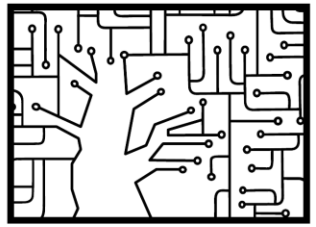
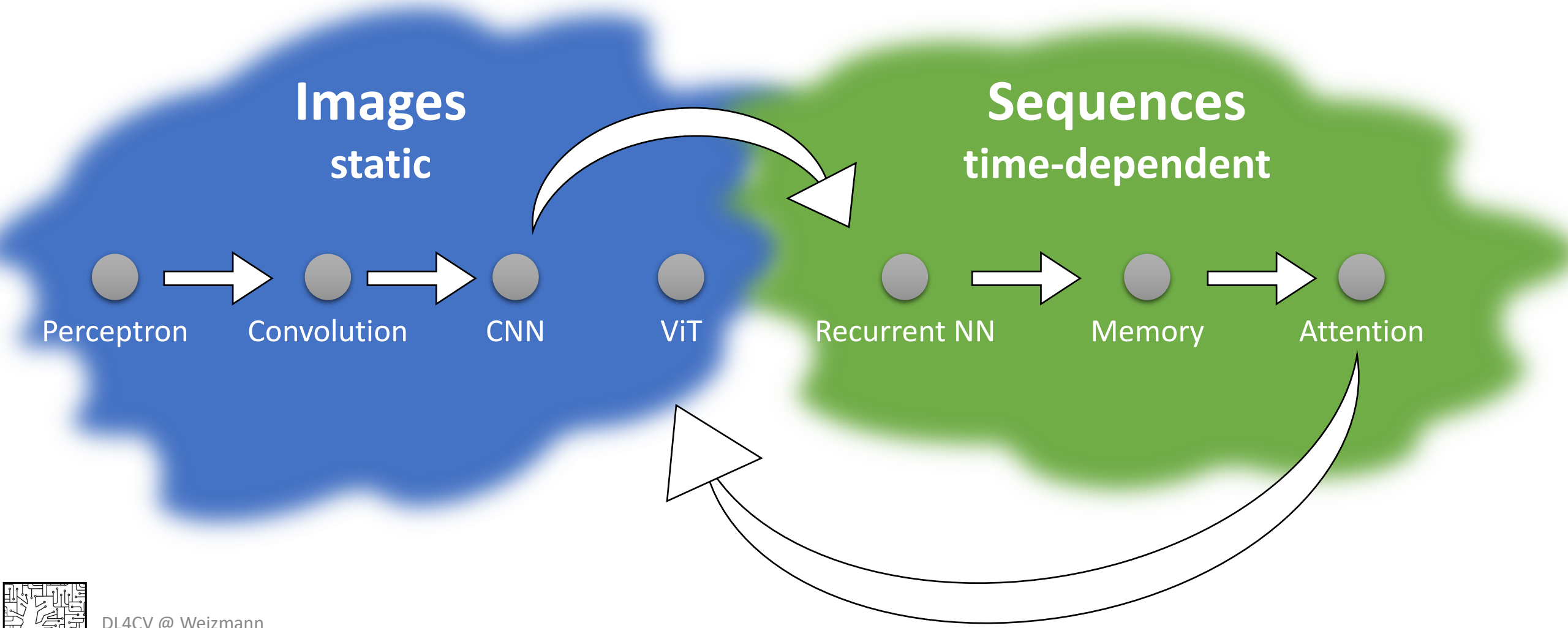
Deep Learning for Computer Vision:
From Sequences to ViT

Shai Bagon



Agenda


Images
static
Sequences
time-dependent
Perceptron
Convolution
CNN
ViT
Recurrent NN
Memory
Attention

Self-Attention Layer

Input:

Input: $\boldsymbol{X}$ (shape: $N_x \times D_x$)

Layer's Parameters:

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

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

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

Compute:

Query: $\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_q$ (shape: $N_x \times D_q$)

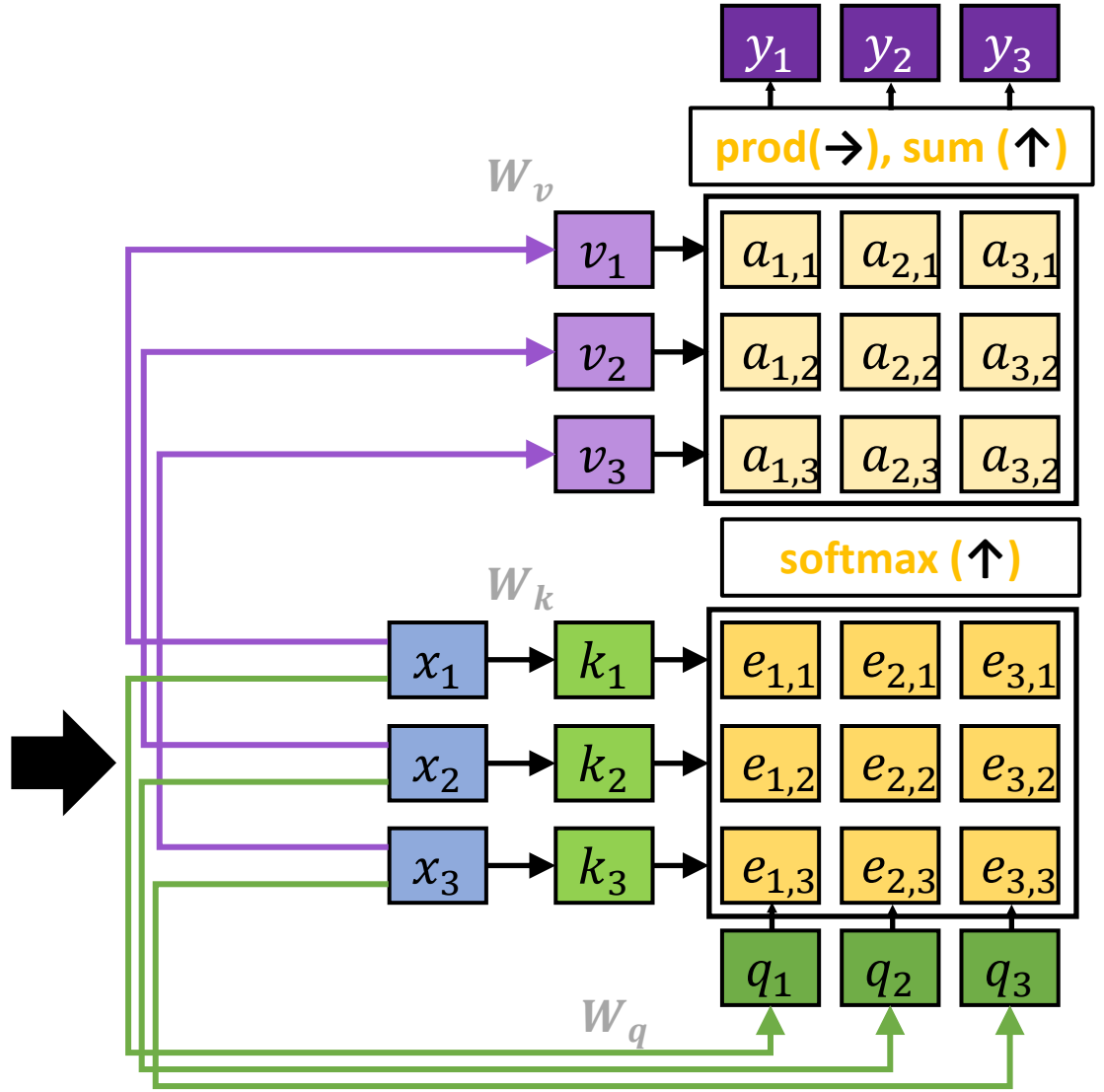
Keys: $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_k$ (shape: $N_x \times D_q$)

Values: $\mathbf{V} = \mathbf{X}\boldsymbol{W}_v$ (shape: $N_x \times D_v$)

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

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

Outputs: $\mathbf{Y} = \mathbf{A}\mathbf{V}$ (shape: $N_x \times D_v$)



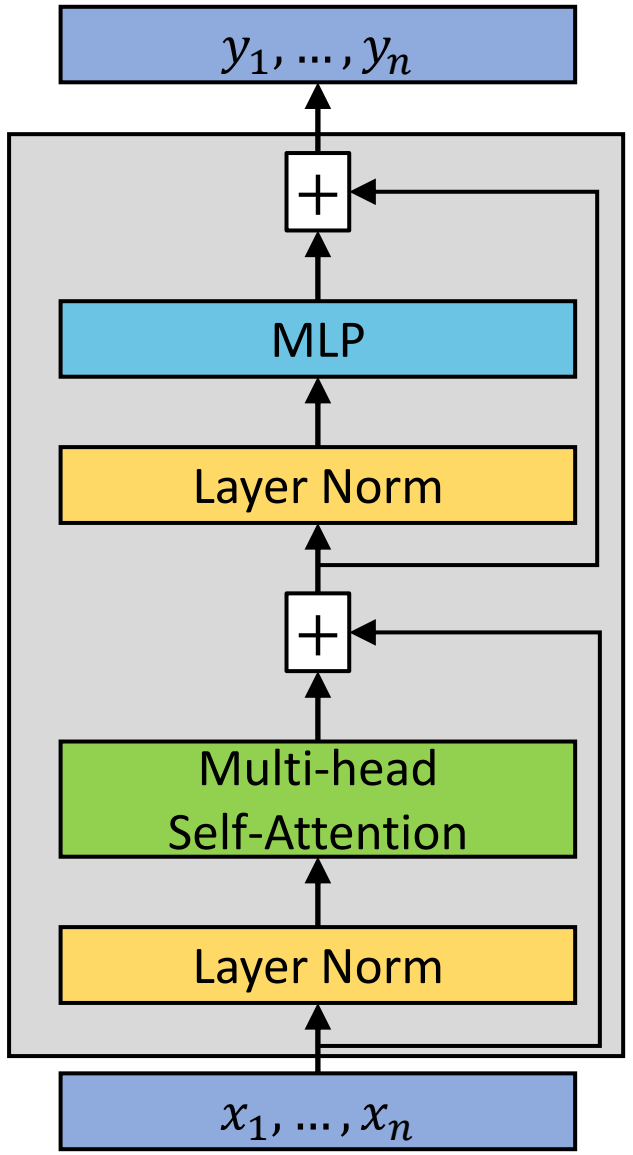
Slide credit: Justin Johnson (EECS-498-007, UMich)

Transformer Layer

Input: $x_1, \ldots, x_n$ (n tokens in D dimensions)
Output: $y_1, \ldots, y_n$ (n tokens in D dimensions)

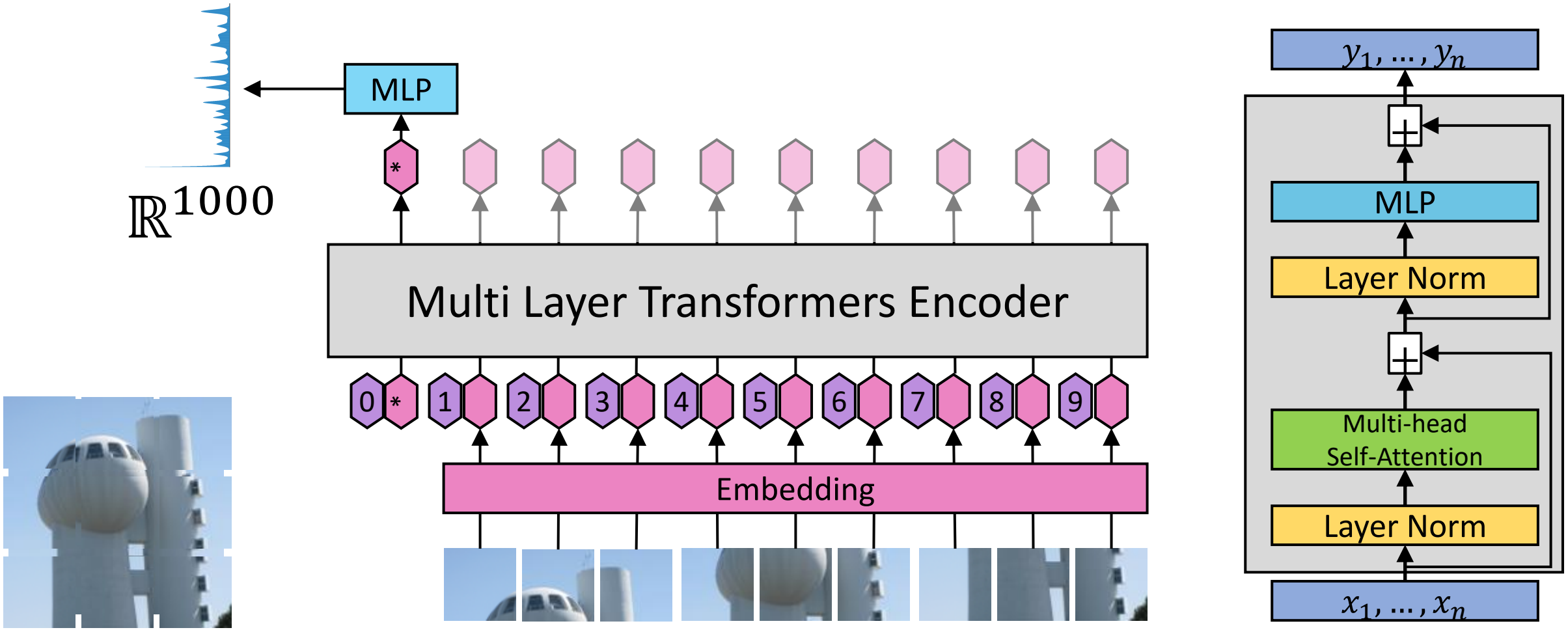
Highly scalable

Highly parallelizable



Slide credit: Justin Johnson (EECS-498-007, UMich)

Vision Transformers (ViT)

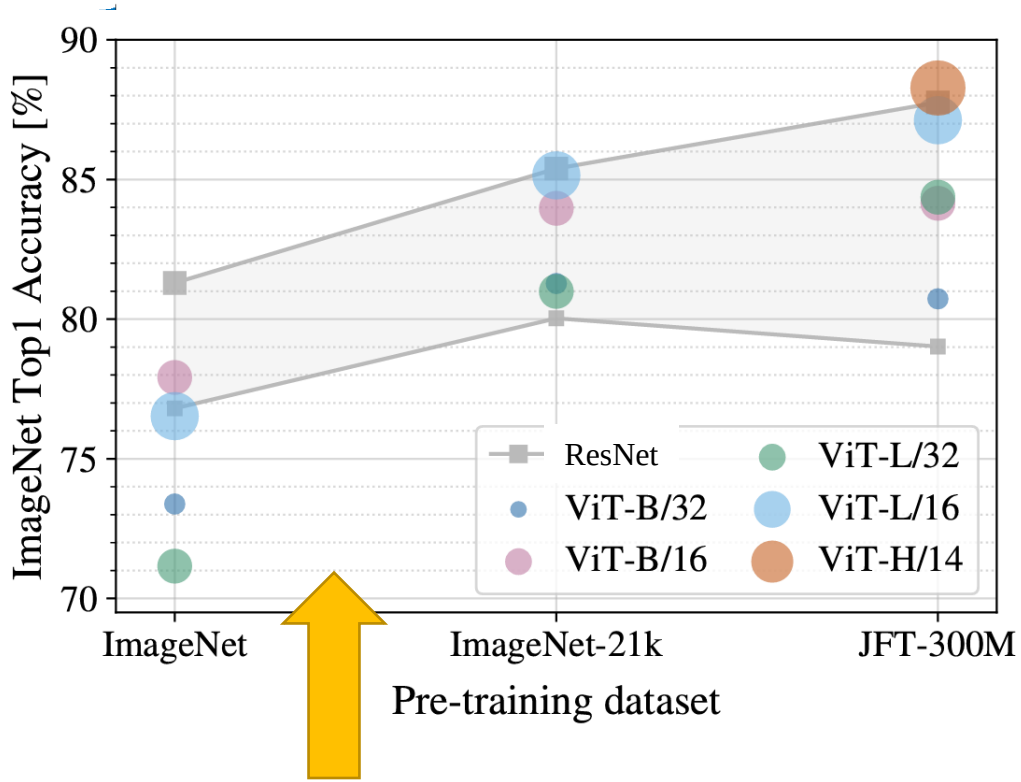


Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J. and Houlsby N. "**An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale**" (ICLR 2021)

Vision Transformers (ViT)

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

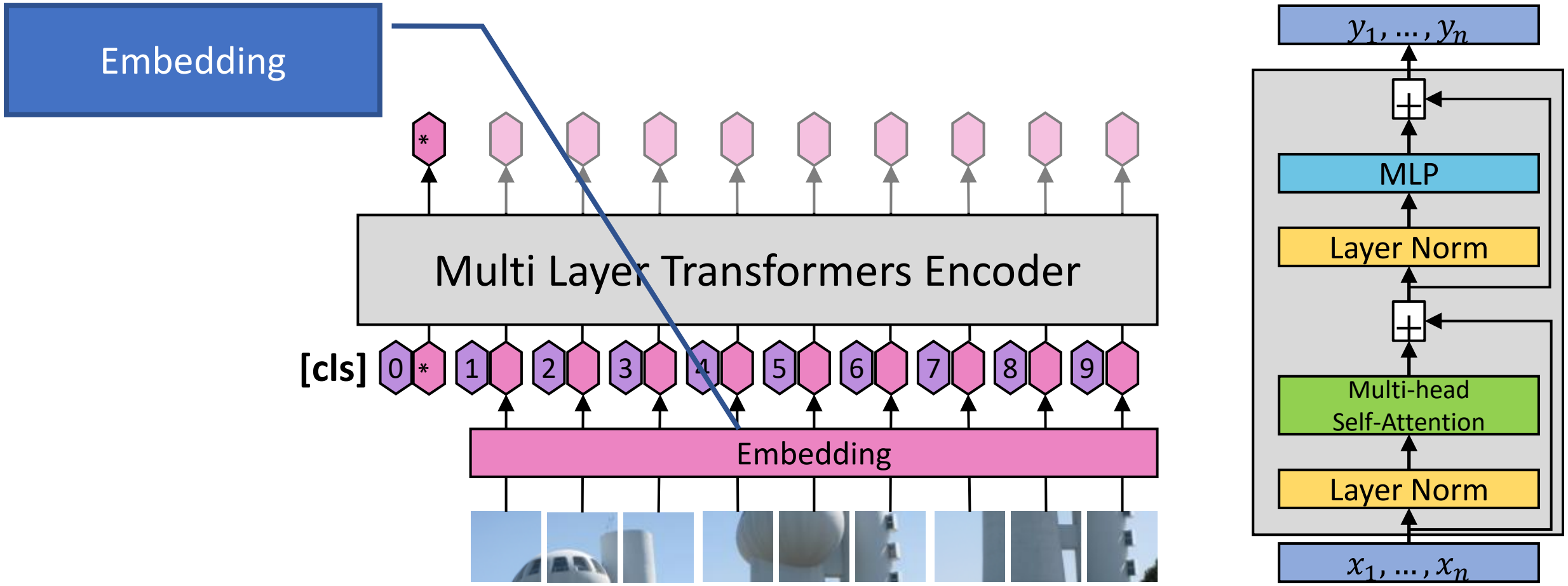
Naming convention: ViT-B/16



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)

