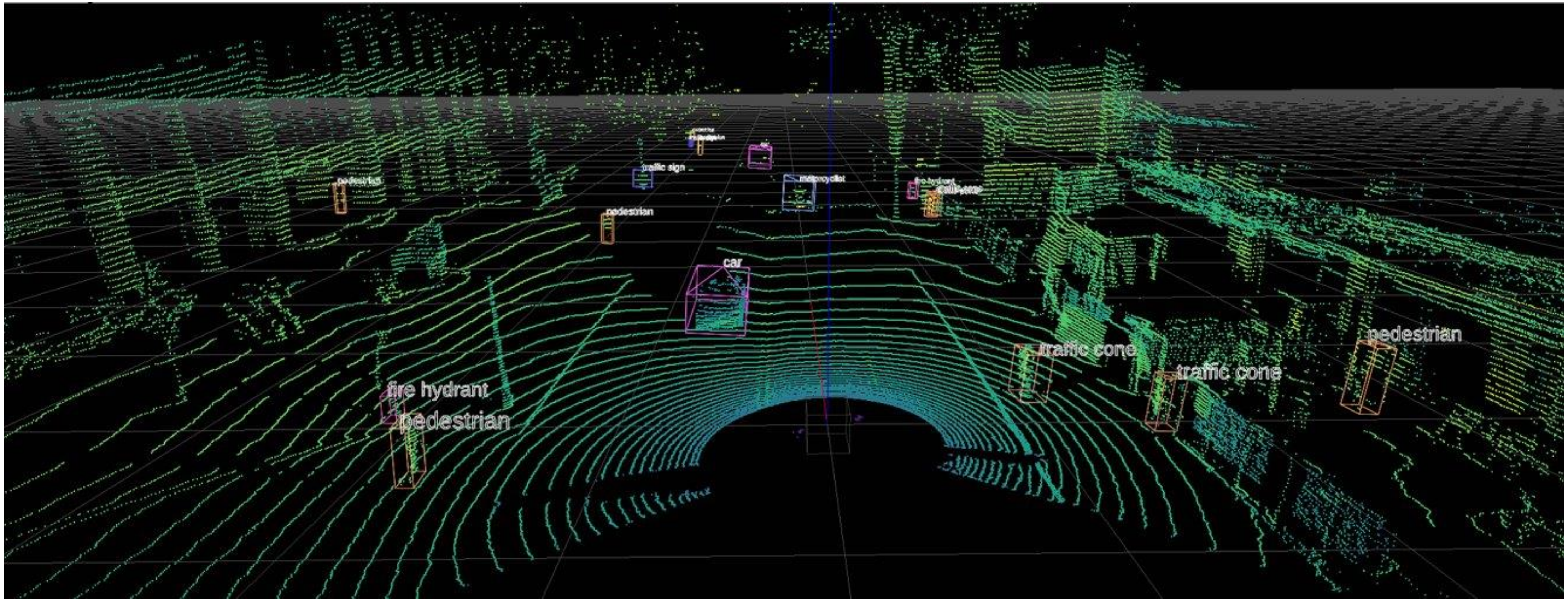
Transformers in Computer Vision

William Guimont-Martin

Transformers + Point Clouds



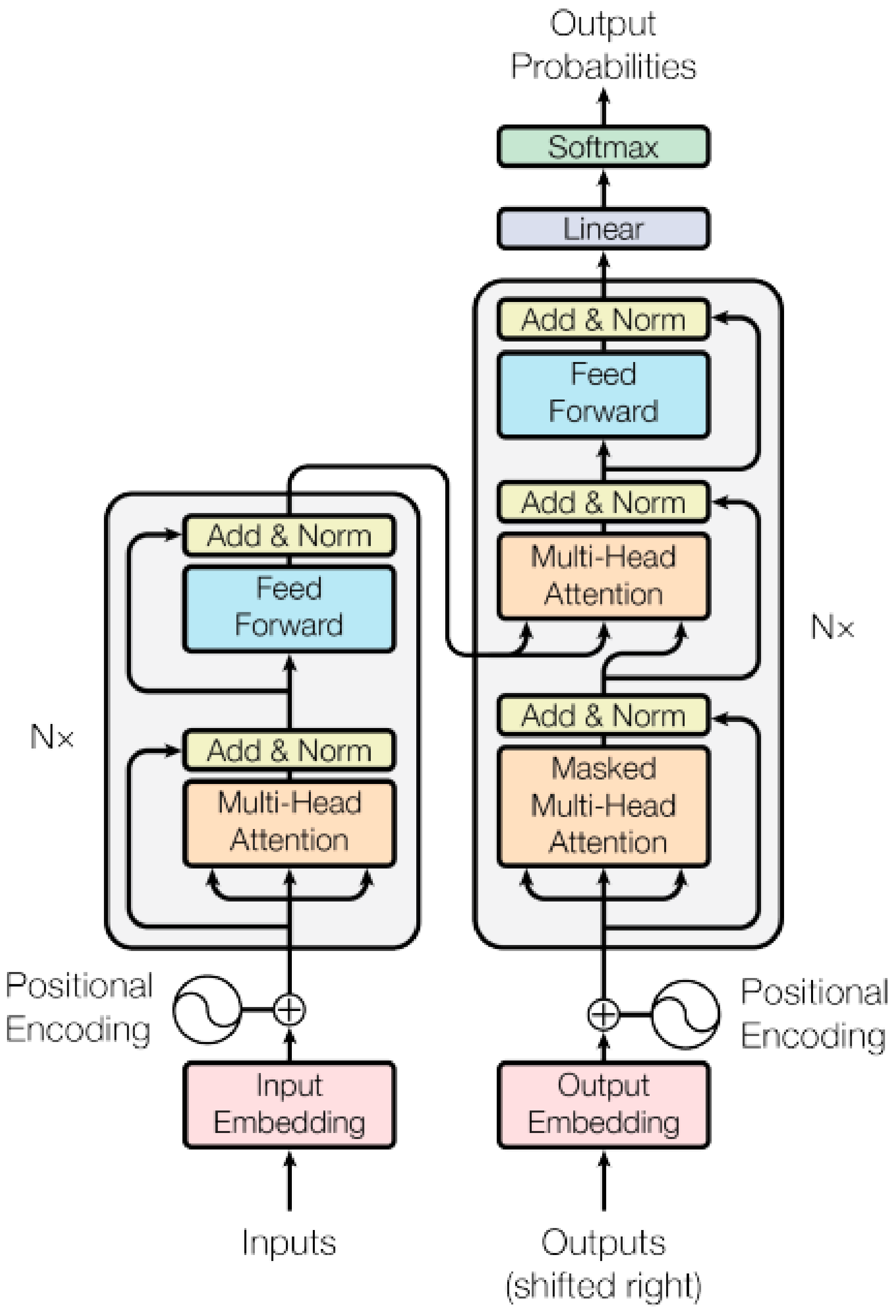


Figure 1: The Transformer - model architecture.

Transformers in NLP

- Attention Is All You Need (2017)
- Revolution in NLP
  - GPT-3 (Generative Pre-trained **Transformer** 3)
    - 175 billion parameters
    - 499 billion tokens
  - BERT (Bidirectional Encoder Representations from **Transformers**)
    - 110 million parameters

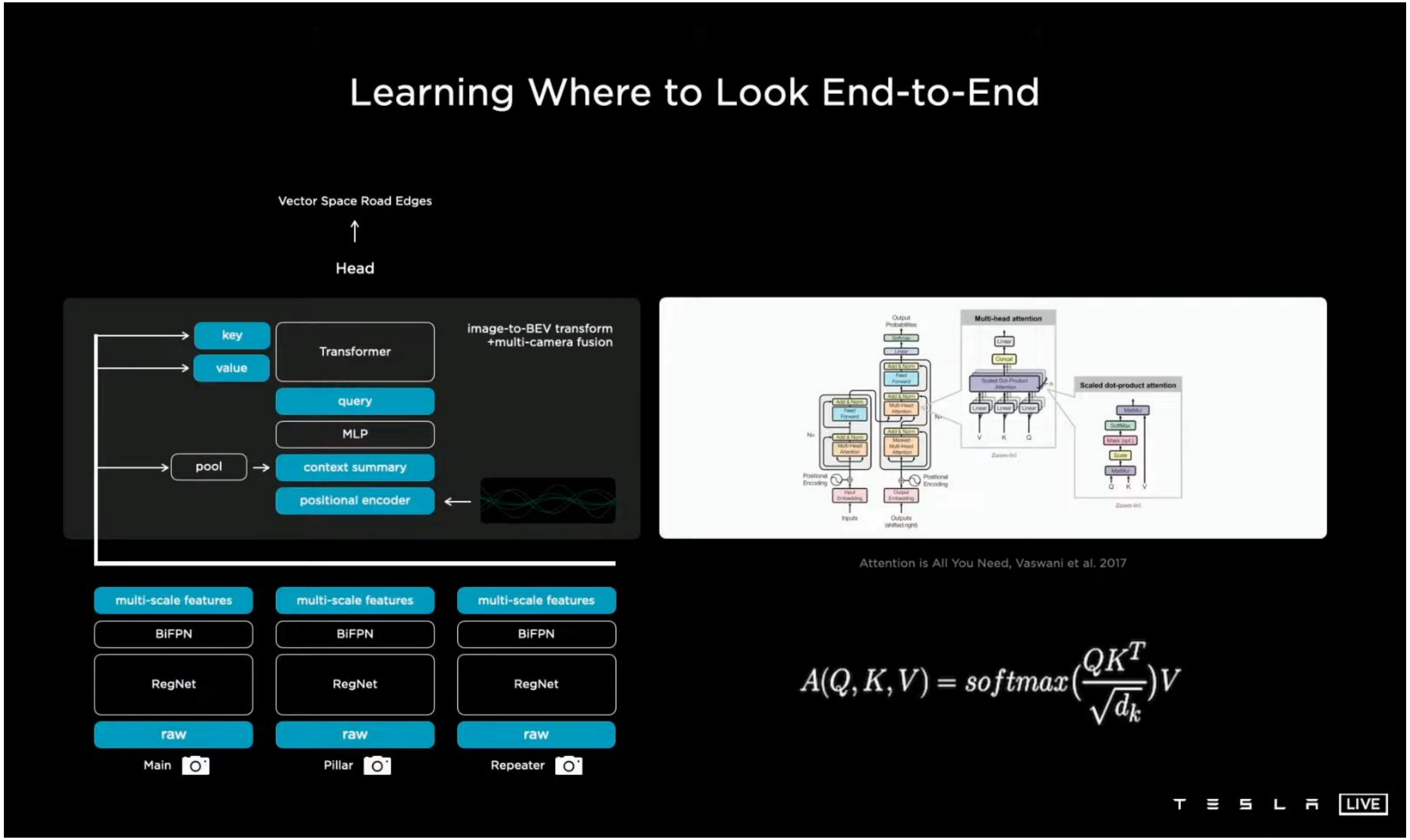
Transformers in Object Detection

- Domination of transformers
- Top-8 models use transformers for "Object Detection on COCO test-dev"

