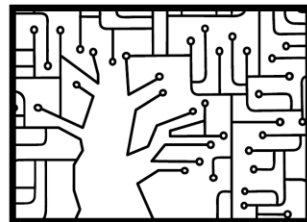


Deep Learning for Computer Vision: From Sequences to ViT

Shai Bagon

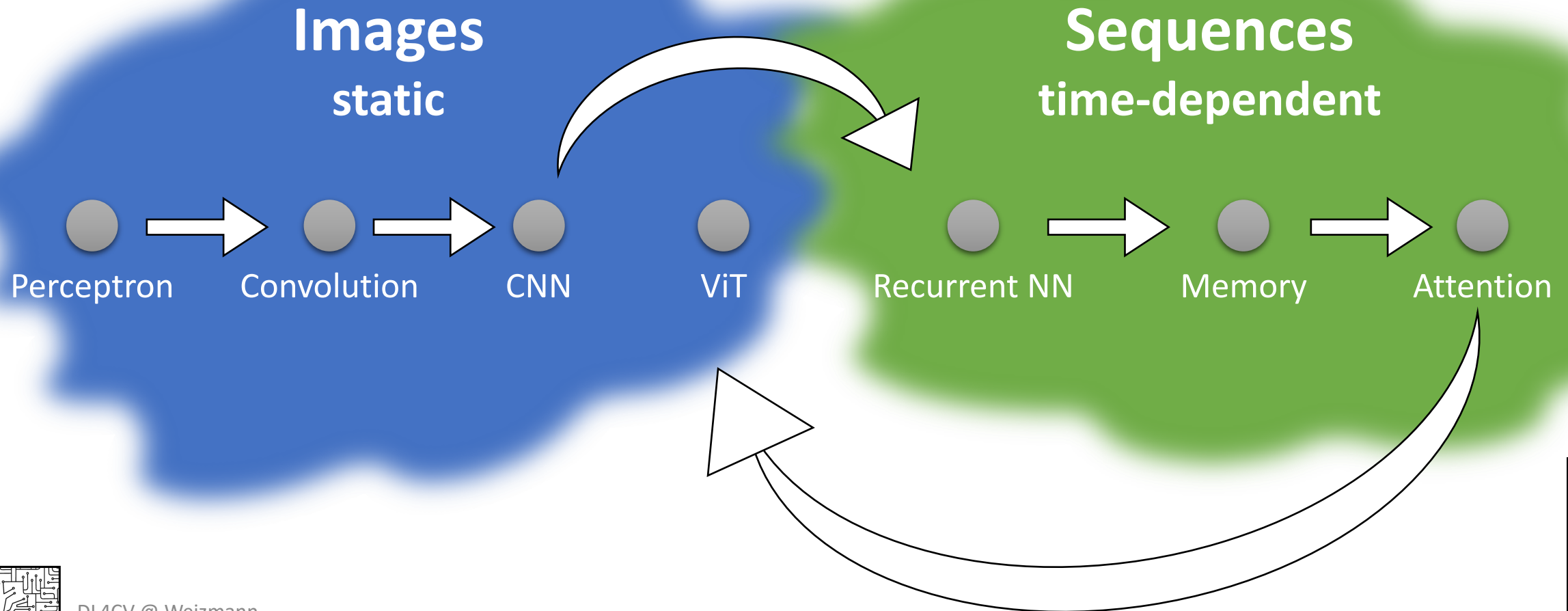


WAIC

100
DAYS IN HELL

Bring Them Home Now!

Agenda



Self-Attention Layer

Input:

Input: X (shape: $N_x \times D_x$)

Layer's Parameters:

$X \rightarrow Q$: W_q (shape: $D_x \times D_q$)

$X \rightarrow K$: W_k (shape: $D_x \times D_q$)

$X \rightarrow V$: W_v (shape: $D_x \times D_v$)

Compute:

Query: $Q = XW_q$ (shape: $N_x \times D_q$)

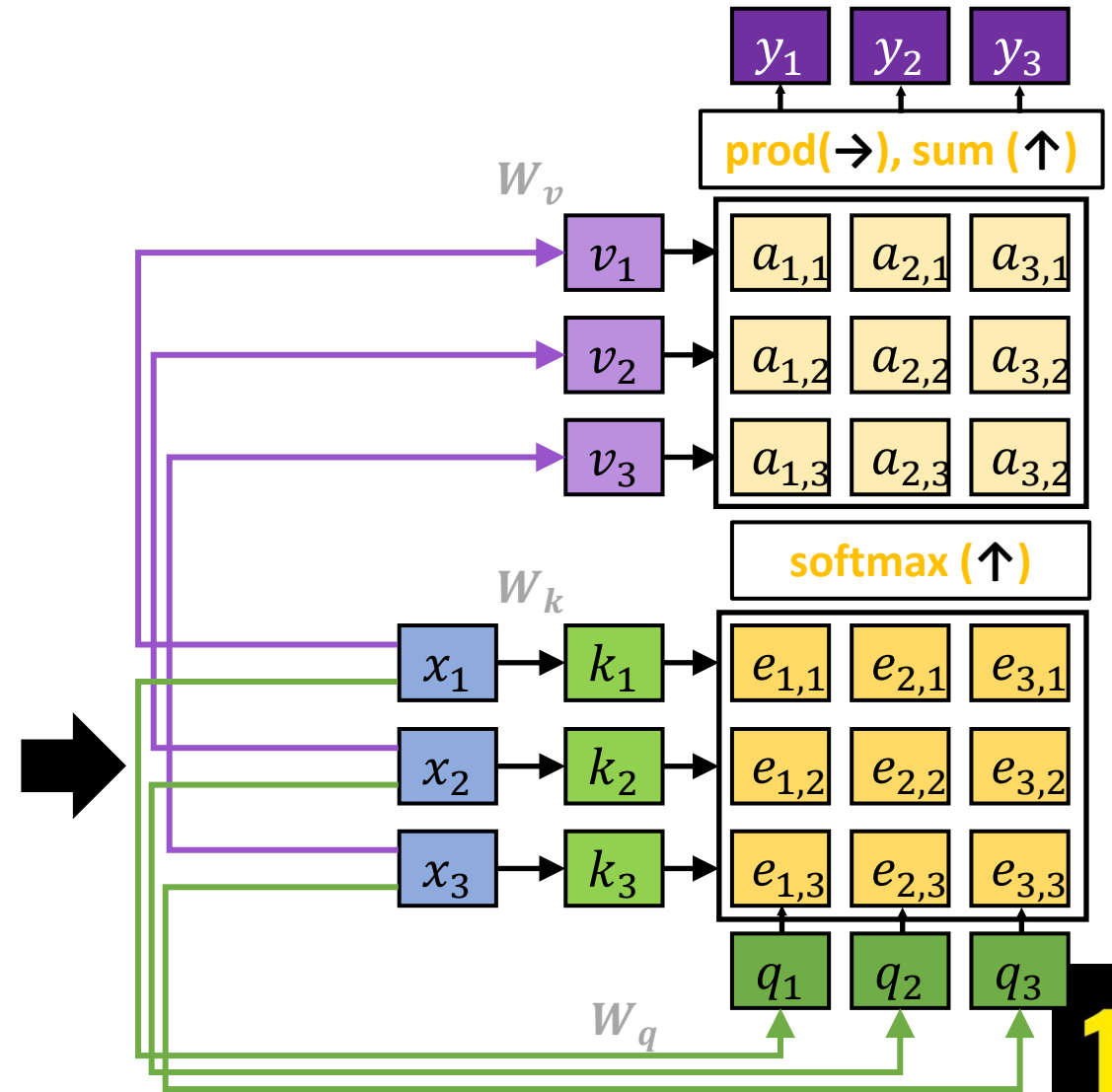
Keys: $K = XW_k$ (shape: $N_x \times D_q$)

Values: $V = XW_v$ (shape: $N_x \times D_v$)

Similarities: $E = QK^T / \sqrt{D_q}$ (shape: $N_x \times N_x$)

Attention: $A = \text{softmax}(E; \uparrow)$ (shape: $N_x \times N_x$)

Outputs: $Y = AV$ (shape: $N_x \times D_v$)



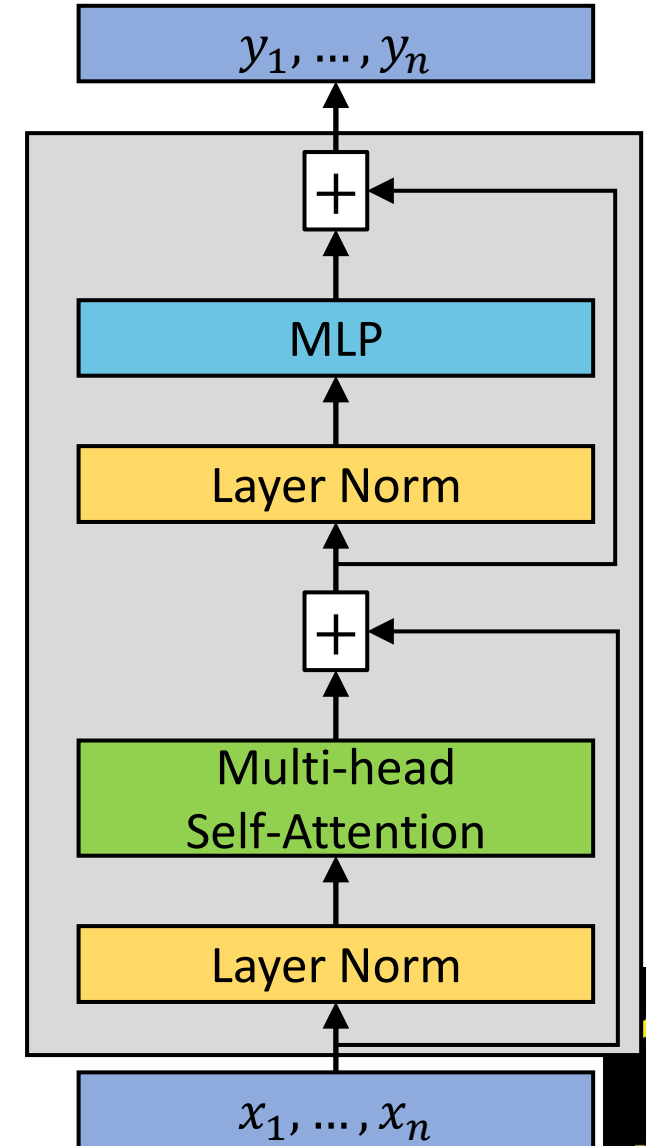
Transformer Layer

Input: x_1, \dots, x_n (n tokens in D dimensions)

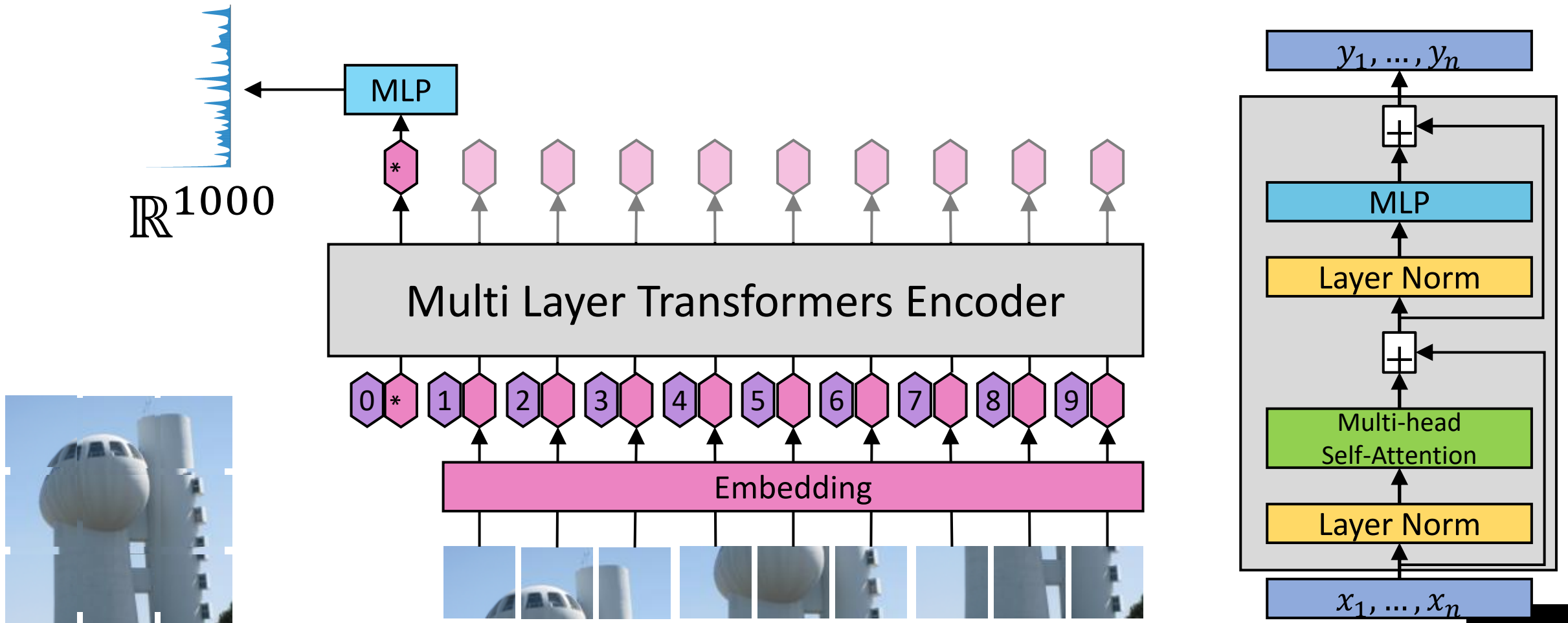
Output: y_1, \dots, y_n (n tokens in D dimensions)

Highly scalable

Highly parallelizable




Vision Transformers (ViT)





Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J. and Houlsby N. [“An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”](#) (ICLR 2021)

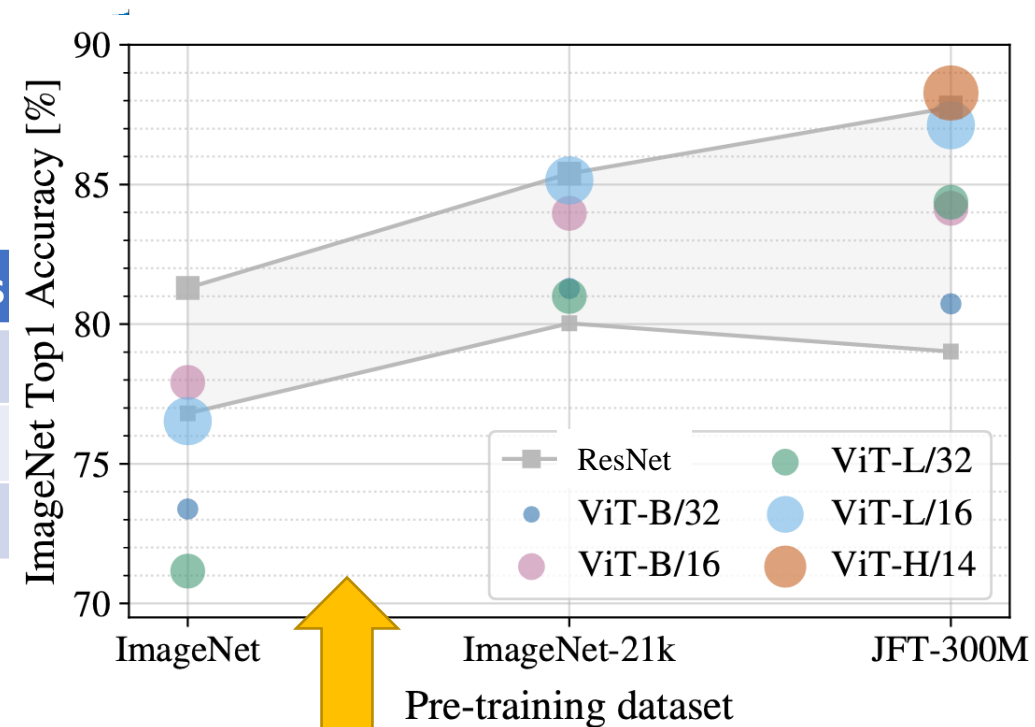
Vision Transformers (ViT)



Model	Layers	Width (D)	#Heads	#Params
ViT-Base	12	768	12	86M
ViT-Large	24	1024	16	307M
ViT-Huge	32	1280	16	632M