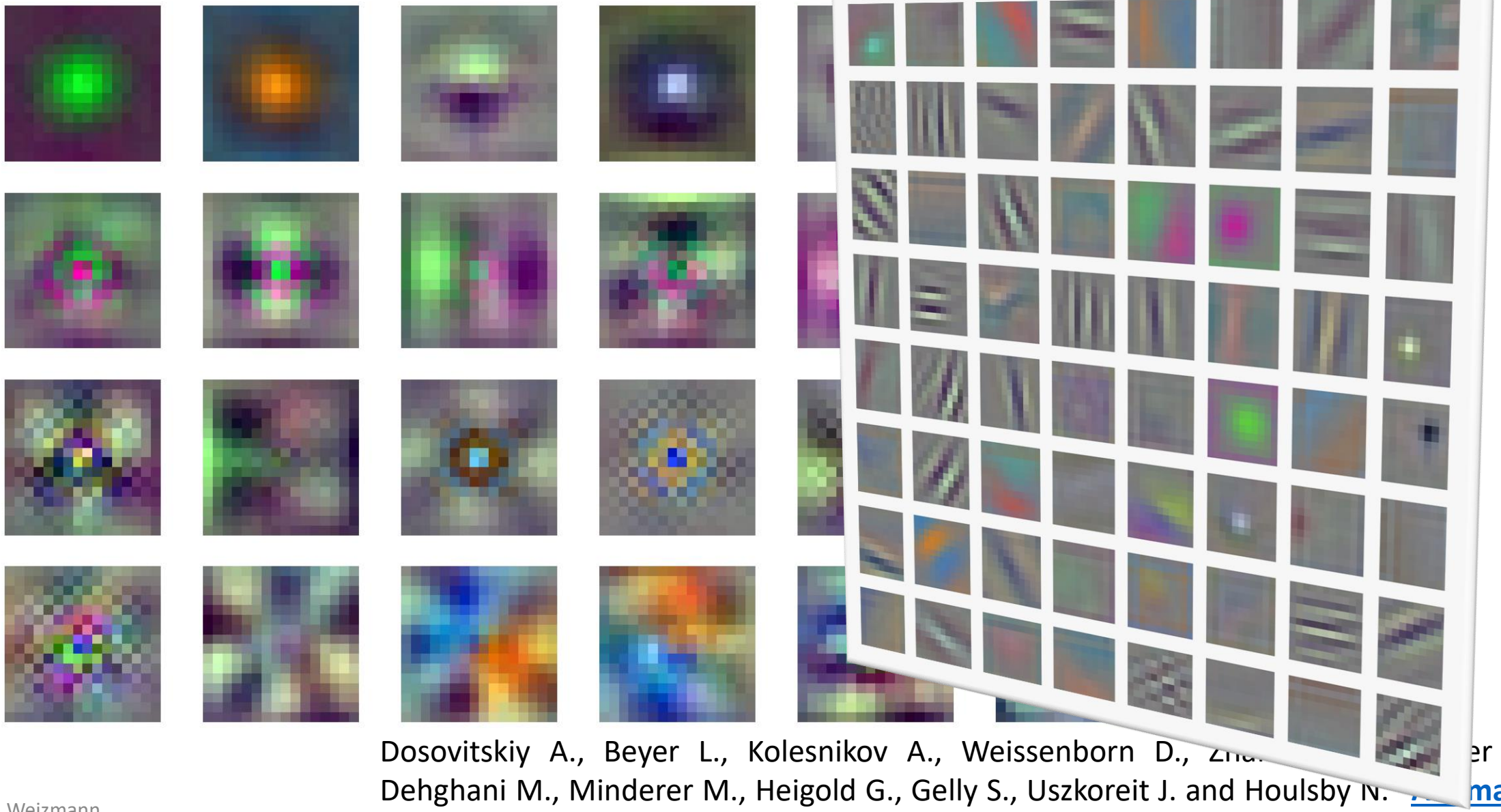
Vision Transformers (ViT) - Properties

Vision Transformers (ViT)



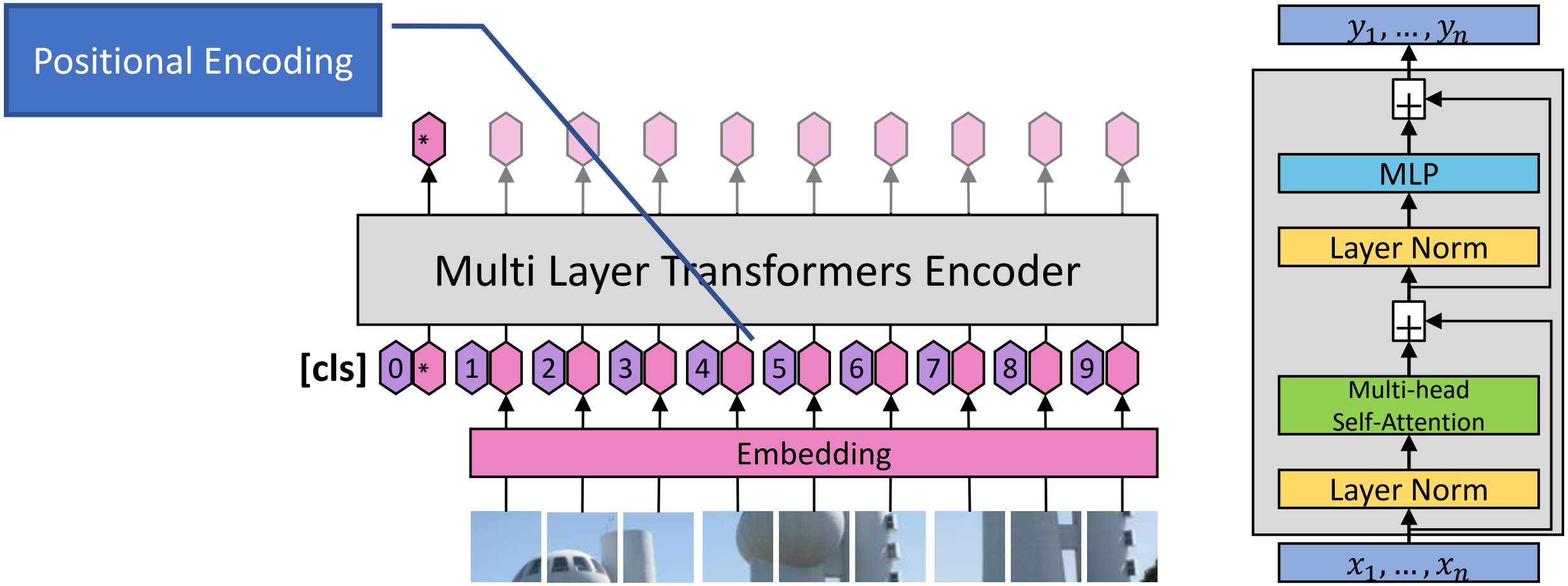
Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J. and Houlsby N. "**An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale**" (ICLR 2021)

Vision Transformers (ViT)



Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zha… …er T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J. and Houlsby N. "A… …mage is Worth 16x16 Words: Transformers for Image Recognition at Scale" (ICLR 2021)

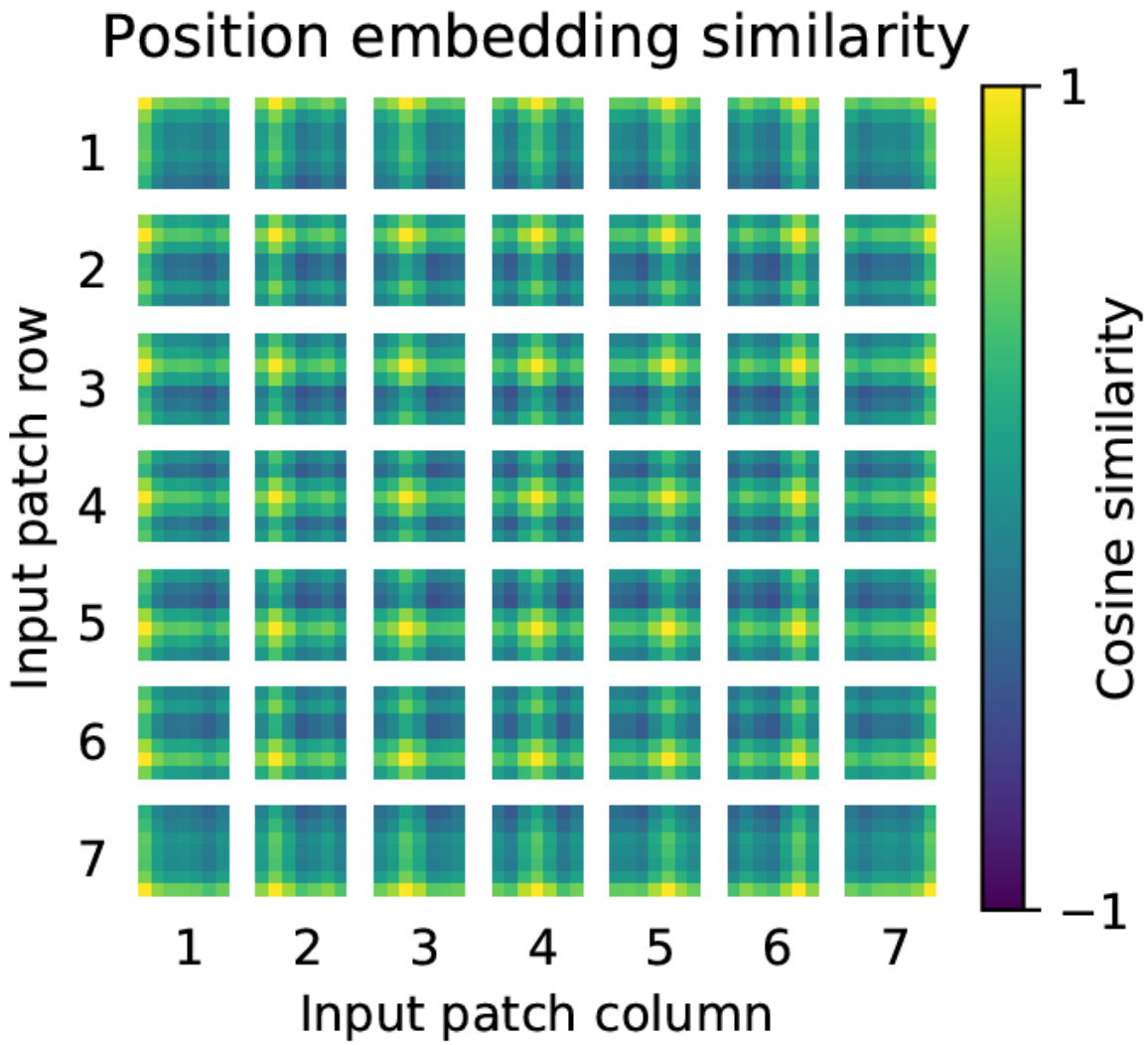
Vision Transformers (ViT)



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)

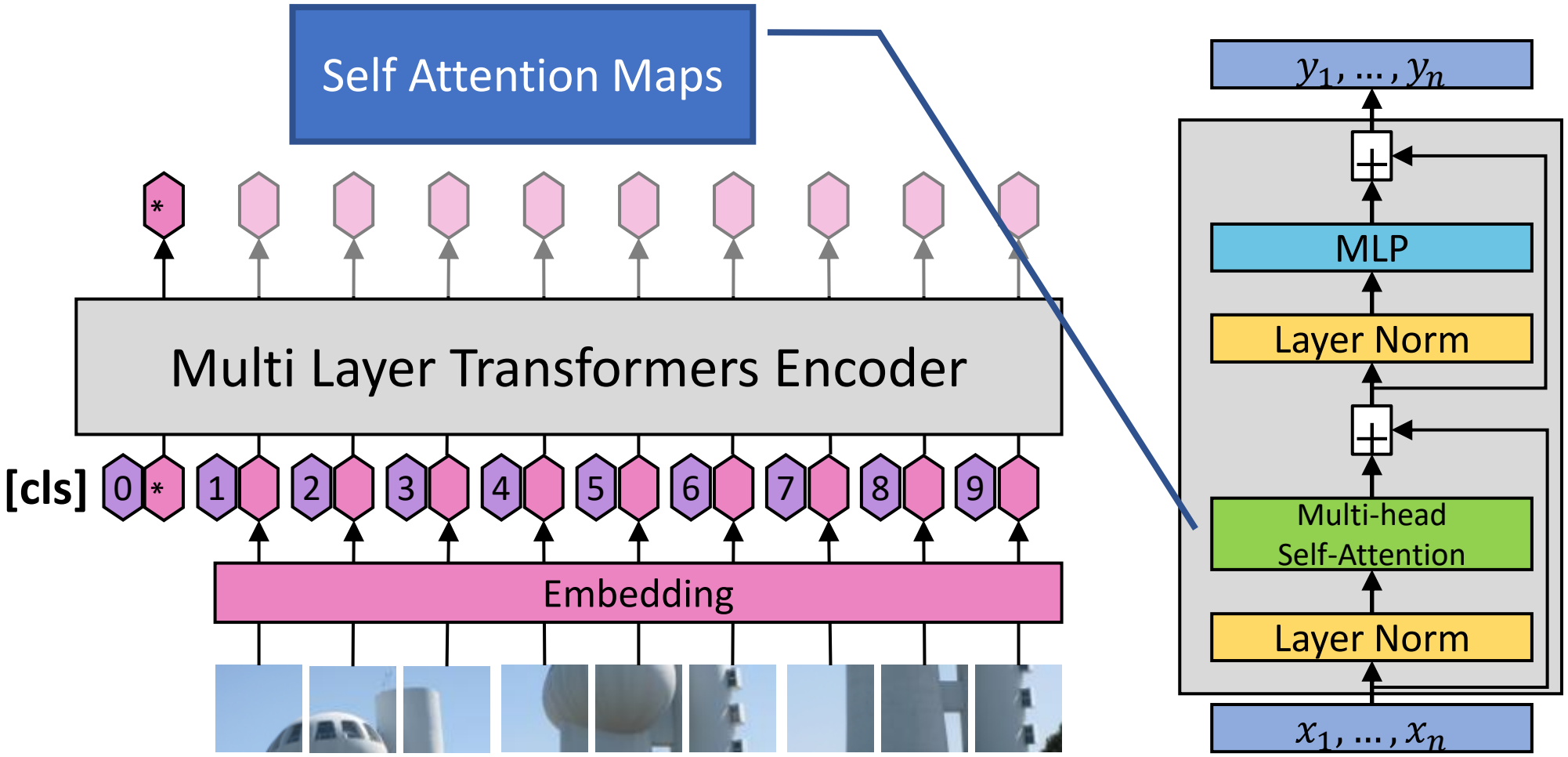
Vision Transformers (ViT)

