An Image Is Worth 393 Areas:

Training image Transformers with Area-Attention

Osher Tidhar

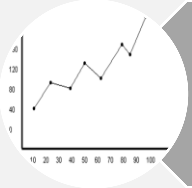
Yoav Kurtz

Transformer

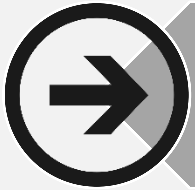
Agenda

Motivation

Our Method



Results



Further Steps

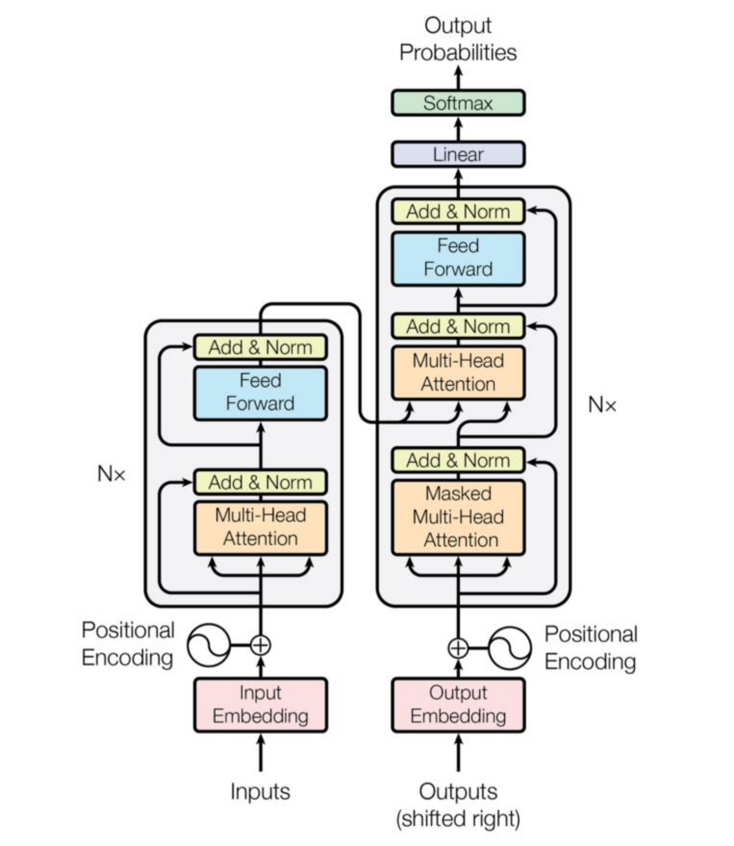
Transformers

- Model of choice for NLP problems.

OpenAI

GPT-3, an autoregressive language model with 175 billion parameters

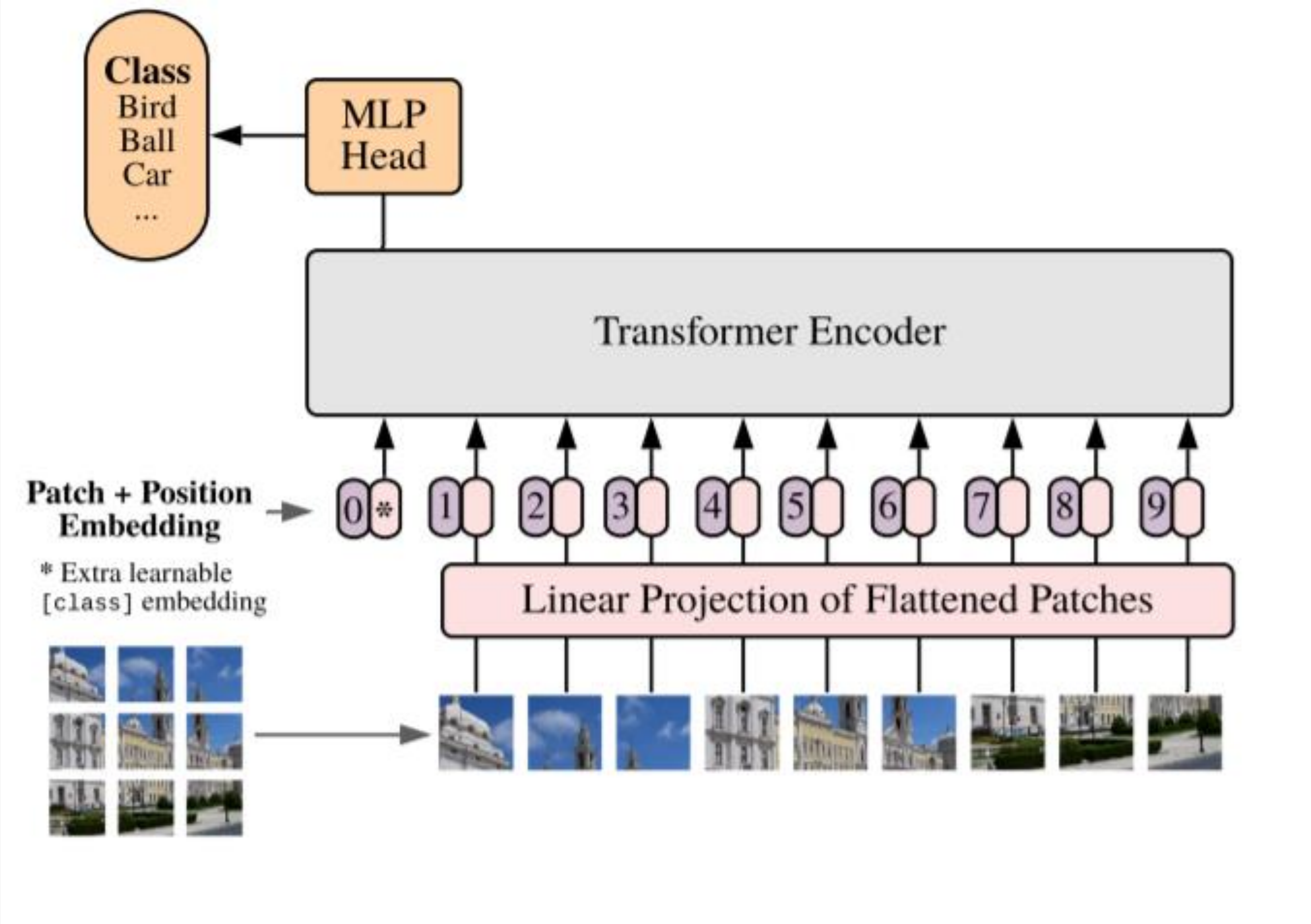
- Recently migrated to computer vision.