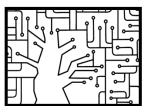



Naming convention: ViT-B/16



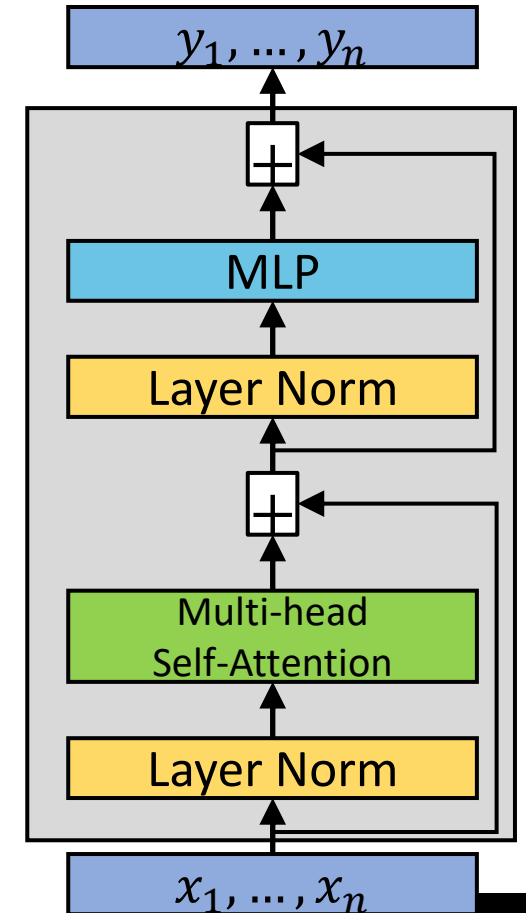
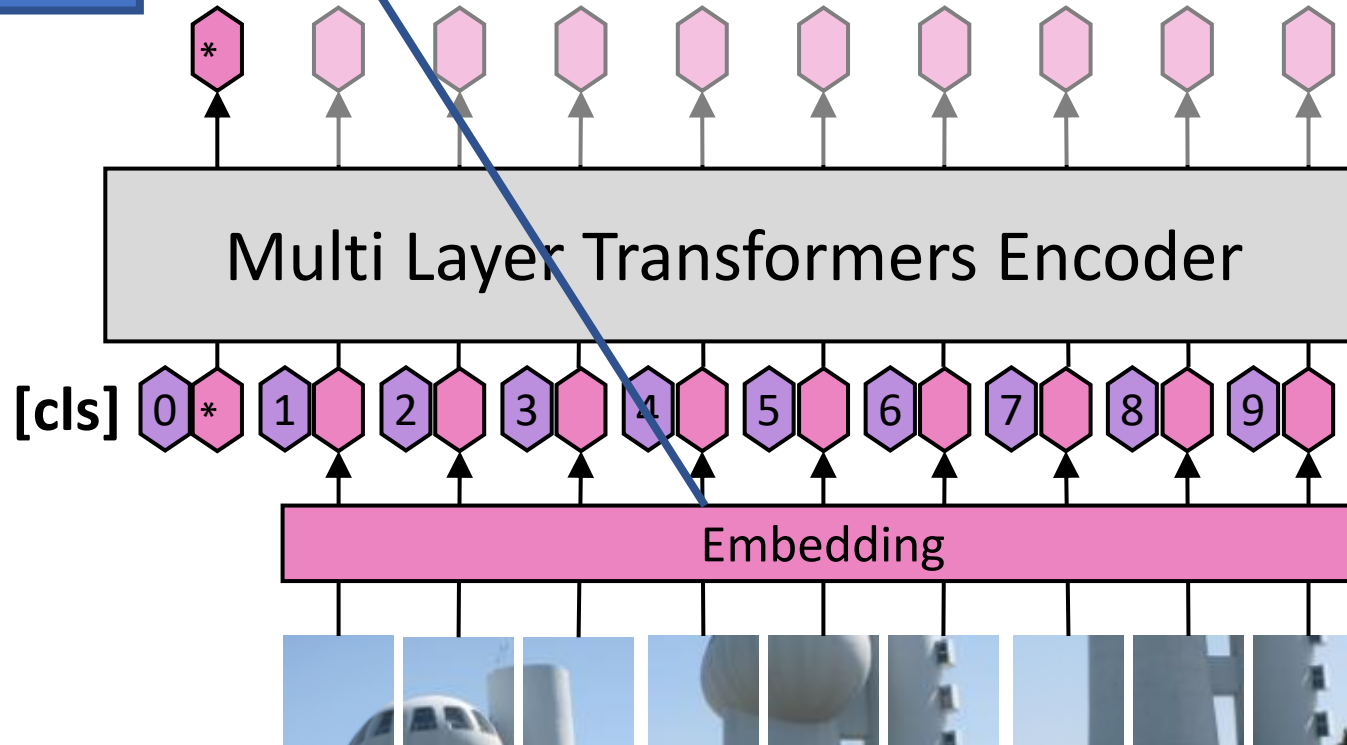
Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J. and Houlsby N. [“An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”](#) (ICLR 2021)

Vision Transformers (ViT) - Properties



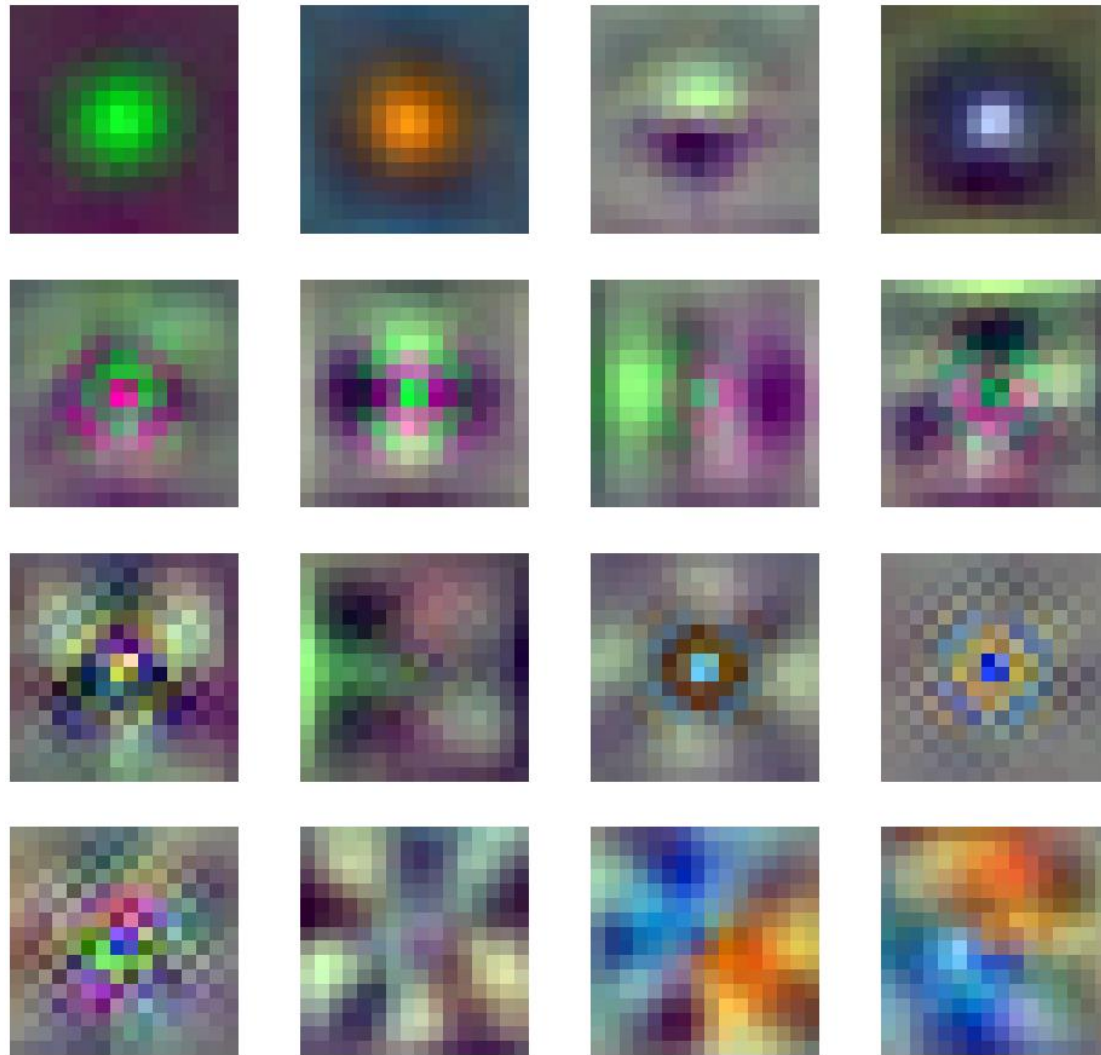
Vision Transformers (ViT)

Embedding

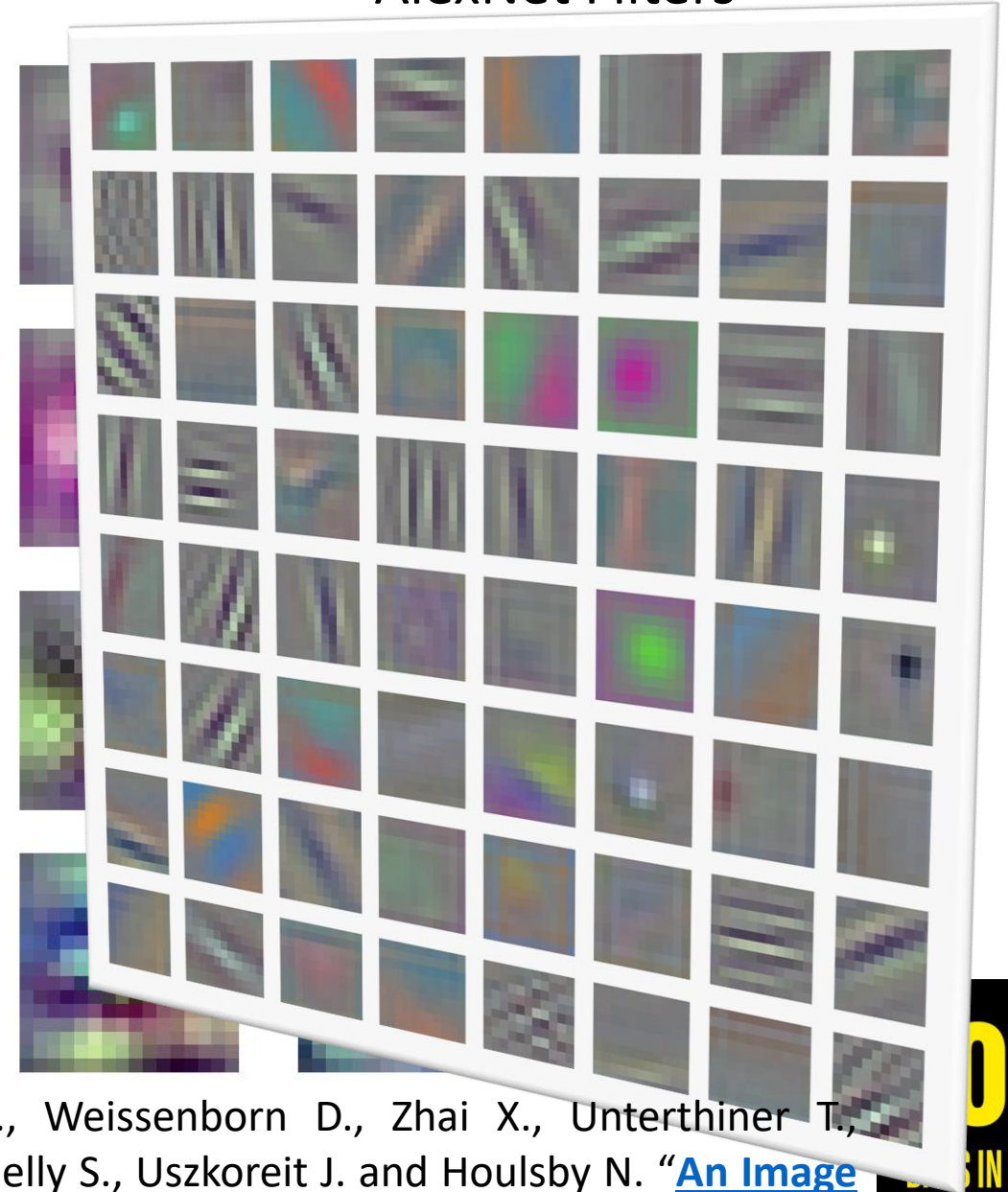


Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J. and Houlsby N. [“An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”](#) (ICLR 2021)

Vision Transformers (ViT)



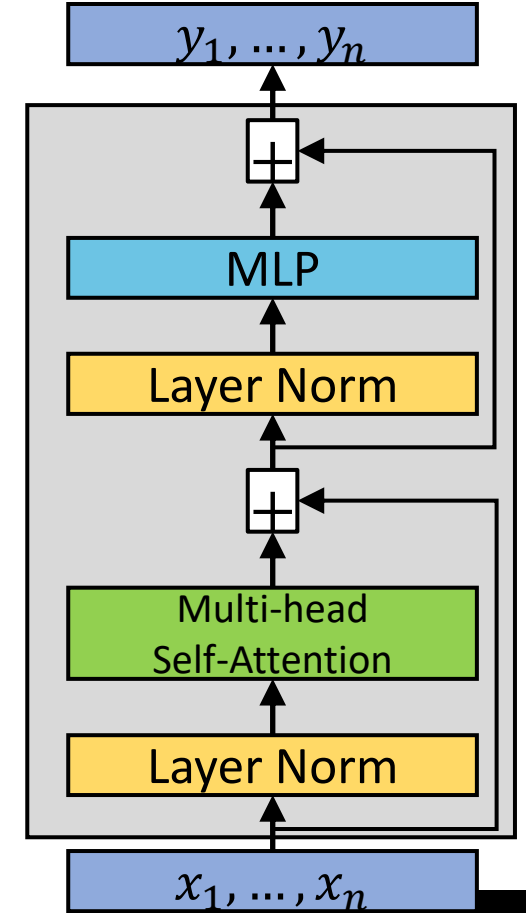
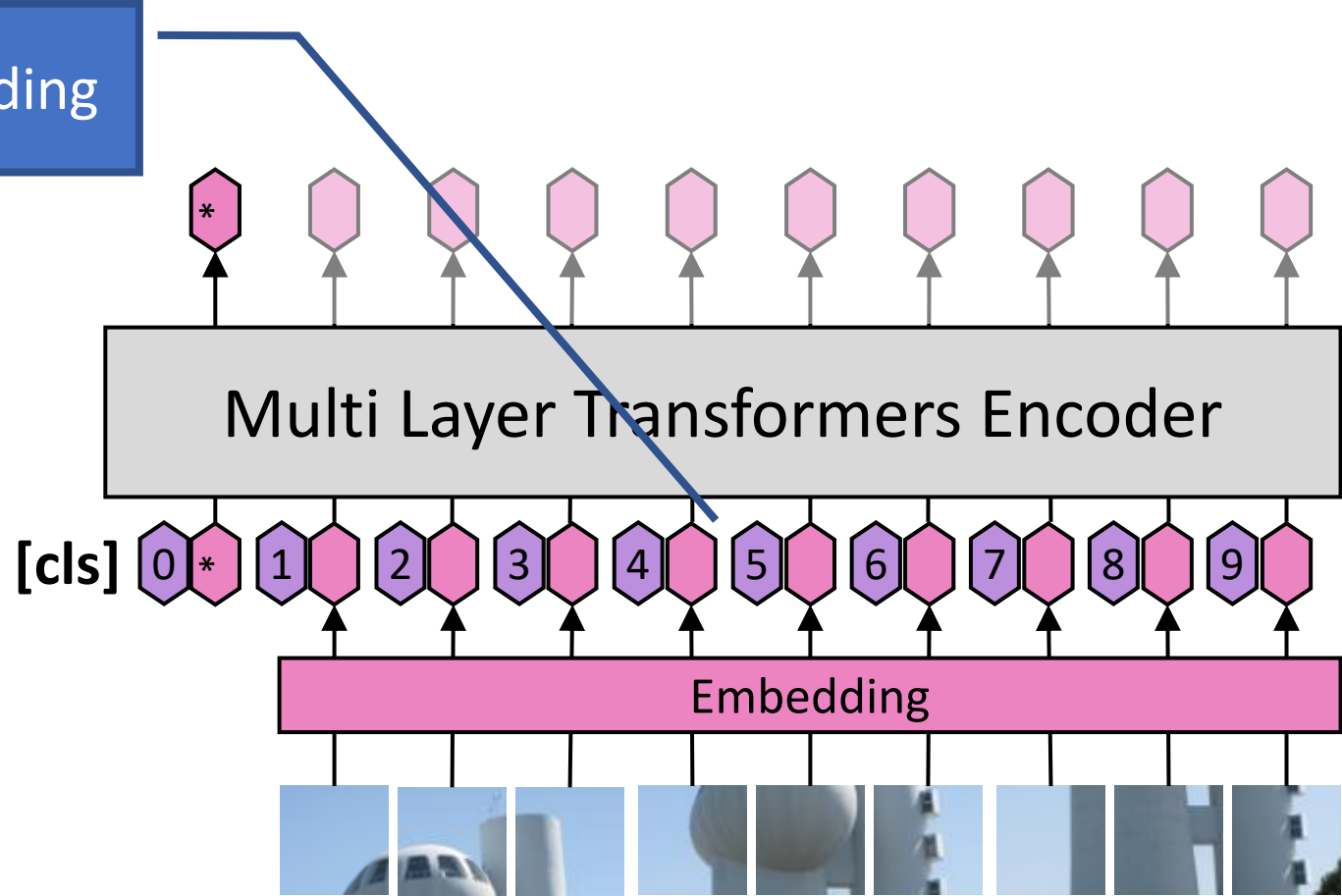
AlexNet Filters



Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J. and Houlsby N. "[An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale](#)" (ICLR 2021)

Vision Transformers (ViT)

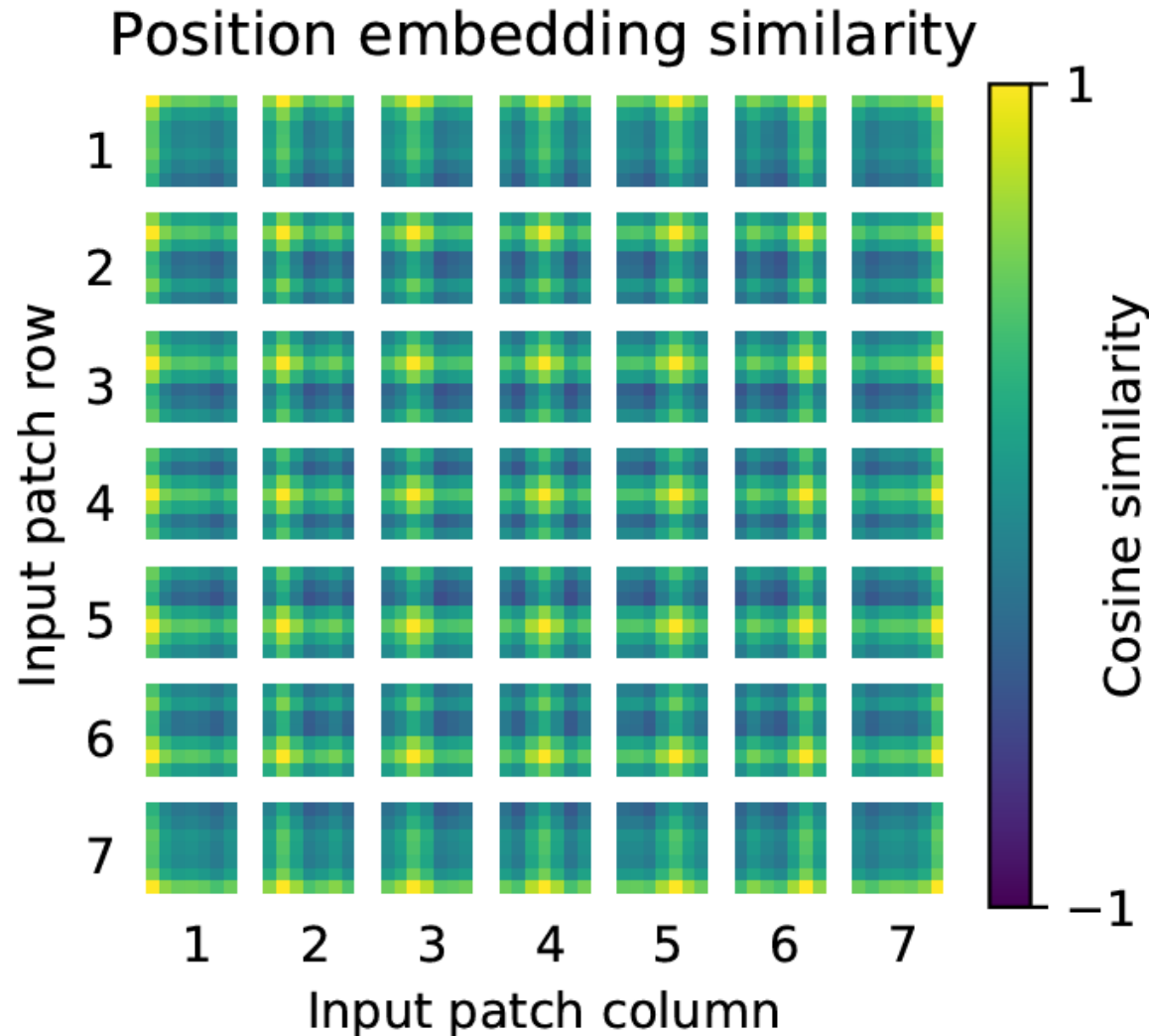
Positional Encoding



Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J. and Houlsby N. [“An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”](#) (ICLR 2021)