Rank	Model	box AP	AP50	AP75	APS	APM	APL	Extra Training Data	Paper	Code	Result	Year	Tags
1	**DyHead** (Swin-L, multi scale, self-training)	60.6	78.5	66.6	43.9	64.0	74.2	✓	Dynamic Head: Unifying Object Detection Heads with Attentions			2021	multiscale Swin-Transformer
2	**Dual-Swin-B** (HTC, multi-scale)	59.3						×	CBNetV2: A Composite Backbone Network Architecture for Object Detection			2021	multiscale
3	**Focal-L** (DyHead, multi-scale)	58.9						×	Focal Self-attention for Local-Global Interactions in Vision Transformers			2021	multiscale Focal-Transformer
4	**DyHead** (Swin-L, multi scale)	58.7	77.1	64.5	41.7	62.0	72.8	×	Dynamic Head: Unifying Object Detection Heads with Attentions			2021	
5	**Swin-L** (HTC++, multi scale)	58.7						×	Swin Transformer: Hierarchical Vision Transformer using Shifted Windows			2021	multiscale Swin-Transformer
6	**Dual-Swin-B** (HTC)	58.6						×	CBNetV2: A Composite Backbone Network Architecture for Object Detection			2021	single scale
7	**Focal-L** (HTC++, multi-scale)	58.4						×	Focal Self-attention for Local-Global Interactions in Vision Transformers			2021	
8	**Swin-L** (HTC++, single scale)	57.7						×	Swin Transformer: Hierarchical Vision Transformer			2021	single scale Swin-Transformer

Papers with Code, Screenshot taken Jul. 12th 2021

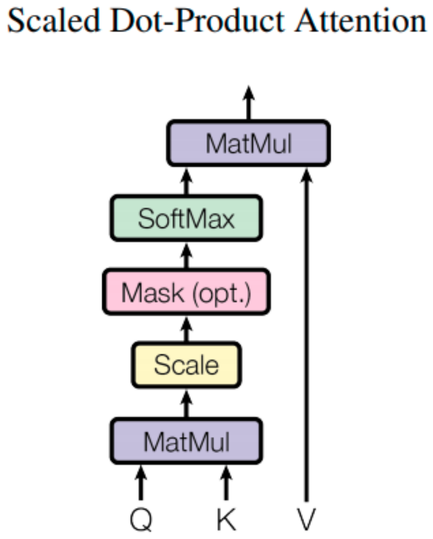
Tesla AI Day

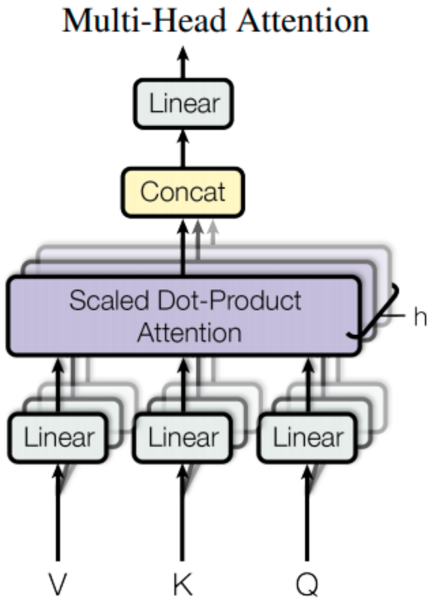


https://youtu.be/j0z4FweCy4M

A Quick Review on QKV Attention

- Query
- Key
- Value





Transformers

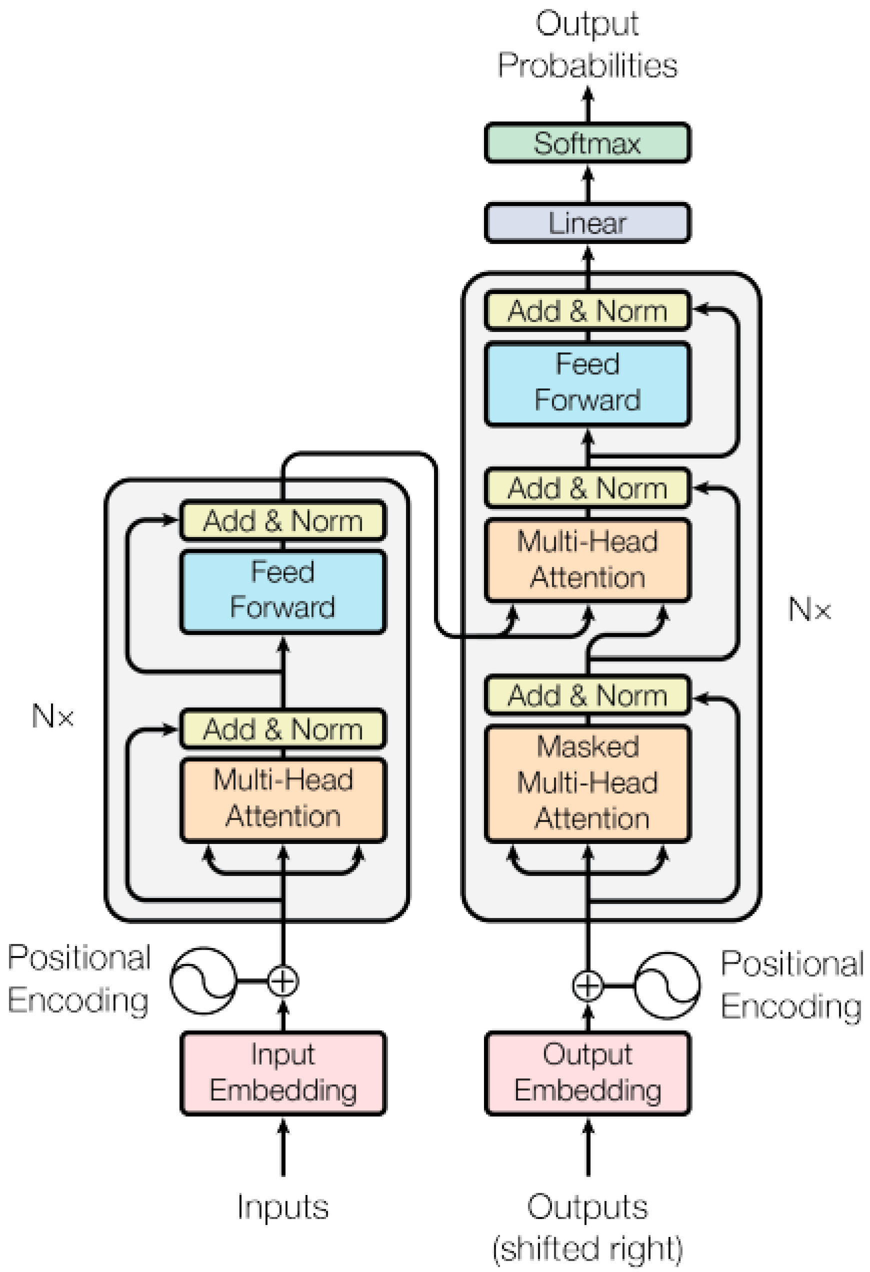
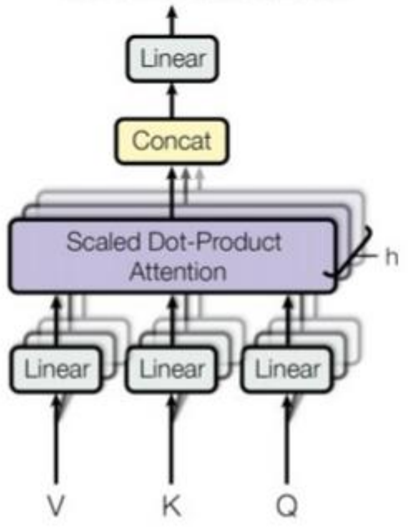
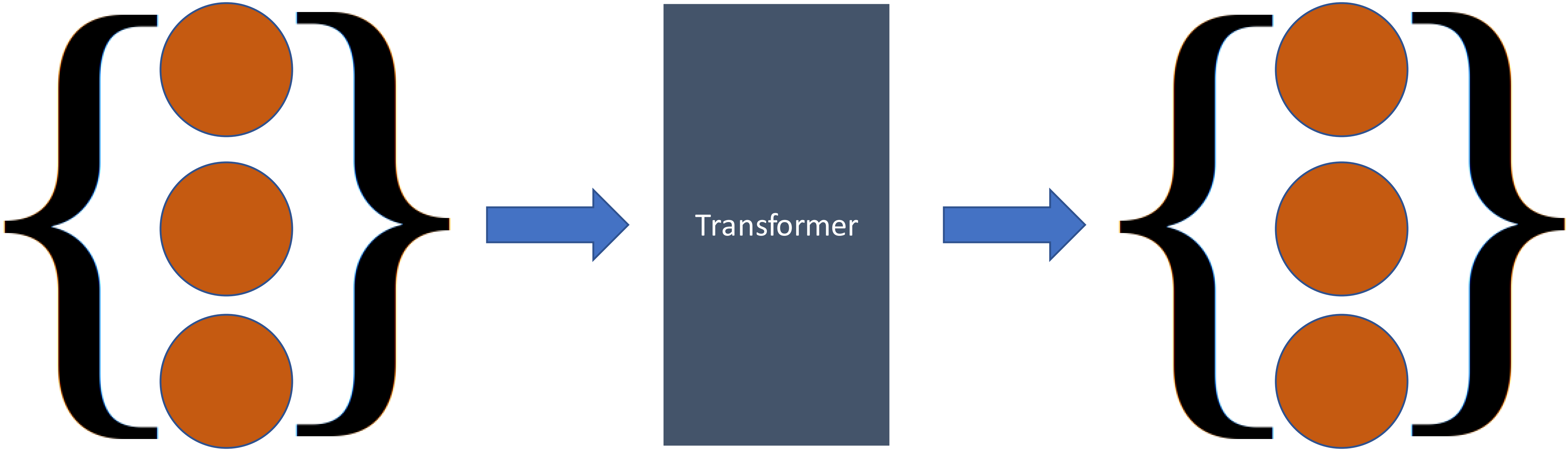


Figure 1: The Transformer - model architecture.

