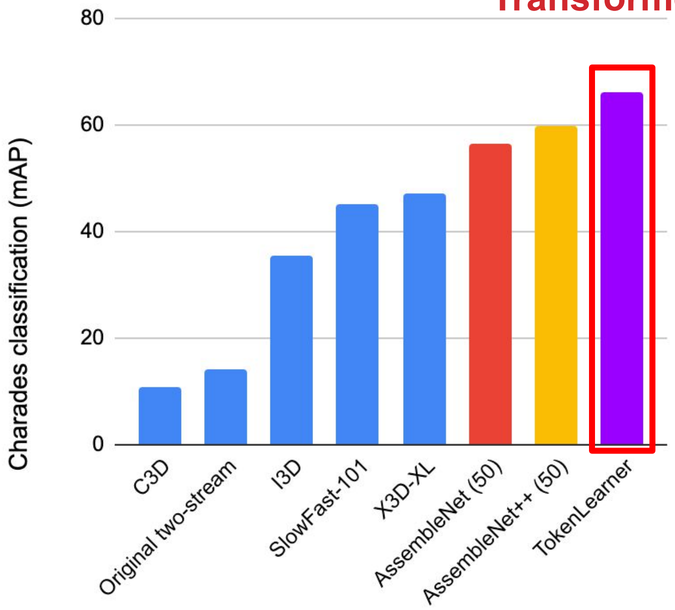
- Compared against prior works, including ViViT
- 85.4% on Kinetics400 is the new SOTA.

Kinetics400 results

Method	Accuracy	GFLOPS
ViViT-L/16	82.8	1446
ViViT-L/16 320	83.5	3992
ViViT-H/14	84.8	3981
ViViT-L/16 (our run)	83.4	1446
TokenLearner 16at12 + L/16	83.5	766
TokenLearner 8at18 + L/16	84.5	1105
TokenLearner 16at18+ L/14	84.7	1621
TokenLearner 16at18+ L/10	85.4	4076

TokenLearner within ViViT

TokenLearner within Bottleneck Transformer



Method	Top-1 accuracy	total GFLOPS
R(2+1)D [39]	73.9	304 × 115
SlowFast 16x8, R101+NL [13]	79.8	234 × 30
TimeSformer-L [3]	80.7	2380 × 3
ViViT-L/16 [2]	82.8	1446 × 12
Swin-L [24]	83.1	604 × 12
Swin-L (384) [24]	84.6	2107 × 12
Swin-L (384) [24]	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

Comparison to SOTA (Kinetics 400)

Method	Top-1	Top-5	total GFLOPS
SlowFast 16x8, R101+NL [13]	81.8	95.1	234 × 30
X3D-XL [12]	81.9	95.5	48 × 30
TimeSformer-HR [3]	82.4	96.0	1703 × 3
ViViT-L/16 [2]	84.3	96.2	1446 × 12
Swin-B [24]	84.0	96.5	282 × 12
Swin-L (384) [24]	85.9	97.1	2107 × 12
Swin-L (384) [24]	86.1	97.3	2107 × 50
TokenLearner 16at12 (L/16)	84.4	96.0	766 × 12
TokenLearner 8at18 (L/16)	86.0	97.0	1105 × 12
TokenLearner 16at18 (L/10)	86.1	97.0	4076 × 12
TokenLearner 16at18 w. Fuser (L/10)	**86.3**	97.0	4100 × 12

Comparison to SOTA (Kinetics 600)