ViT-B/32:
7x7 patches

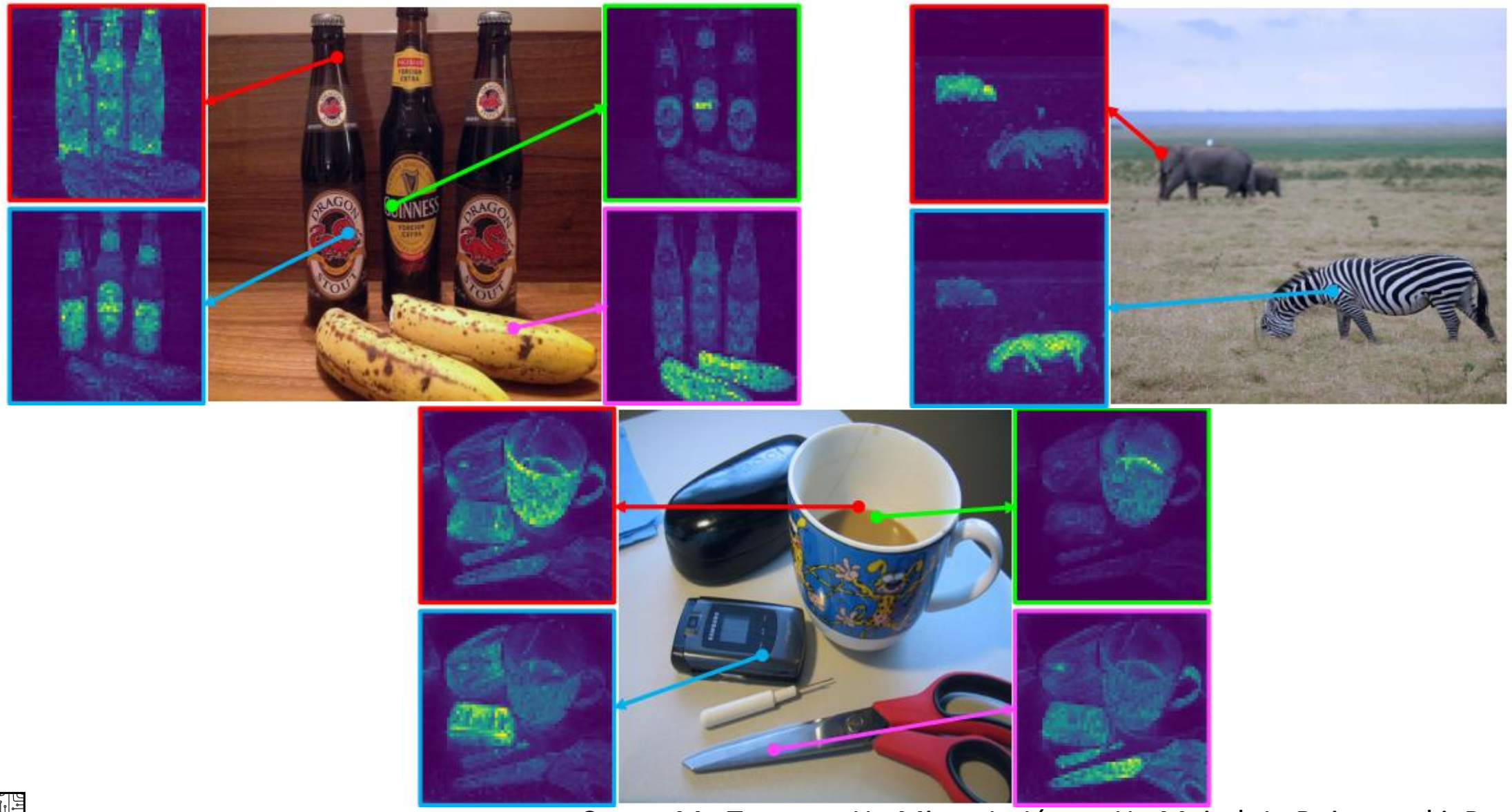


Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J. and Houlsby N. "**An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale**" (ICLR 2021)

Vision Transformers (ViT)

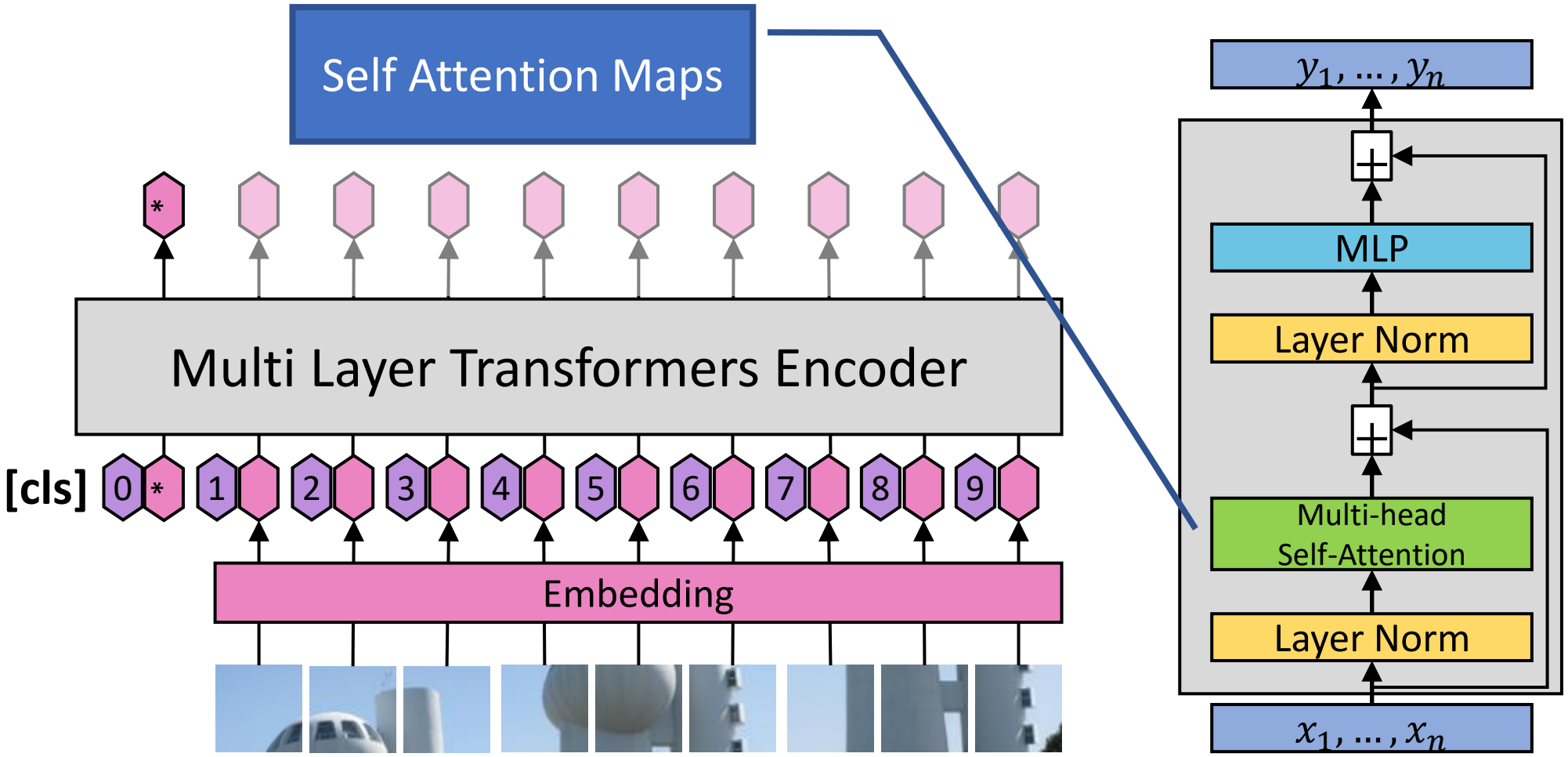


Vision Transformers (ViT)

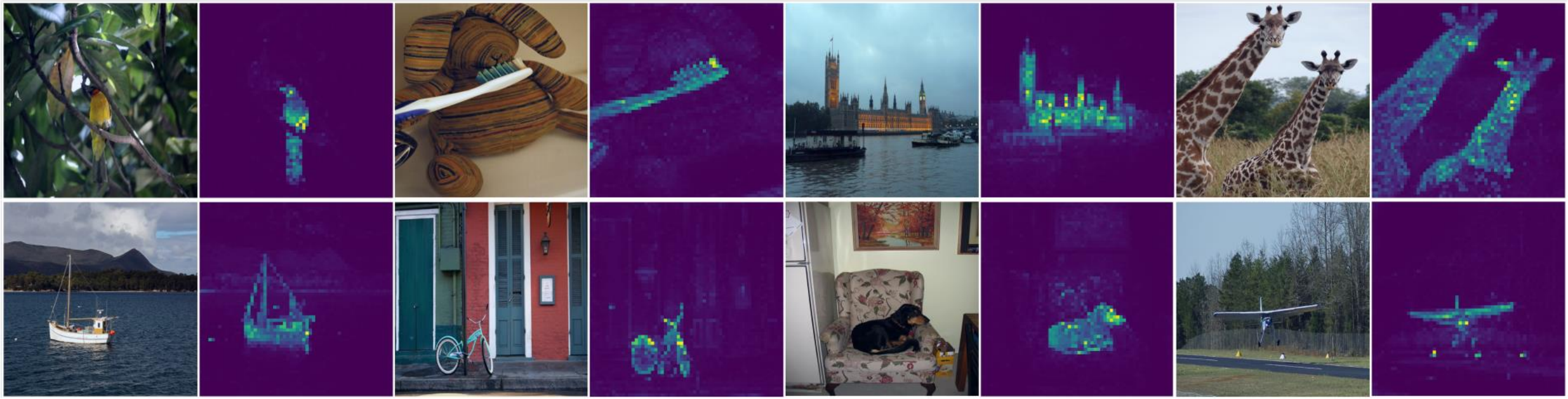


Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P. and Joulin, A., "Emerging Properties in Self-Supervised Vision Transformers". ICCV (2021)

Vision Transformers (ViT)

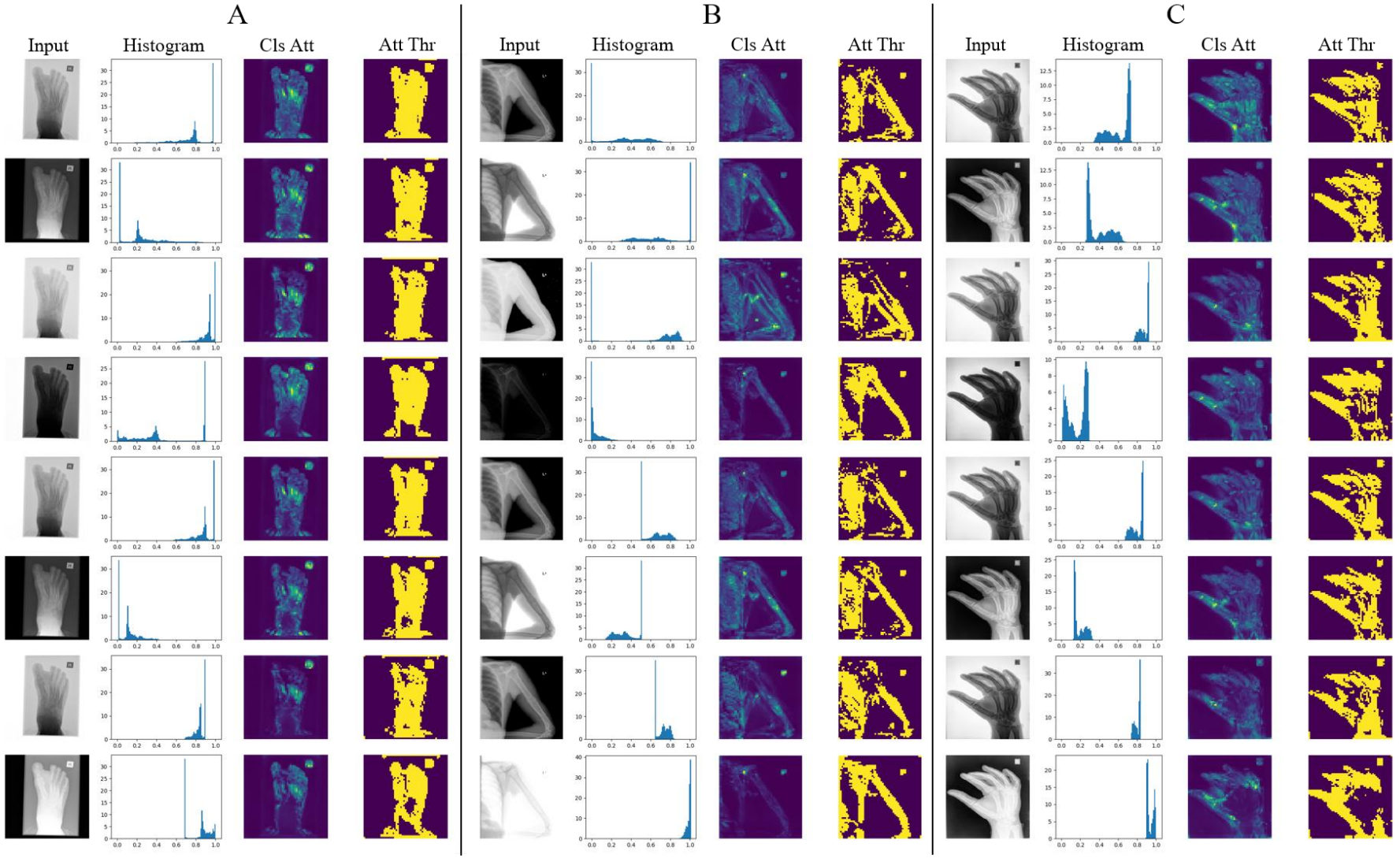


Vision Transformers (ViT)



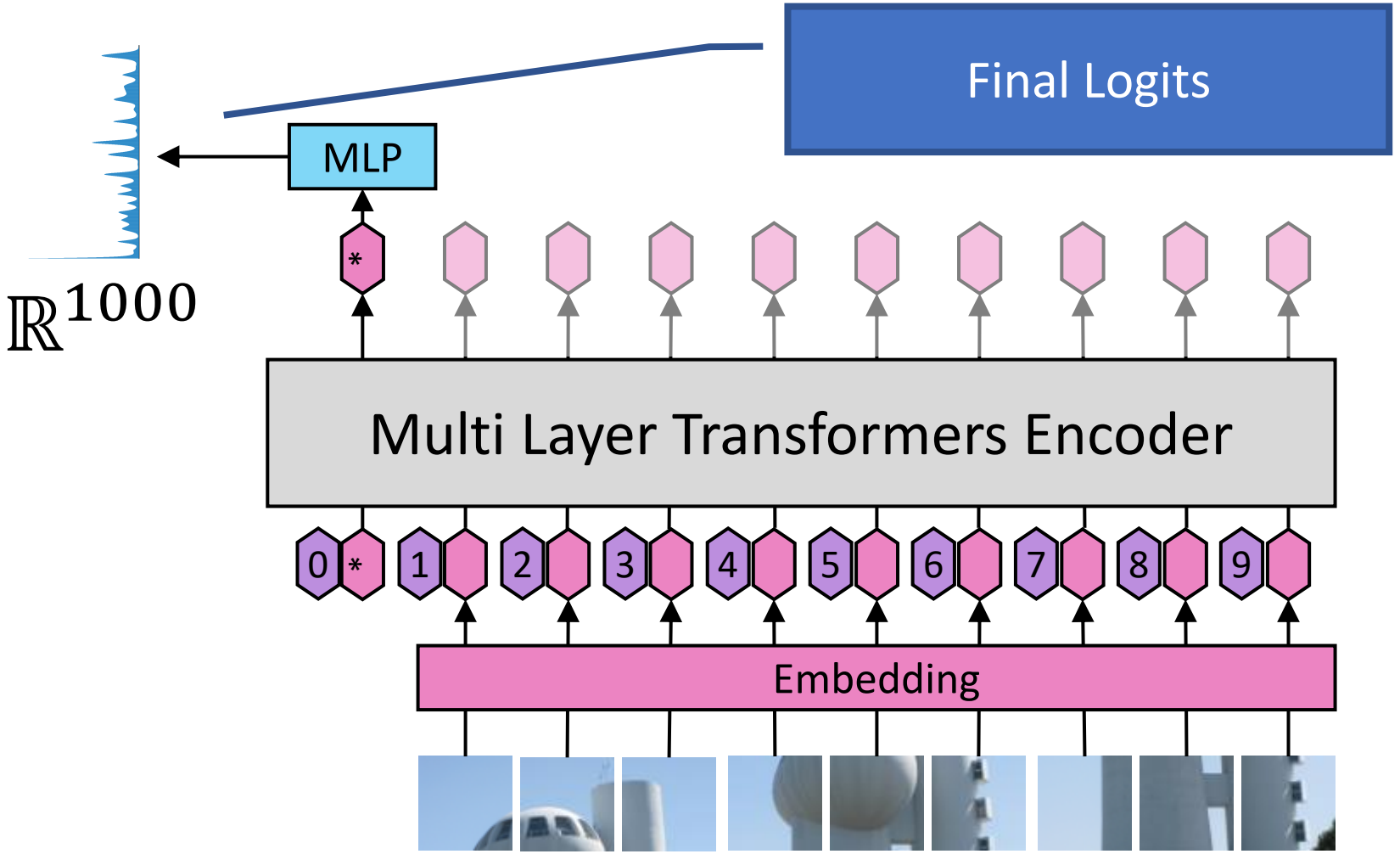
Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P. and Joulin, A., "Emerging Properties in Self-Supervised Vision Transformers". ICCV (2021)

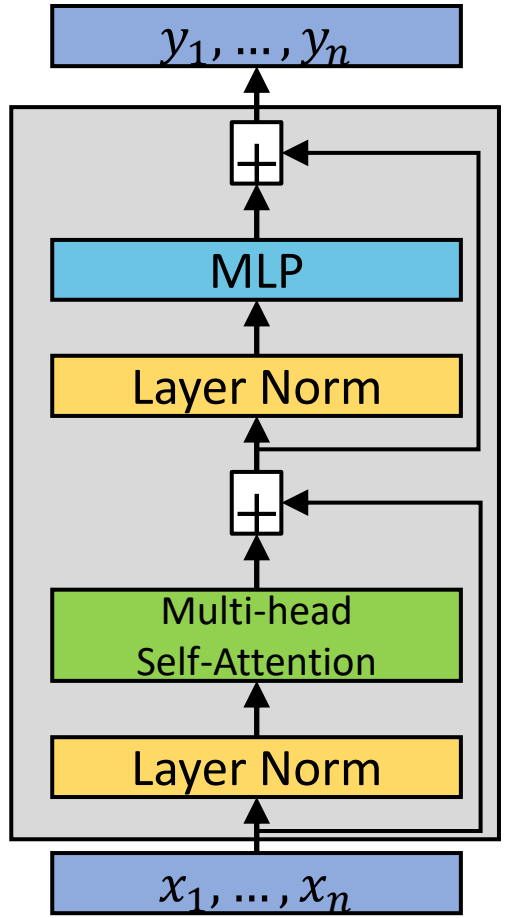
Vision Transformers (ViT)



Tal Zimbalist , Keren Peri-Hanania, Ronnie Rosen, Yaron Caspi, Bar Rinott, Carmel Zeltser Dekel, Eyal Bercovich, Yonina C. Eldar, Shai Bagon "**Detecting Bone Lesions in X-Ray Under Diverse Acquisition Conditions**" (Under review)

Vision Transformers (ViT)