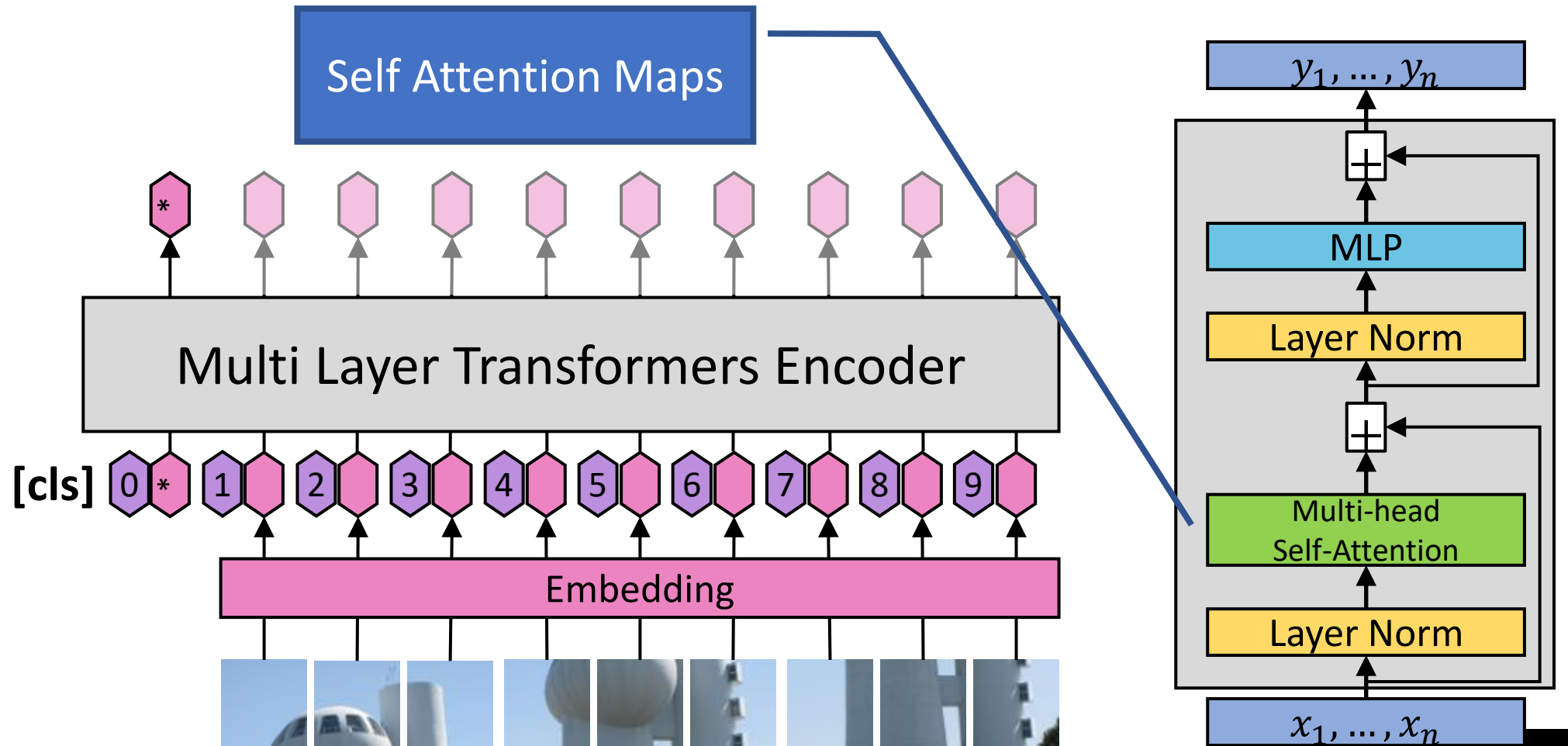
Vision Transformers (ViT)

ViT-B/32:
7x7 patches

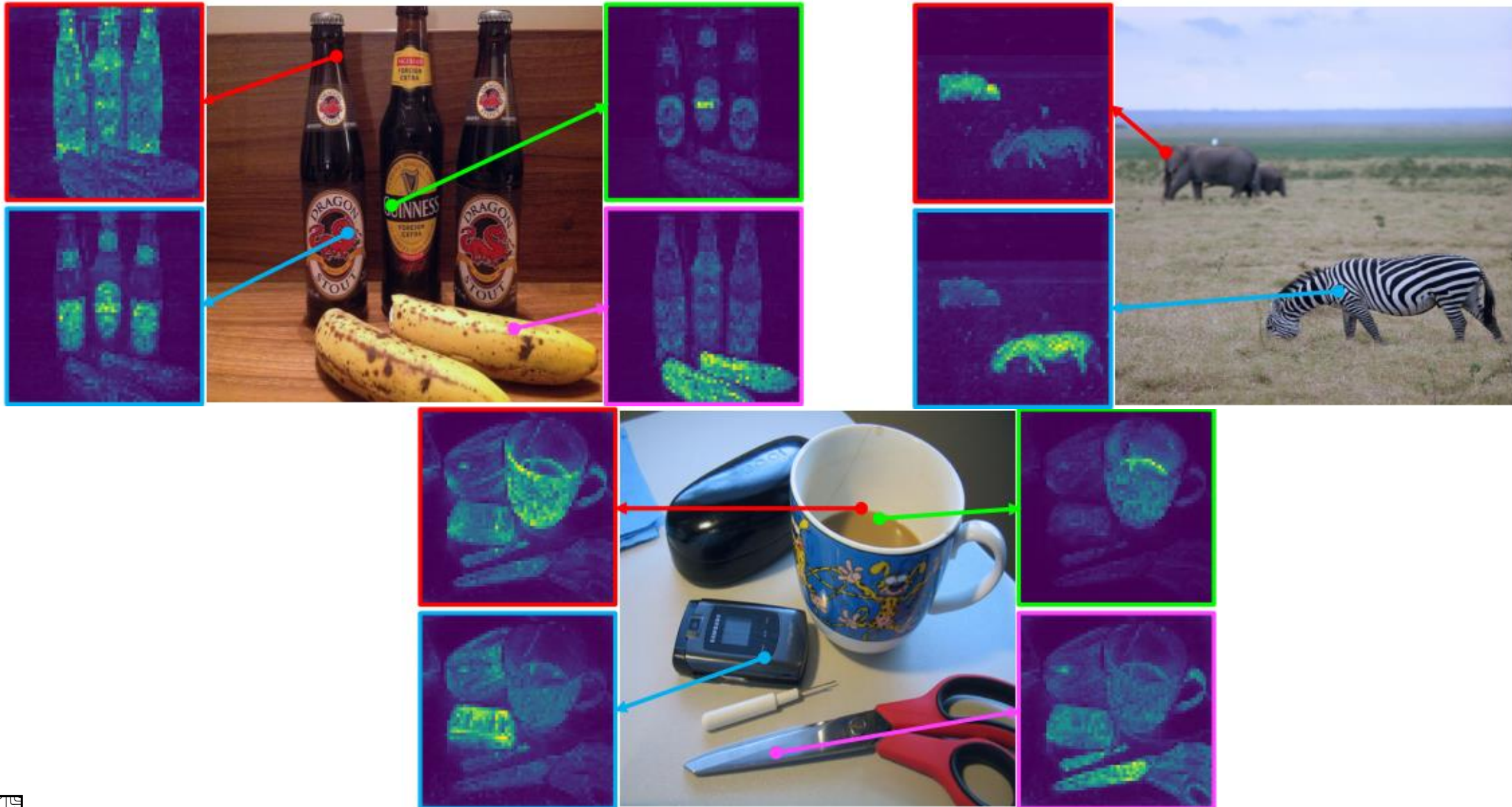


Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J. and Houlsby N. [“An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”](#) (ICLR 2021)

Vision Transformers (ViT)

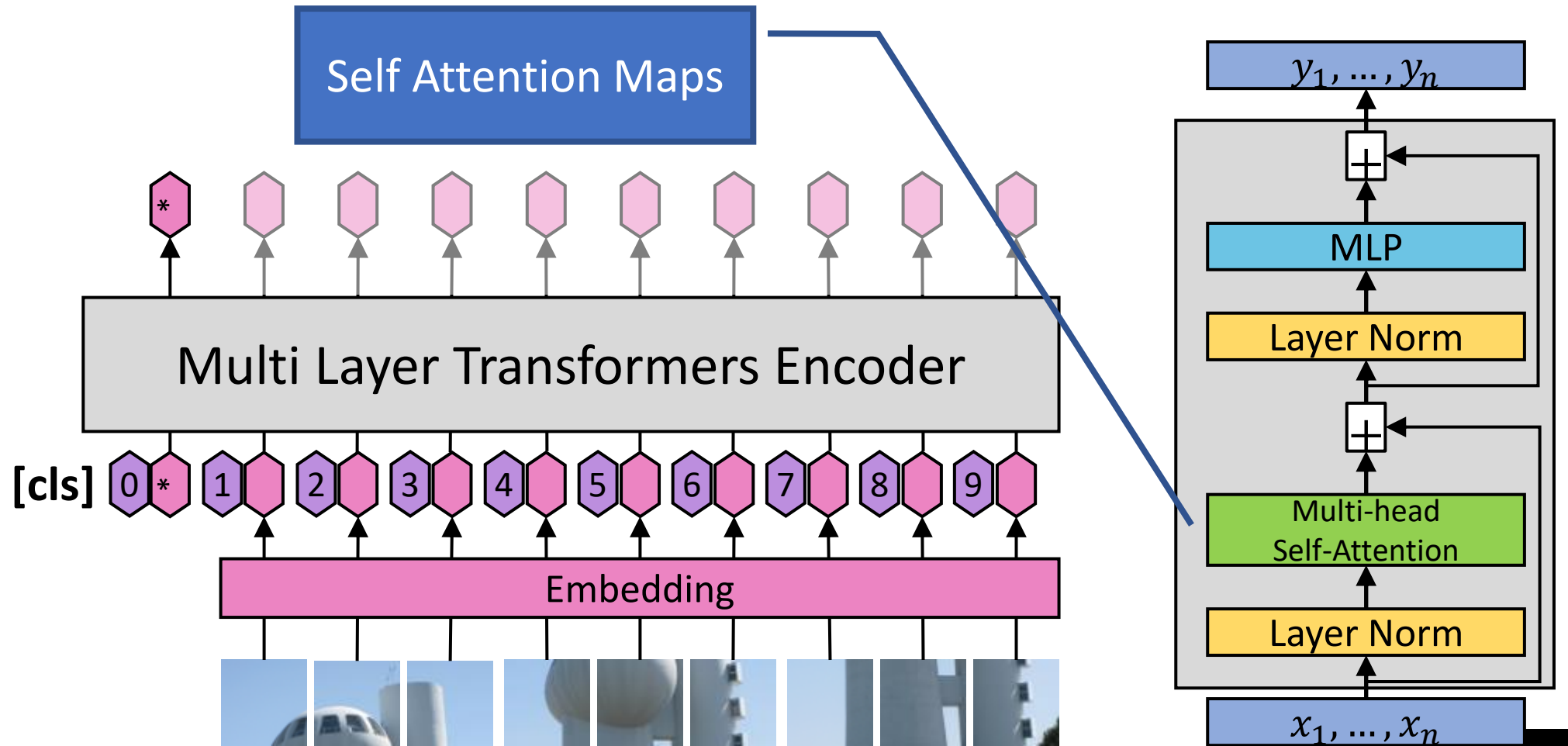


Vision Transformers (ViT)

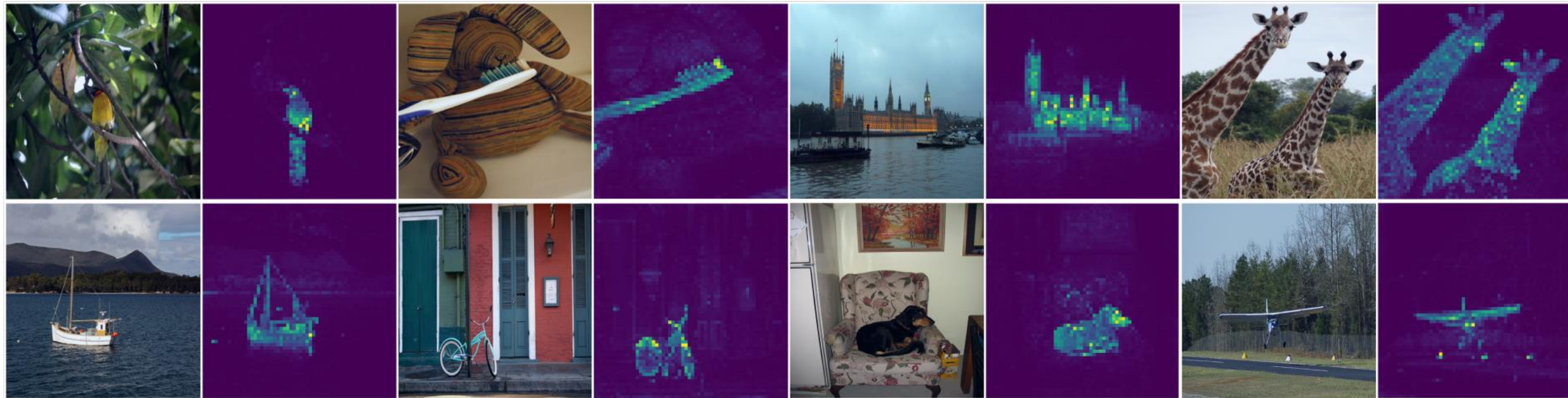


Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P. and Joulin, A.,
“[Emerging Properties in Self-Supervised Vision Transformers](#)”. ICCV (2021)

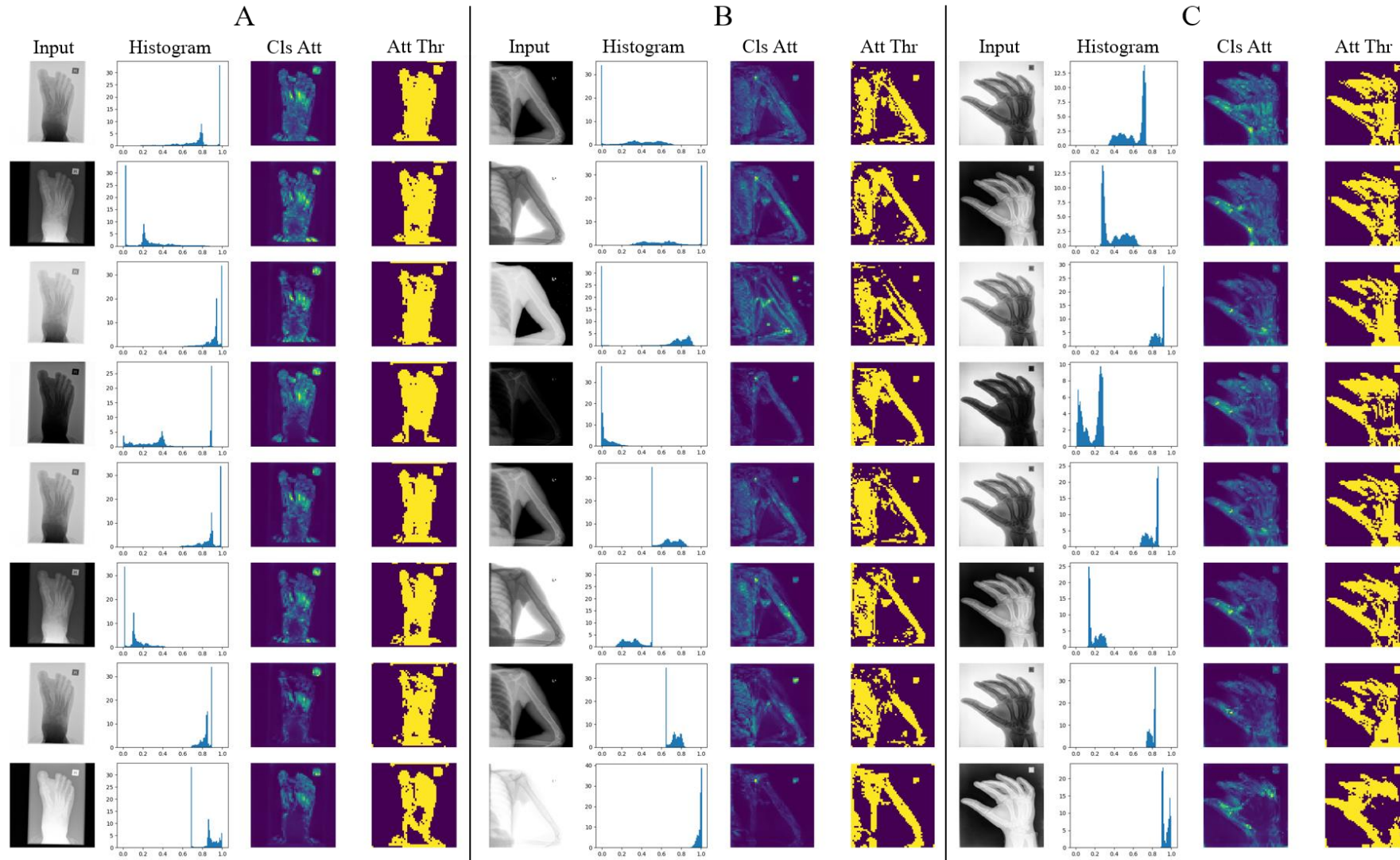
Vision Transformers (ViT)