Multi-Head Attention

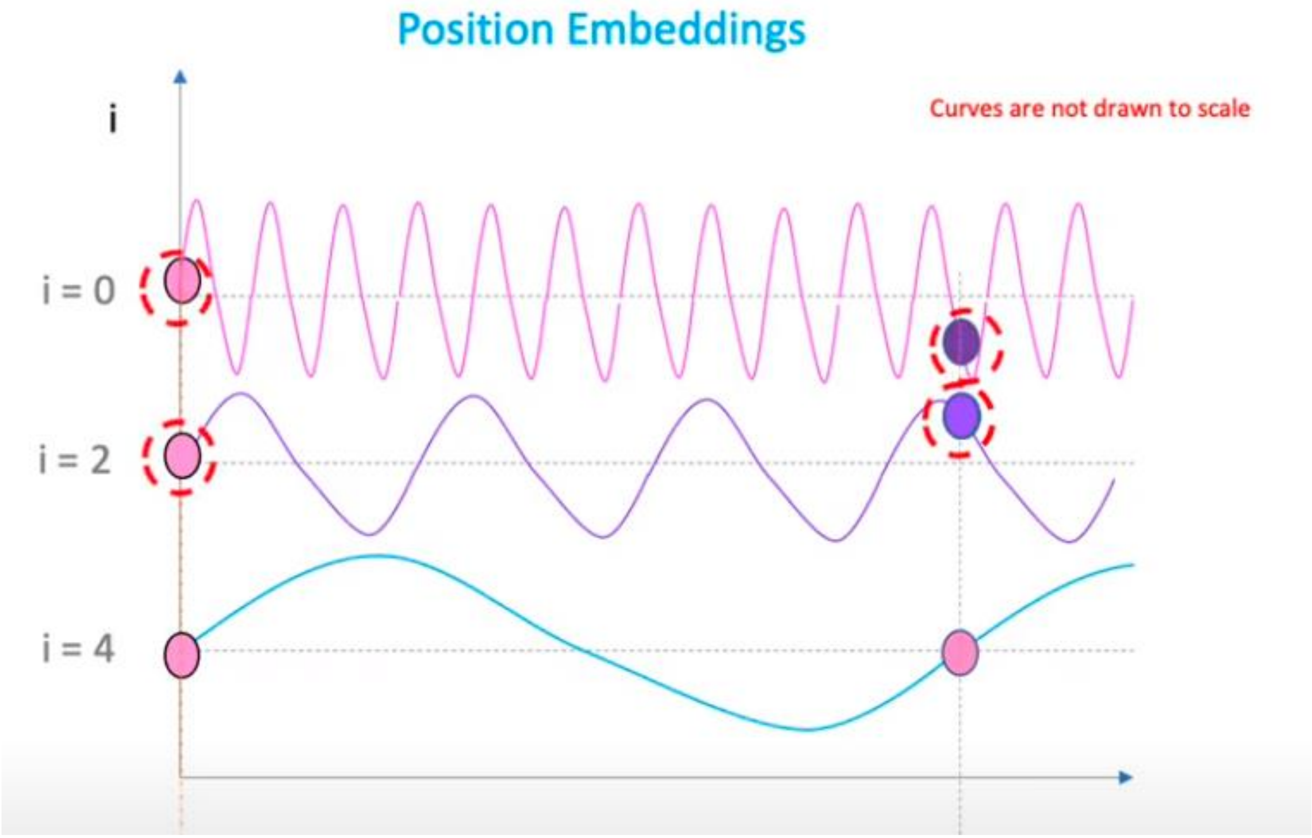


Transformers



Positional encoding

- Fourier positional enconding
- Rethinking Positional Encoding in Language Pre-training (Ke, He, Liu, 2020)



https://www.youtube.com/watch?v=dichIcUZfOw&ab_channel=Hedu-MathofIntelligence

Complexity and Path Length

The quick brown fox jumps over the lazy dog

The quick brown fox jumps over the lazy dog

Table 1: Maximum path lengths, per-layer complexity and minimum number of sequential operations for different layer types. n is the sequence length, d is the representation dimension, k is the kernel size of convolutions and r the size of the neighborhood in restricted self-attention.

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	O(n/k)
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

Attention Is All You Need (2017)

Transformers
in Computer Vision

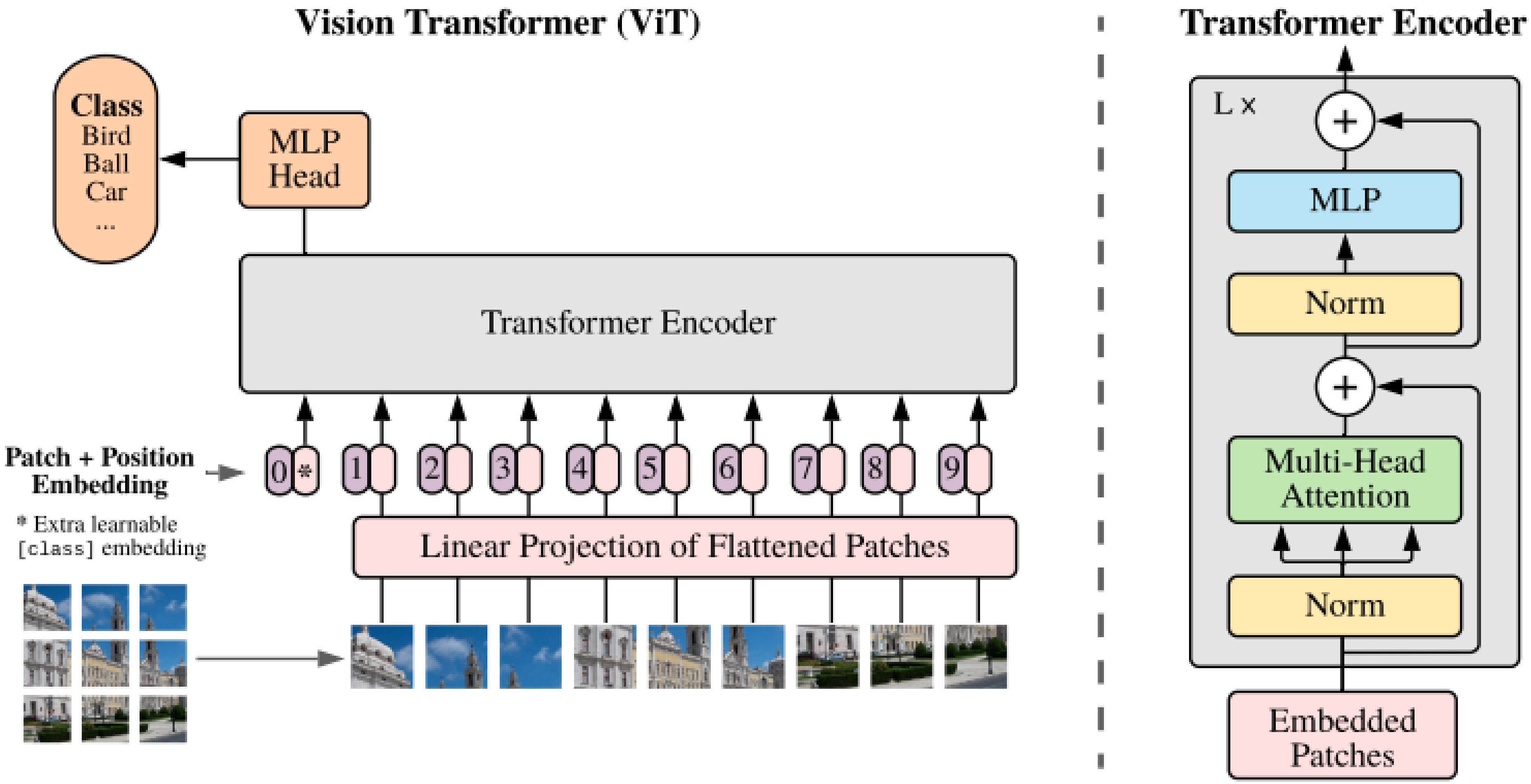
Transformers in CV

- Patch-based
  - **ViT** (classification)
  - SWIN Transformer (classification, detection, panoptic)
- Query-based
  - **DETR** (classification, detection, panoptic)
  - Deformable DETR (classification, detection, panoptic)
- Perceiver

Patch-based

Avoid the Quadratic

Vision Transformer (ViT)