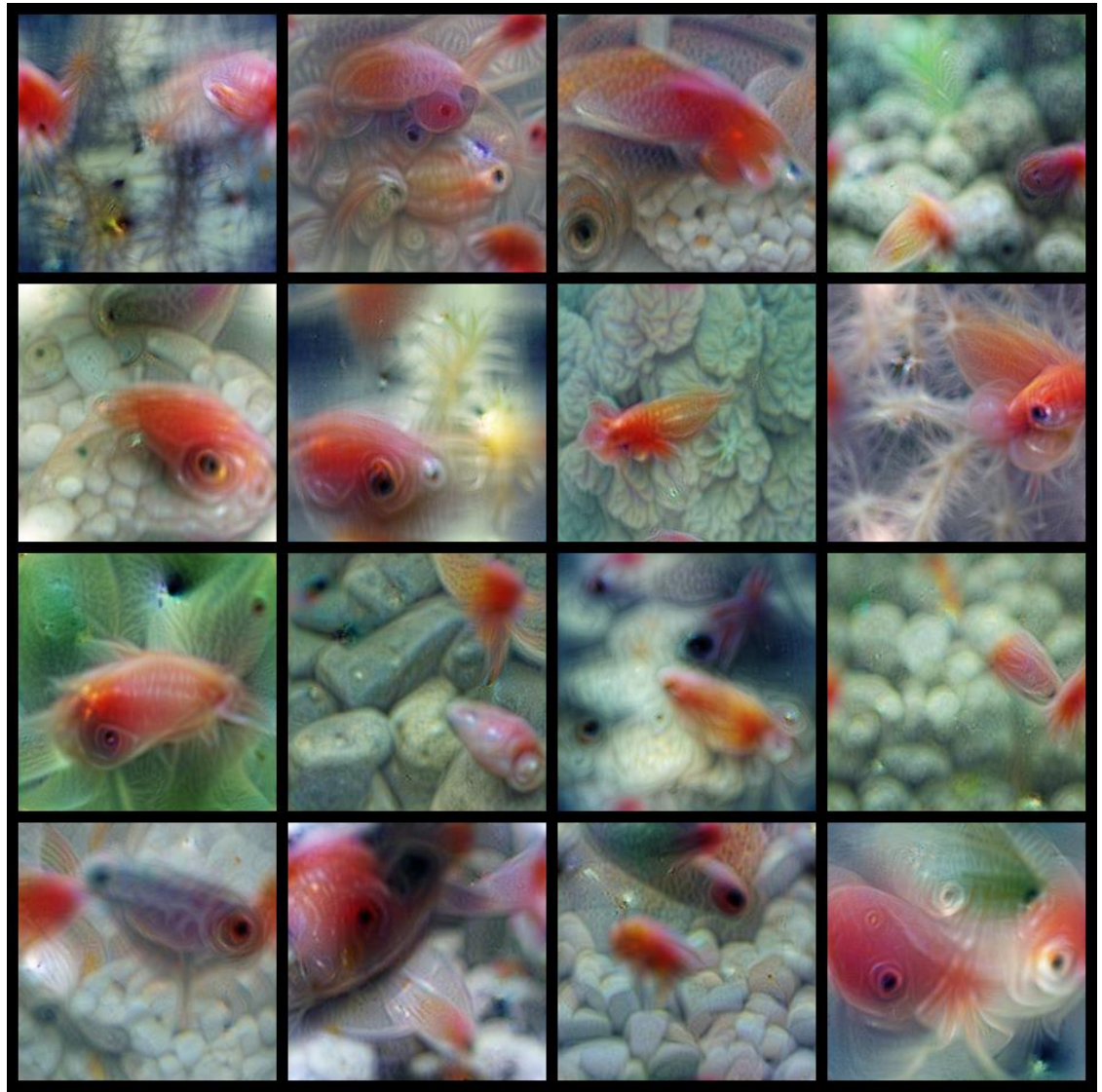




Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J. and Houlsby N. "**An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale**" (ICLR 2021)

Vision Transformers (ViT)

logits for class “gold fish”



Vision Transformers (ViT)

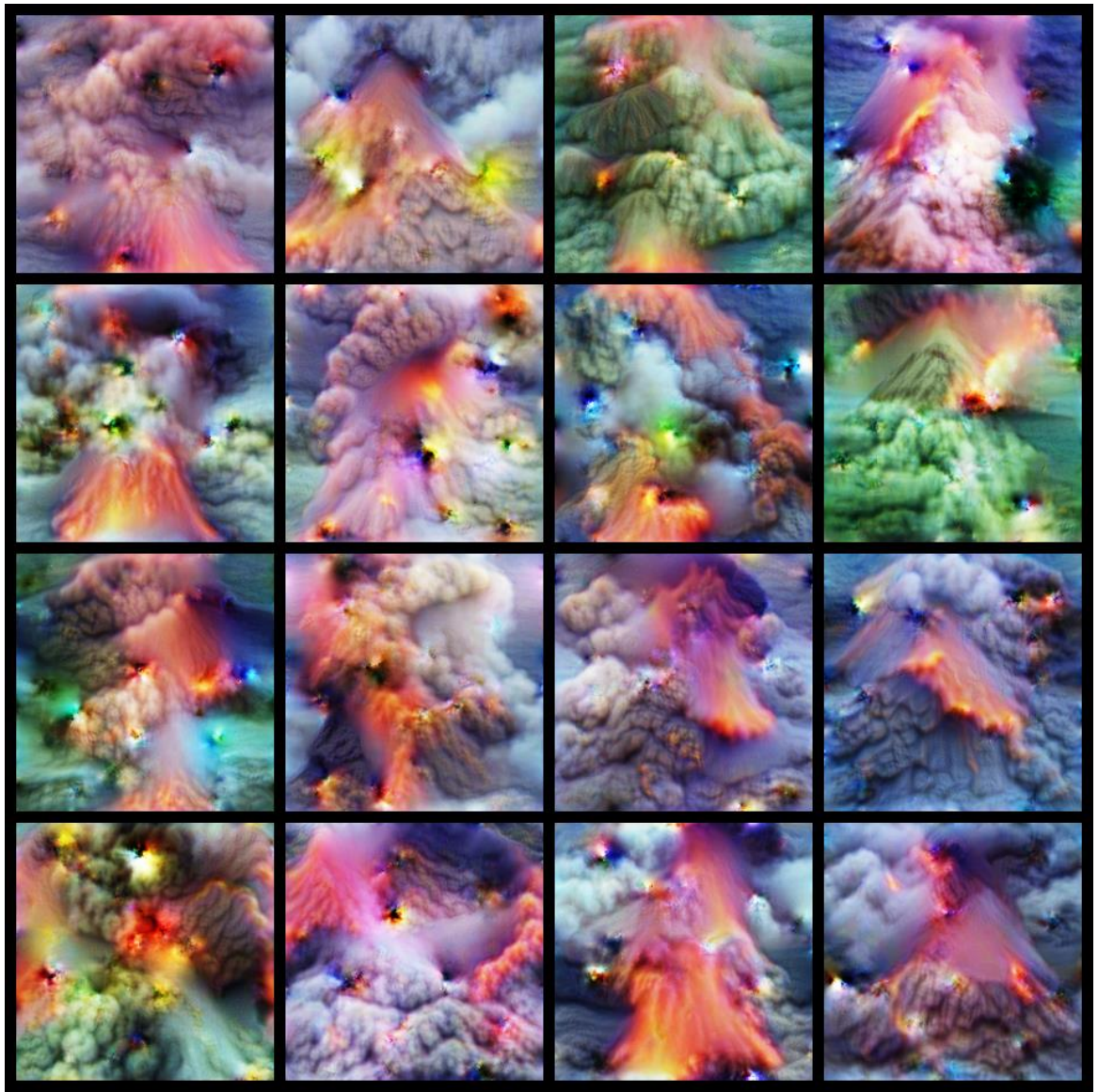
logits for class “bald eagle”

Vision Transformers (ViT)

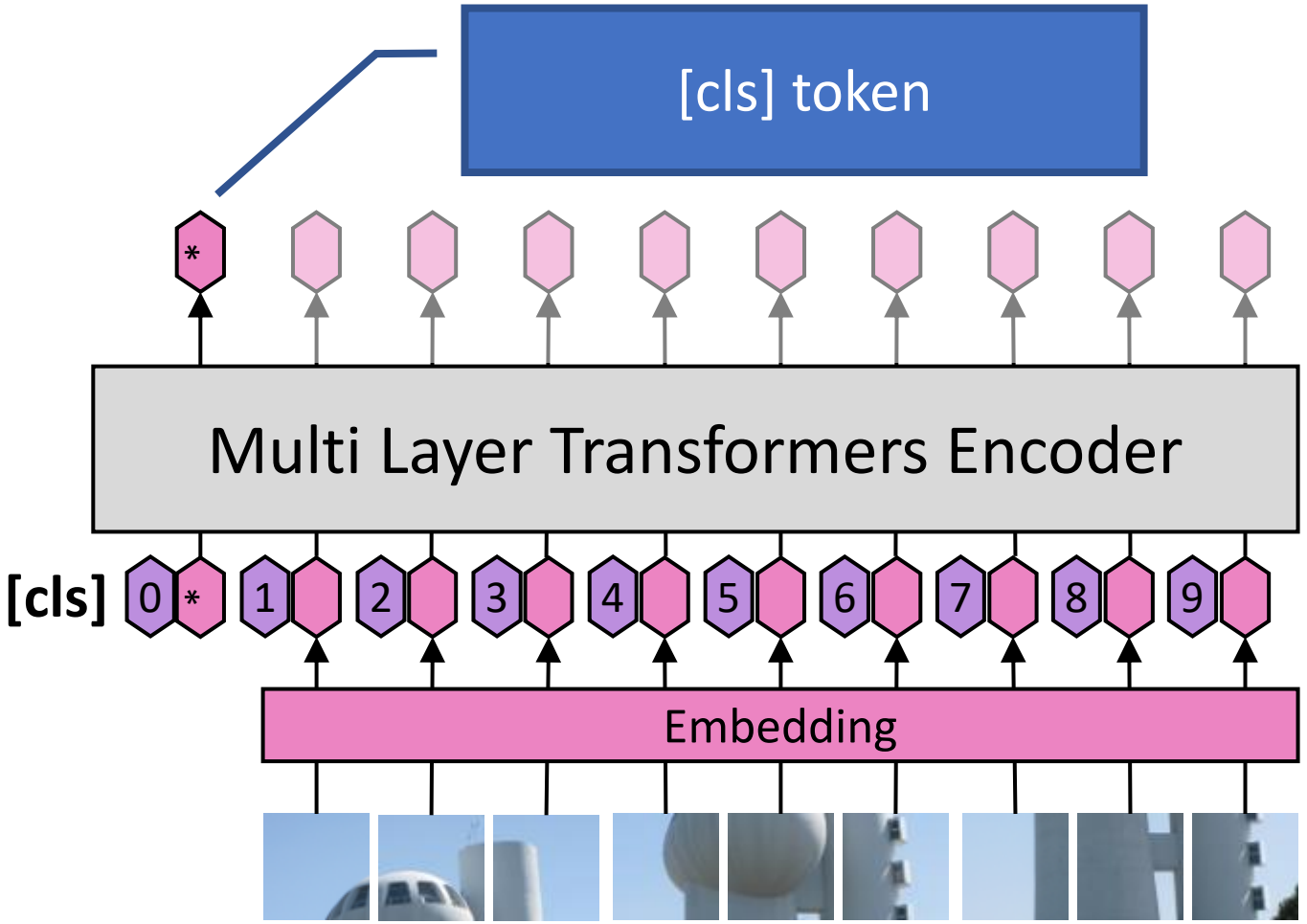
logits for class “curly coated retriever”

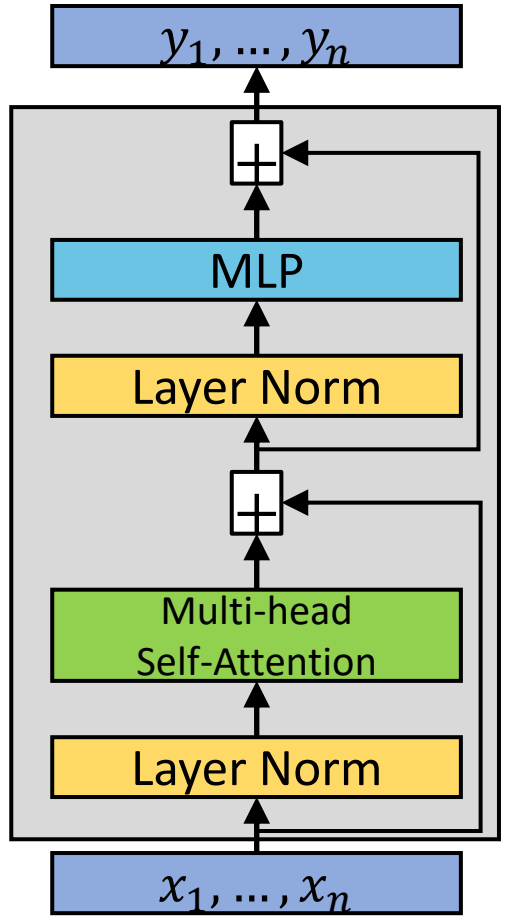
Vision Transformers (ViT)

logits for class “volcano”



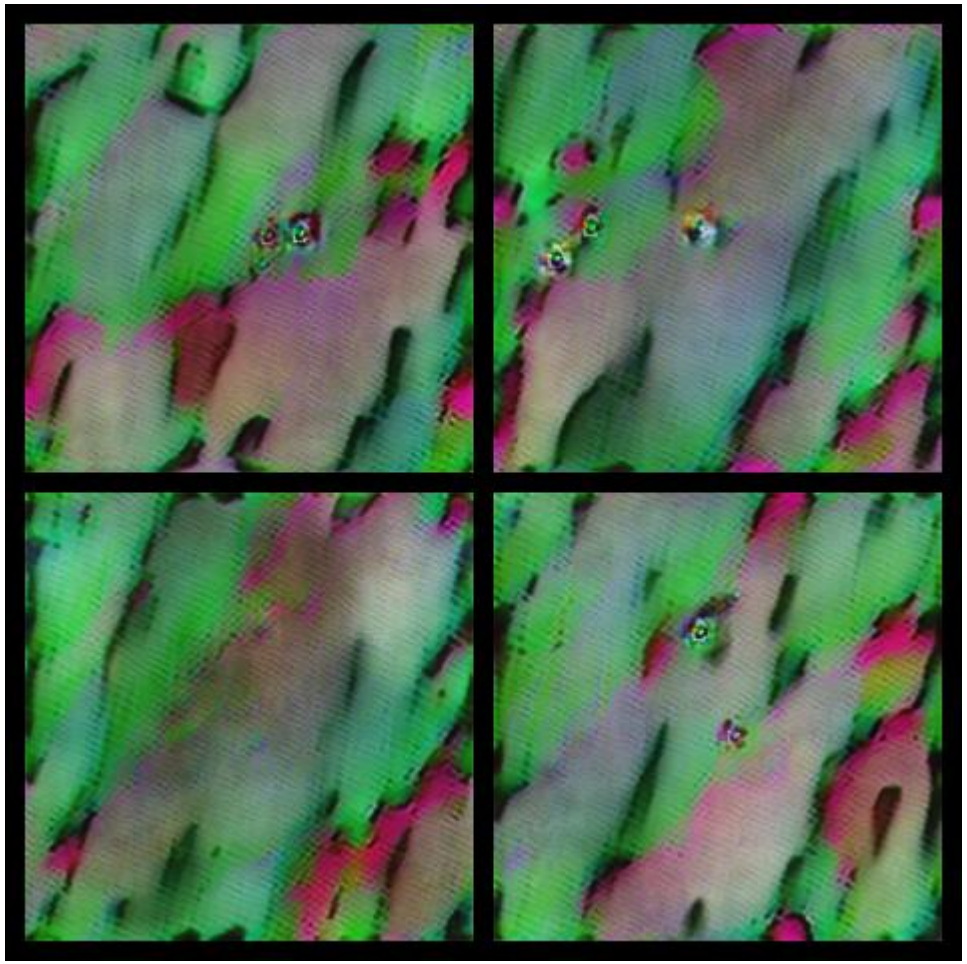
Vision Transformers (ViT)





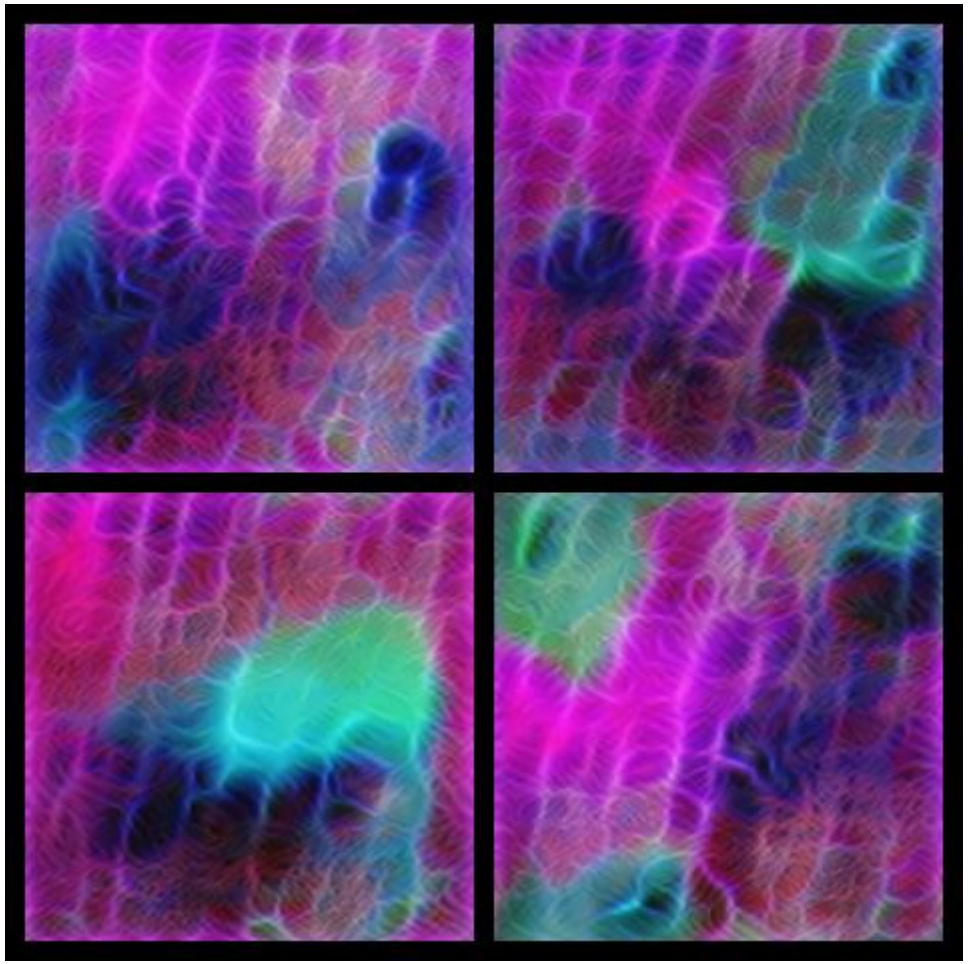
Vision Transformers (ViT)