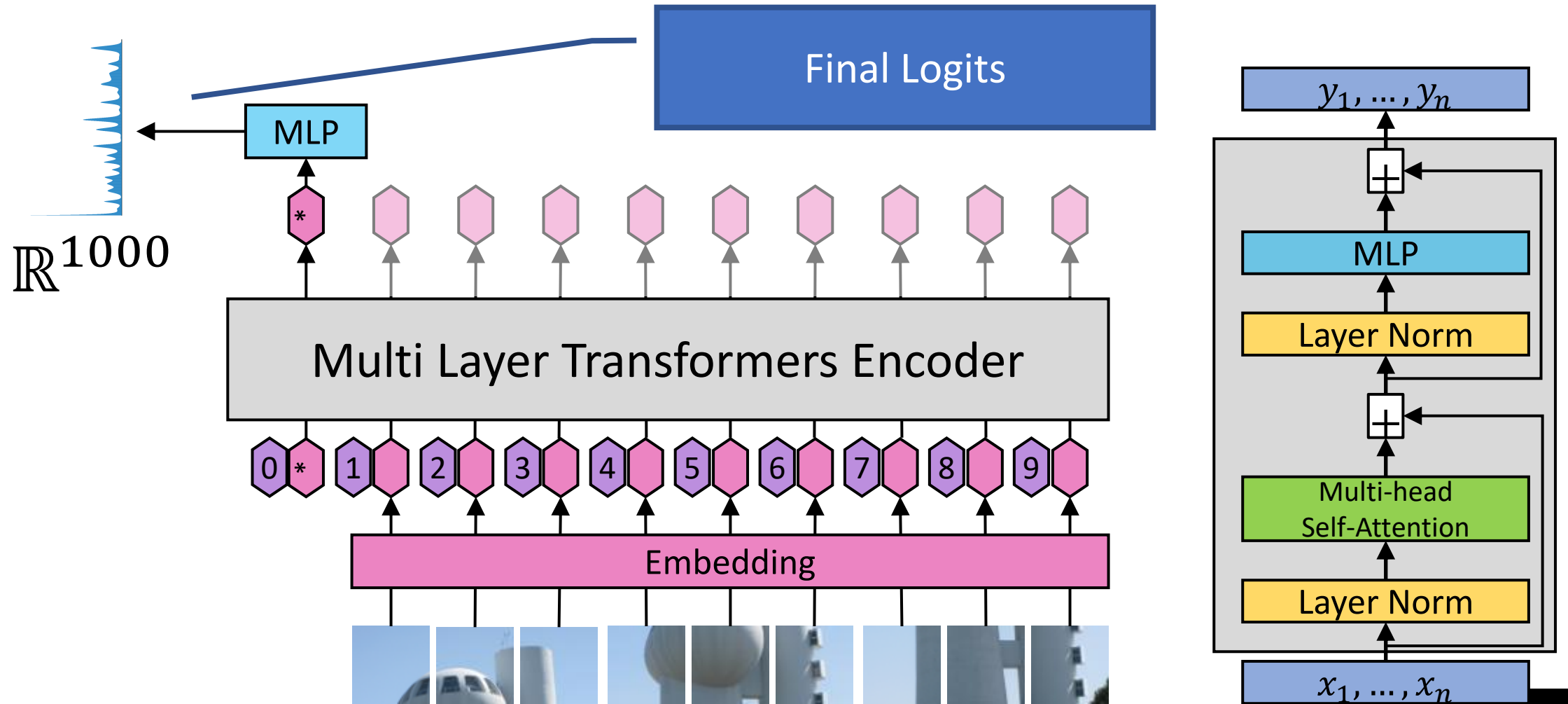
Vision Transformers (ViT)



Vision Transformers (ViT)



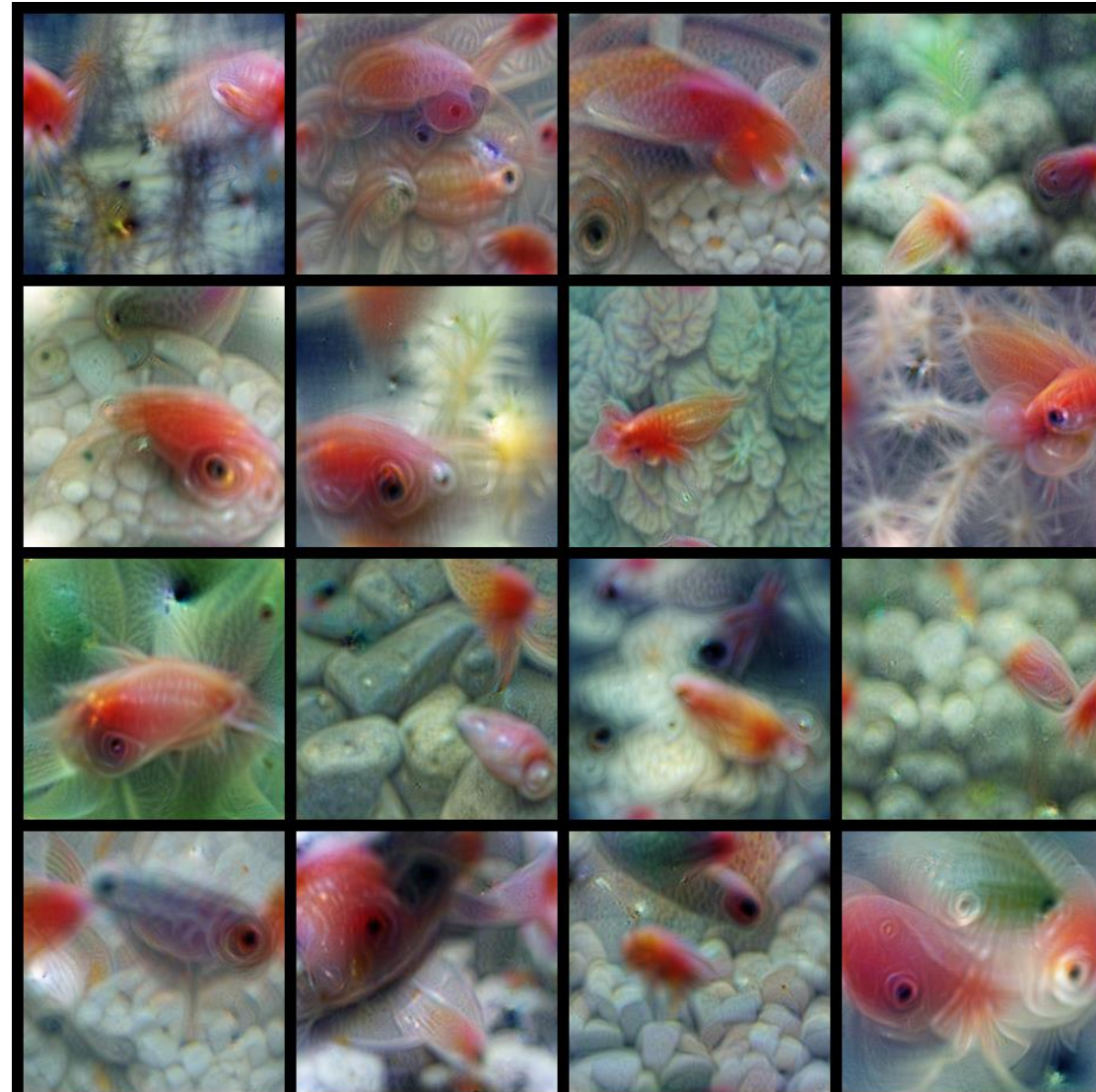
Vision Transformers (ViT)



Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J. and Houlsby N. [“An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”](#) (ICLR 2021)

Vision Transformers (ViT)

logits for class “gold fish”



Vision Transformers (ViT)

logits for class “bald eagle”



Vision Transformers (ViT)

logits for class “curly coated retriever”

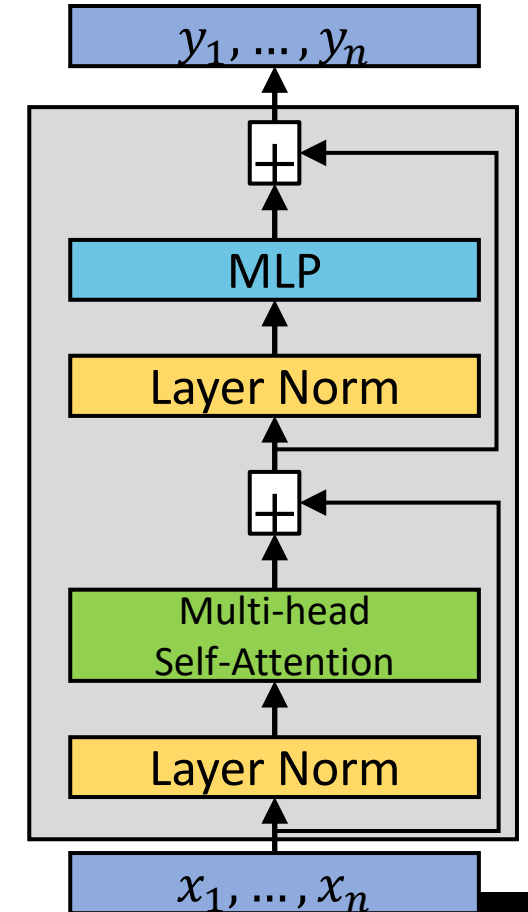
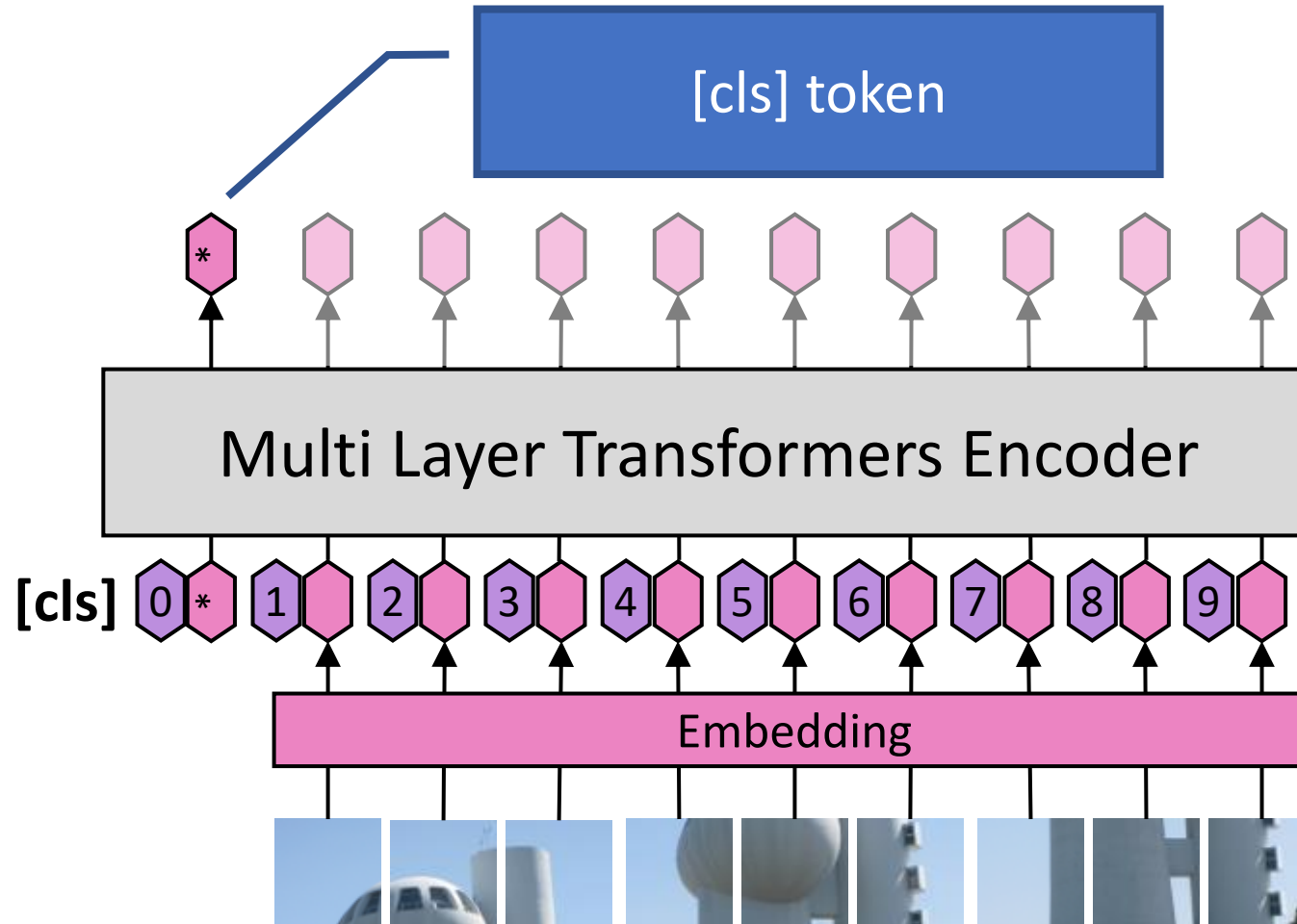


Vision Transformers (ViT)

logits for class “volcano”

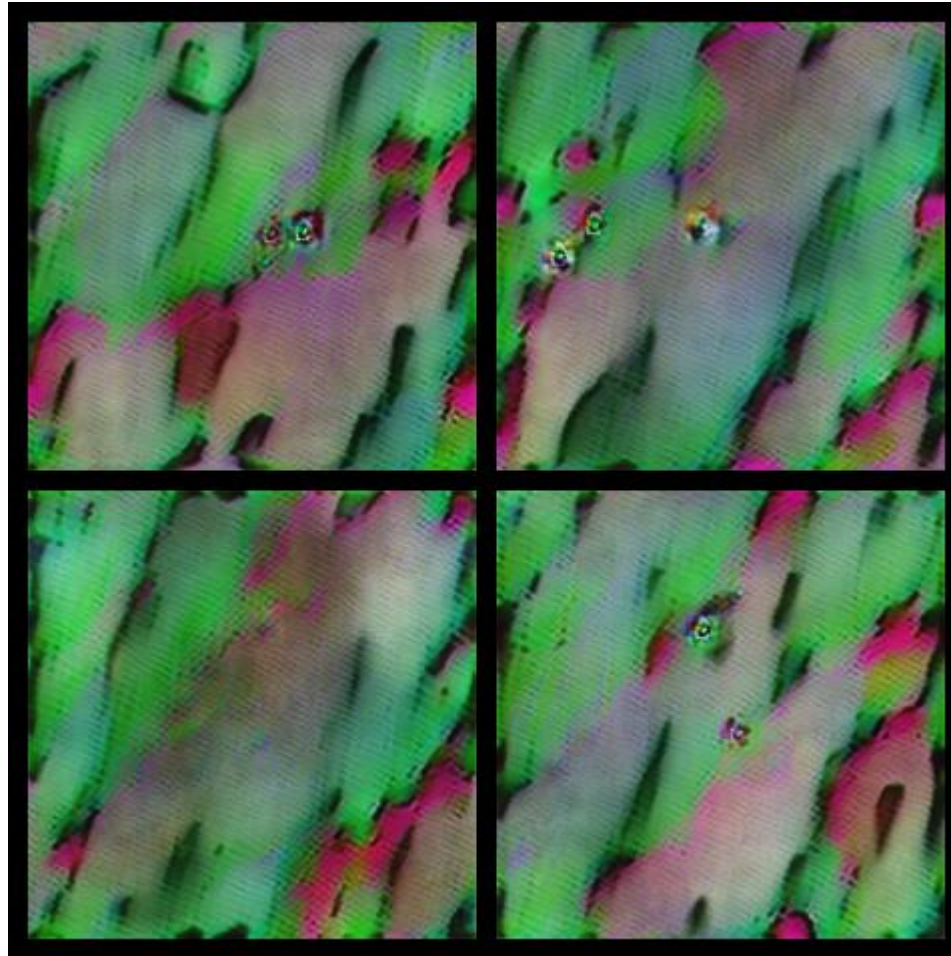


Vision Transformers (ViT)



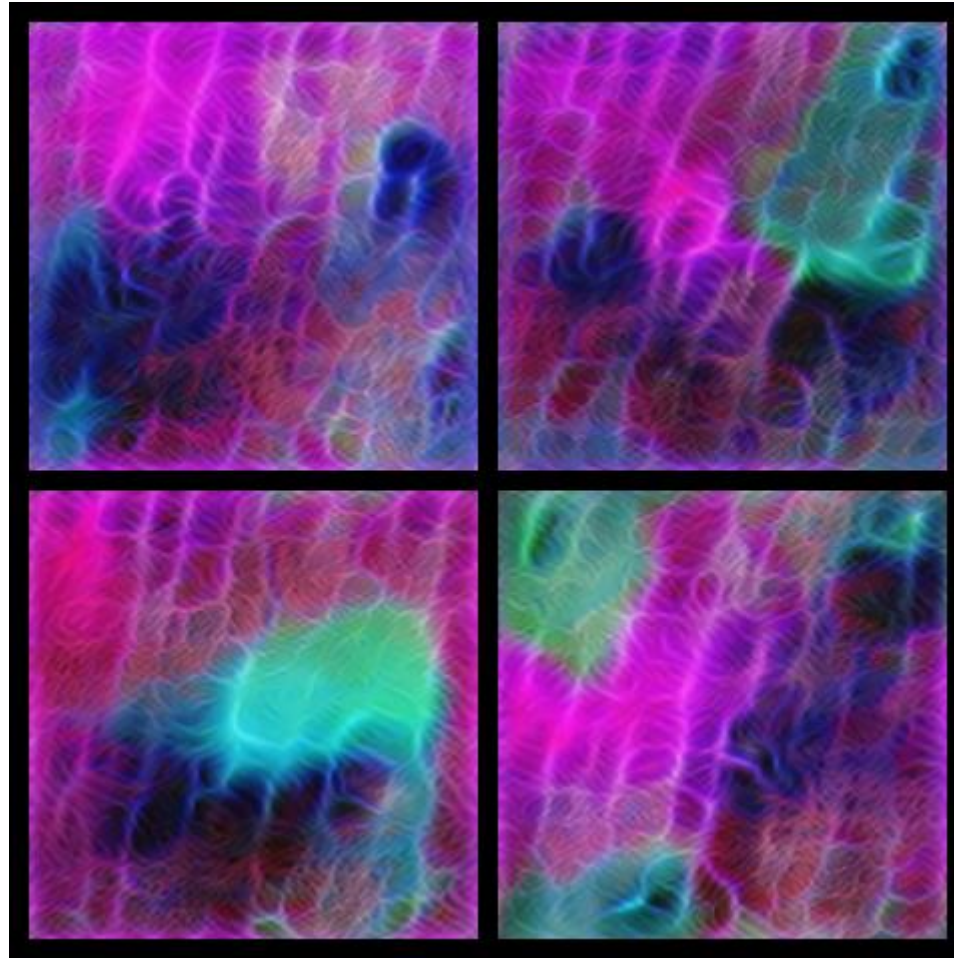
Vision Transformers (ViT)