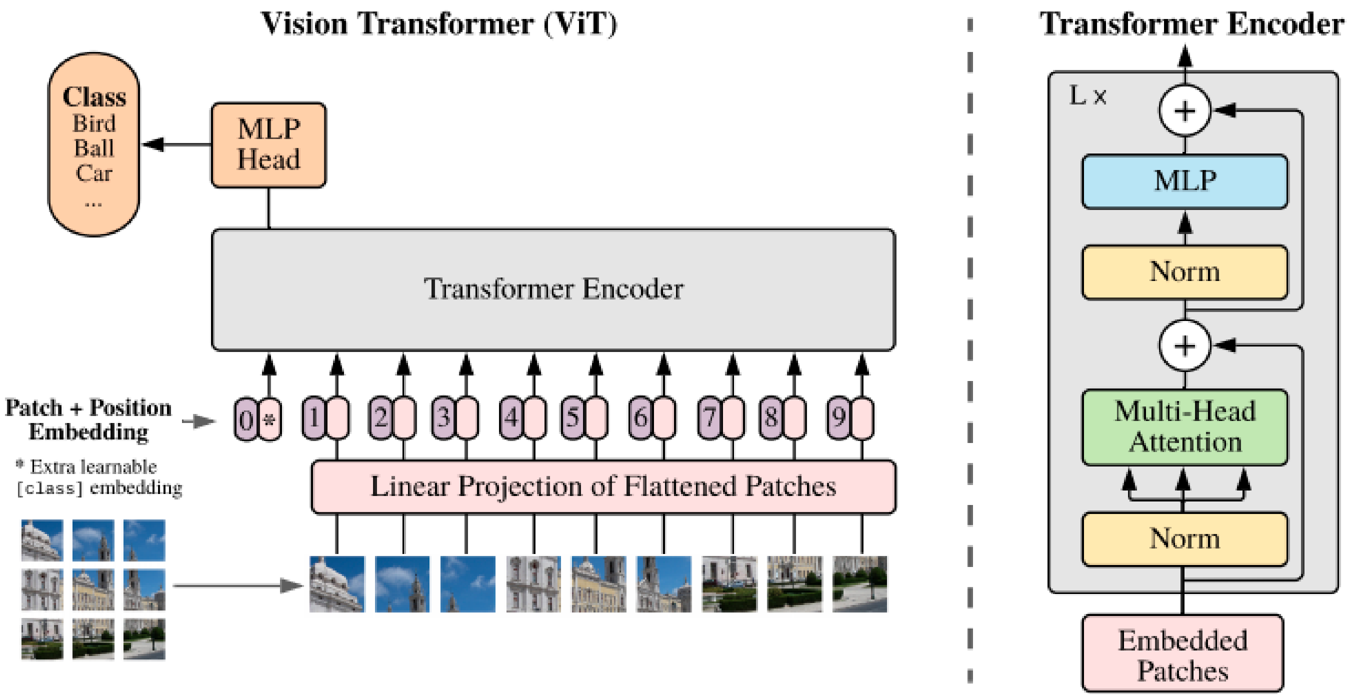


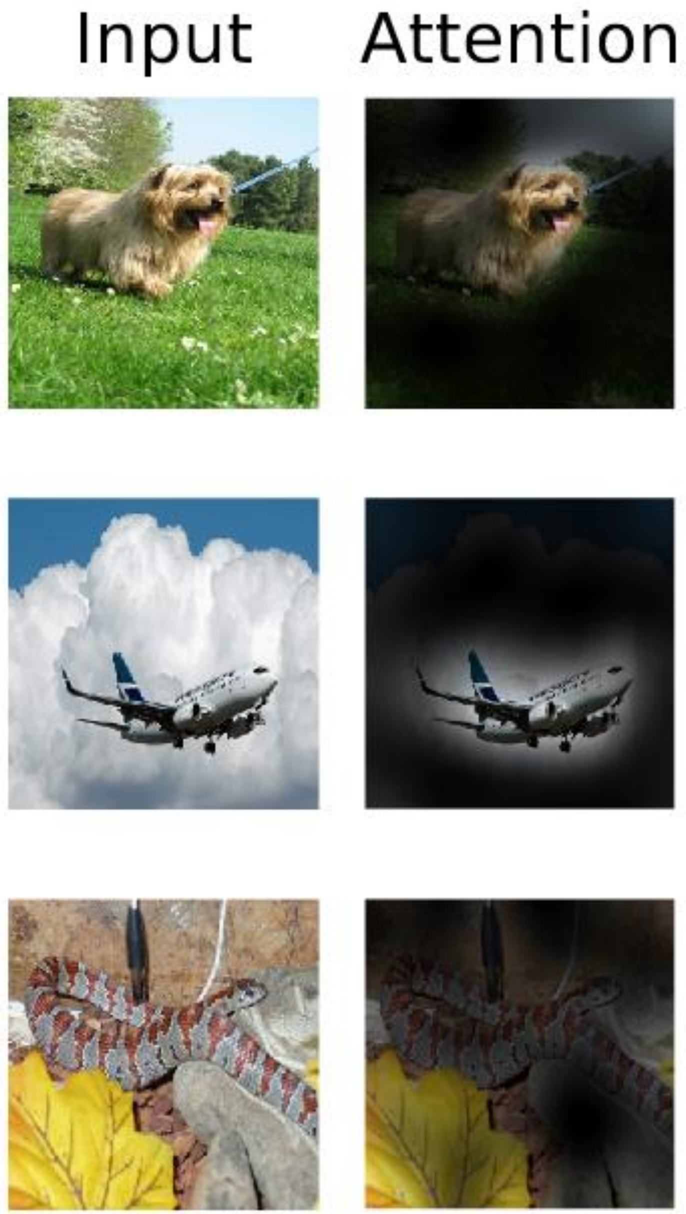
https://ai.googleblog.com/2020/12/transformers-for-image-recognition-at.html

An image is worth 16x16 words : Transformers for image recognition at scale, 2021

Attention

- Attention from CLS to input image





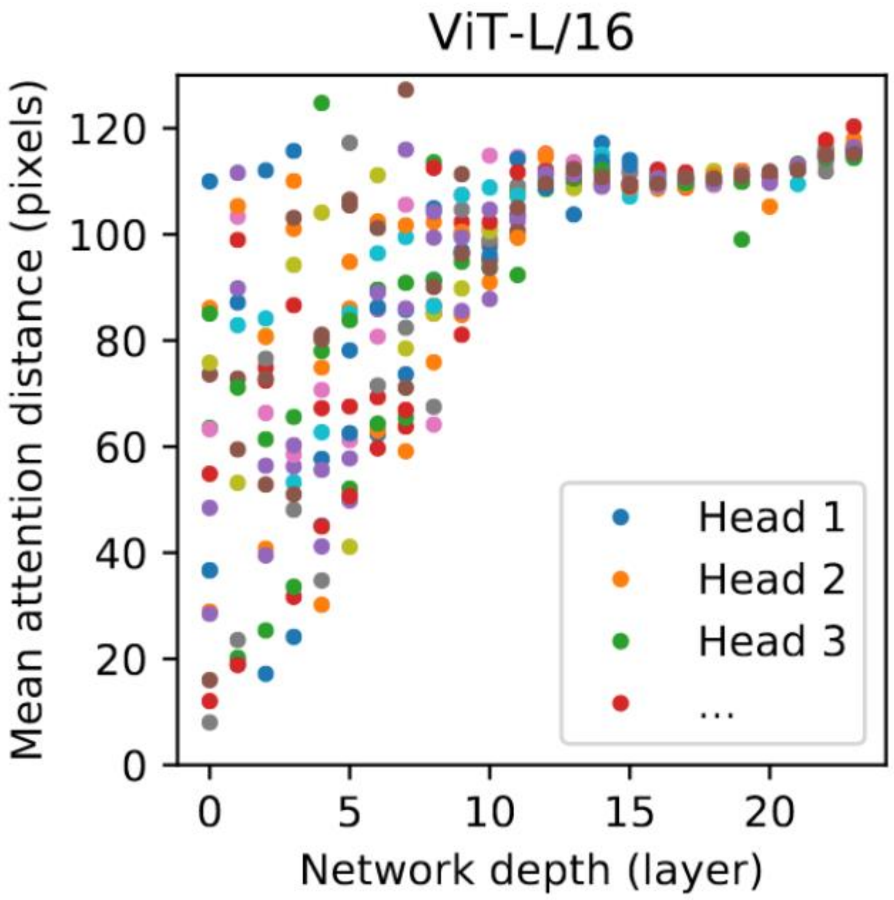
Results

- TPUv3-core-days
- 14x14 patches vs 16x16 (tradeoff compute-precision)

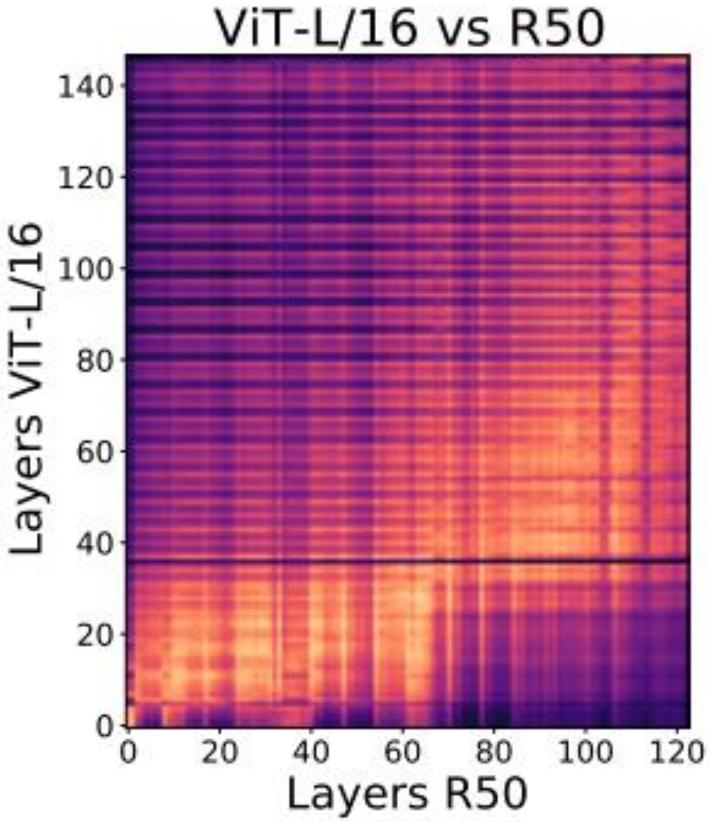
	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