Transformer Overview

Vision Transformer[1]

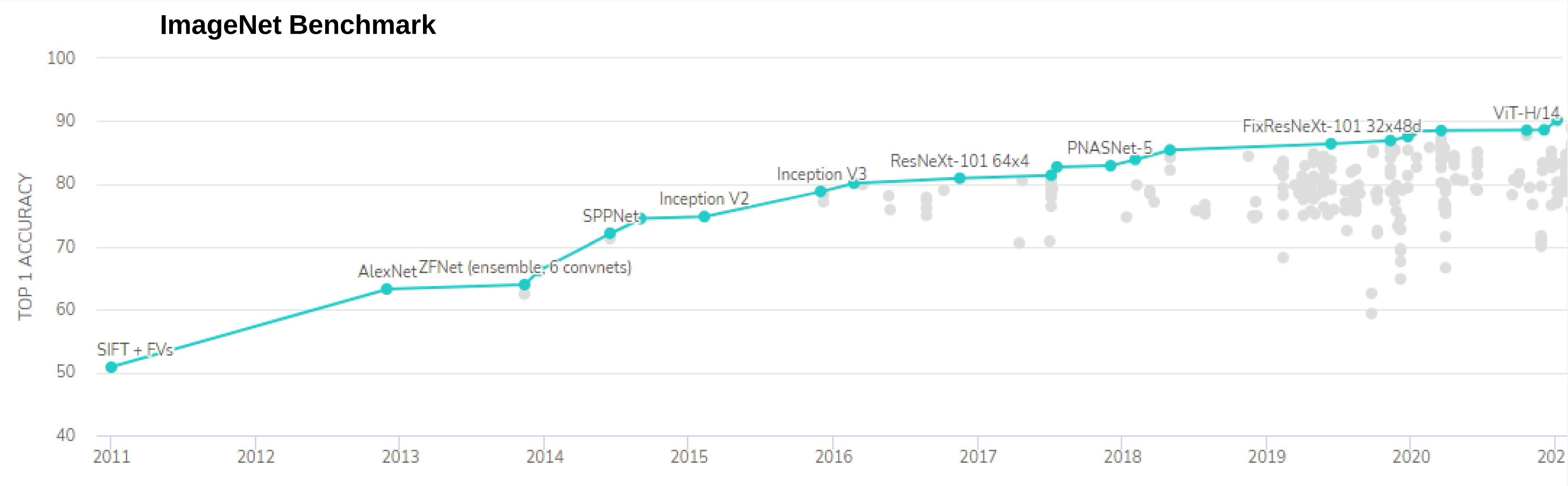
- Motivation
- Our Method
- Results
- Further Steps



[1]An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Transformers

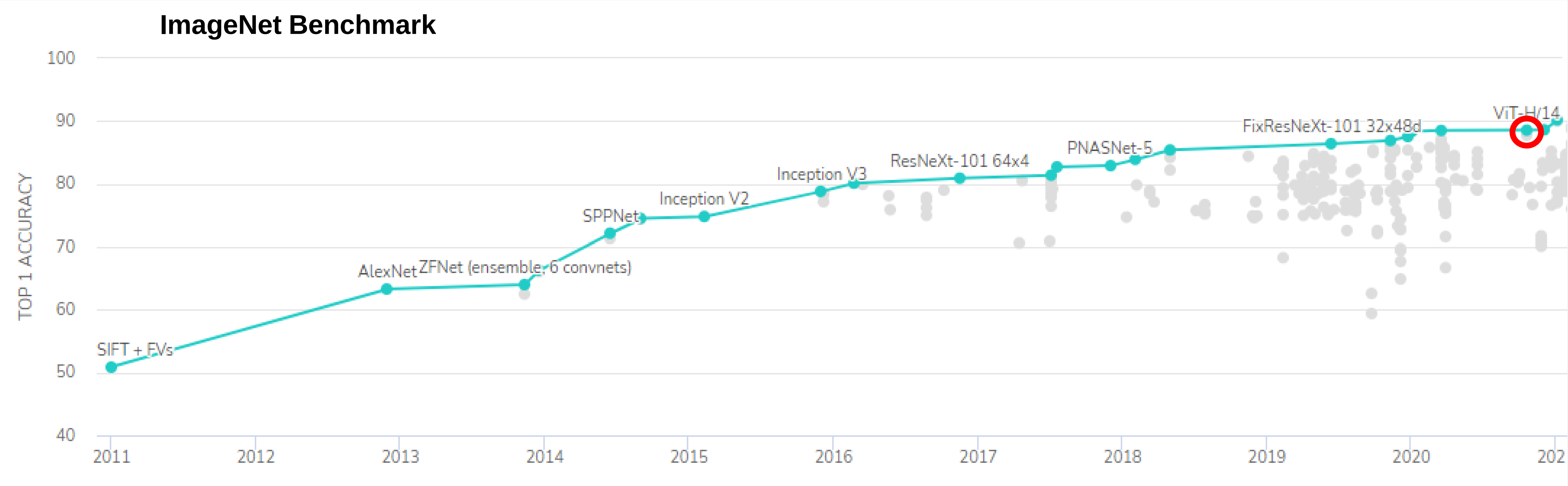
- Model of choice for NLP problems.
- Recently migrated to vision, showing **competitive**[1,2] results.