[cls] token layer=1



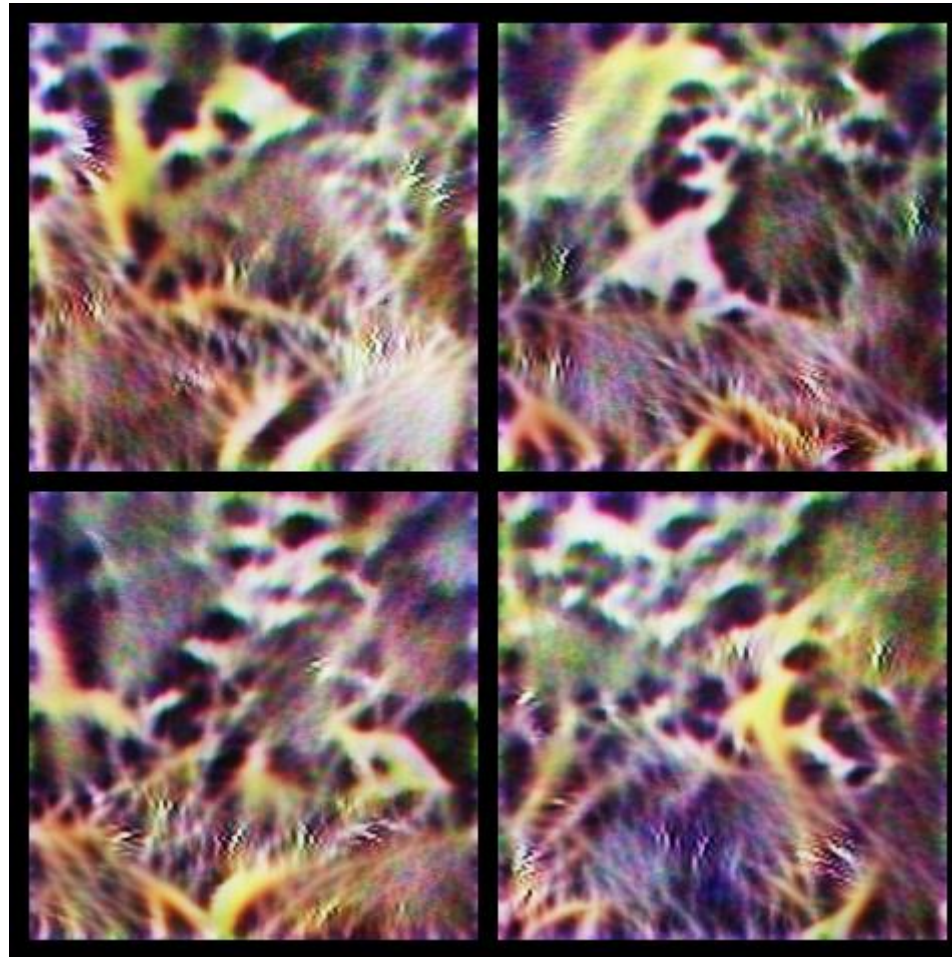
Vision Transformers (ViT)

[cls] token layer=3



Vision Transformers (ViT)

[cls] token layer=5

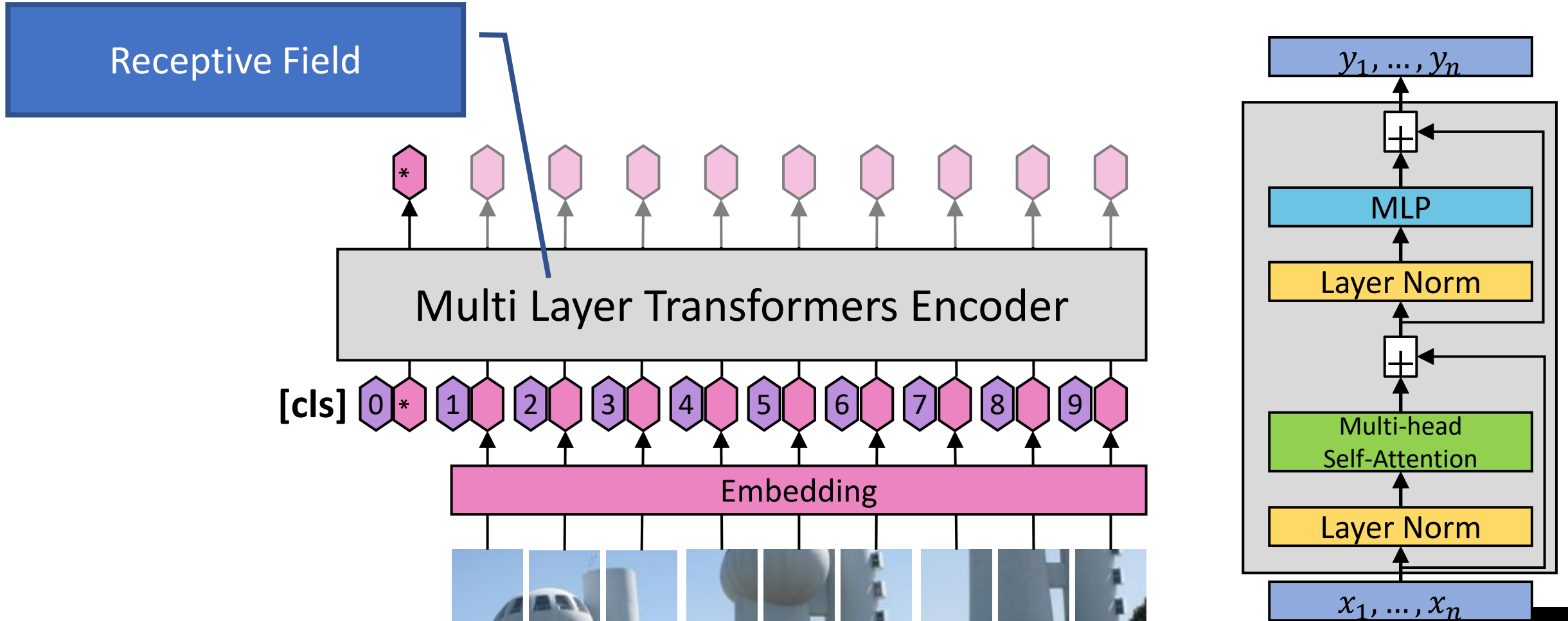


Vision Transformers (ViT)

[cls] token layer=10

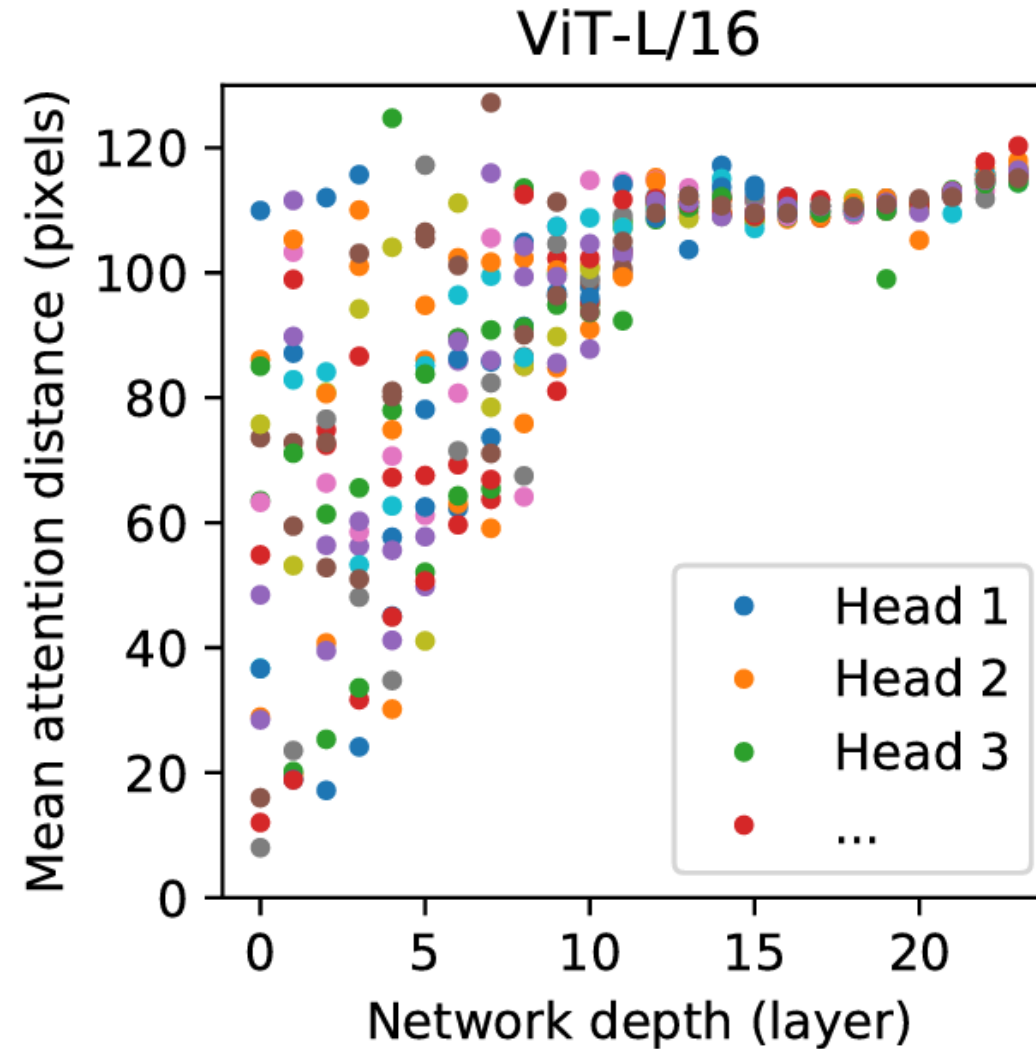


Vision Transformers (ViT)



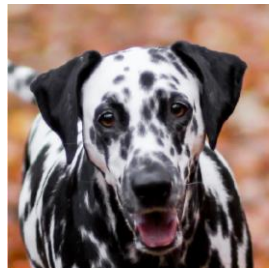
Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J. and Houlsby N. [“An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”](#) (ICLR 2021)

Vision Transformers (ViT)



Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J. and Houlsby N. [“An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”](#) (ICLR 2021)

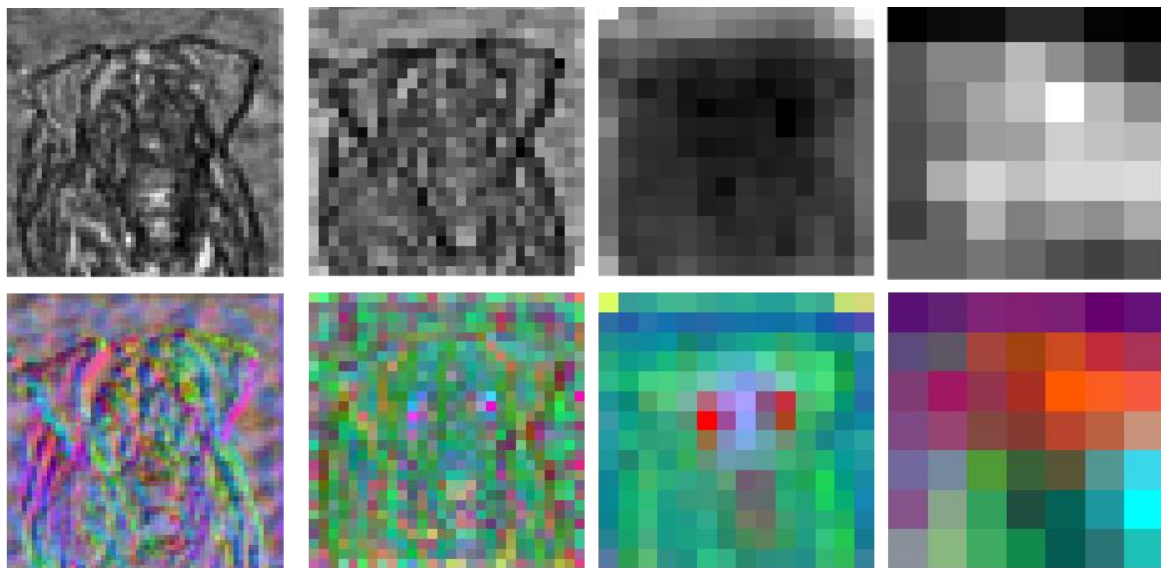
DL4CV: CNN vs. ViT



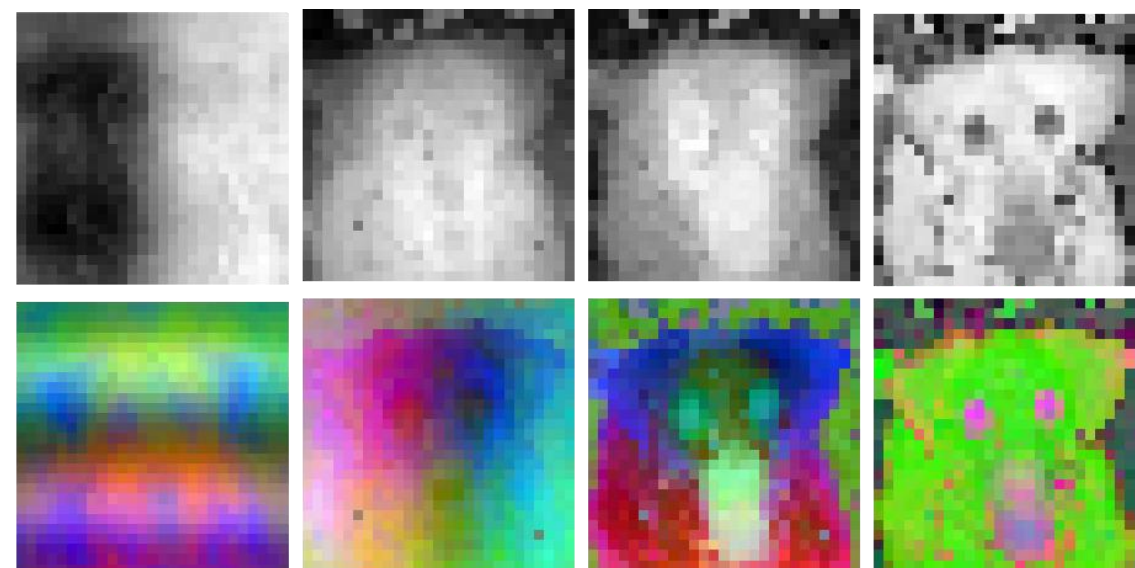
ResNet50

ViT

shallow → deep



shallow → deep



DL4CV: CNN vs. ViT

CNN

ViT



1. Maintain 2D structure logic



2. Shift equivariant



3. Consider only local correlations



4. Hierarchically growing field of view



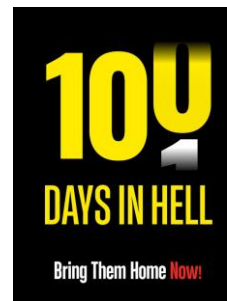
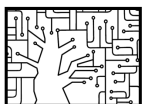
5. Hierarchically progressing complexity



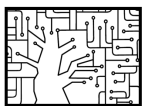
6. Reasonable amount of params



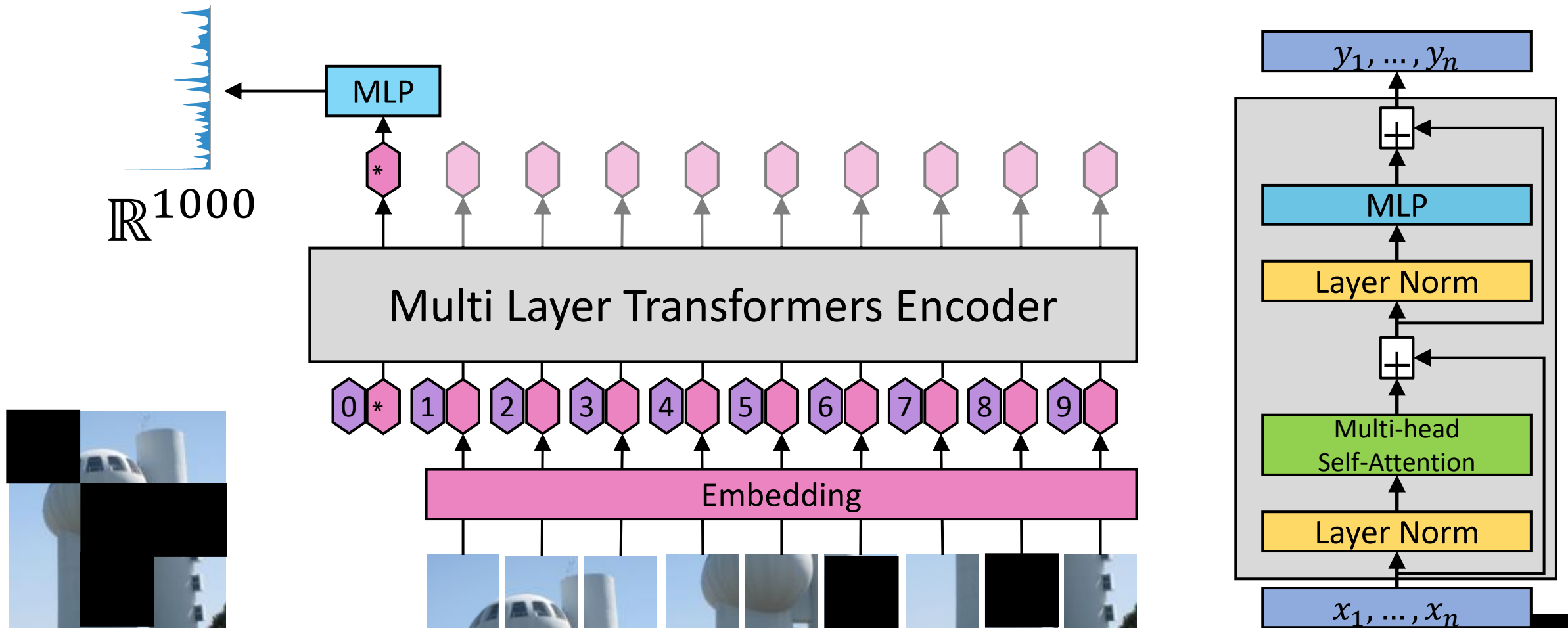
7. Global representation



Vision Transformers (ViT) – More Properties



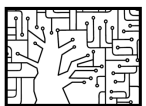
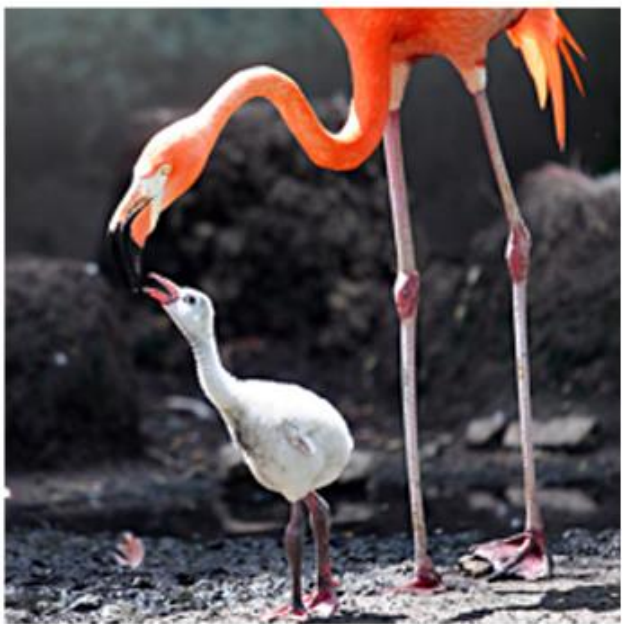
Vision Transformers (ViT) – Robustness