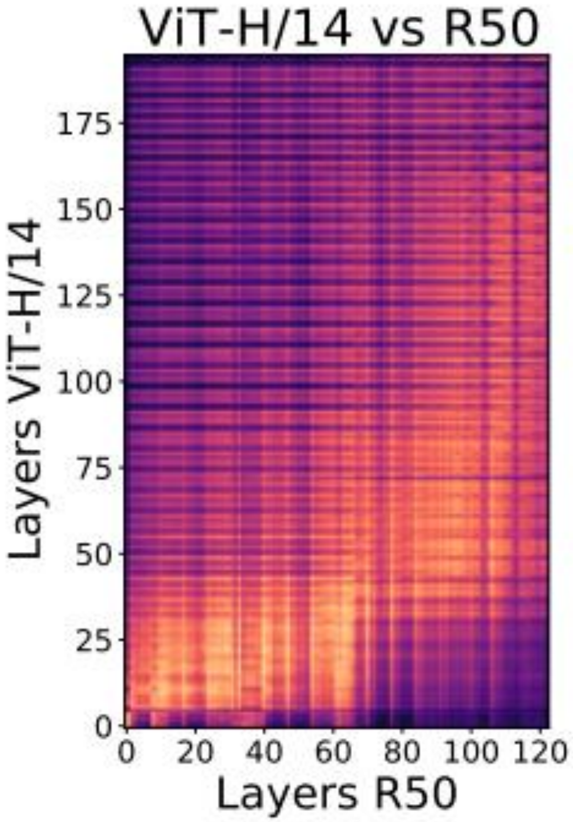
Attention Distance

- Mean attention distance ~ receptive field
- More flexible than CNN



Transformer vs CNN

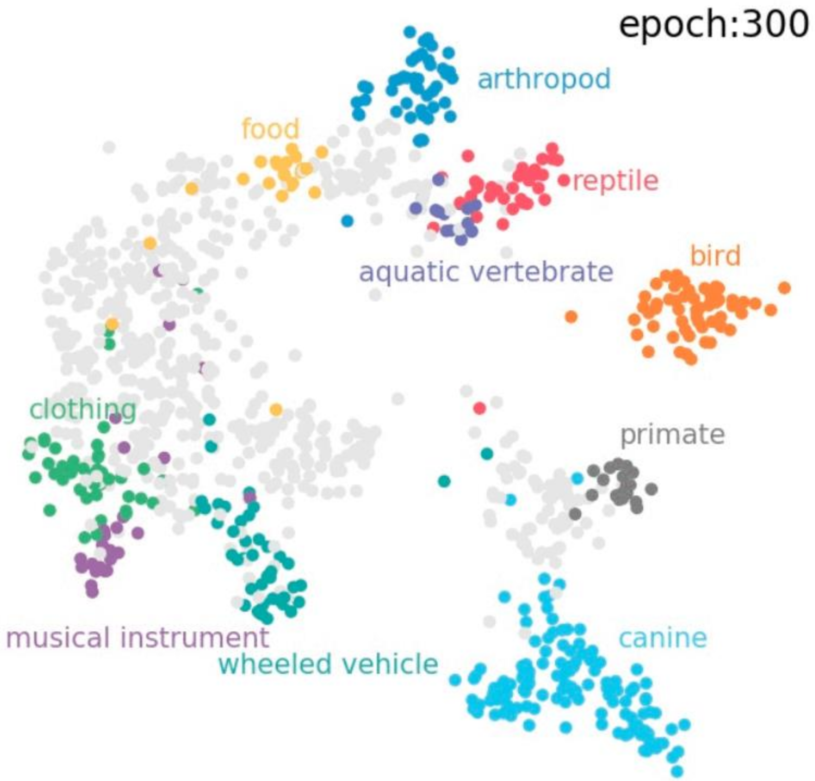


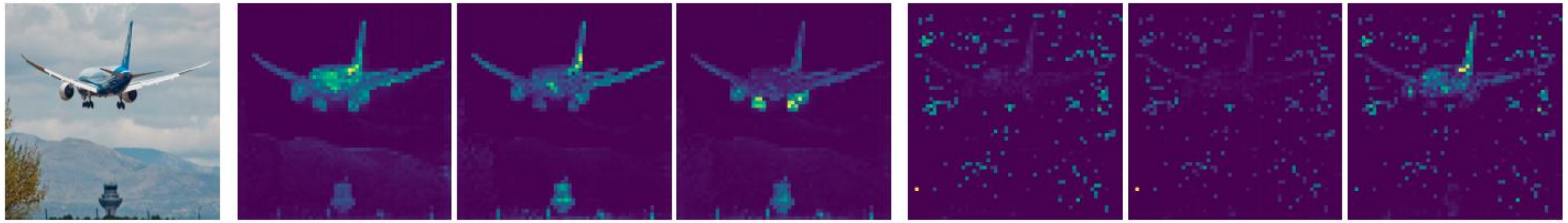


Do Vision Transformers See Like Convolutional Neural Networks? (2021)

DINO

- **Self-supervised learning**
- KNN classification
- Attention maps
  - Attention maps are better in SSL
  - Supervised stops learning when good on task

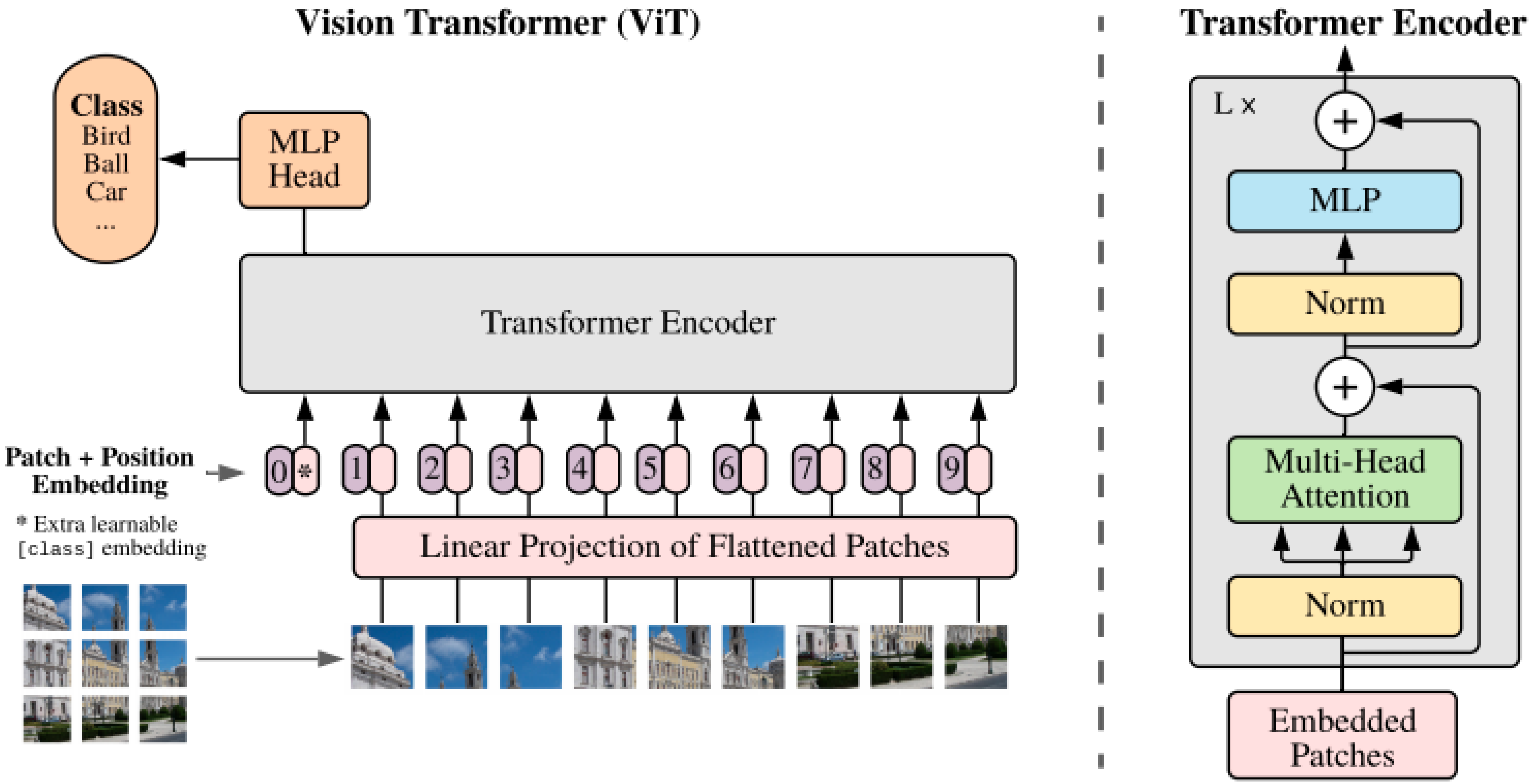




[Emerging Properties in Self-Supervised Vision Transformers](https://ai.facebook.com/blog/dino-paws-computer-vision-with-self-supervised-transformers-and-10x-more-efficient-training)

https://ai.facebook.com/blog/dino-paws-computer-vision-with-self-supervised-transformers-and-10x-more-efficient-training

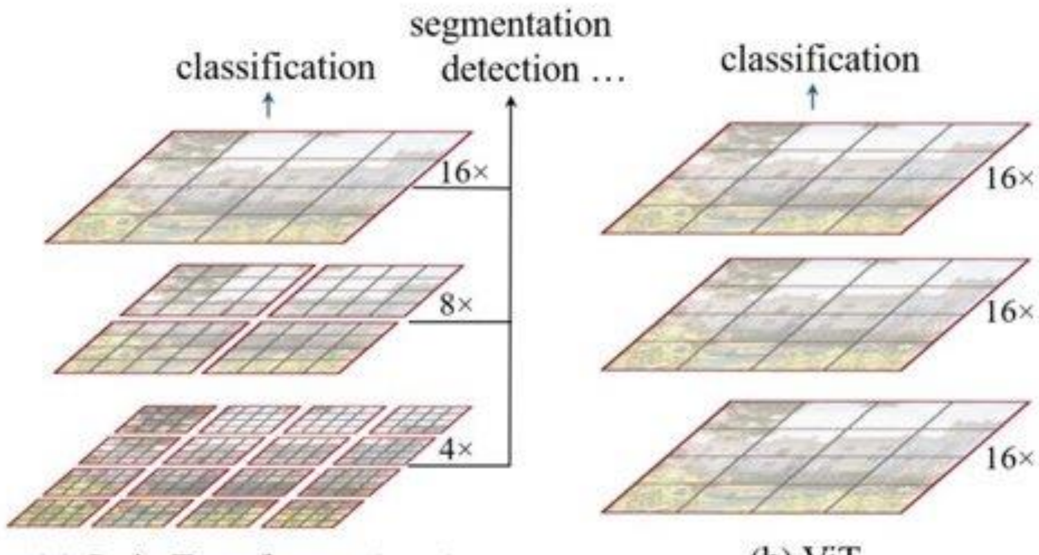
Vision Transformer (ViT)