[cls] token layer=1



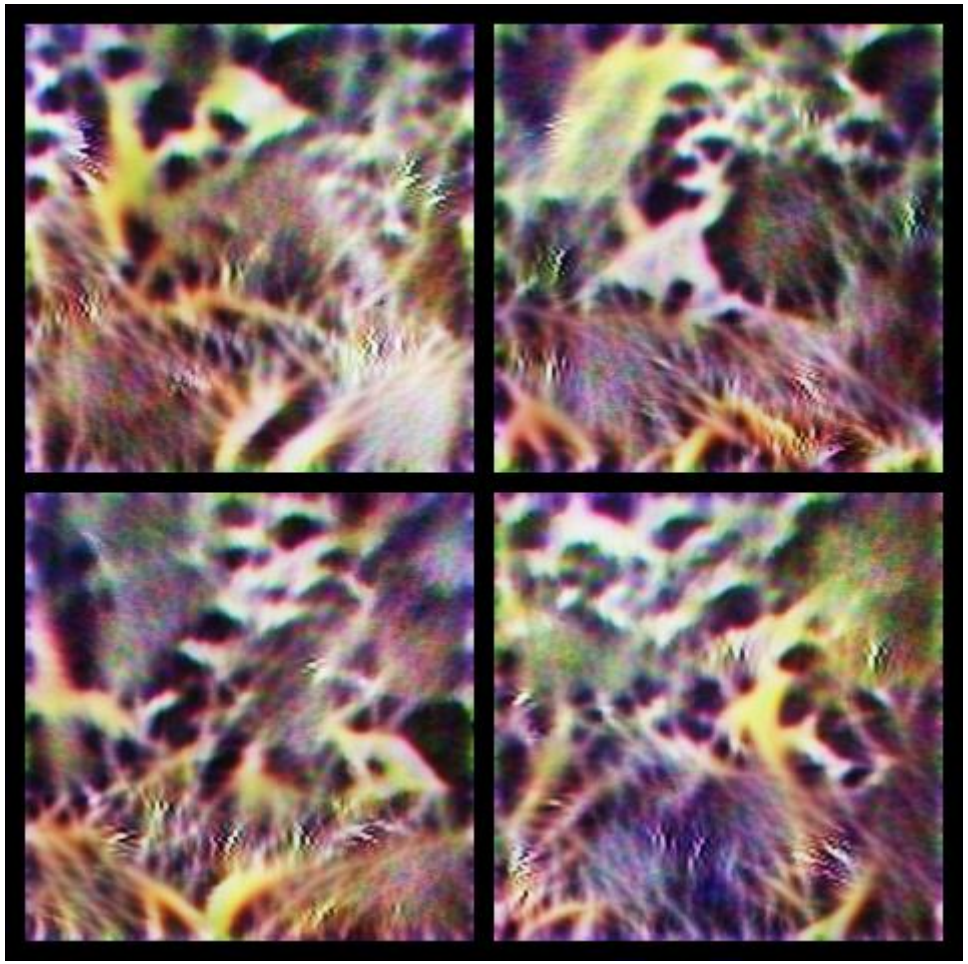
Vision Transformers (ViT)

[cls] token layer=3



Vision Transformers (ViT)

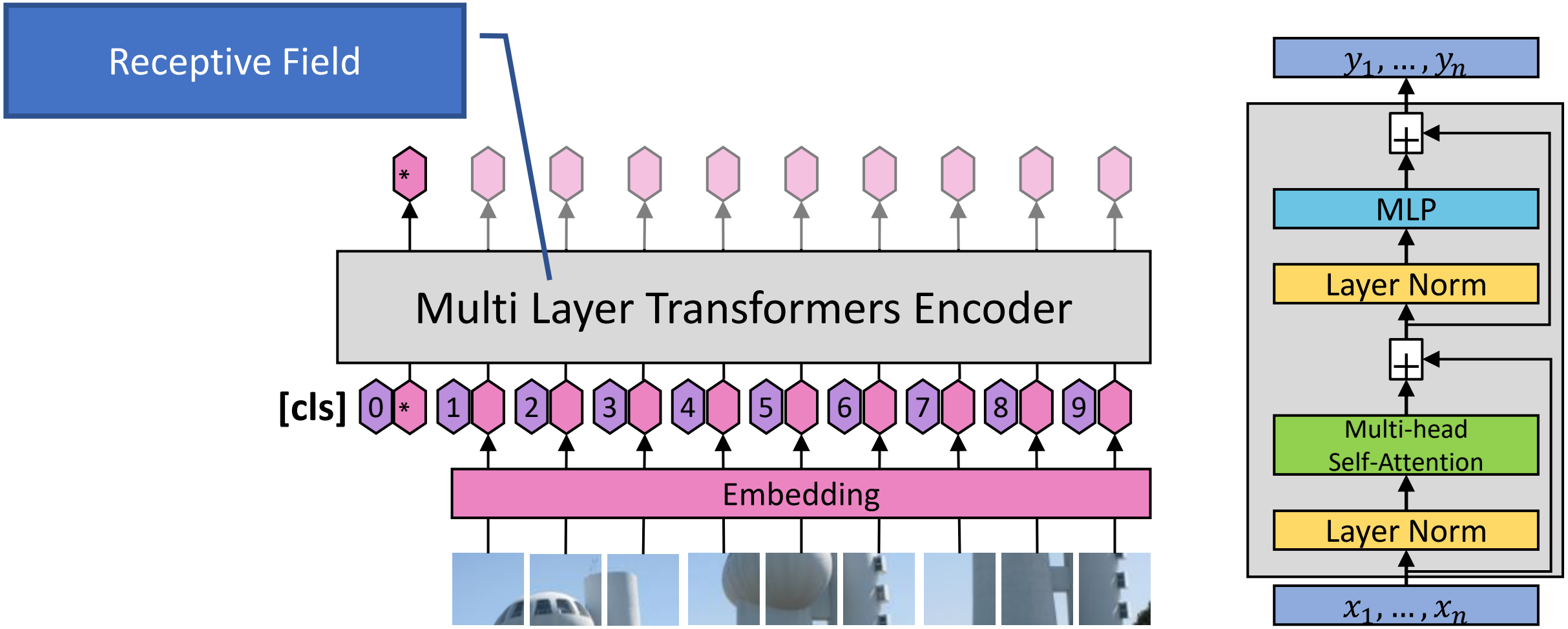
[cls] token layer=5



Vision Transformers (ViT)

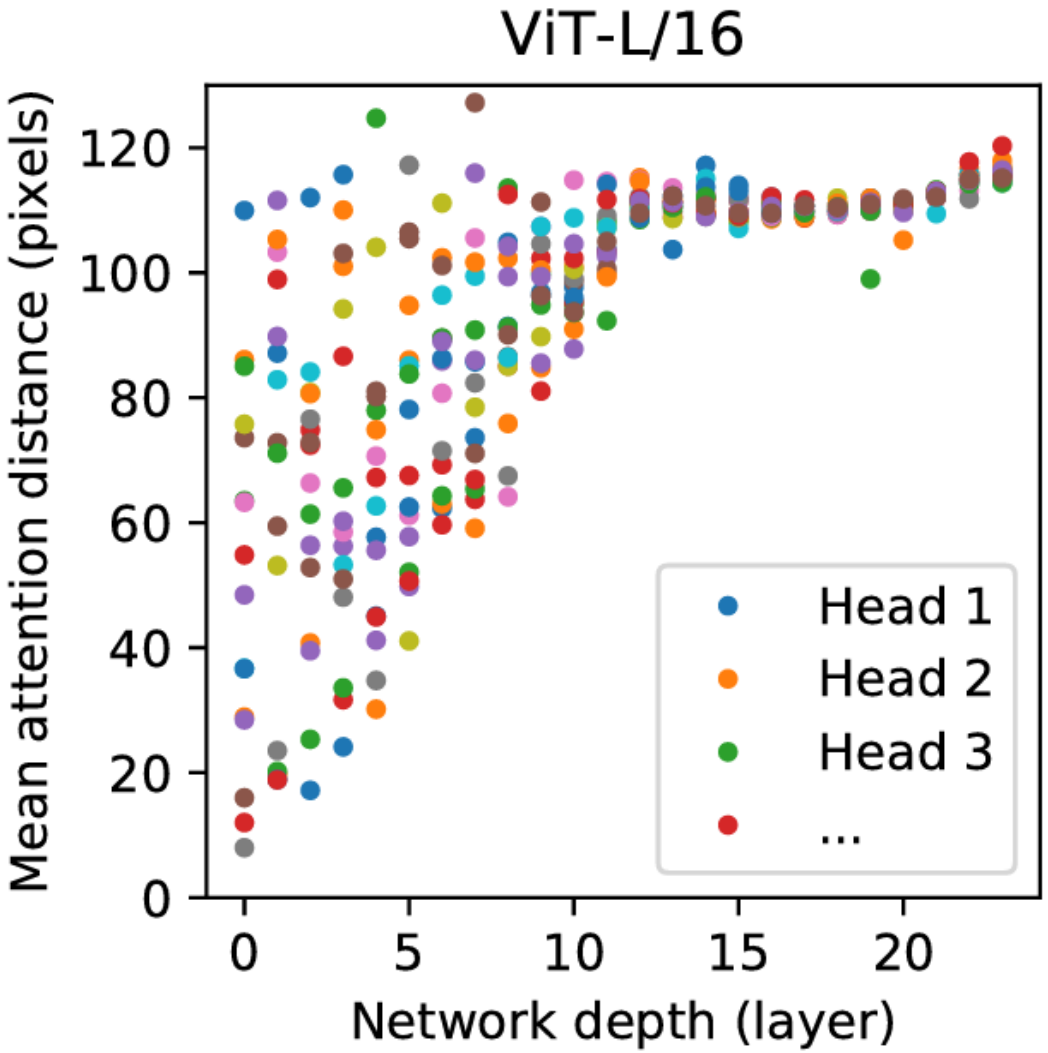
[cls] token layer=10

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)

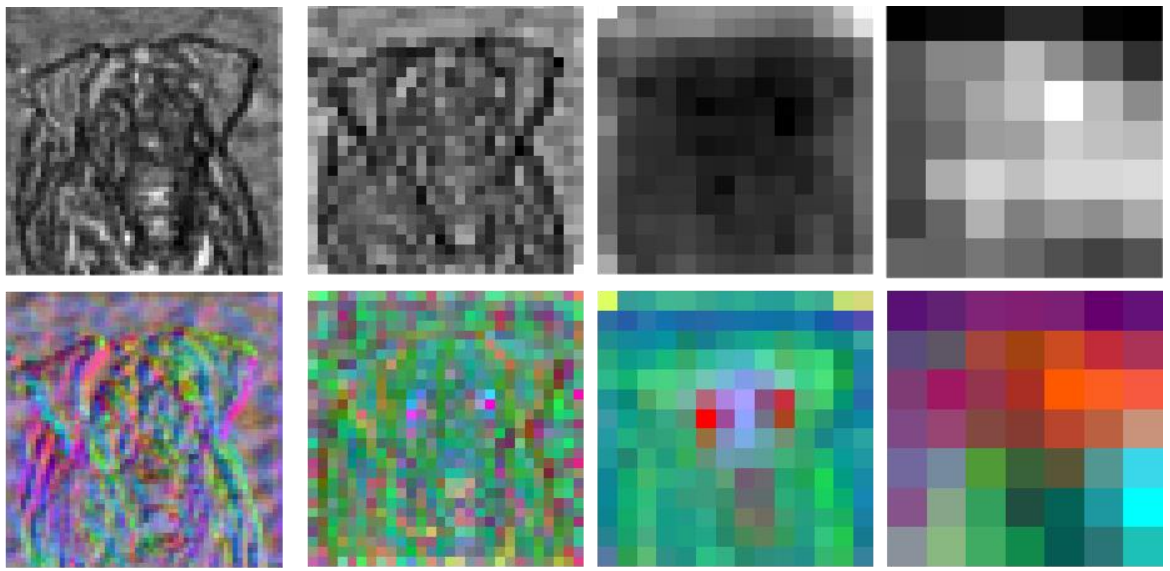


Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J. and Houlsby N. "**An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale**" (ICLR 2021)

DL4CV: CNN vs. ViT

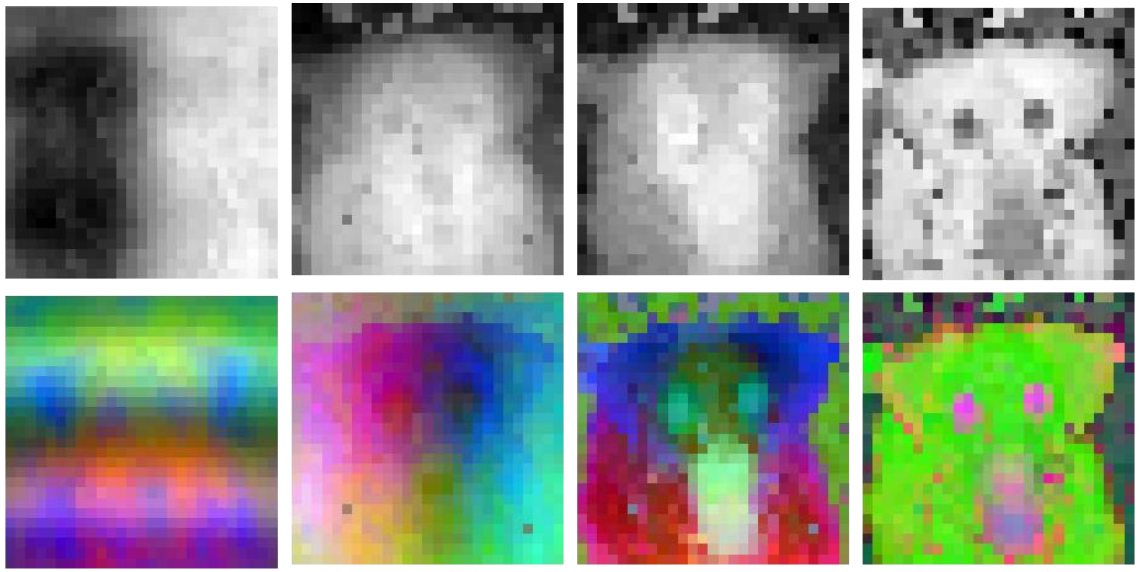
ResNet50

shallow → deep



ViT

shallow → deep



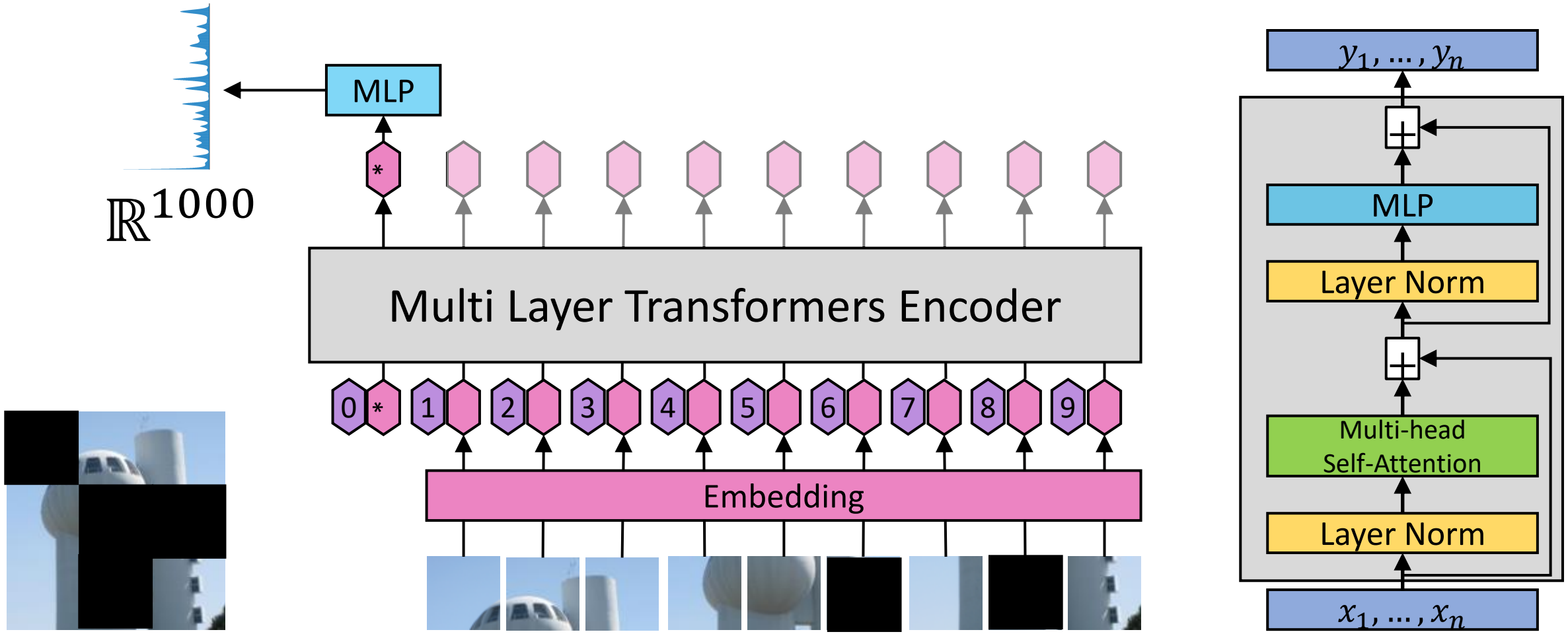
Amir, S., Gandelsman, Y., Bagon, S., Dekel, T.
"**Deep ViT Features as Dense Visual Descriptors**". ECCVW (2022)