[1] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
[2] Training data-efficient image transformers & distillation through attention

Transformers

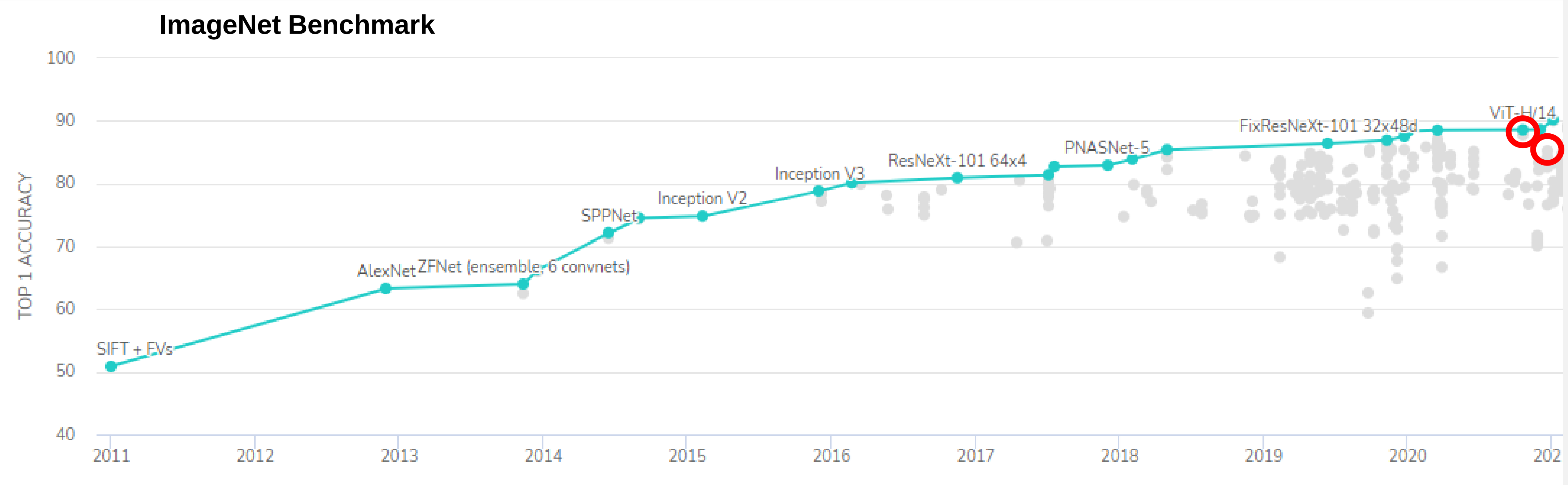
- Model of choice for NLP problems.
- Recently migrated to vision, showing **competitive**[1,2] results.



[1] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
[2] Training data-efficient image transformers & distillation through attention

Transformers

- Model of choice for NLP problems.
- Recently migrated to vision, showing **competitive**[1,2] results.

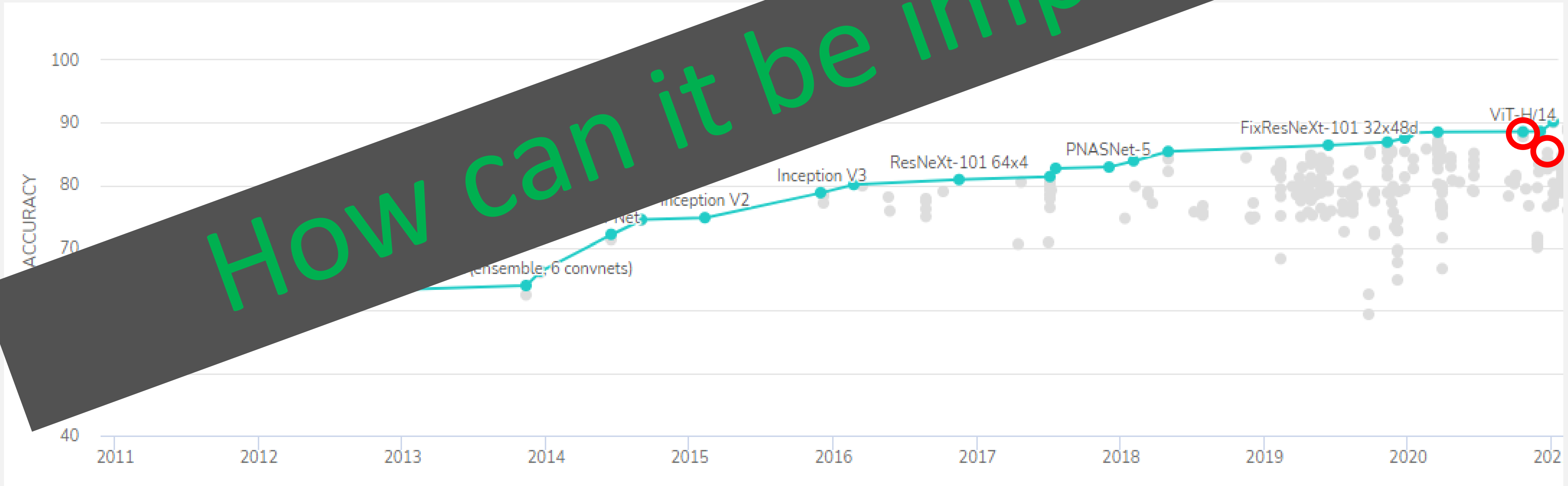


[1]An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
[2]Training data-efficient image transformers & distillation through attention