Naseer, M.M., Ranasinghe, K., Khan, S.H., Hayat, M., Shahbaz Khan, F. and Yang, M.H.,
Intriguing properties of vision transformers. (NeurIPS 2021)

Vision Transformers (ViT) – Robustness

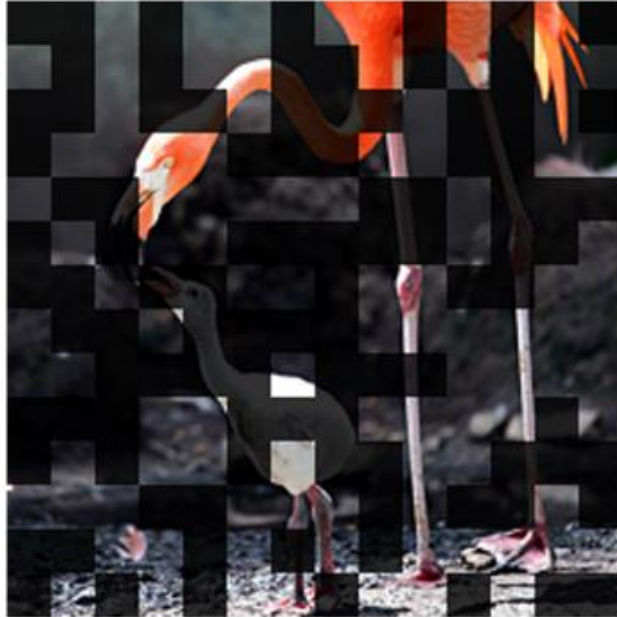
Original Image



Vision Transformers (ViT) – Robustness

Original Image

Random PatchDrop



Vision Transformers (ViT) – Robustness

“Patch Drop” is good for:

- Efficiency
- Augmentation
- Shows “redundancy” in images (videos?)

Vision Transformers (ViT) – Properties

Additional properties:

- “Data hungry” model – benefit larger train set/aug
- Relative robustness to adversarial examples
- Shape/texture



(a) Texture image

81.4%	Indian elephant
10.3%	indri
8.2%	black swan



(b) Content image

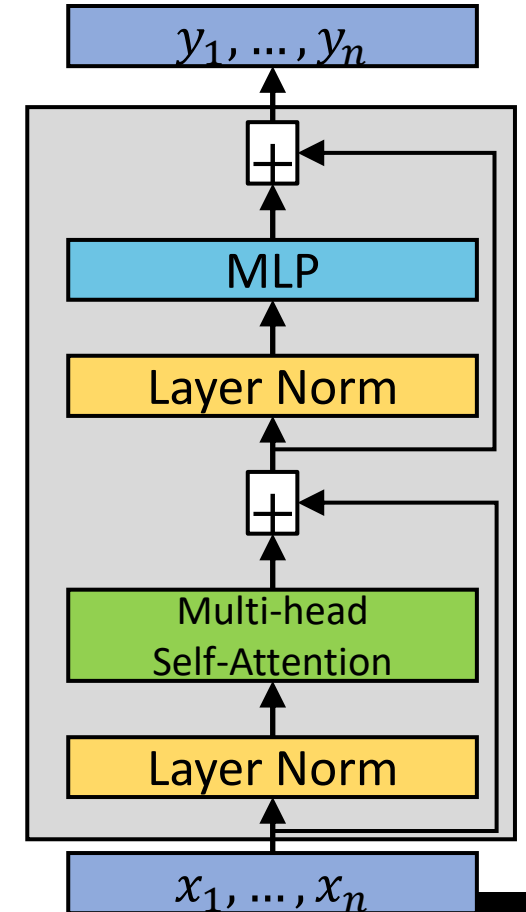
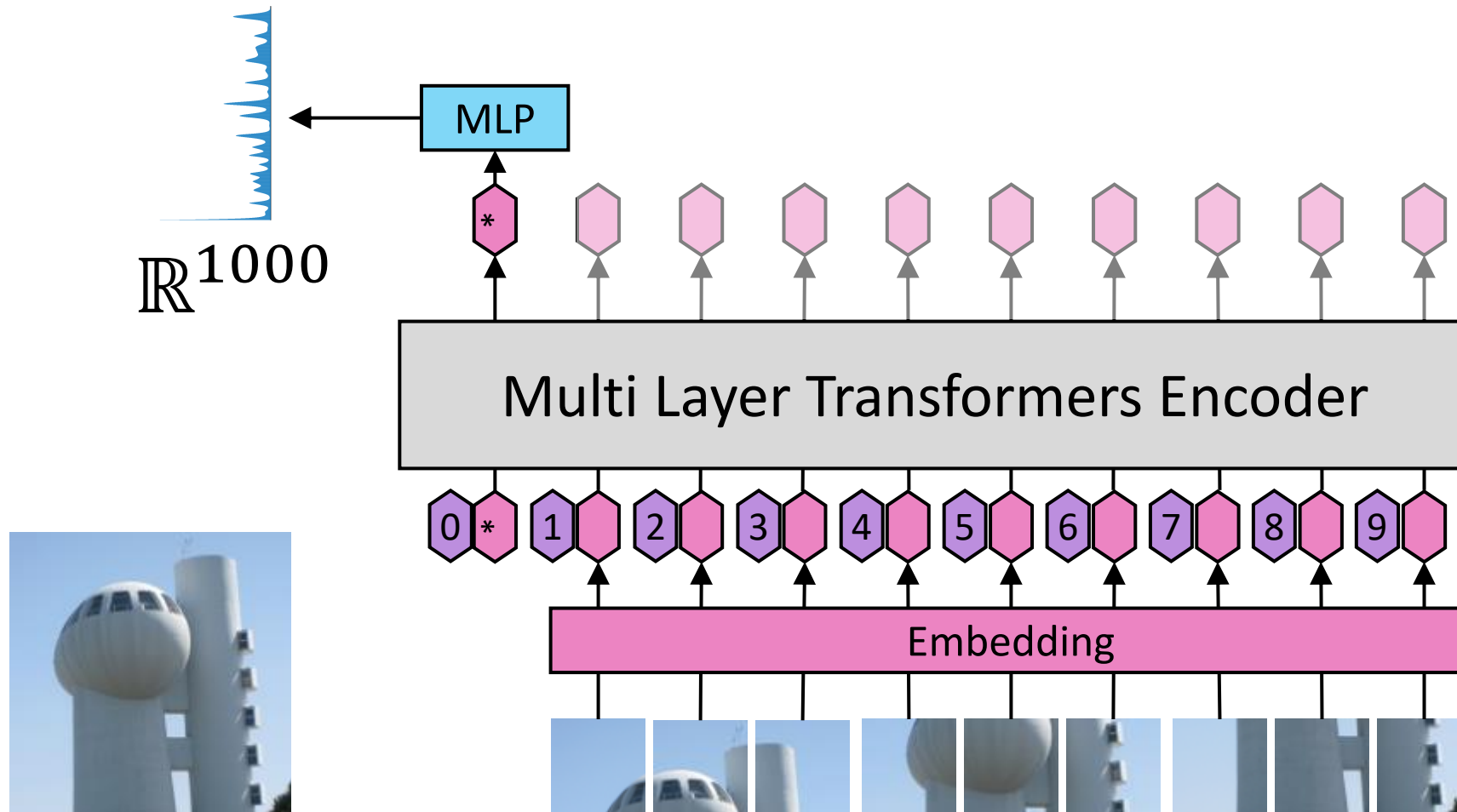
71.1%	tabby cat
17.3%	grey fox
3.3%	Siamese cat



(c) Texture-shape cue conflict

63.9%	Indian elephant	H.,
26.4%	indri	?)
9.6%	black swan	

Vision Transformers (ViT)



Vision Transformers (ViT)



Complexity

CNN

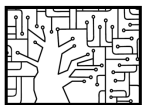
$$y_i = \sum_k x_{i+k} w_k$$

$$O(n)$$

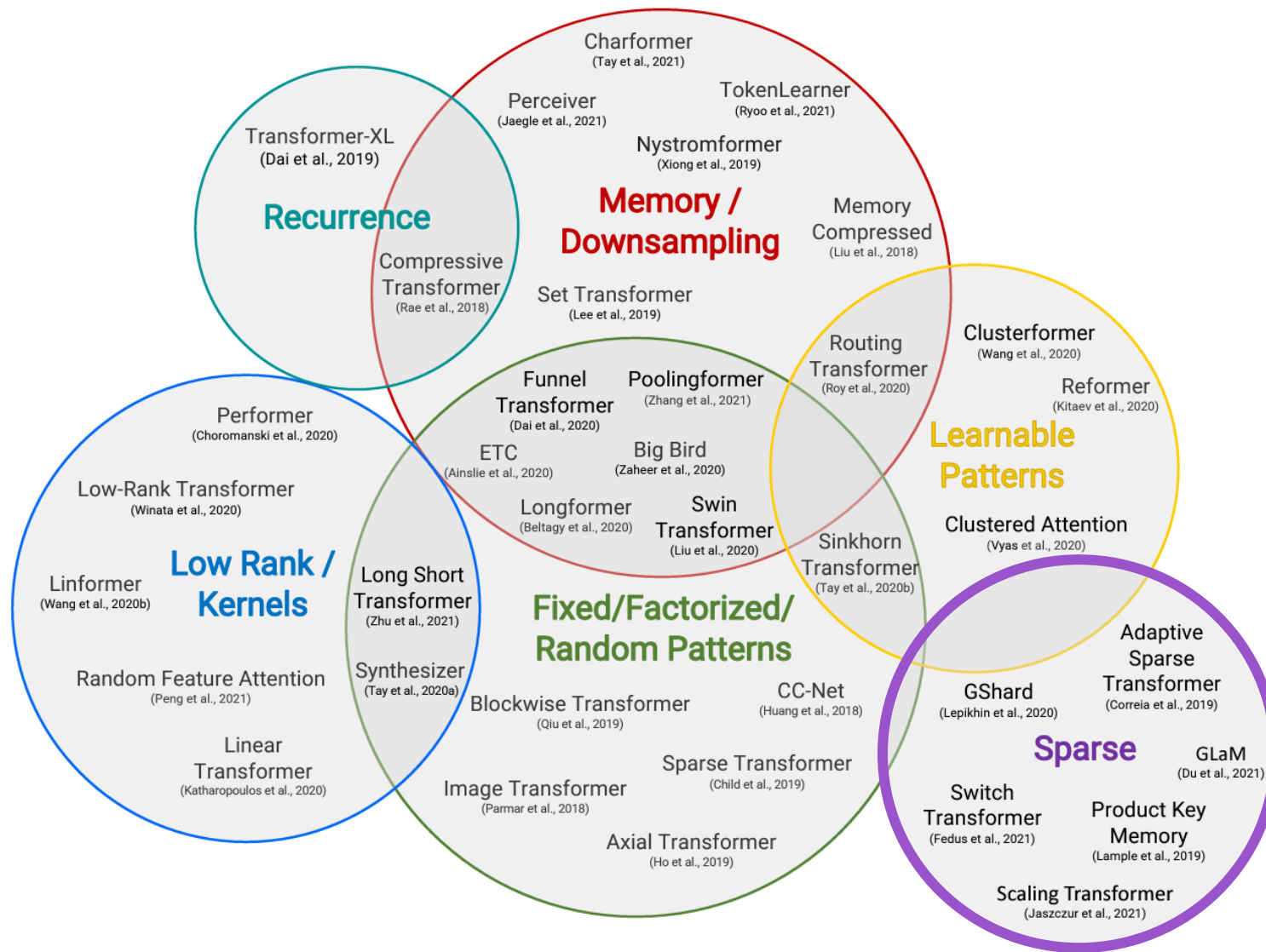
Self-Attention

$$y_i = \sum_j \sigma(q_i k_j) v_j$$

$$O(n^2)$$



Efficient Transformers

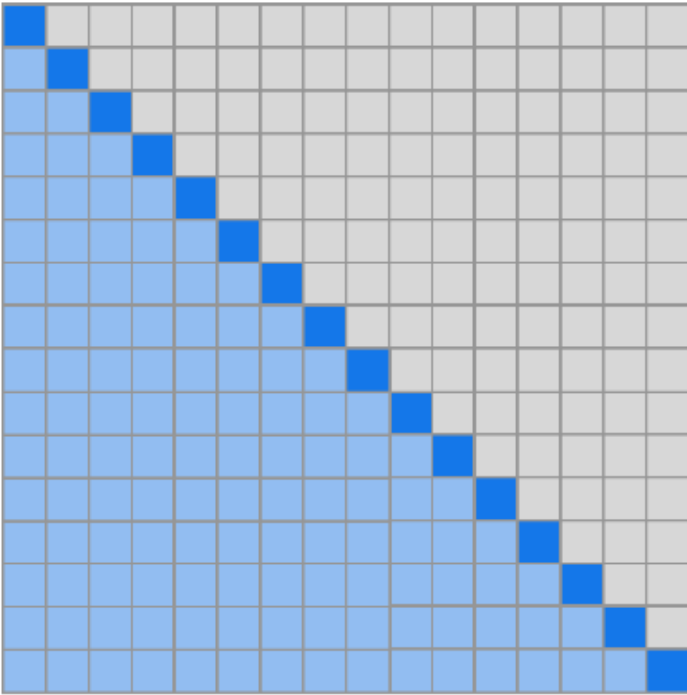


Tay, Y., Dehghani, M., Bahri, D. and M.

["Efficient transformers: A survey"](#)