https://ai.googleblog.com/2020/12/transformers-for-image-recognition-at.html

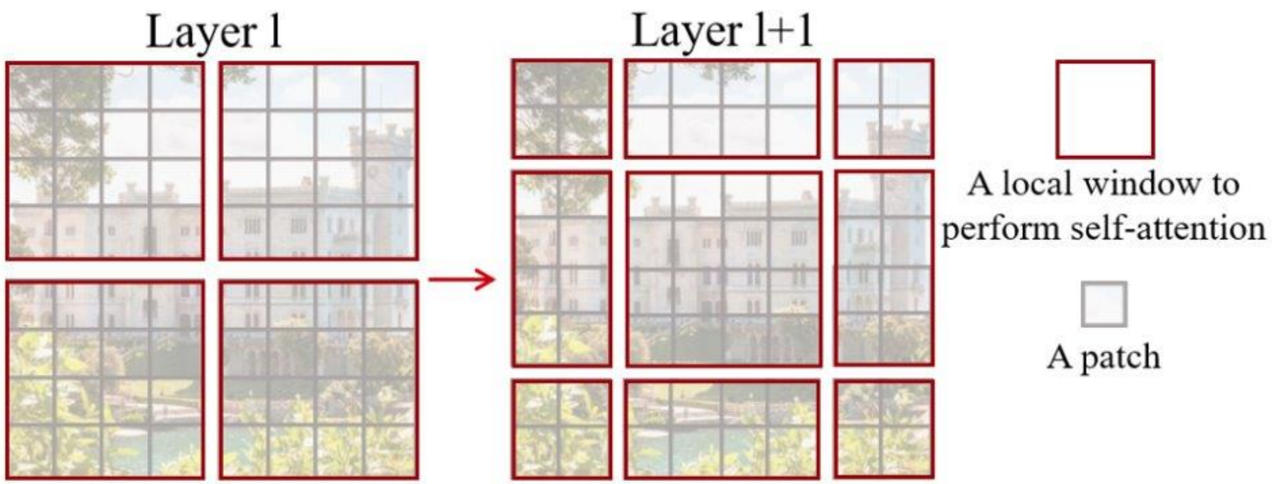
An image is worth 16x16 words : Transformers for image recognition at scale, 2021

SWIN Transformer – A New Backbone



(a) Swin Transformer (ours)

(b) ViT

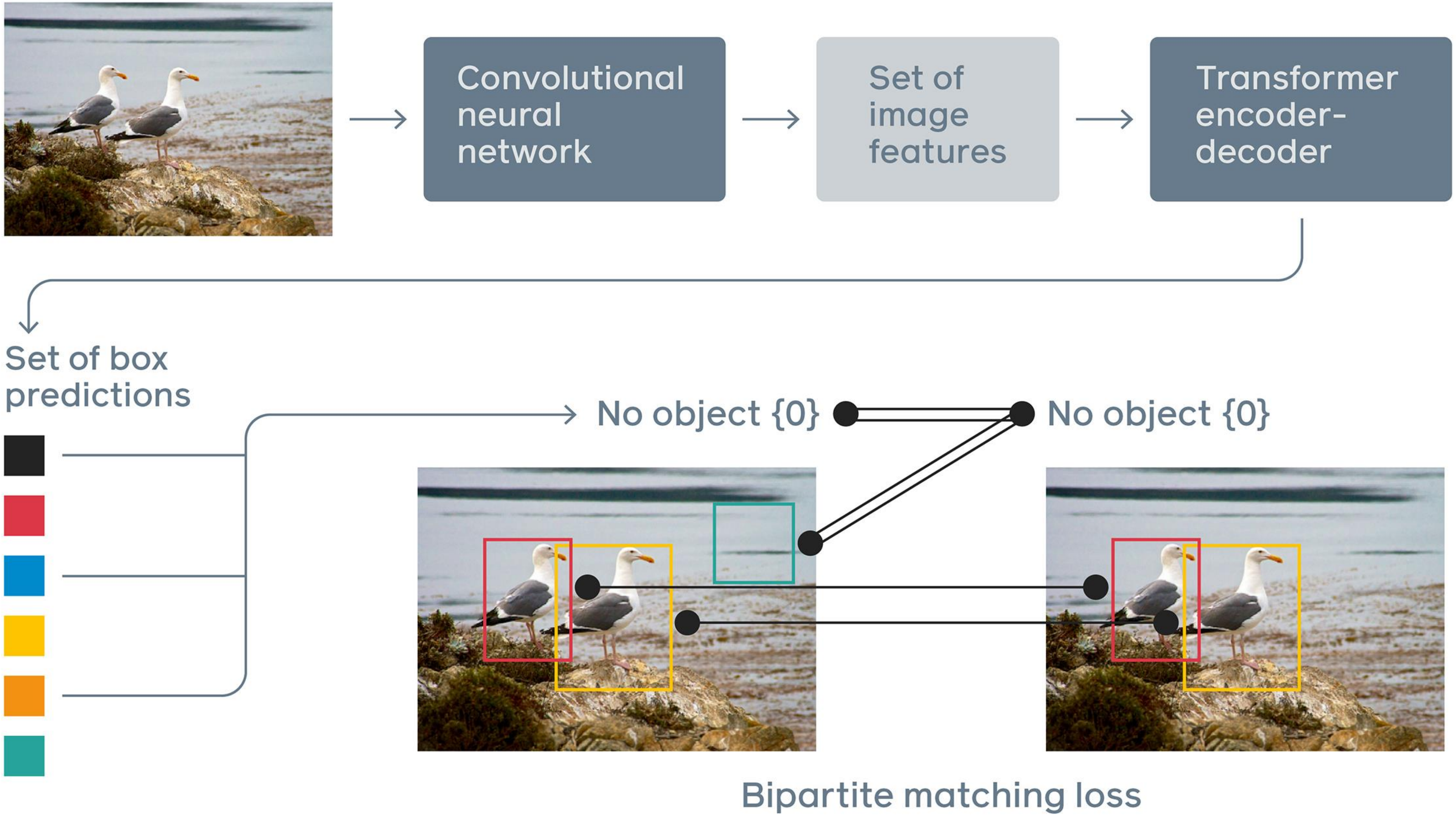


[Swin Transformer: Hierarchical Vision Transformer using Shifted Windows]()

Query-based

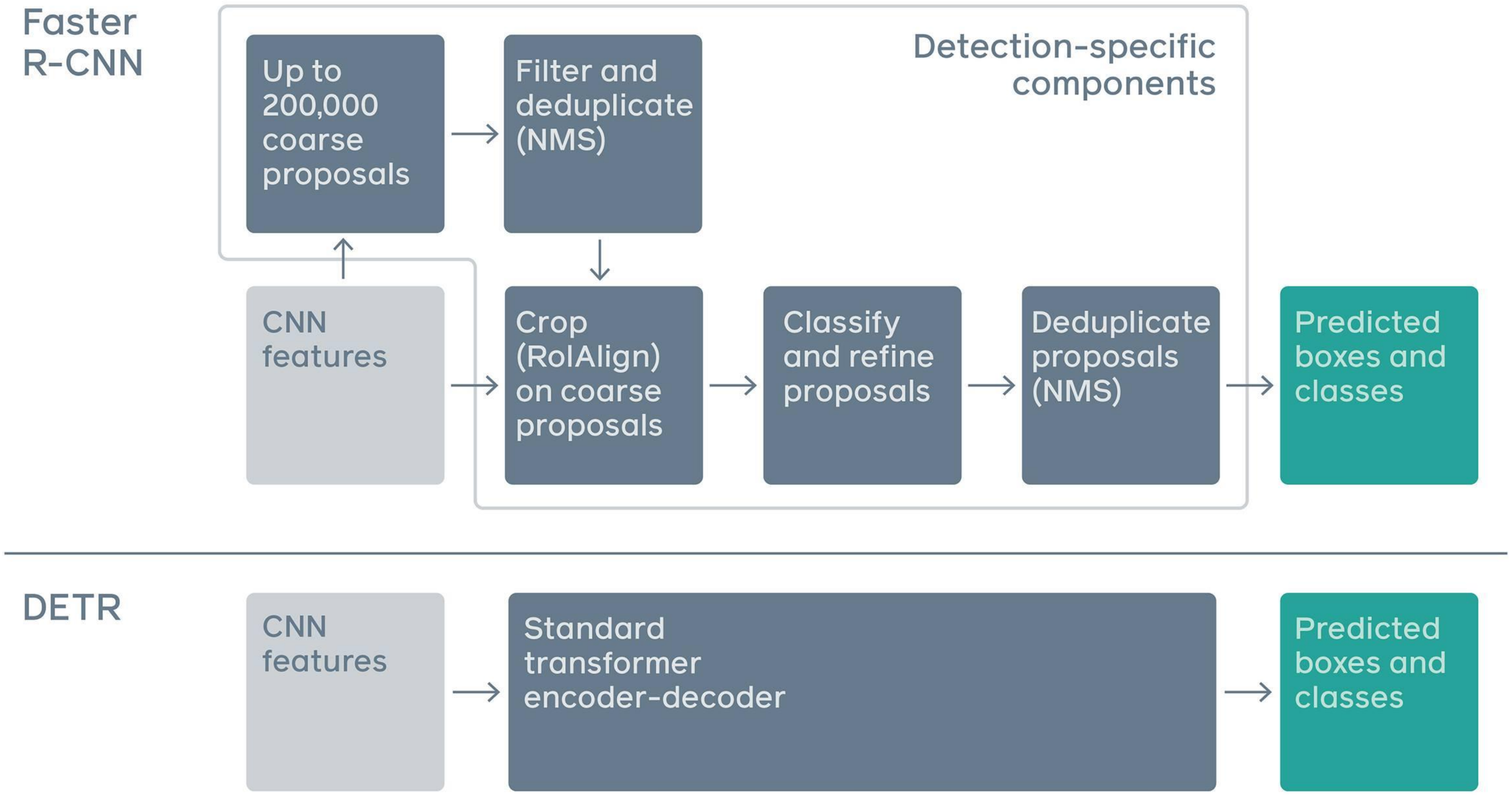
Asking the real questions

DETR — **DE**tection **TR**ansformer



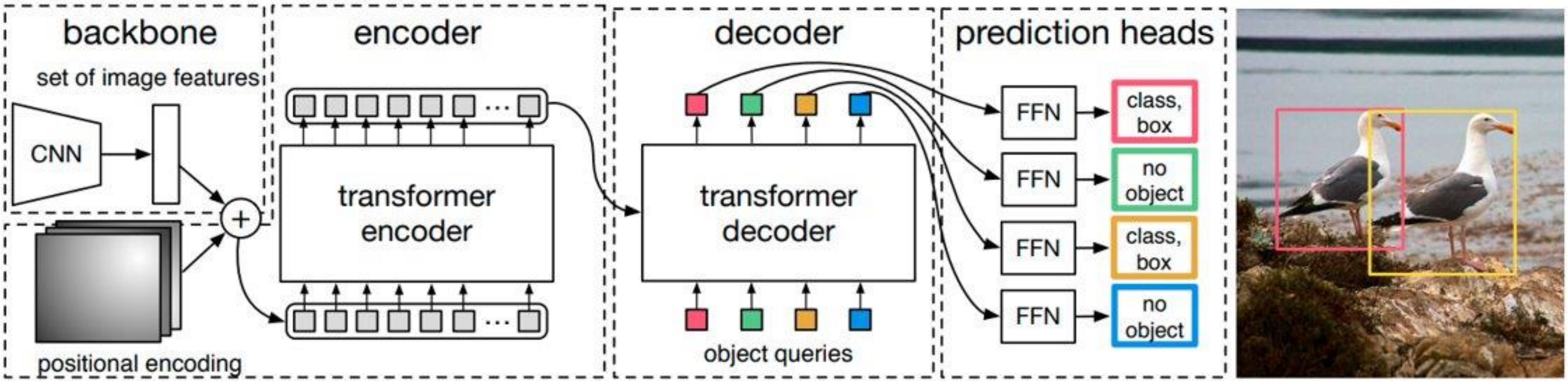
https://ai.facebook.com/blog/end-to-end-object-detection-with-transformers/