Transformers

- Model of choice for NLP problems.
- Recently migrated to vision, showing **com**

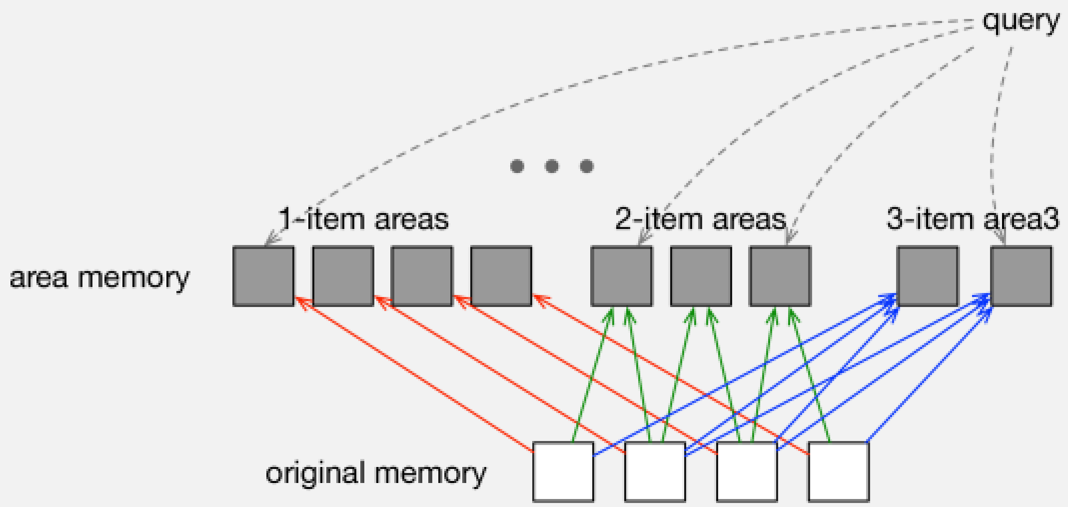
How can it be improved?



[1]An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
[2]Training data-efficient image transformers & distillation through attention

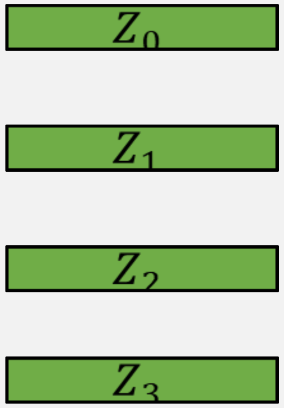
Area-Attention[1]

- Attending group of items in the memory that are structurally adjacent.
- Model can attend to combinations of items.



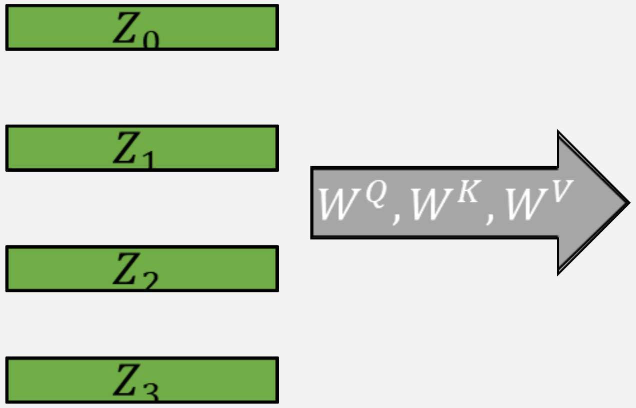
[1]Area Attention

Area-Attention[1]