Efficient Transformers – Sparse Attention



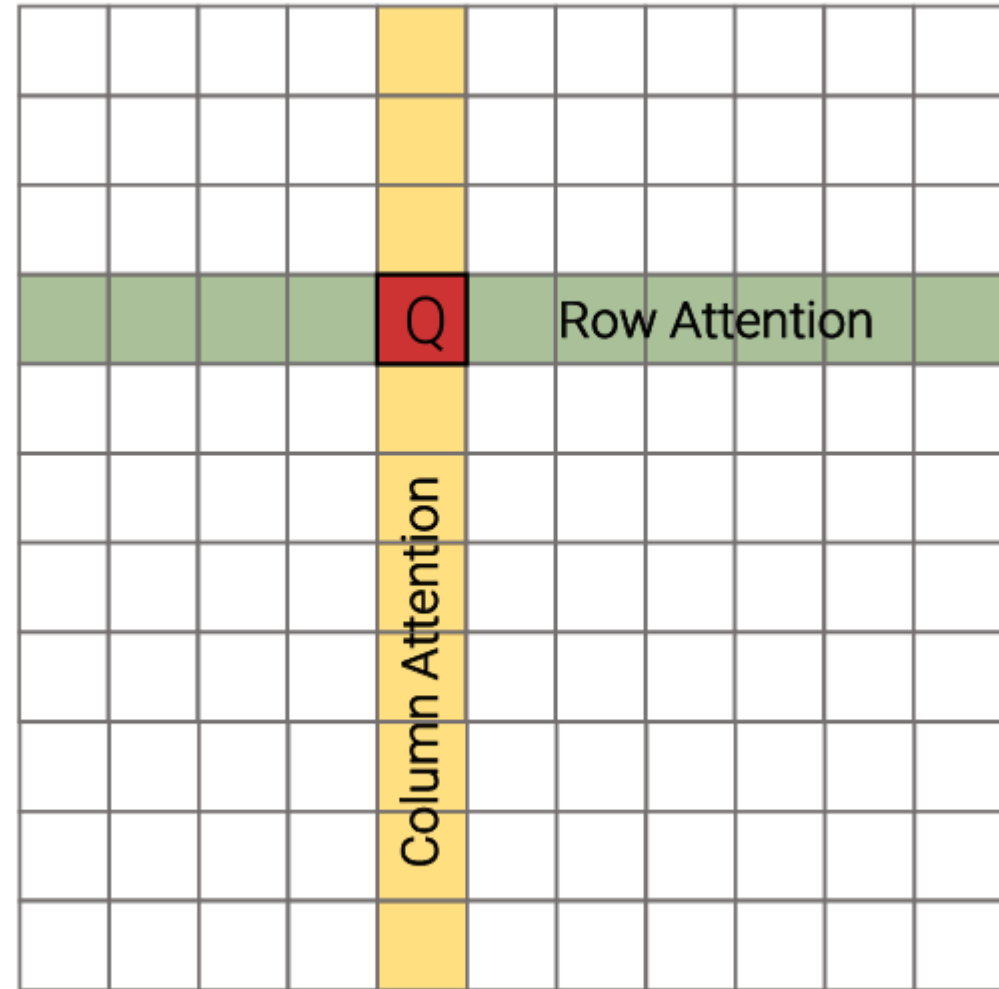
(a) Transformer

$$O(n\sqrt{n})$$

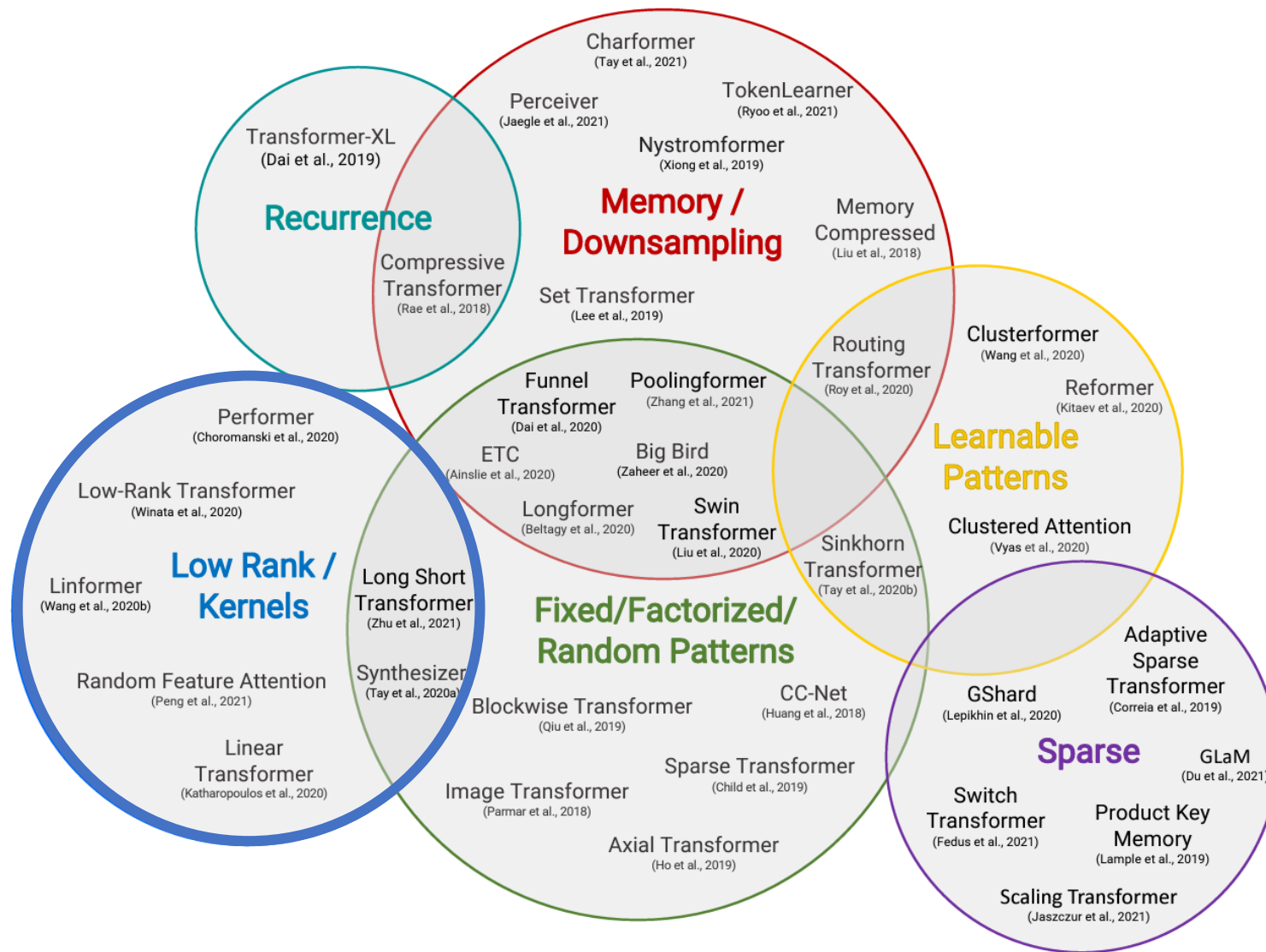
Efficient Transformers – Sparse Attention

Axial attention for 2D data

$$O(n\sqrt{n})$$



Efficient Transformers



Tay, Y., Dehghani, M., Bahri, D. and M. ...
["Efficient transformers: A survey"](#)

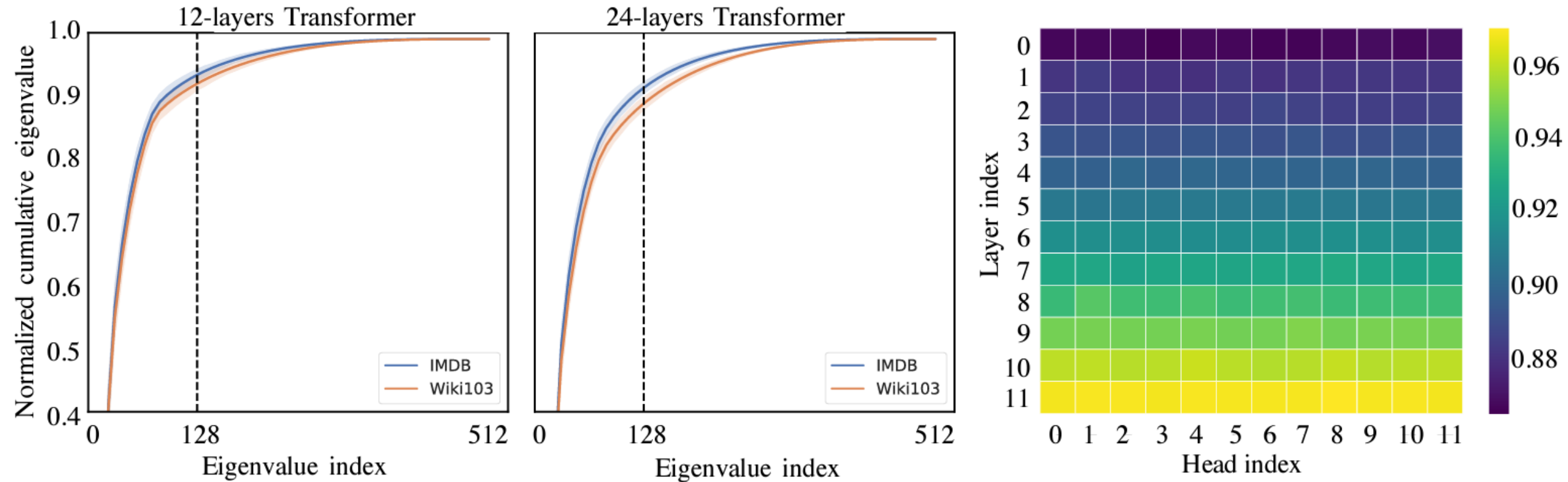
Efficient Transformers – Low Rank

$$Y = \sigma(QK^T)V$$

$$Y = \sigma \left(\begin{array}{c} \text{blue box } Q \\ \text{blue box } K^T \end{array} \right) \begin{array}{c} \text{blue box } V \end{array} \in \mathbb{R}^{n \times d}, n \gg d$$

$$\begin{array}{c} \text{blue box } Q \\ \text{blue box } K^T V \end{array}$$

Efficient Transformers – Low Rank



Efficient Transformers – Low Rank

$$\begin{matrix} Q & K & V \end{matrix} \in \mathbb{R}^{n \times d}$$

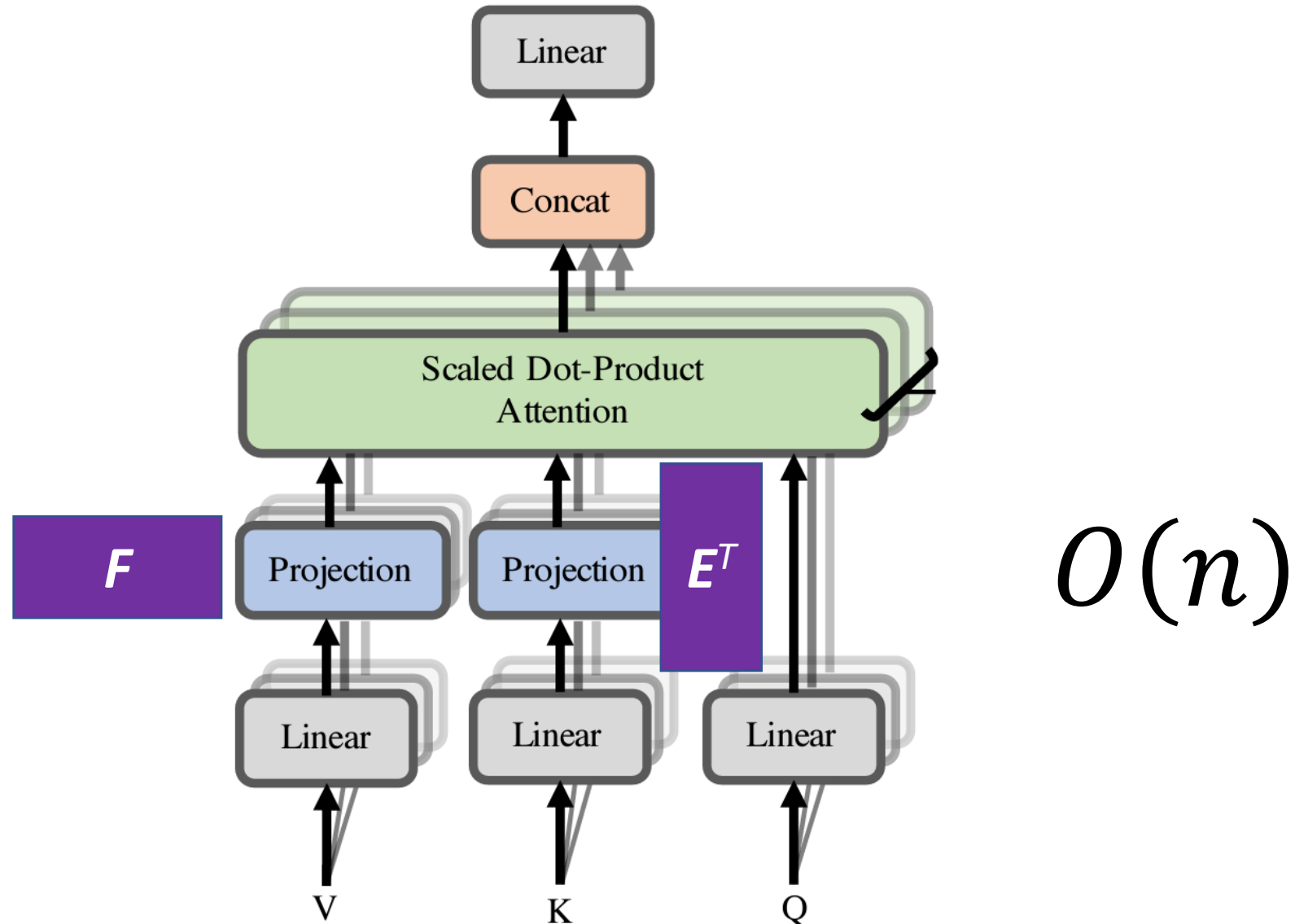
$$\begin{matrix} E \\ F \end{matrix} \in \mathbb{R}^{k \times n}$$

$$Y = \sigma(QK^T)V$$

$$Y = \sigma \left(\begin{matrix} Q & K^T \\ E^T & V \end{matrix} \right) F$$

$$Y = \sigma \left(\begin{matrix} Q & K^T E^T \\ FV \end{matrix} \right) \quad \begin{matrix} k \times d \\ n \times k \end{matrix}$$

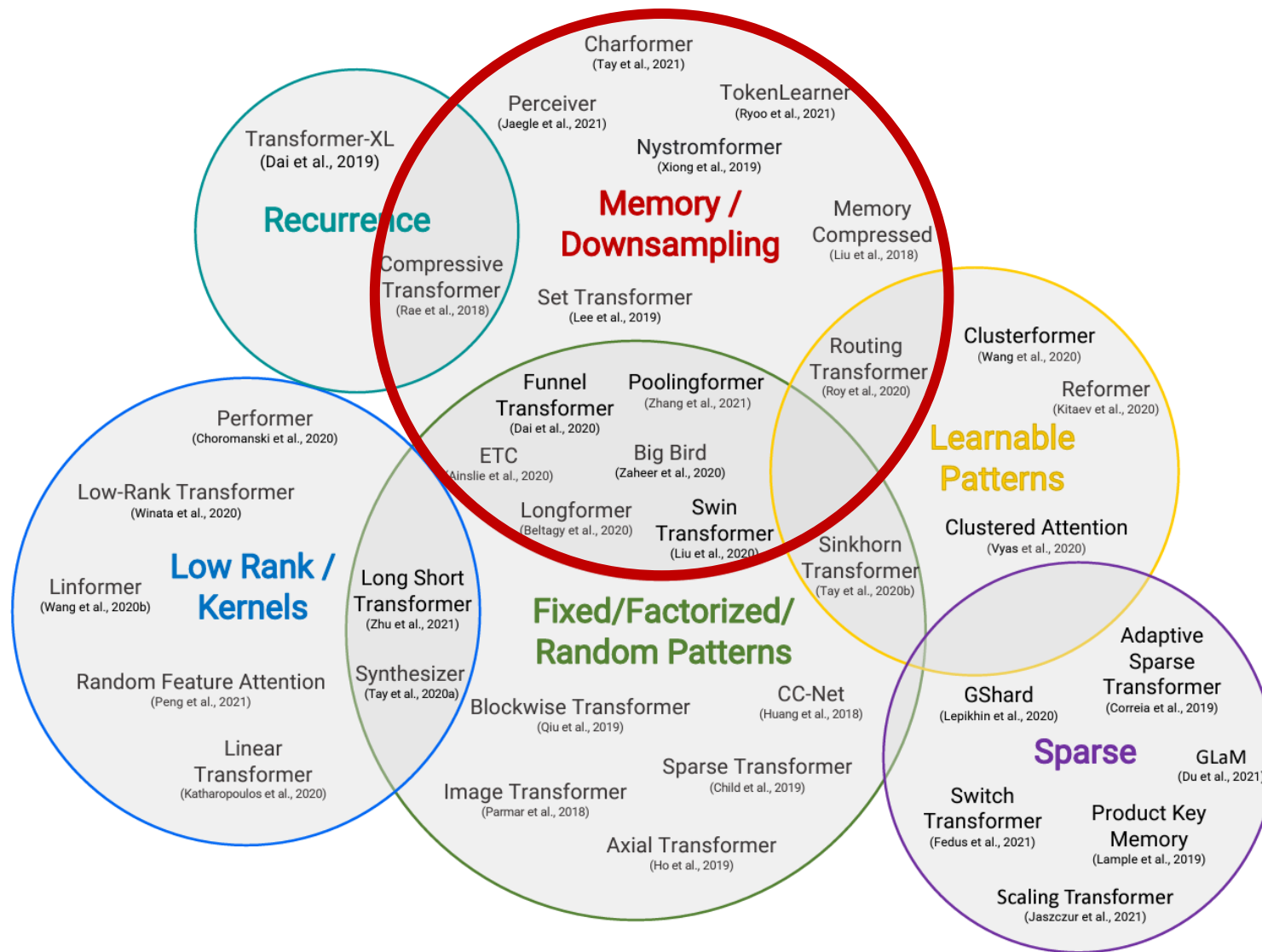
Efficient Transformers – Low Rank



Wang, S., Li, B.Z., Khabsa, M., Fang, H. and Ma, H.

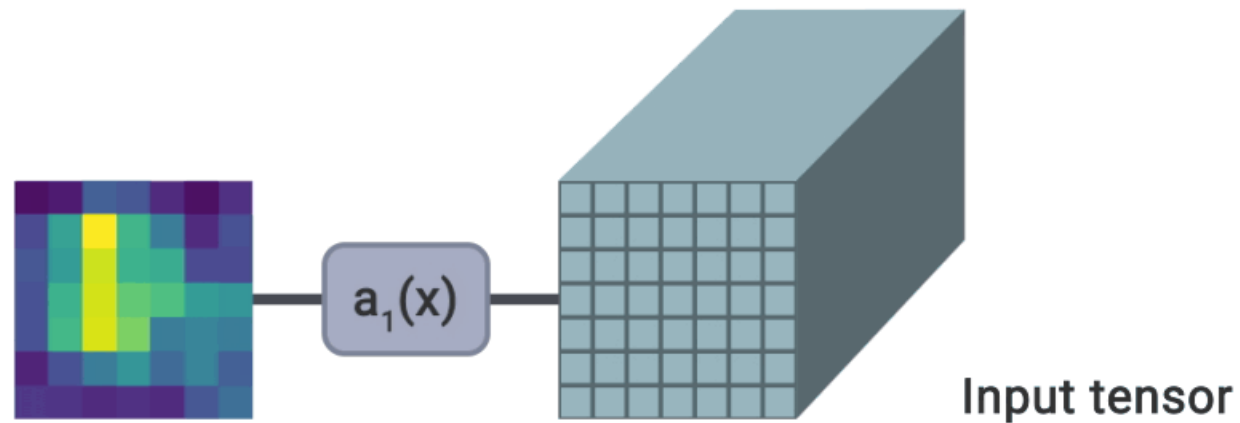
”[Linformer: Self-attention with linear complexity](#)” (arXiv 2020)

Efficient Transformers

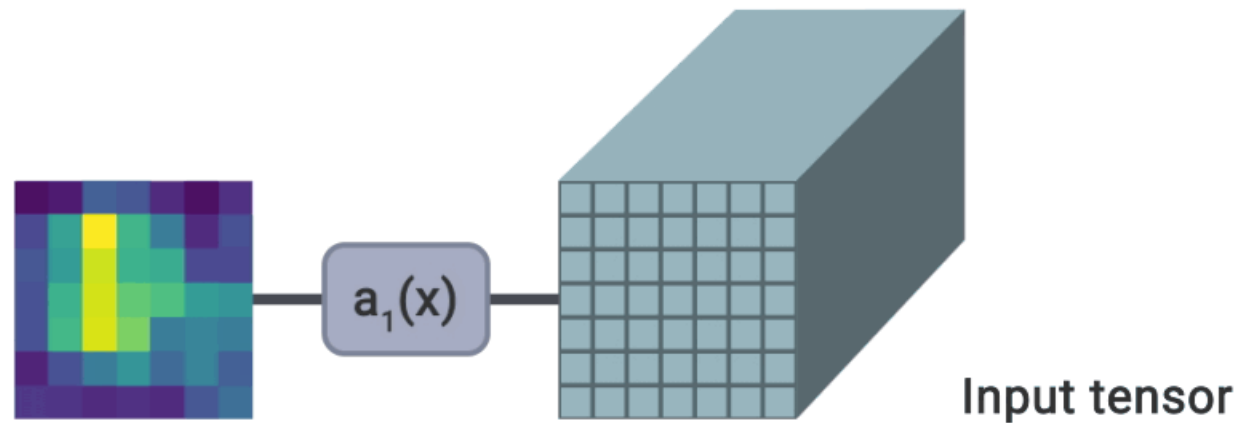


Tay, Y., Dehghani, M., Bahri, D. and Metzler, D. **“Efficient transformers: A survey”**. (2022).