DETR vs Faster R-CNN

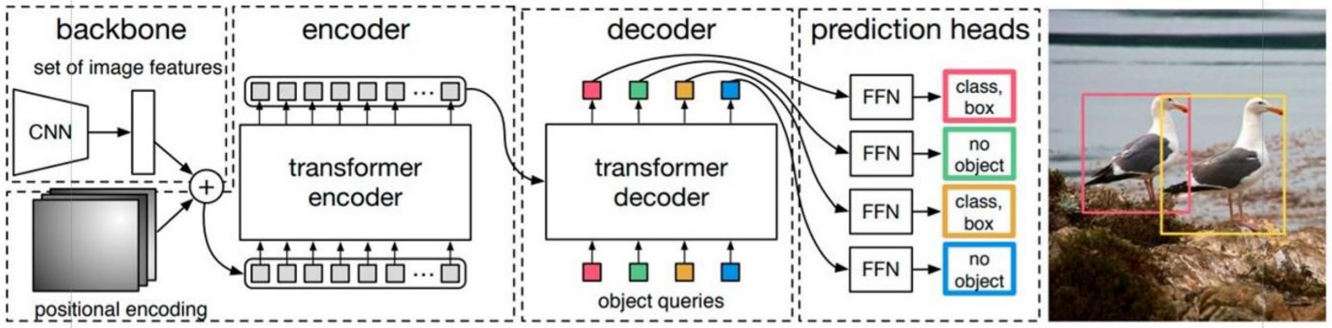


https://ai.facebook.com/blog/end-to-end-object-detection-with-transformers/

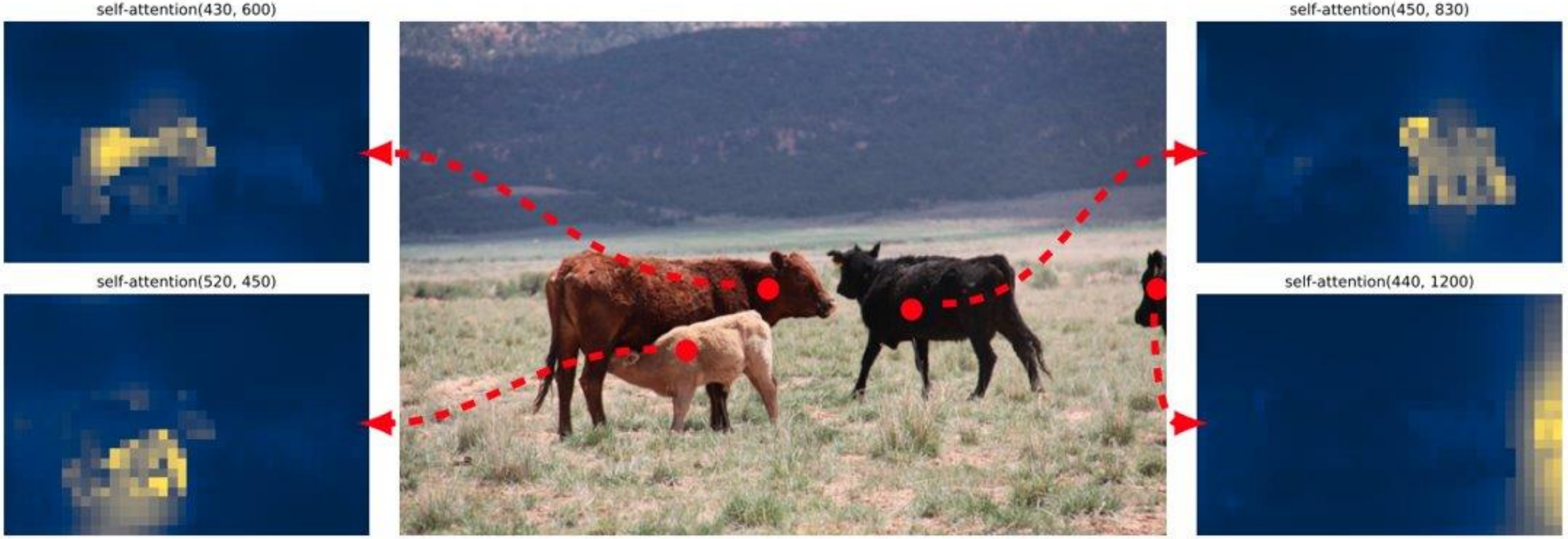
DETR Architecture



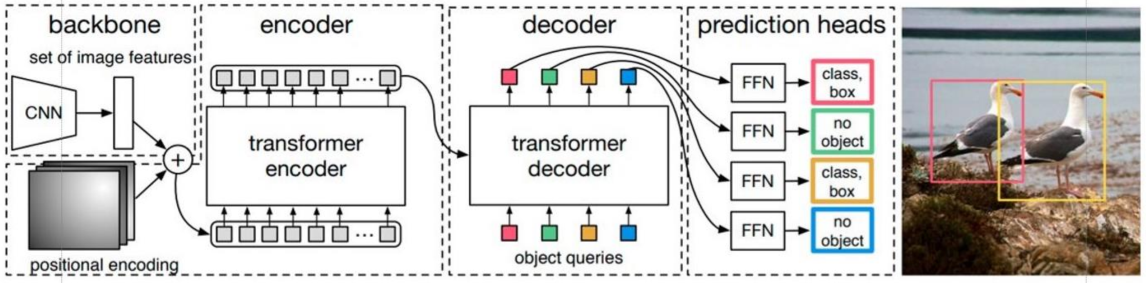
Encoder attention

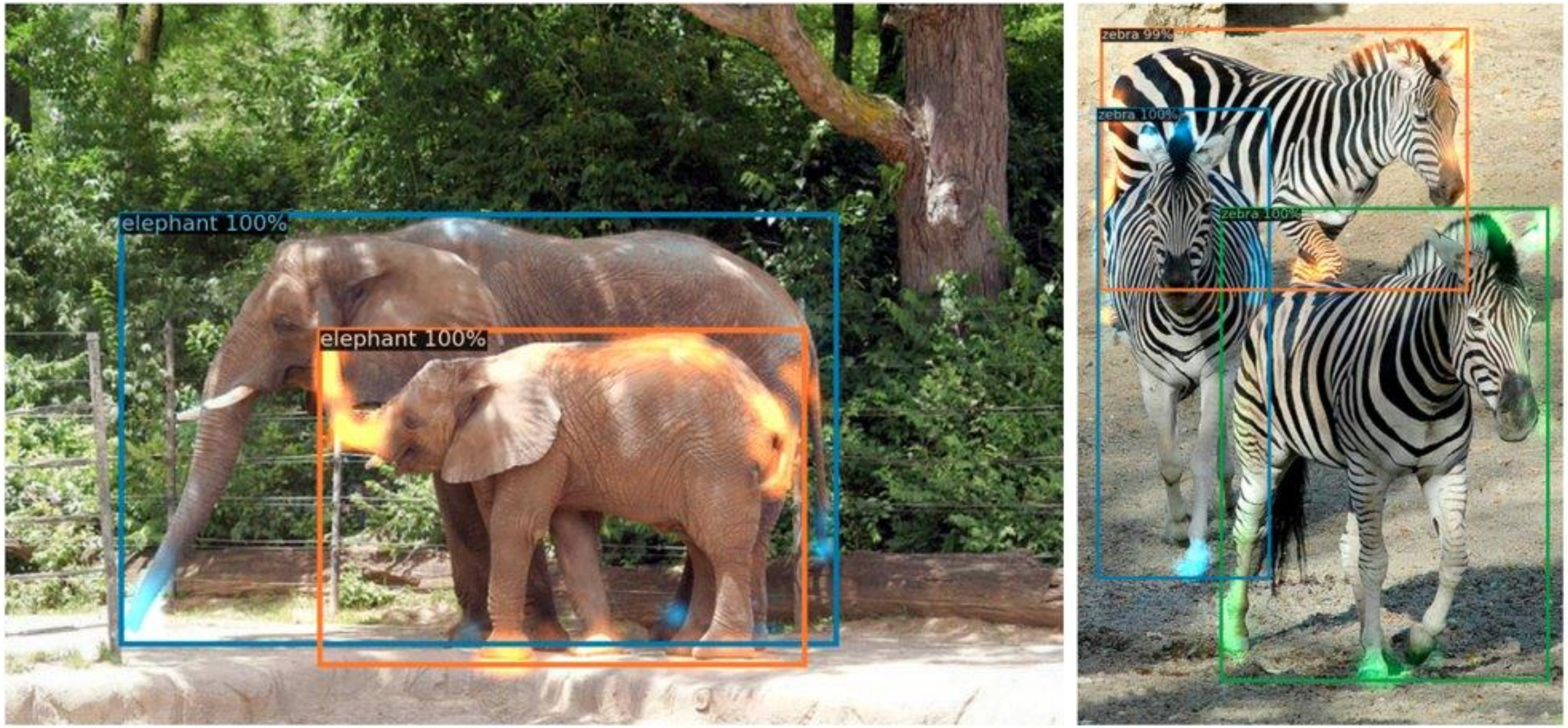


- Attention map of the last encoder layer
- Trained on bounding boxes



Decoder attention scores





Results

Model	GFLOPS/FPS	#params	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RetinaNet+ [1]	205/18	38M	41.1	60.4	43.7	25.6	44.8	53.6
Faster RCNN-FPN+	180/26	42M	42.0	62.1	45.5	26.6	45.4	53.4
DETR	86/28	41M	42.0	62.4	44.2	20.5	45.8	61.1
DETR-DC5	187/12	41M	43.3	63.1	45.9	22.5	47.3	61.1