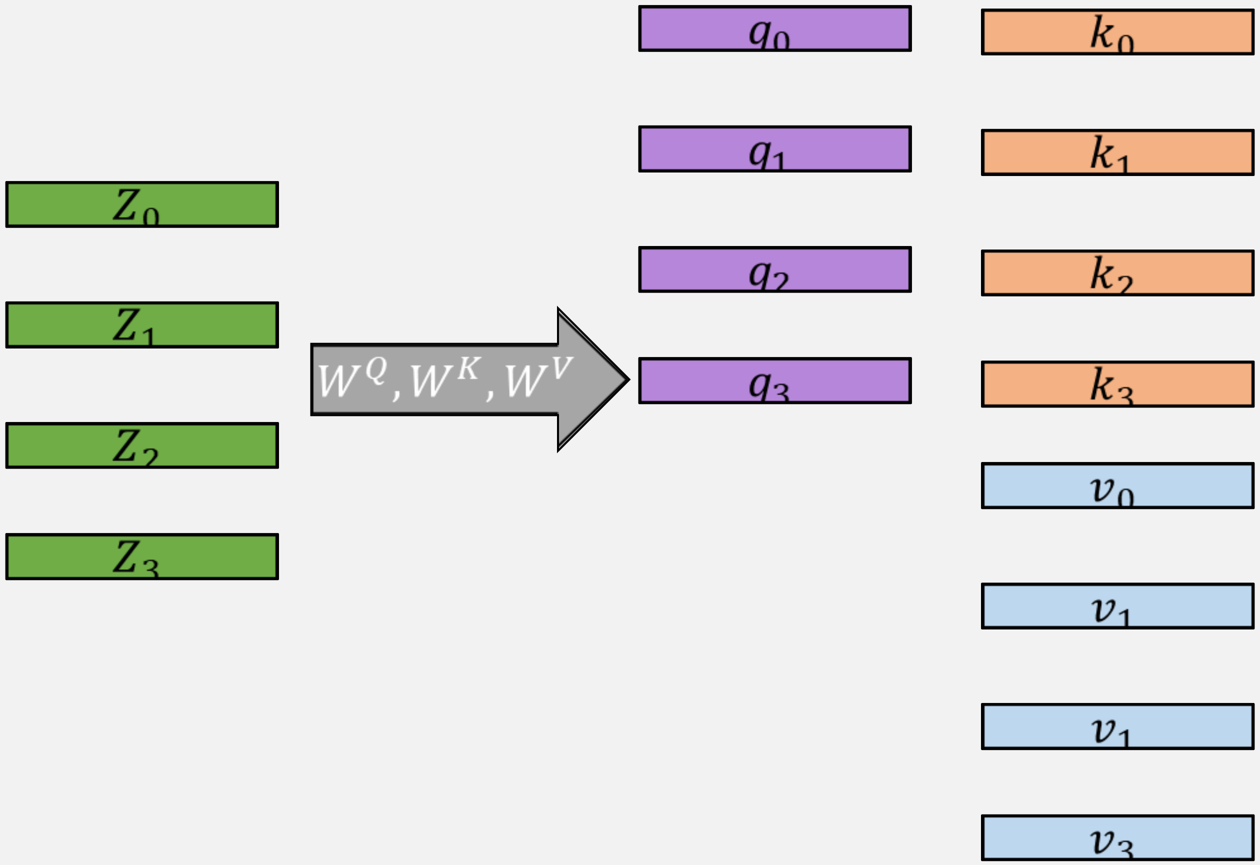
[1]Area Attention

Area-Attention[1]



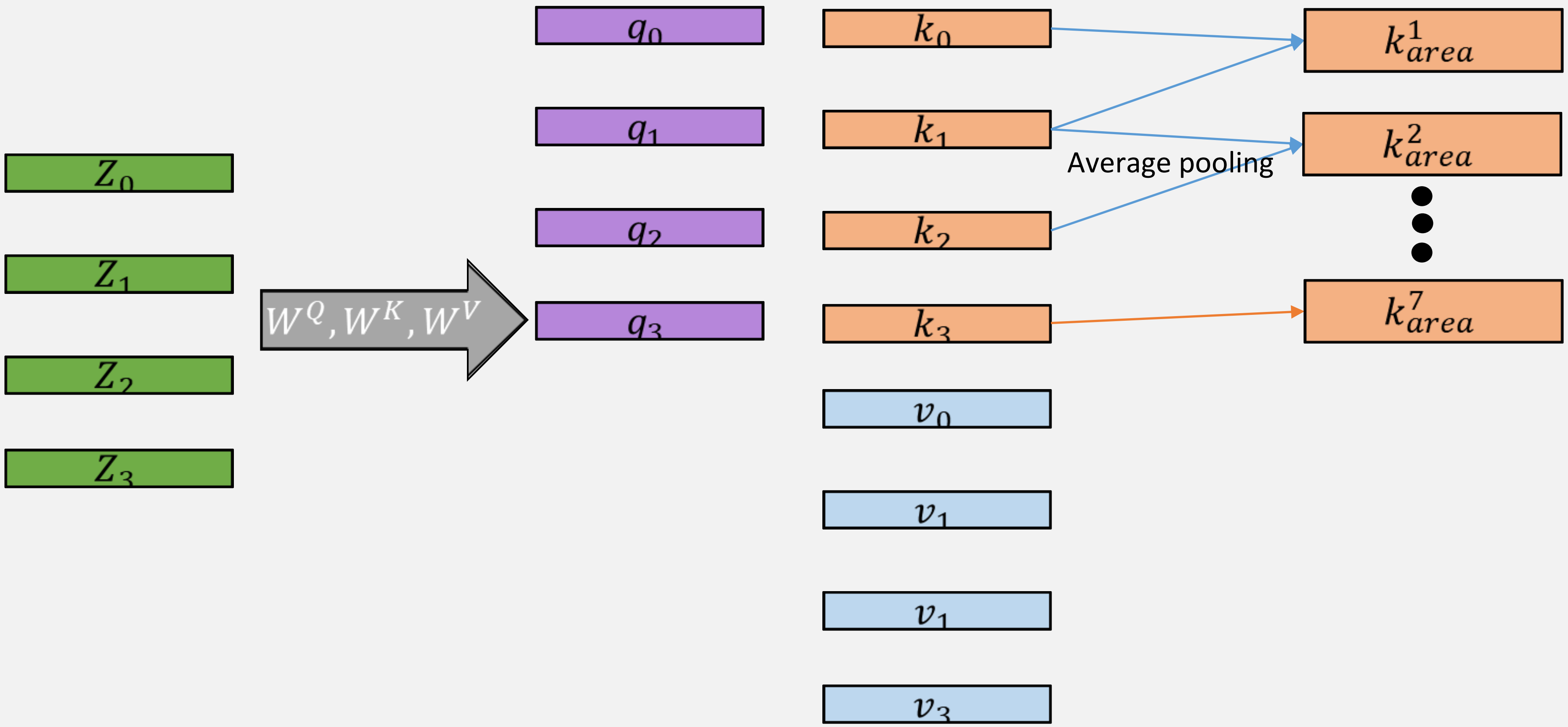
[1]Area Attention

Area-Attention[1]