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

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

Random PatchDrop

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

“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

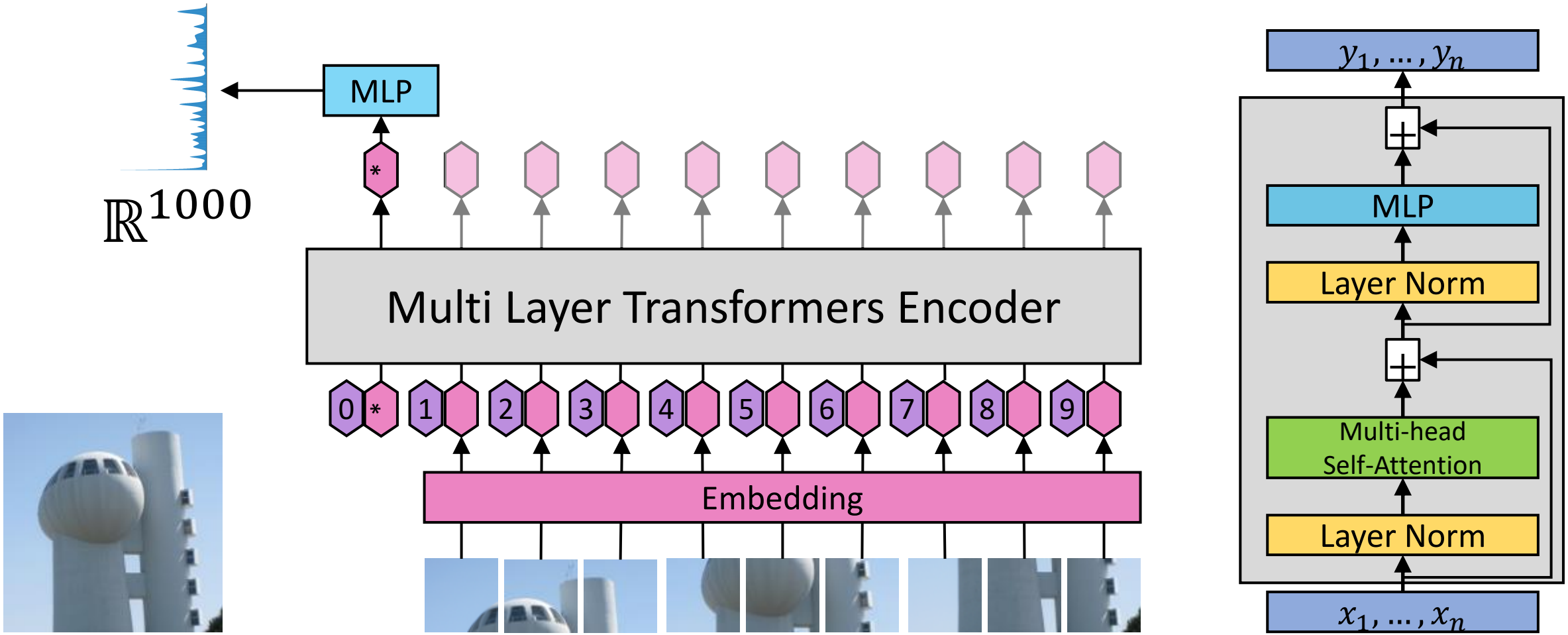
71.1%	**tabby cat**
17.3%	grey fox
3.3%	Siamese cat

(c) Texture-shape cue conflict

63.9%	**Indian elephant**
26.4%	indri
9.6%	black swan

Na *l Yang, M.H.,* *eurIPS 2021)*

Vision Transformers (ViT)



Vision Transformers (ViT)

Complexity

CNN

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

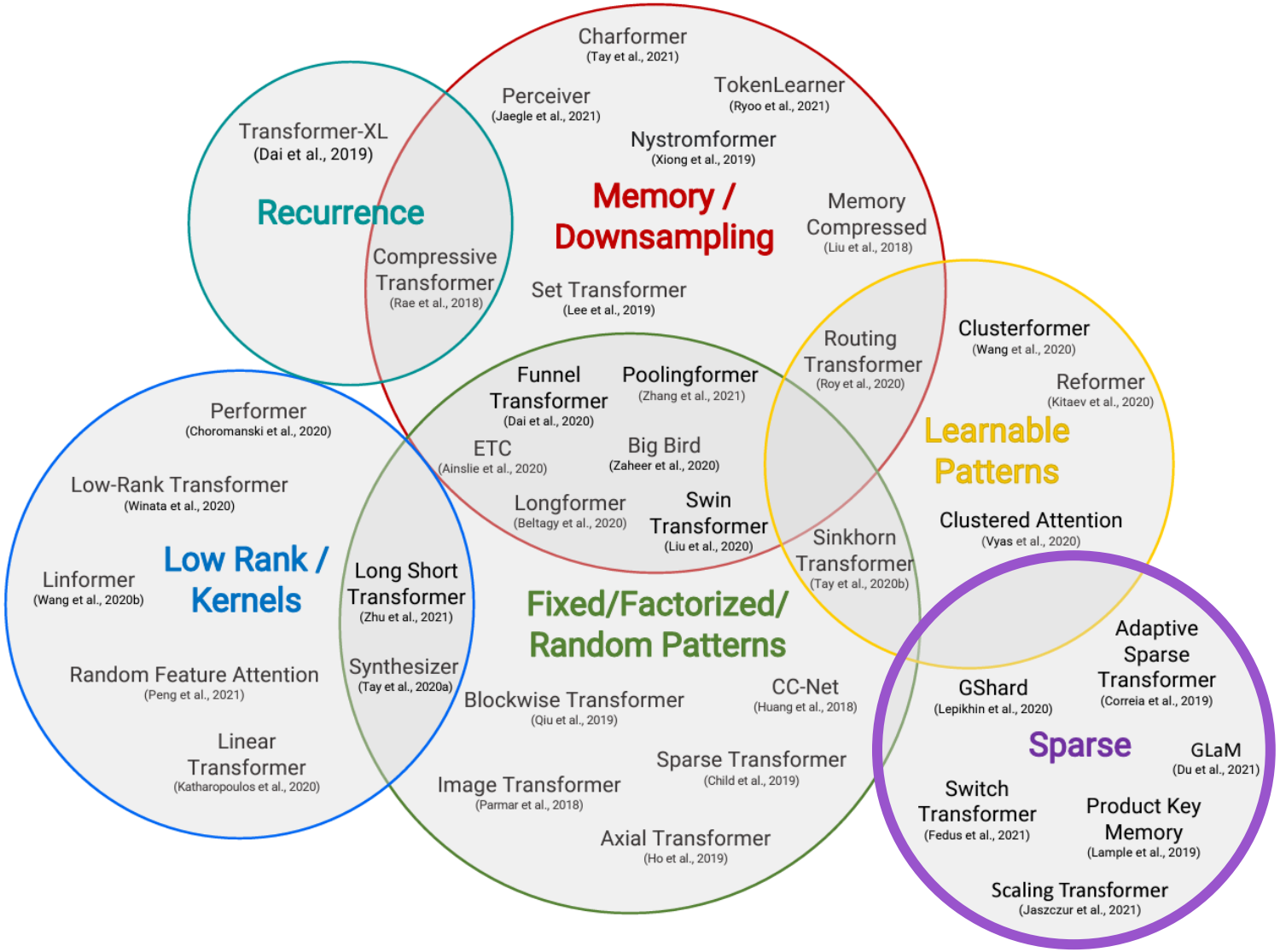
$$O(n)$$

Self-Attention

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

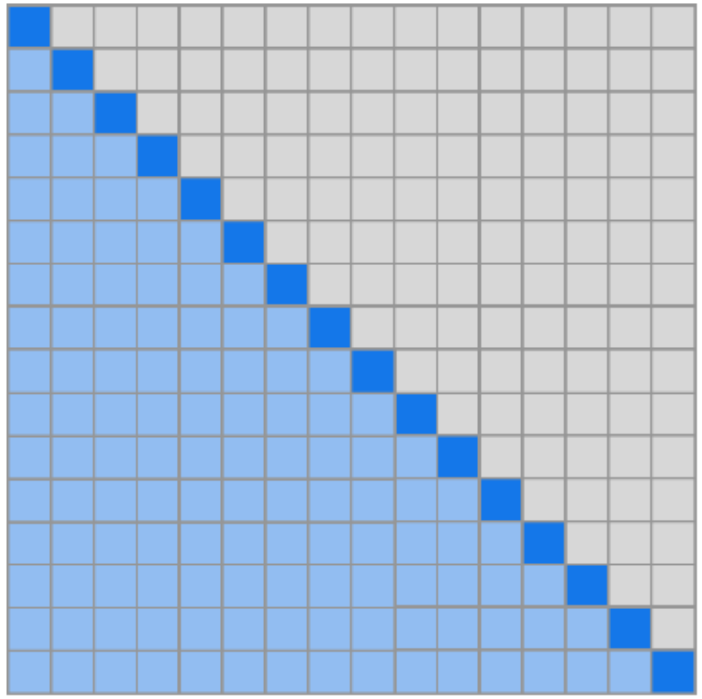
$$O(n^2)$$

Efficient Transformers



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

Efficient Transformers – Sparse Attention



(a) Transformer

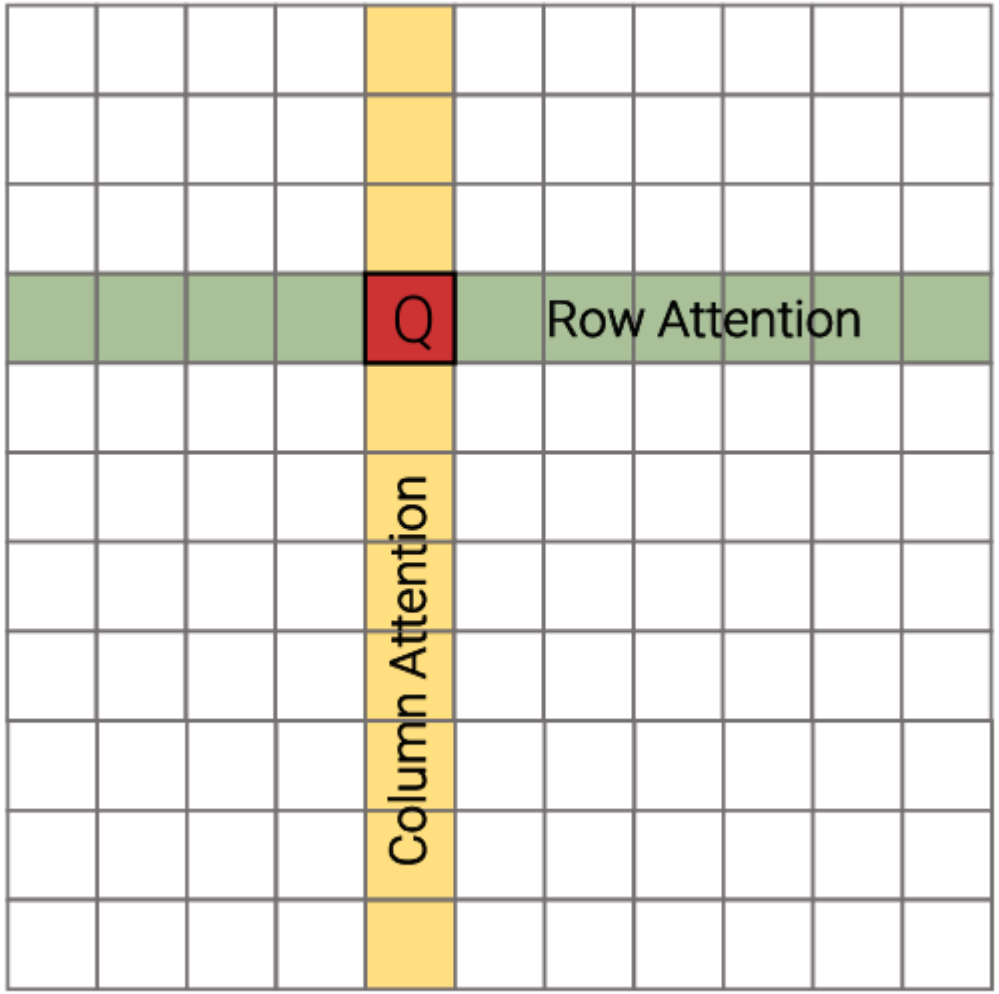
$$O\left(n\sqrt{n}\right)$$

Child, R., Gray, S., Radford, A. and Sutskever, I. "**Generating long sequences with sparse transformers**". (arXiv 2019)

Efficient Transformers – Sparse Attention

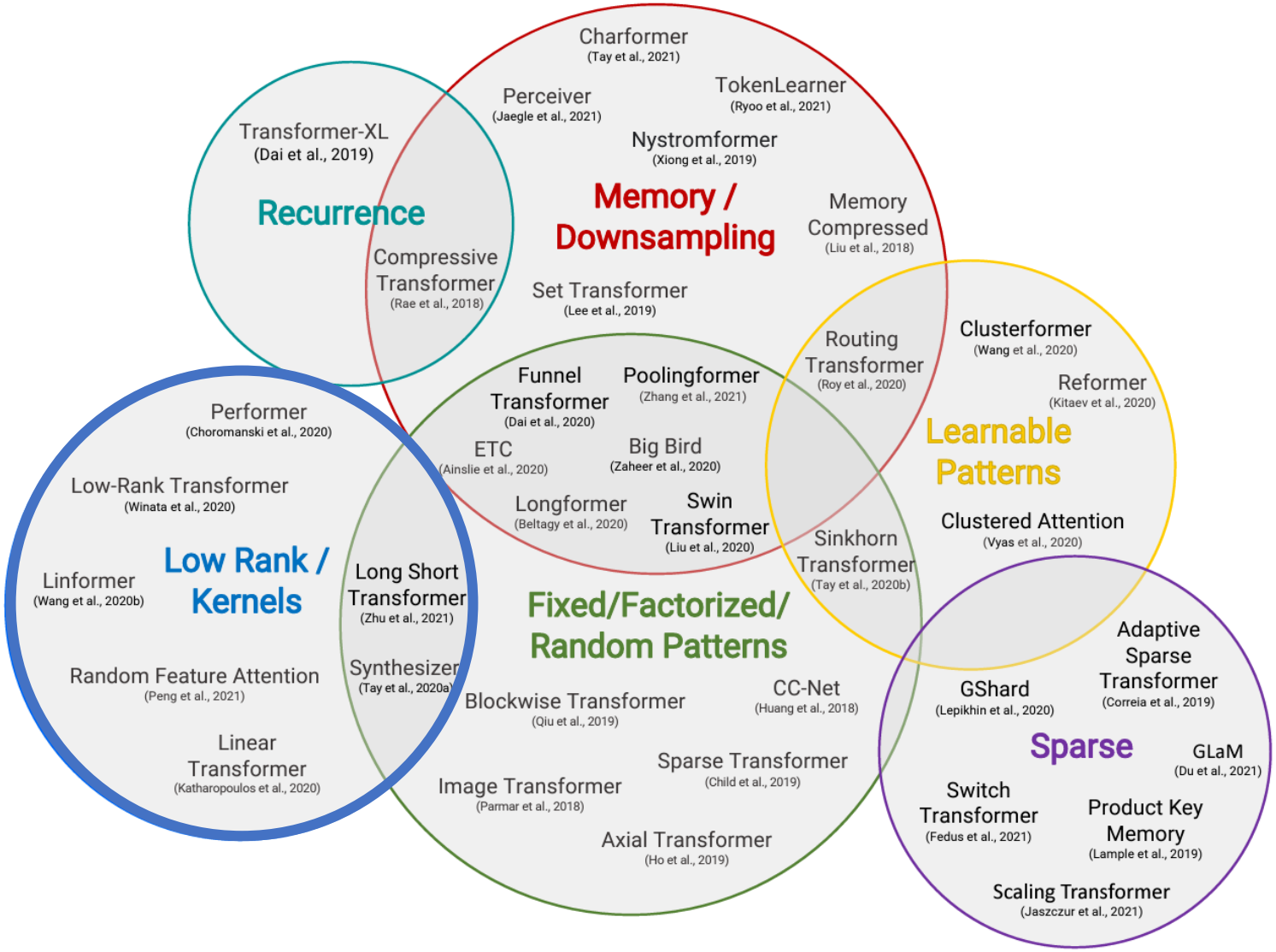
Axial attention for 2D data

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



Ho, J., Kalchbrenner, N., Weissenborn, D. and Salimans, T. "**Axial attention in multidimensional transformers**" (arXiv 2019)