Perceiver

Another way to see

Perceiver

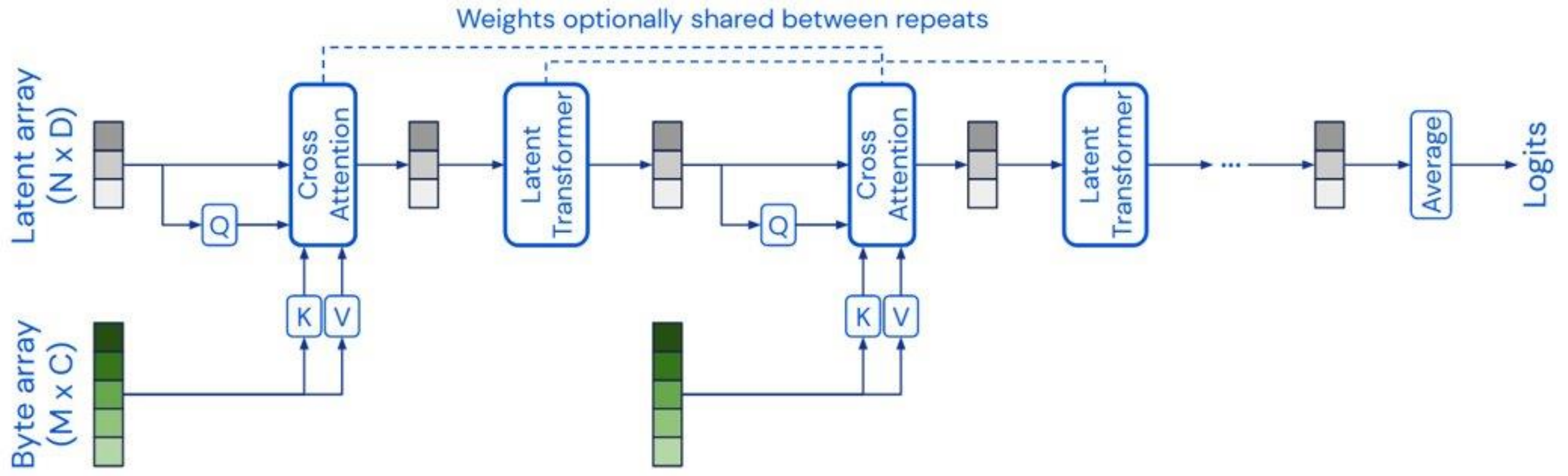
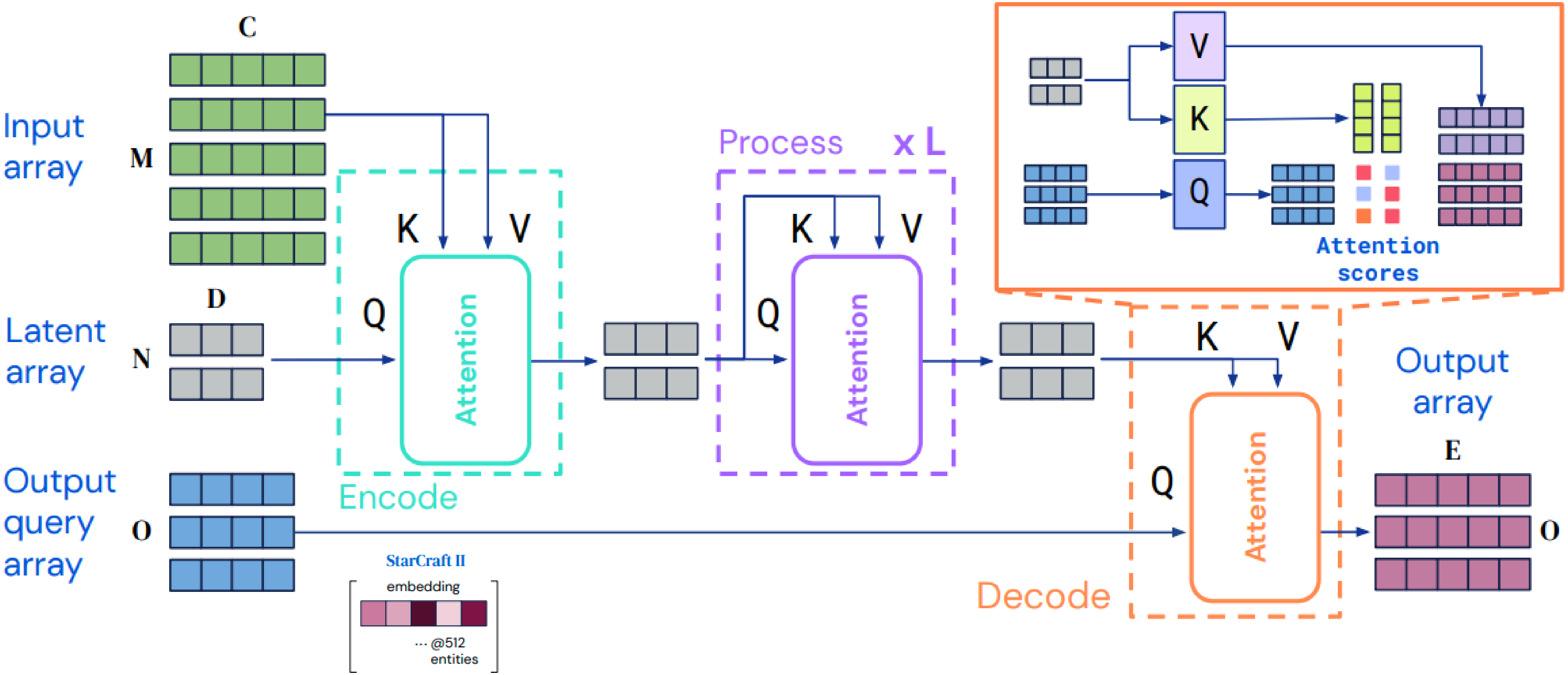


Figure 1. The Perceiver is an architecture based on attentional principles that scales to high-dimensional inputs such as images, videos, audio, point-clouds, and multimodal combinations without making domain-specific assumptions. The Perceiver uses a cross-attention module to project an high-dimensional input byte array to a fixed-dimensional latent bottleneck (the number of input indices M is much larger than the number of latent indices N) before processing it using a deep stack of Transformer-style self-attention blocks in the latent space. The Perceiver iteratively attends to the input byte array by alternating cross-attention and latent self-attention blocks.

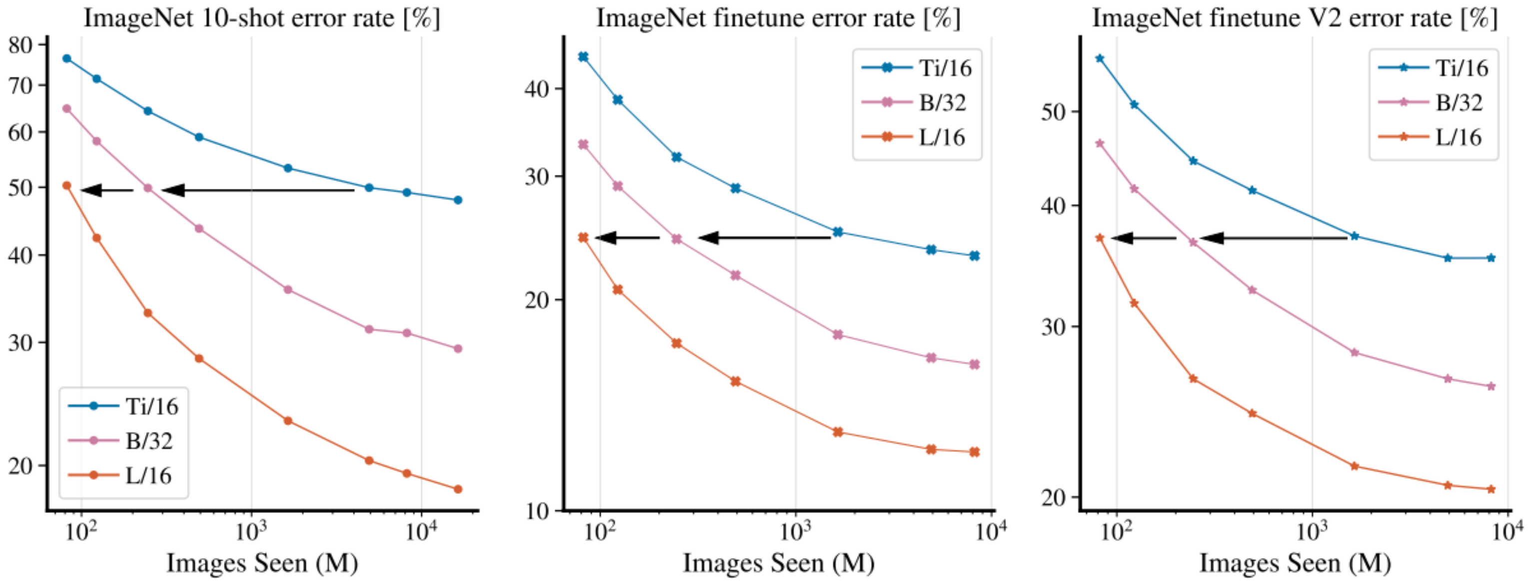
PerceiverIO



[Perceiver IO: A General Architecture for Structured Inputs & Outputs](), (2021)

Conclusion

- Transformers revolutioned NLP
  - The revolution started in CV
- Weaker inductive biases than CNN
  - Possibly better with enough data
- Scale very well



Scaling Vision Transformers

Useful Links

- ViT @ Google AI Blog
- SWIN @ arXiv
- DETR @ Facebook AI
- DINO @ Facebook AI
- Perceiver @ arXix
- PerceiverIO @ arXiv