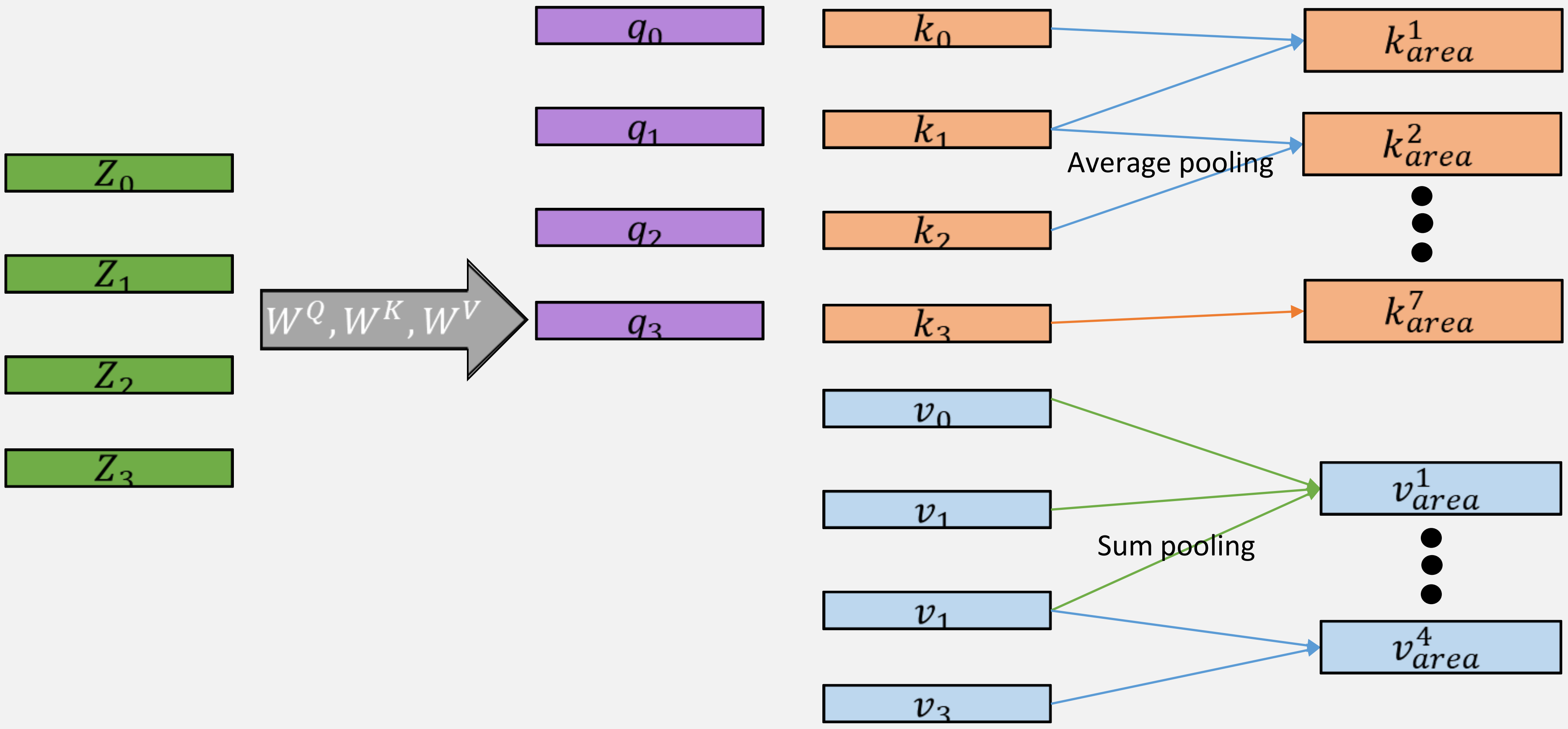
[1]Area Attention

Area-Attention[1]



[1]Area Attention

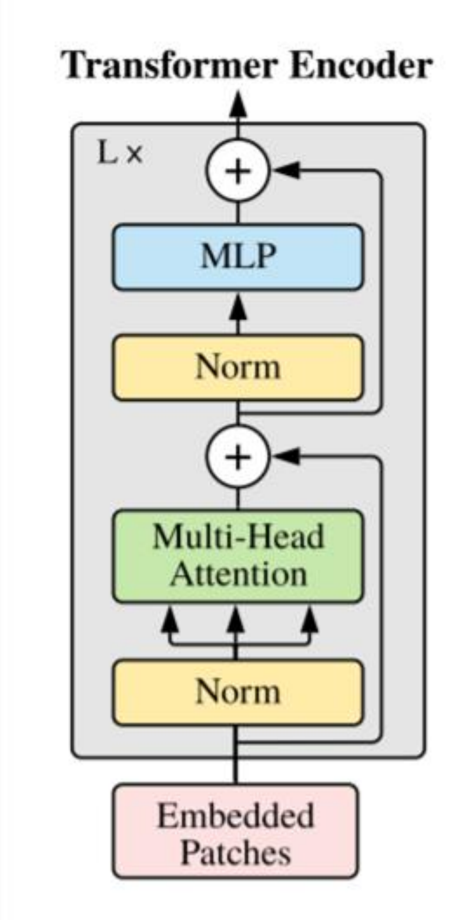
Area-Attention[1]