Charades results

Longer videos

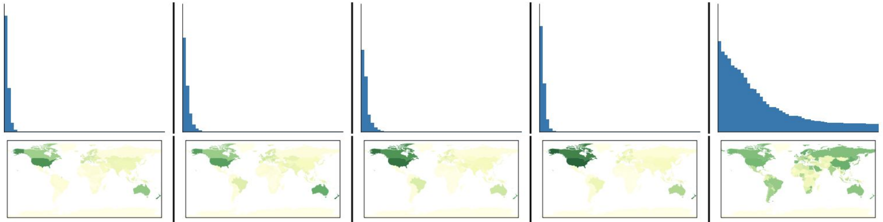
- ~30 seconds
- 360 frames with 12 fps

Method	Input	Pre-train	mAP
I3D [5]	RGB	Kinetics	32.9
I3D from [39]	RGB	Kinetics	35.5
I3D + Non-local [39]	RGB	Kinetics	37.5
EvaNet [25]	RGB	Kinetics	38.1
STRG [40]	RGB	Kinetics	39.7
LFB-101 [42]	RGB	Kinetics	42.5
SGFB-101 [19]	RGB	Kinetics	44.3
SlowFast-101 [12]	RGB+RGB	Kinetics	45.2
AssembleNet-50 [29]	RGB+Flow	None	47.0
Multiscale ViT [10]	RGB	Kinetics	47.7
AssembleNet-101 [29]	RGB+Flow	Kinetics	58.6
AssembleNet++ [28] (w/o object)	RGB+Flow	None	55.0
MoViNets [22]	RGB	None	63.2
Backbone (X(2+1)D-M)	RGB	None	62.7
Ours	RGB	None	**66.3**

AViD dataset results

Anonymous Videos from Diverse Countries

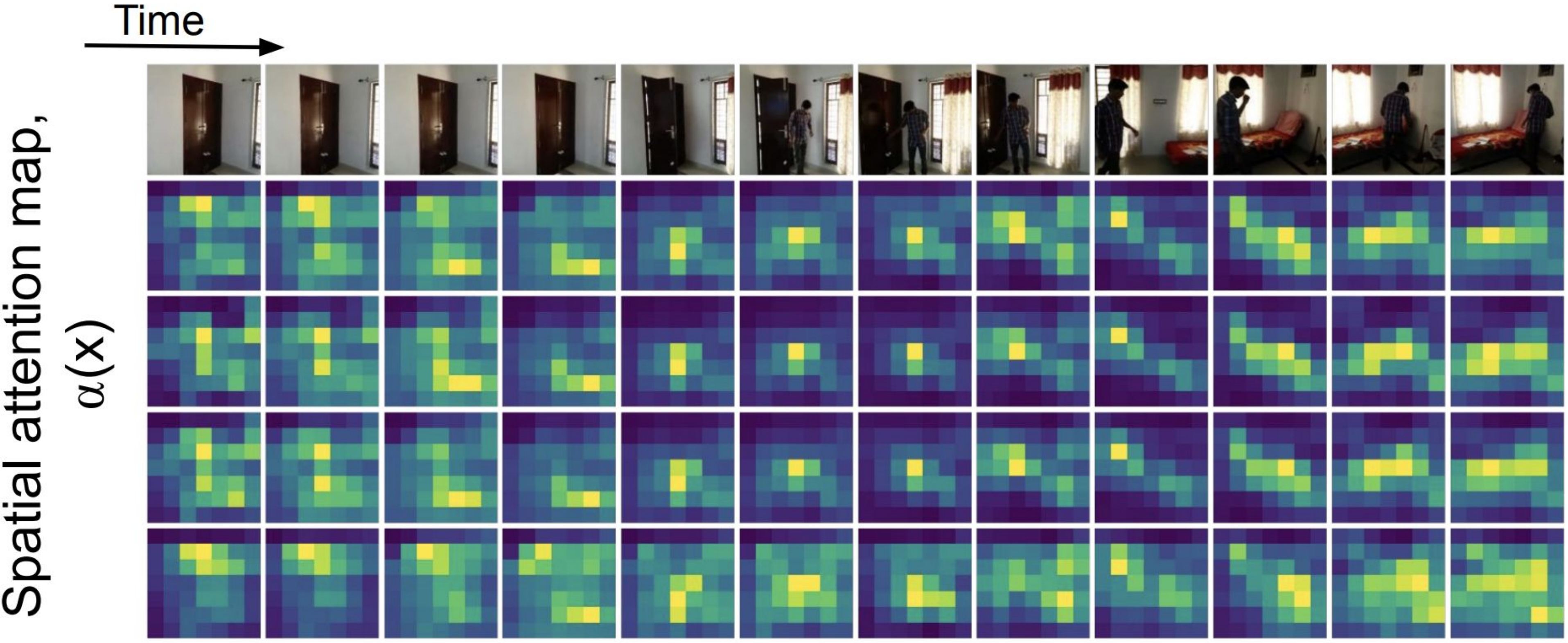
- 467k videos and 887 action classes
- 3-15 sec per video