Efficient Transformers



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

Efficient Transformers – Low Rank

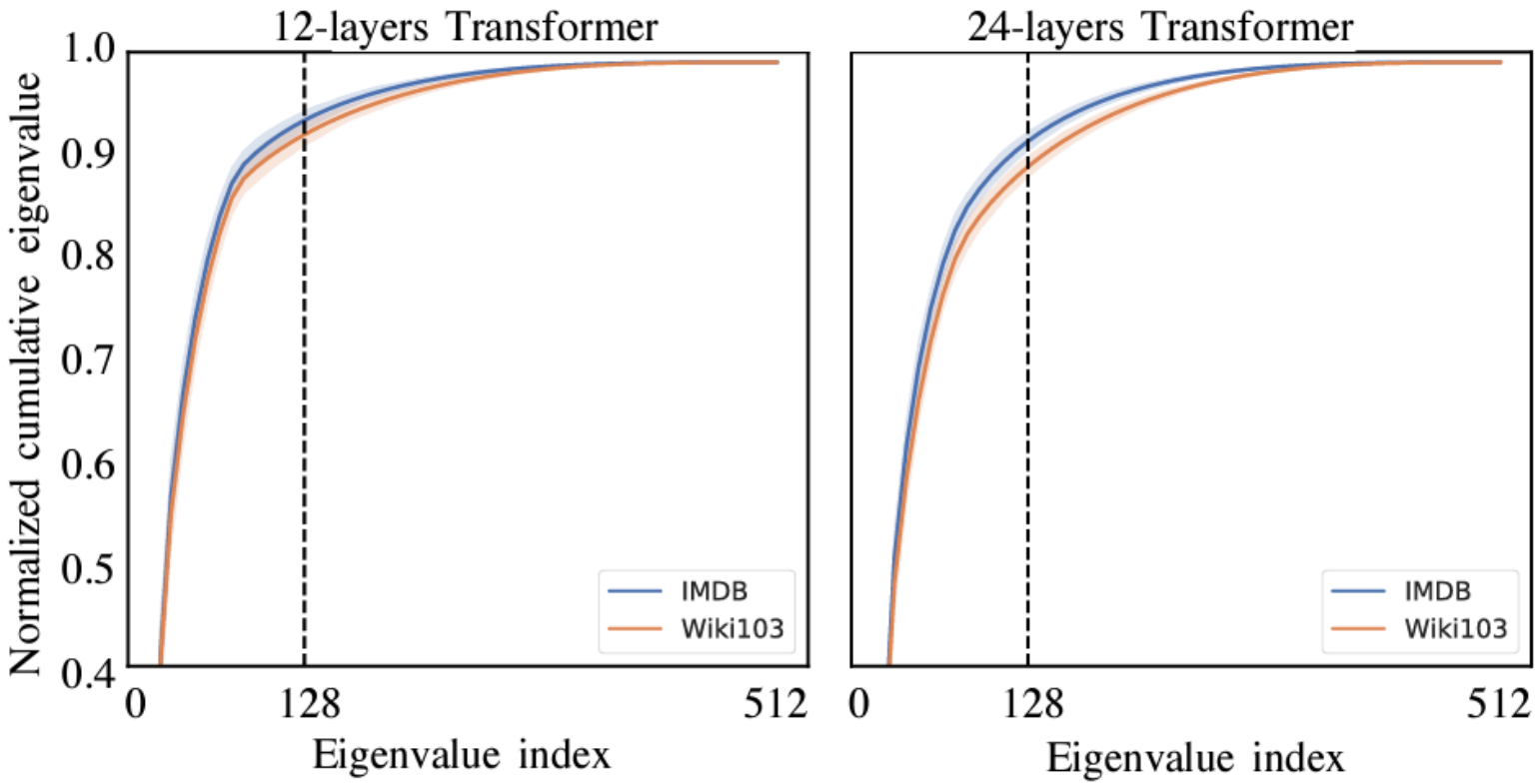
$$Y = \sigma(QK^T)V$$

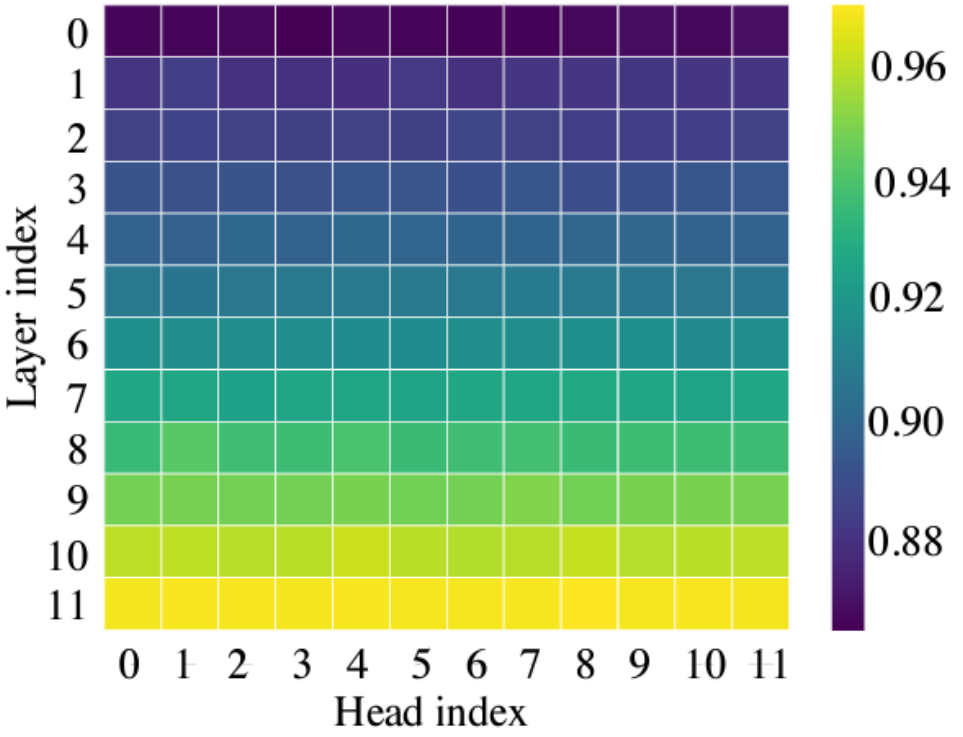
$$Y = \sigma(\boxed{Q}\ \boxed{K^T}\ )\ \boxed{V} \quad \in \mathbb{R}^{n \times d}, n \gg d$$

$$\boxed{Q}\ \boxed{K^T V}$$

Wang, S., Li, B.Z., Khabsa, M., Fang, H. and Ma, H. ”**Linformer: Self-attention with linear complexity**” (arXiv 2020)

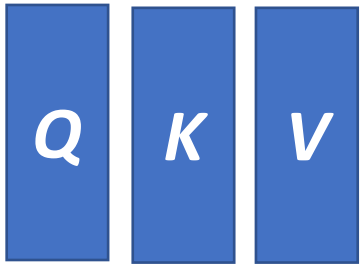
Efficient Transformers – Low Rank



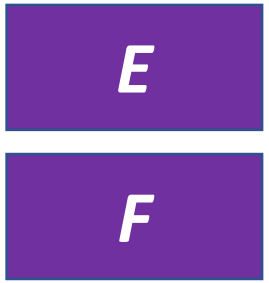


Wang, S., Li, B.Z., Khabsa, M., Fang, H. and Ma, H. ”**Linformer: Self-attention with linear complexity**” (arXiv 2020)

Efficient Transformers – Low Rank



$\in \mathbb{R}^{n \times d}$



$\in \mathbb{R}^{k \times n}$

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

$$Y = \sigma(\ Q\ \ K^T\ \ E^T\ \ F\ )$$

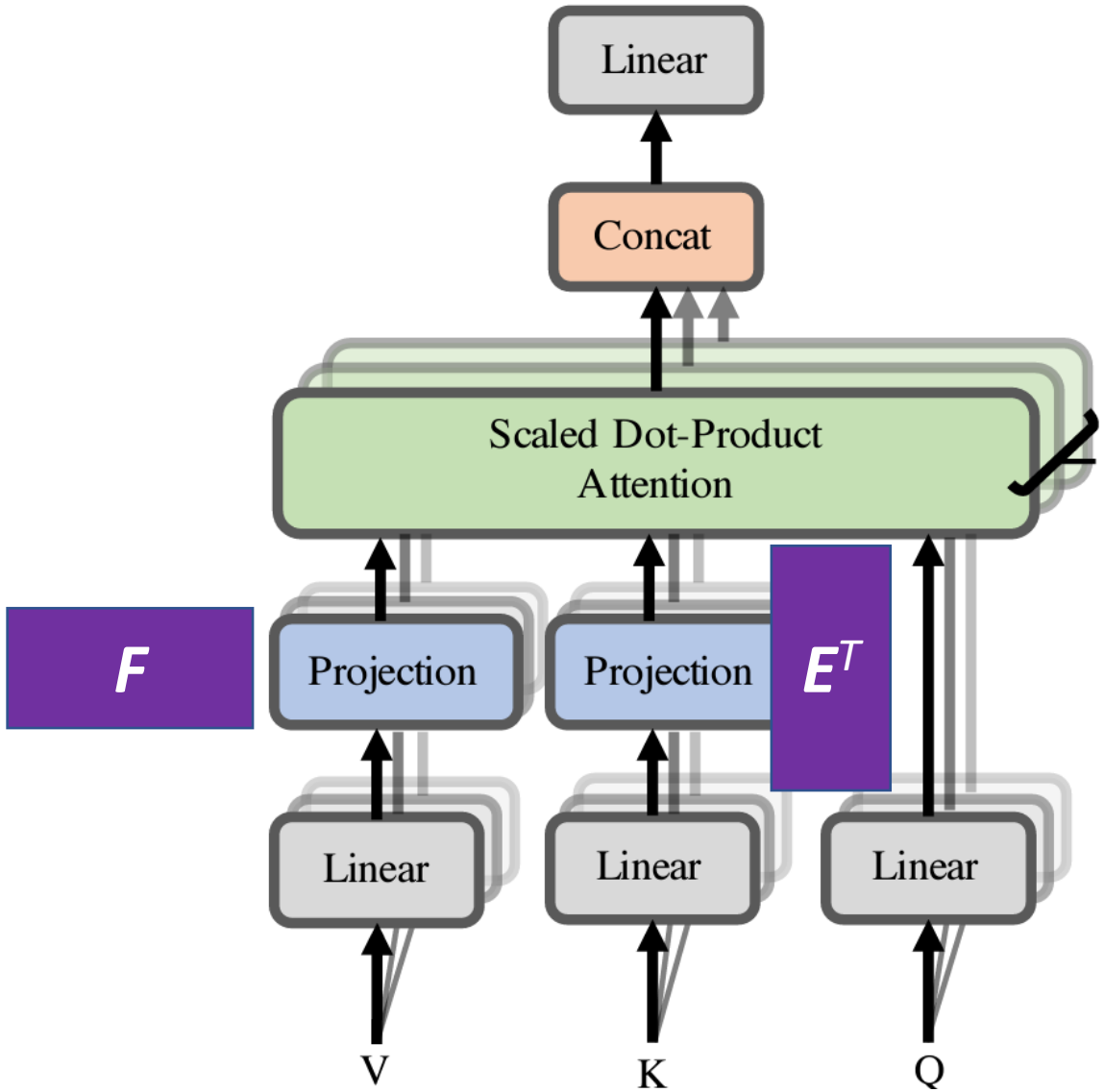
$$Y = \sigma(\ Q\ \ K^T E^T\ )\ FV$$

$k \times d$

$n \times k$

Wang, S., Li, B.Z., Khabsa, M., Fang, H. and Ma, H. ”**Linformer: Self-attention with linear complexity**” (arXiv 2020)

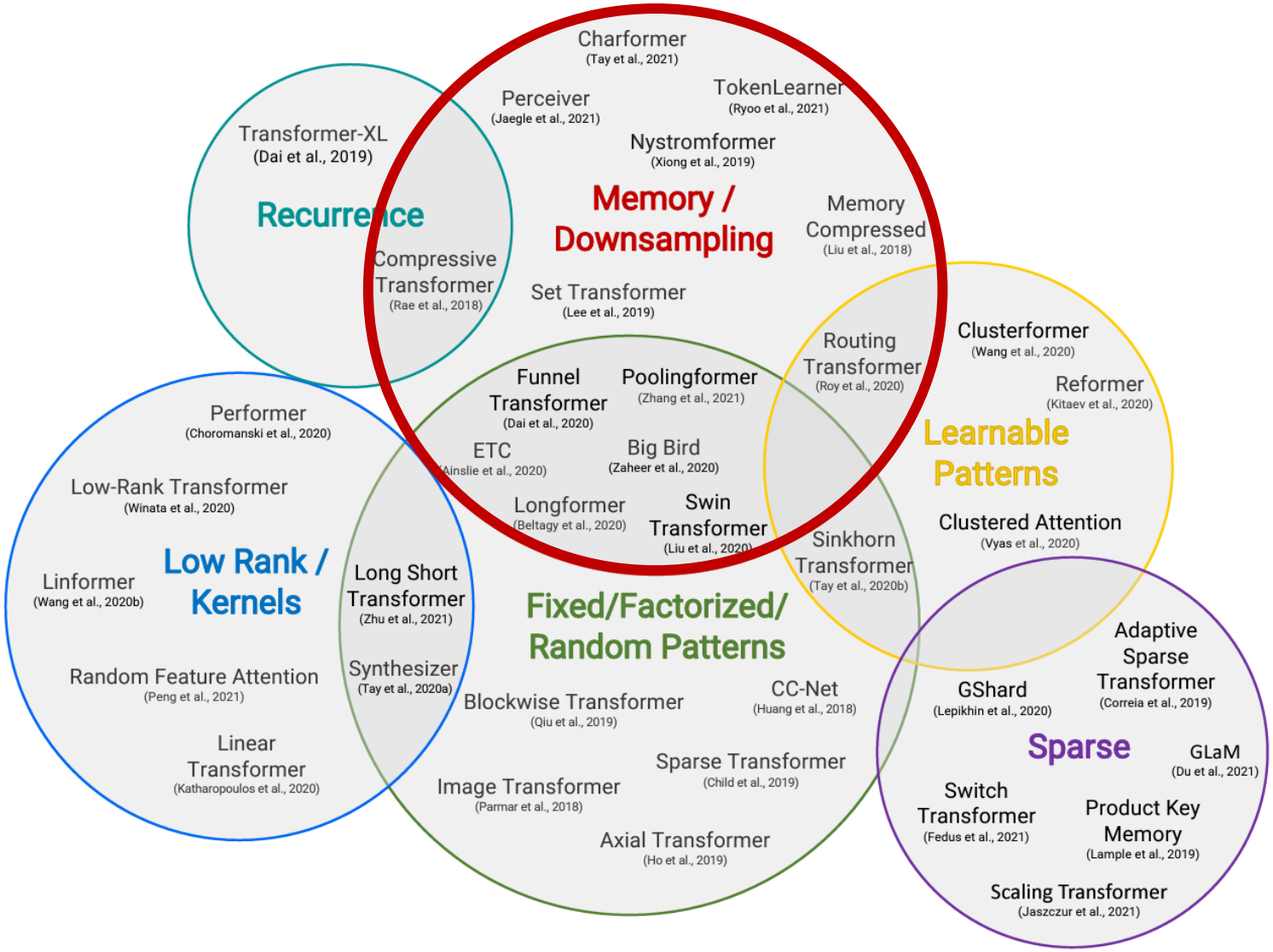
Efficient Transformers – Low Rank



$$O(n)$$

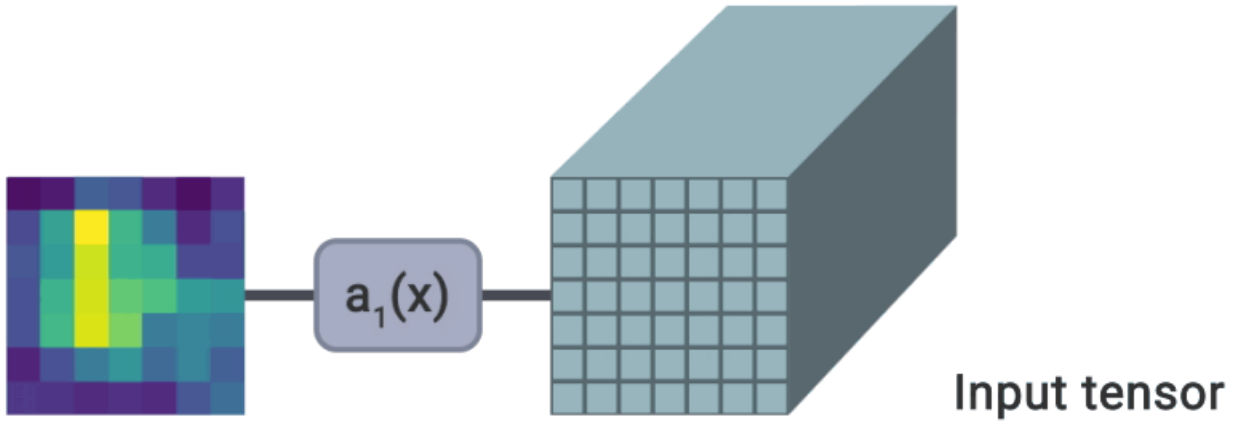
Wang, S., Li, B.Z., Khabsa, M., Fang, H. and Ma, H. ”**Linformer: Self-attention with linear complexity**” (arXiv 2020)

Efficient Transformers



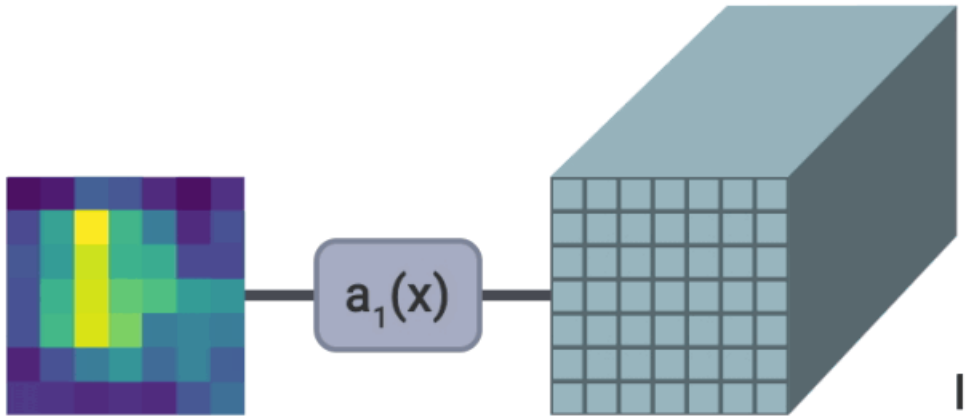
Tay, Y., Dehghani, M., Bahri, D. and Metzler, D. “**Efficient transformers: A survey**”. *(2022)*.