[1]Area Attention

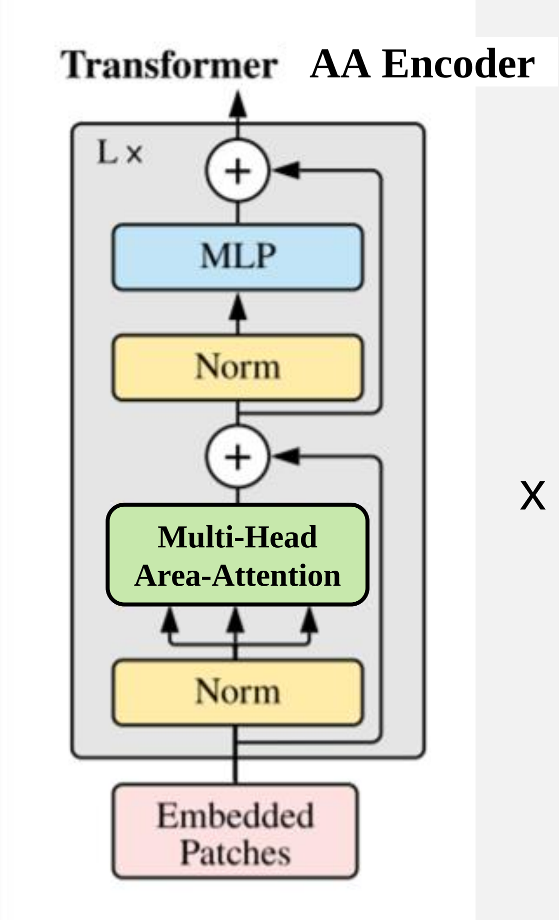
Vision Transformer + Area Attention

- Multi-head self-attention replaced with multi-head area-attention.
- Different AA configurations are tested.



Vision Transformer + Area Attention

- Multi-head self-attention replaced with multi-head area-attention.
- Different AA configurations are tested.



x L_{AA} first blocks

Number of areas in ViT + AA

- In ViT, each image is represented by patches of 16x16 pixels.
- In our model, what is the total number of areas that can be generated?

→ For the following configurations:

(H, W)	(P, P)	max area size
224x224	16x16	2

we got a sequence of length 197:

14x14 patch images + 1 token class.

which corresponds to 393 areas:

197 areas built of 1 element + 196 areas built of a combination of 2 adjacent elements.

Choosing a dataset for our experiments

- Dataset - CIFAR-10

	Train size	Test size	#classes
CIFAR-10	50,000	10,000	10

Choosing models for our experiments

- Original architecture of the Vision-Transformer model:

	embedding	#heads	#layers	#params	training resolution
ViT-small	768	12	12	86M	224

- Models we used: Vision-Transformers small[1] and tiny[2]

	embedding	#heads	#layers	# AA layers	#params	training resolution
ViT-small	384	6	12	n/a	22M	224
ViT-small+ AA	384	6	12	2	22M	224
ViT-tiny	192	3	12	n/a	5M	224
ViT-tiny+ AA	192	3	12	2	5M	224

[1,2] DeiT

Choosing models for our experiments

- Original architecture of the Vision-Transformer model:

	embedding	#heads	#layers	#params	training resolution
ViT-small	768	12	12	86M	224

- Models we used: Vision-Transformers small[1

Same number of Parameters

	embedding	#heads	#l[illegible]	[illegible]	[illegible]params	training resolution
ViT-small	384	6	[illegible]	n/a	22M	224
ViT-small+ AA	384	6	12	2	22M	224
ViT-tiny	192	3	12	n/a	5M	224
ViT-tiny+ AA	192	3	12	2	5M	224

[1,2] DeiT

Accuracy achieved with ViT + AA

- Accuracy of our pretrained weights on CIFAR-10 Testset:
  - ViT-Small model:

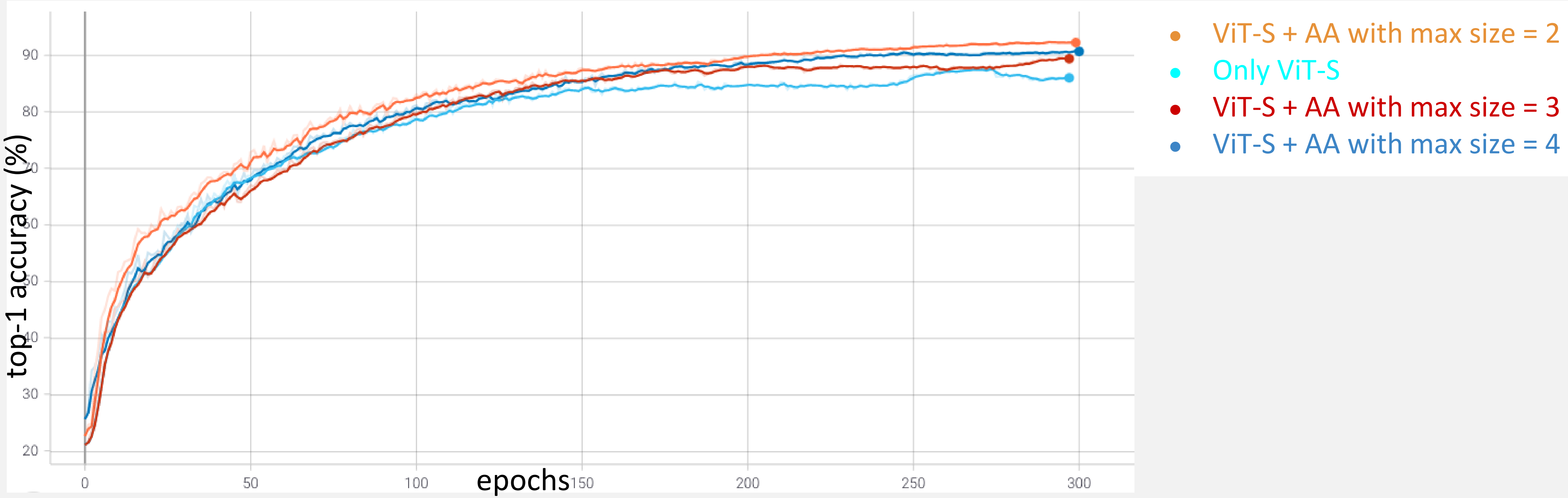
	top-1 acc	top-5 acc	loss
ViT-small + AA with max_size=2	**92.19**	**99.68**	**0.38**
ViT-small + AA with max_size=3	**89.32**	**99.51**	**0.473**
ViT-small + AA with max_size=4	**85.67**	**98.29**	**0.577**
Only ViT-small	90.6	99.47	0.411