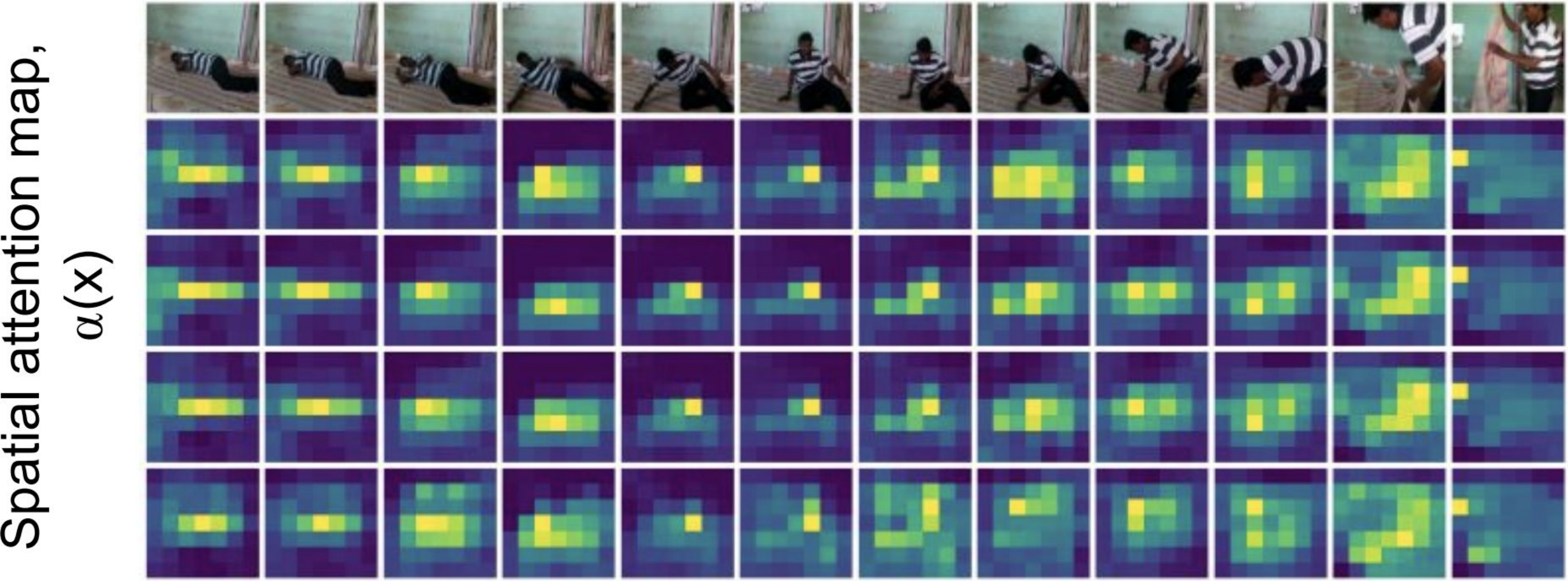
(a) Kinetics-400 (b) Kinetics-600 (c) HVU (d) HACS (e) AViD

Method	Accuracy
I3D [5]	46.5
(2+1)D ResNet-50	46.7
3D ResNet-50	47.9
SlowFast-50 4x4 [12]	48.5
SlowFast-50 8x8 [12]	50.2
SlowFast-101 16x4 [12]	50.8
Backbone (X(2+1)D-M)	48.6
X(2+1)D-M w/ joint space-time module (like [2])	53.1
Ours	**53.8**

Visualizing spatial attention in TokenLearner



Visualizing spatial attention in TokenLearner



Thank you for listening.

Michael Ryoo
mryoo@google.com

AJ Piergiovanni
ajpiergi@

Anurag Arnab
aarnab@

Mostafa Dehghani
dehghani@

Anelia Angelova
anelia@