Efficient Transformers – Down-Sample

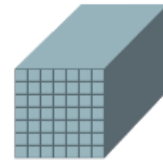


Efficient Transformers – Down-Sample

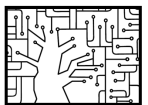


Efficient Transformers – Down-Sample

$O(n)$



Other Transformers for Vision



Shifted Window Transformer (Swin)

Layer 1

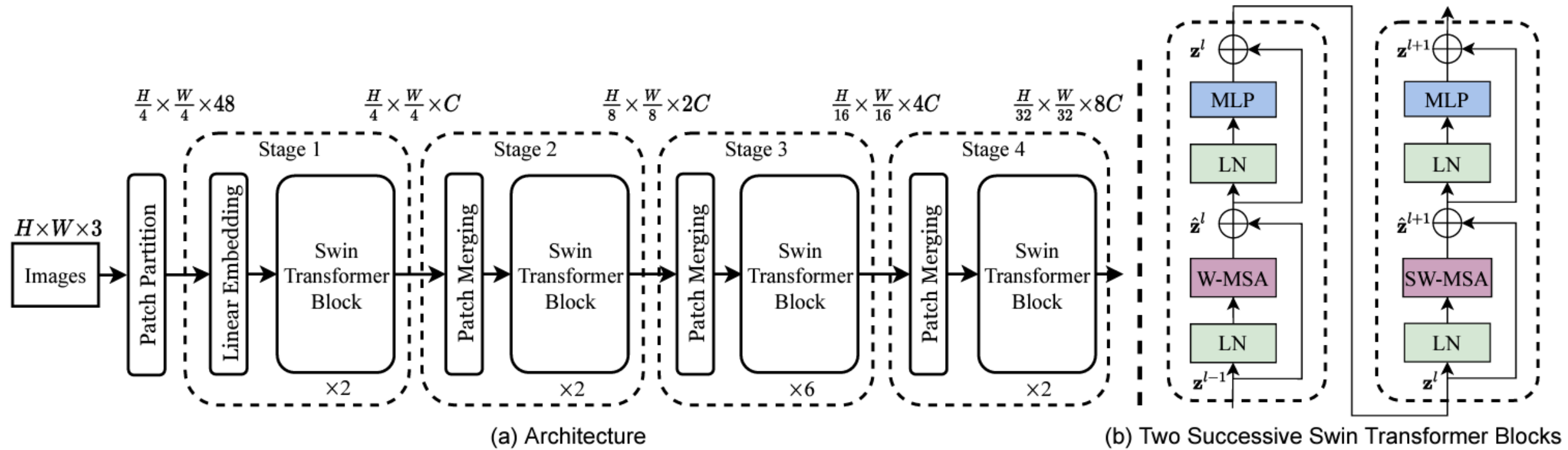


A local window to perform self-attention

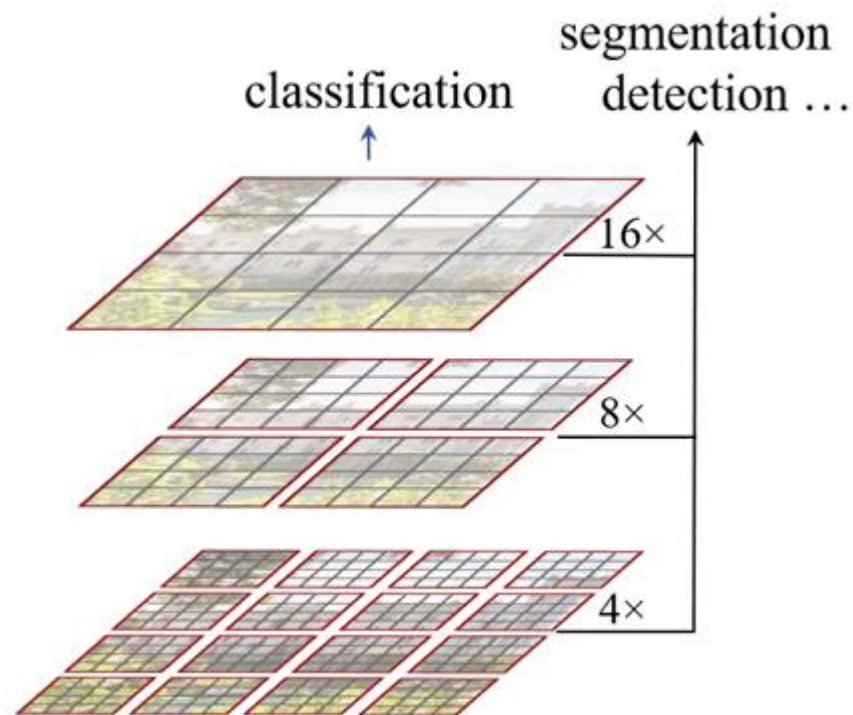


A patch

Shifted Window Transformer (Swin)



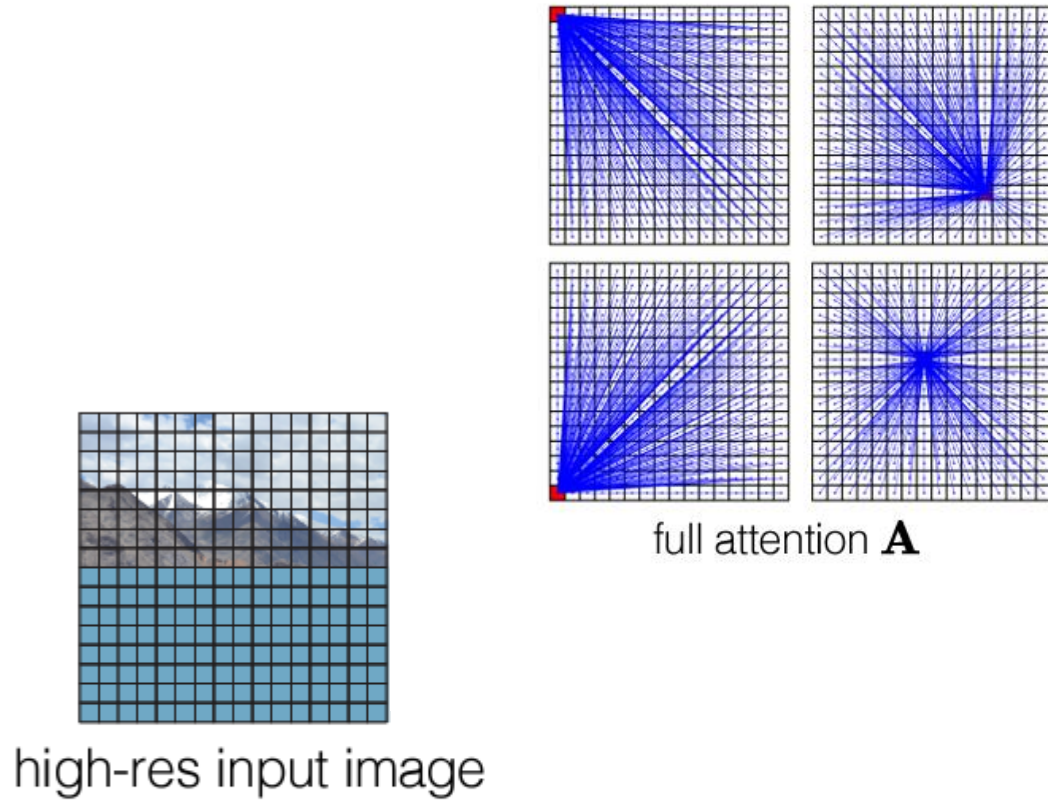
Shifted Window Transformer (Swin)



(a) Regular ImageNet-1K trained models

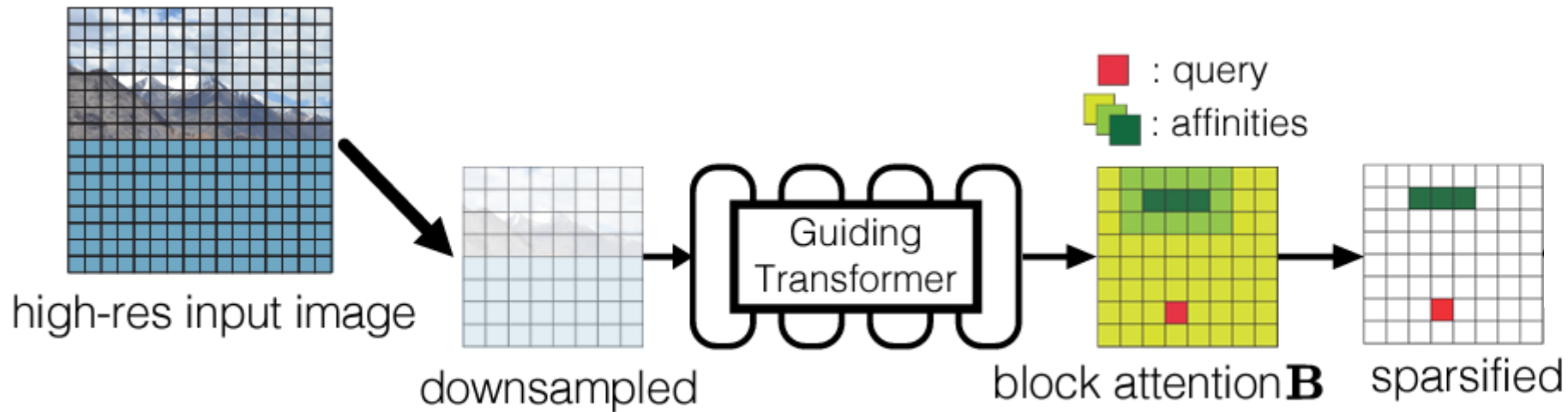
method	image size	#param.	FLOPs	throughput (image / s)	ImageNet top-1 acc.
ViT-B/16 [19]	384^2	86M	55.4G	85.9	77.9
ViT-L/16 [19]	384^2	307M	190.7G	27.3	76.5
Swin-T	224^2	29M	4.5G	755.2	81.3
Swin-S	224^2	50M	8.7G	436.9	83.0
Swin-B	224^2	88M	15.4G	278.1	83.5
Swin-B	384^2	88M	47.0G	84.7	84.5

Multiscale Vision Transformer

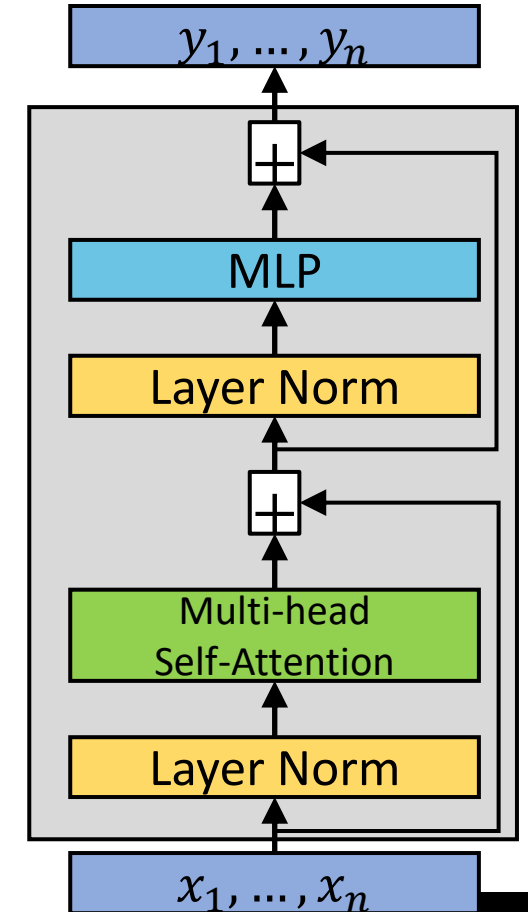
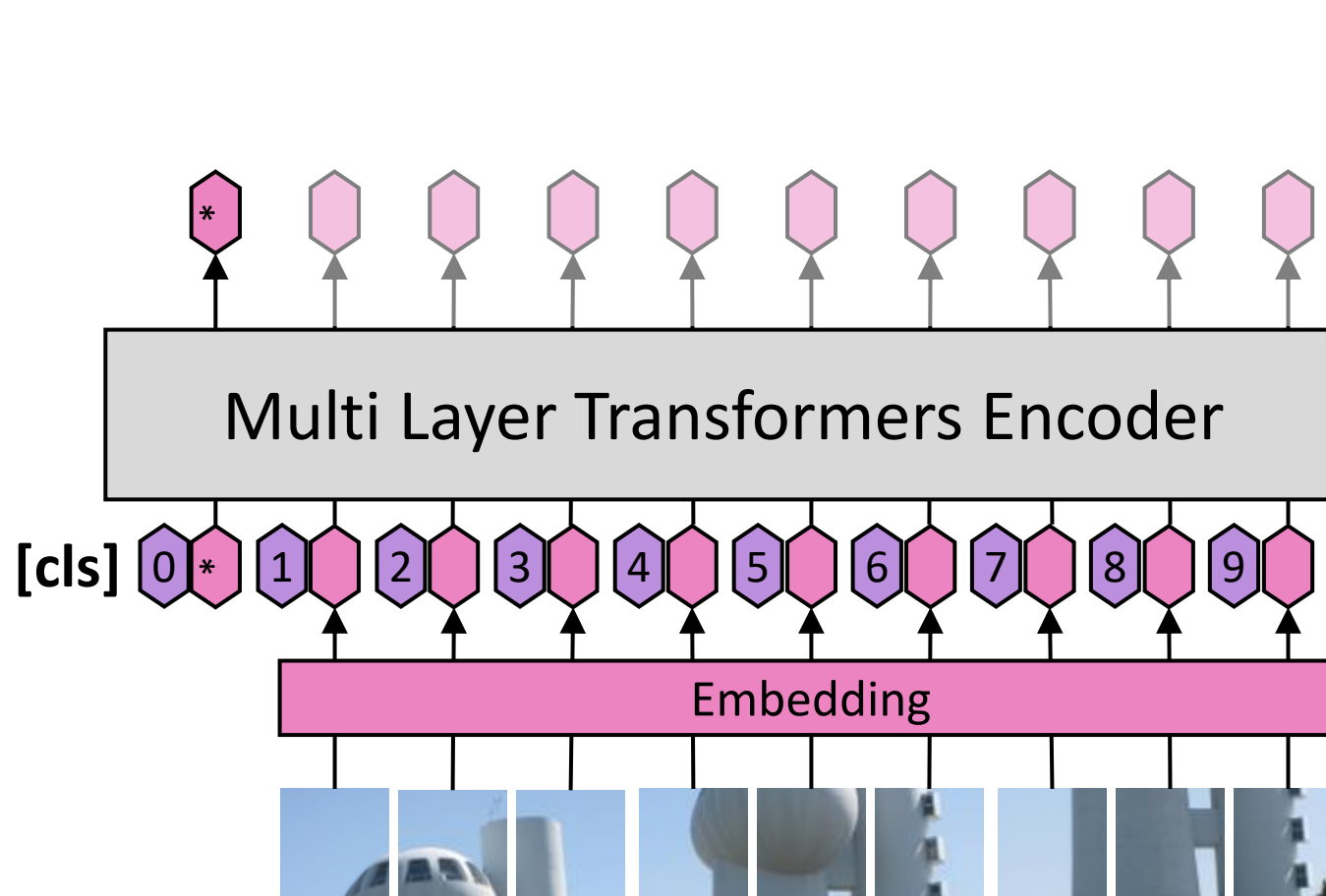


Multiscale Vision Transformer

Use multi-scale to derive smart sparse patterns

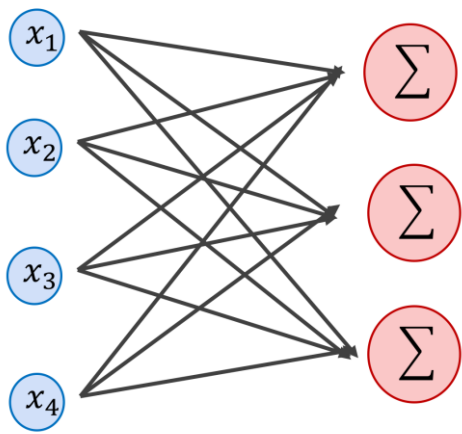


Vision Transformers (ViT)

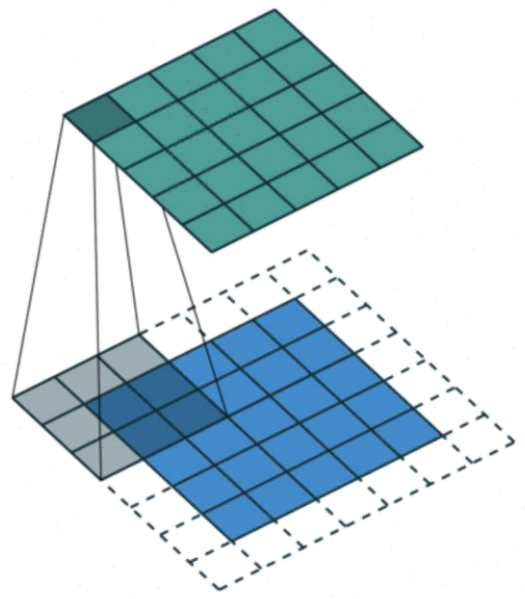


DL4CV: NN Building blocks

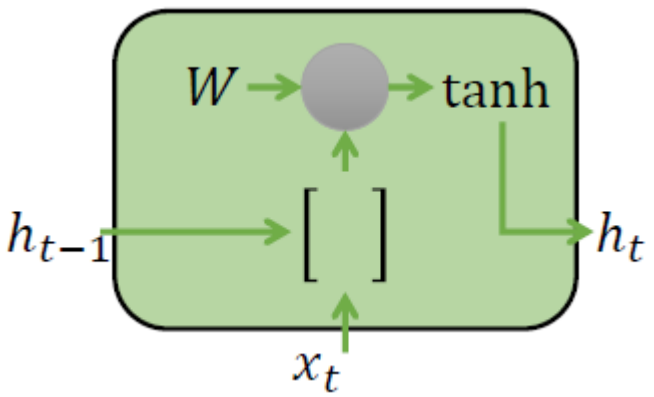
Linear Layer



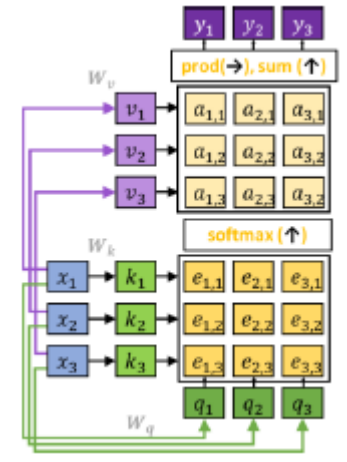
Convolution Layer



Recurrent Layer



Attention Layer



What's next?

Next lecture:

Generative Models (Assaf)