  - ViT-Tiny model:

	top-1 acc	top-5 acc	loss
ViT-tiny + AA with max_size=2	**86.14**	**99.52**	**0.557**
Only ViT-tiny	85.49	99.49	0.576

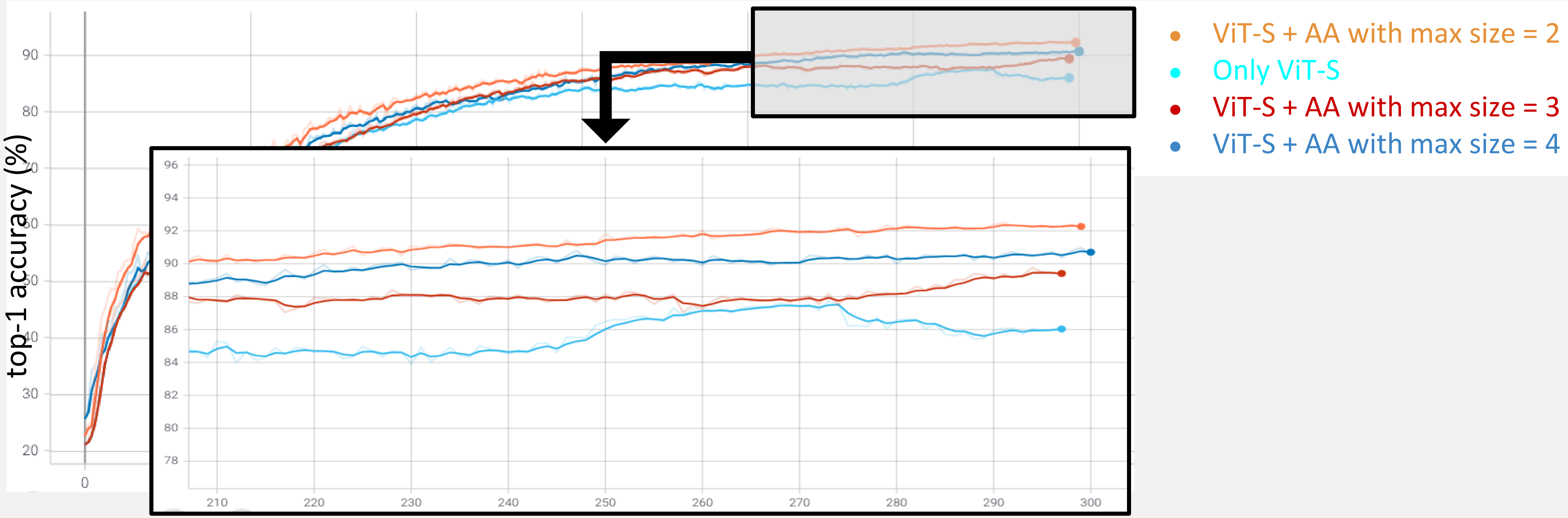
Accuracy achieved with ViT + AA

- Only AA2 configuration achieves better accuracy than only ViT-S.
- Accuracy improves as we decrease the maximum size of an area (for the first 2 blocks).



Accuracy achieved with ViT + AA

- Only AA2 configuration achieves better accuracy than only ViT-S.
- Accuracy improves as we decrease the maximum size of an area (for the first 2 blocks).



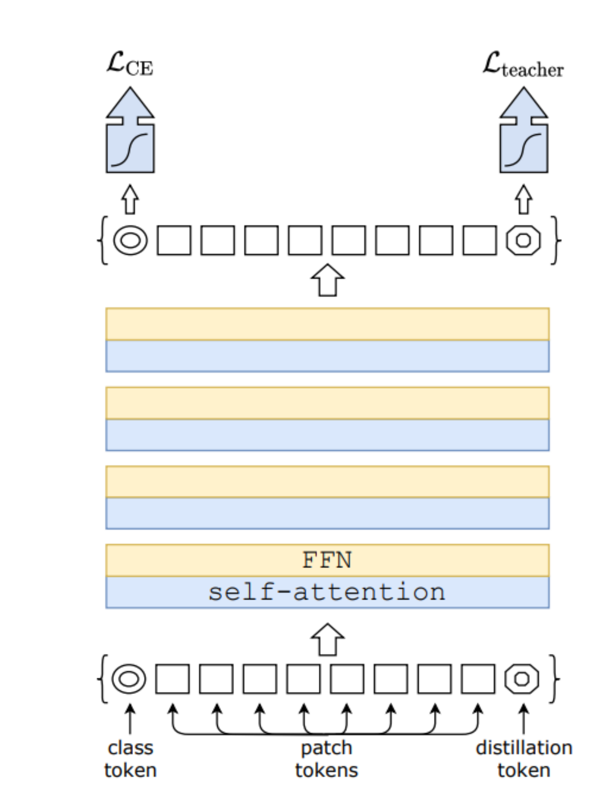
Further steps

- Training AA-ViT using network distillation.

  - Plays the same role as the class token, except that it aims at reproducing the label estimated by the teacher.

  - Both tokens interact in the transformer through attention

  - Achieved results that were competitive with the results of convnets for Imagenet.



Questions?