Efficient Transformers – Down-Sample



Ryoo, M., Piergiovanni, A.J., Arnab, A., Dehghani, M. and Angelova, A. "**Tokenlearner: Adaptive space-time tokenization for videos**" (NeurIPS 2021)

Efficient Transformers – Down-Sample



Ryoo, M., Piergiovanni, A.J., Arnab, A., Dehghani, M. and Angelova, A. "**Tokenlearner: Adaptive space-time tokenization for videos**" (NeurIPS 2021)

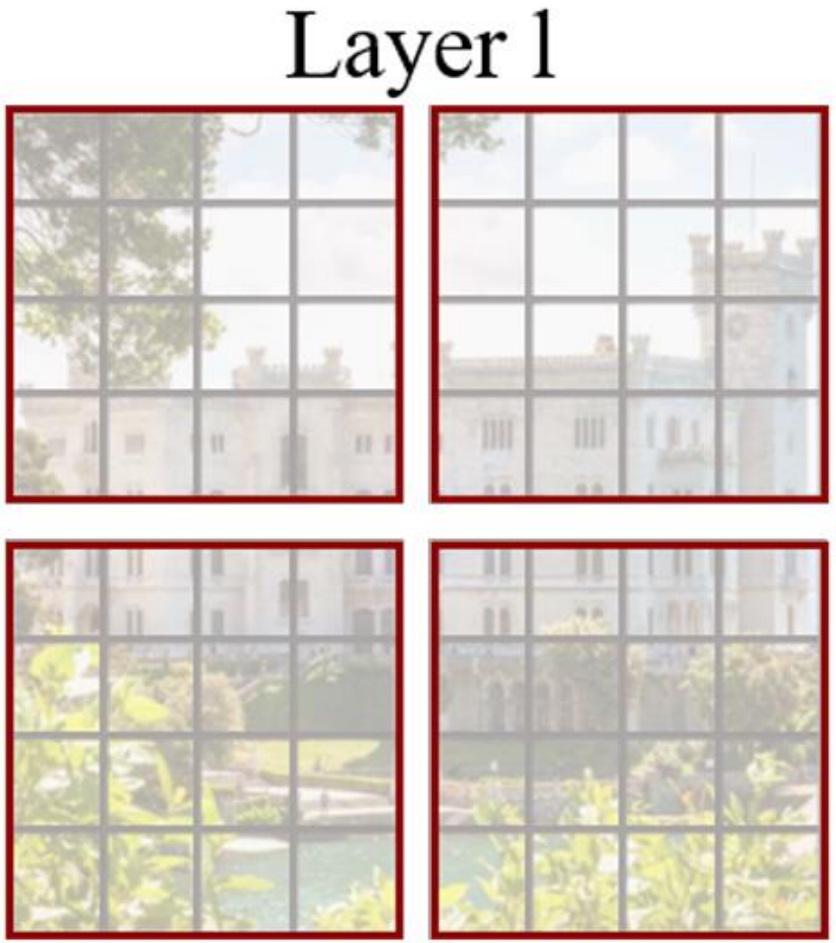
Efficient Transformers – Down-Sample

$O(n)$

Ryoo, M., Piergiovanni, A.J., Arnab, A., Dehghani, M. and Angelova, A. **"Tokenlearner: Adaptive space-time tokenization for videos"** (NeurIPS 2021)

Other Transformers for Vision

Shifted Window Transformer (Swin)

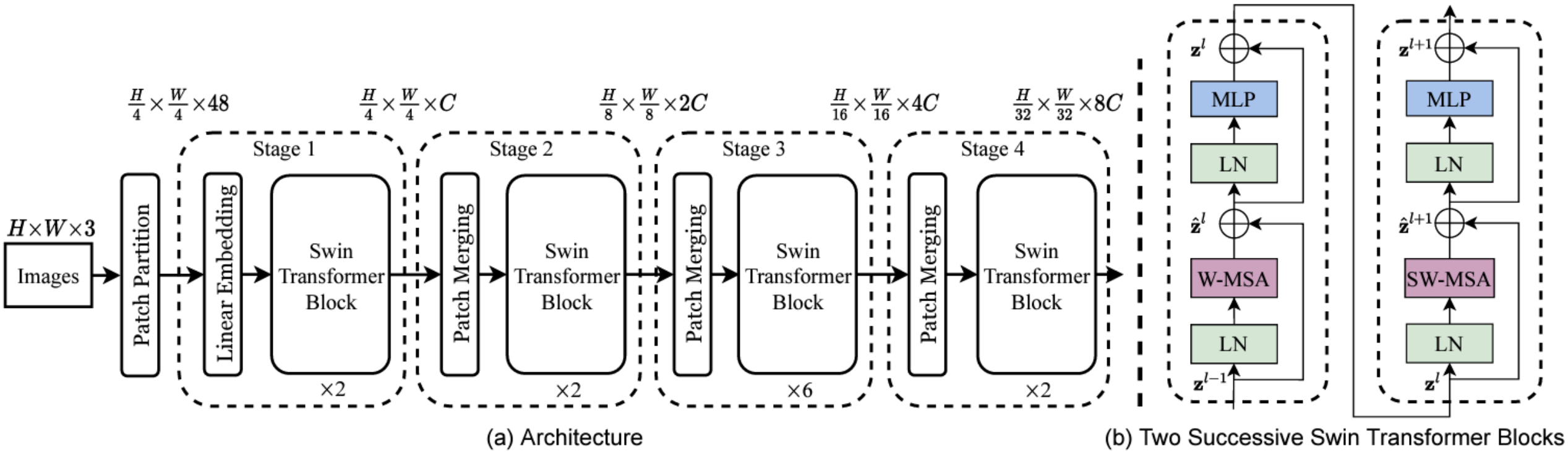


A local window to perform self-attention

A patch

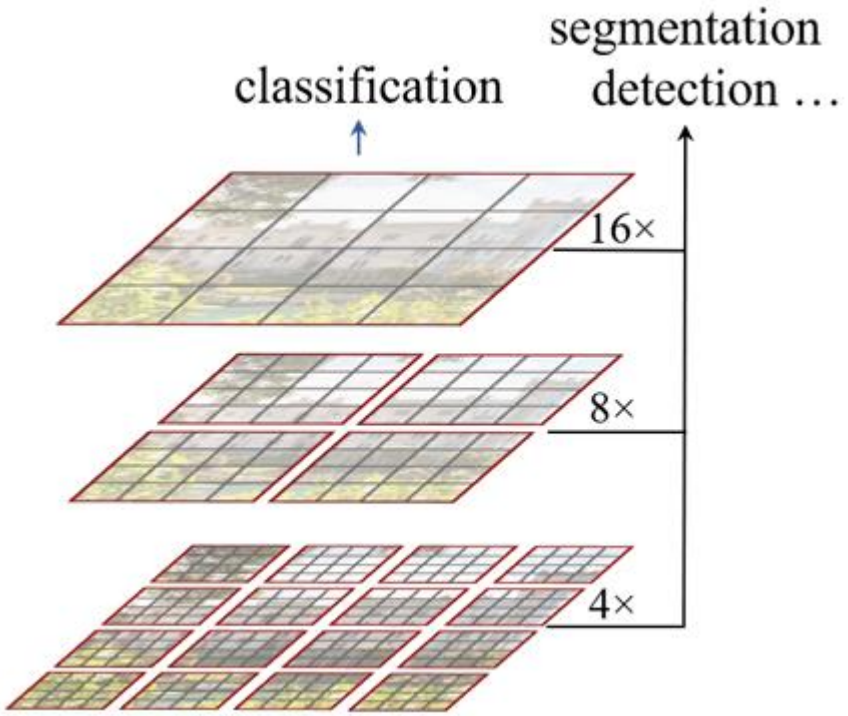
Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B. **Swin transformer: Hierarchical vision transformer using shifted windows**. (ICCV 2021)

Shifted Window Transformer (Swin)



Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B. **Swin transformer: Hierarchical vision transformer using shifted windows**. (ICCV 2021)

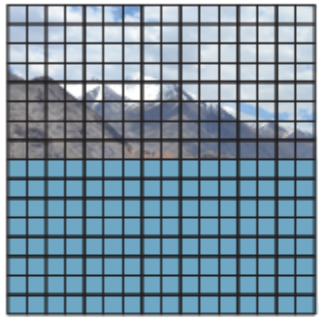
Shifted Window Transformer (Swin)



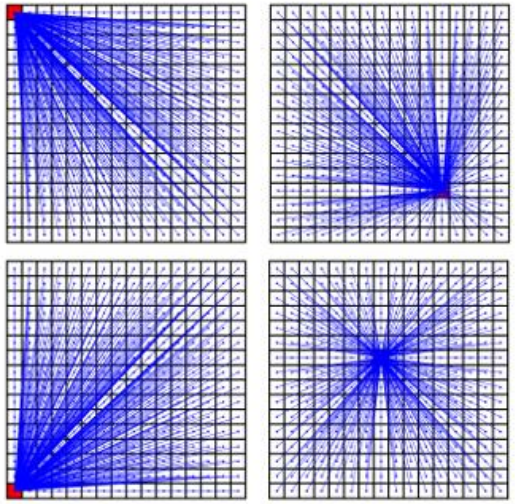
(a) Regular ImageNet-1K trained models					
method	image size	#param.	FLOPs	throughput (image / s)	ImageNet top-1 acc.
ViT-B/16 [19]	384^2	86M	55.4G	85.9	77.9
ViT-L/16 [19]	384^2	307M	190.7G	27.3	76.5
Swin-T	224^2	29M	4.5G	755.2	81.3
Swin-S	224^2	50M	8.7G	436.9	83.0
Swin-B	224^2	88M	15.4G	278.1	83.5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

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B. **Swin transformer: Hierarchical vision transformer using shifted windows**. (ICCV 2021)

Multiscale Vision Transformer



high-res input image

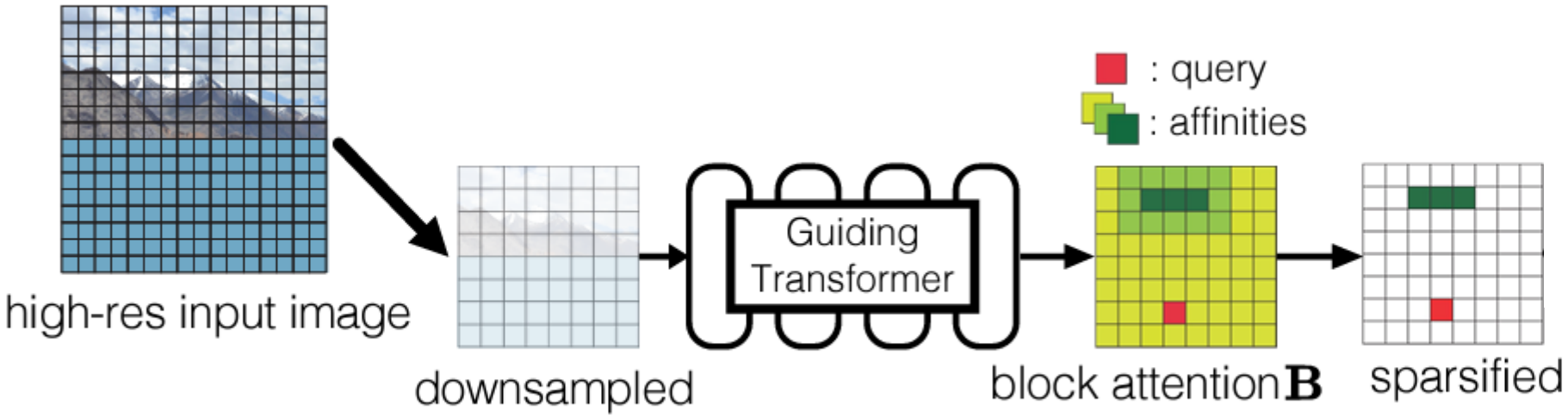


full attention $\mathbf{A}$

Liu, D., Shetty, S., Hinz, T., Fisher, M., Zhang, R., Park, T. and Kalogerakis, E. **ASSET: autoregressive semantic scene editing with transformers at high resolutions**. (TOG 2022)

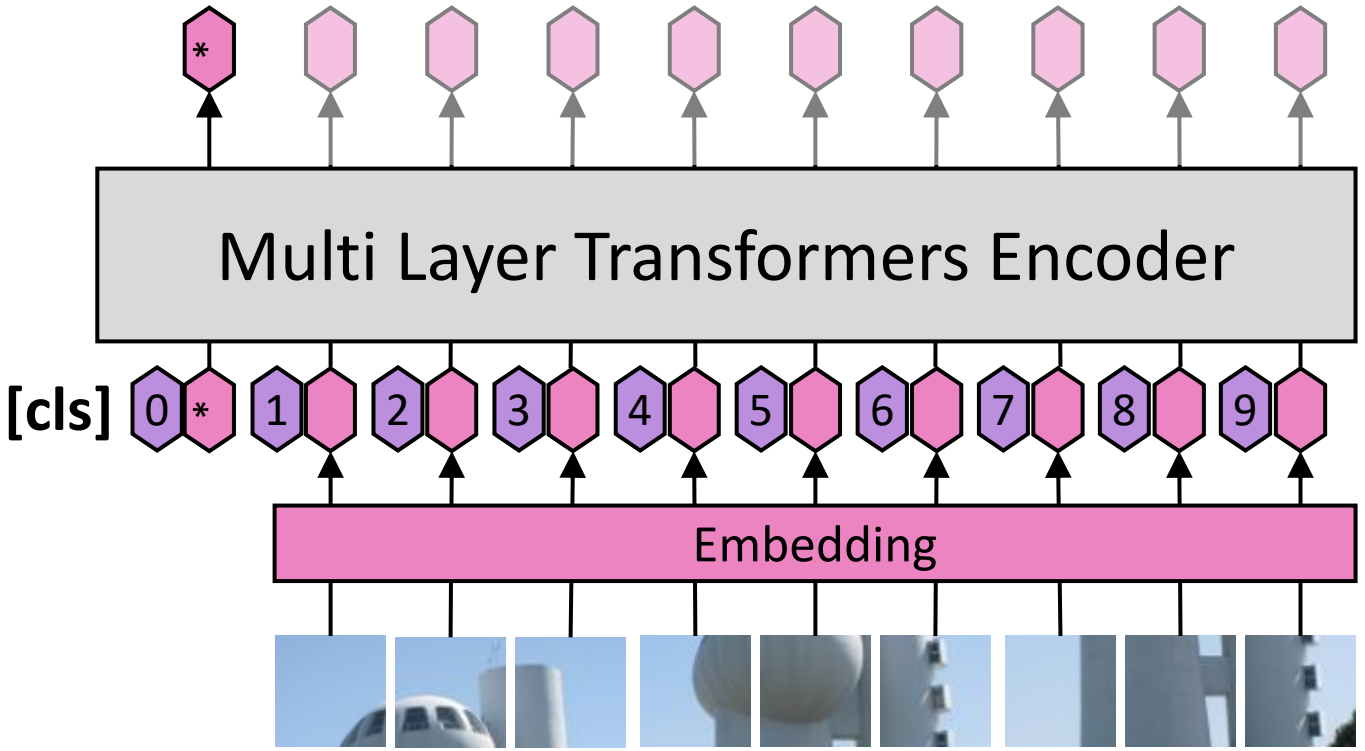
Multiscale Vision Transformer

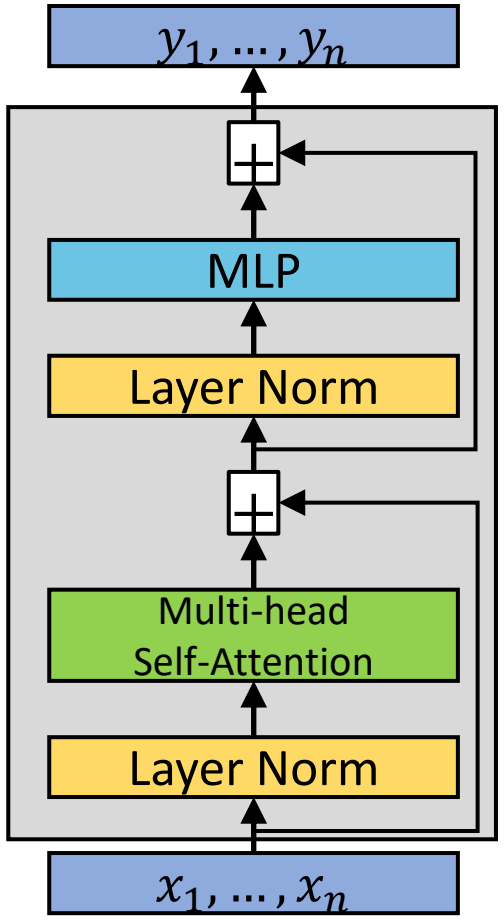
Use multi-scale to derive smart sparse patterns



Liu, D., Shetty, S., Hinz, T., Fisher, M., Zhang, R., Park, T. and Kalogerakis, E. **ASSET: autoregressive semantic scene editing with transformers at high resolutions**. (TOG 2022)

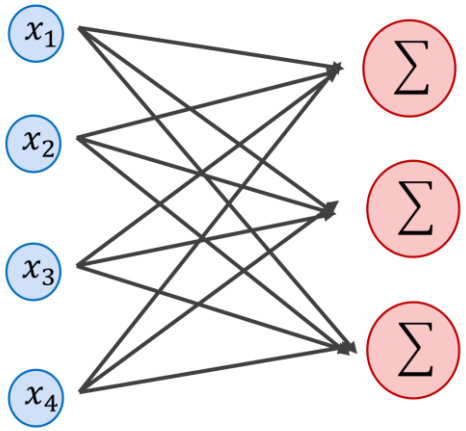
Vision Transformers (ViT)



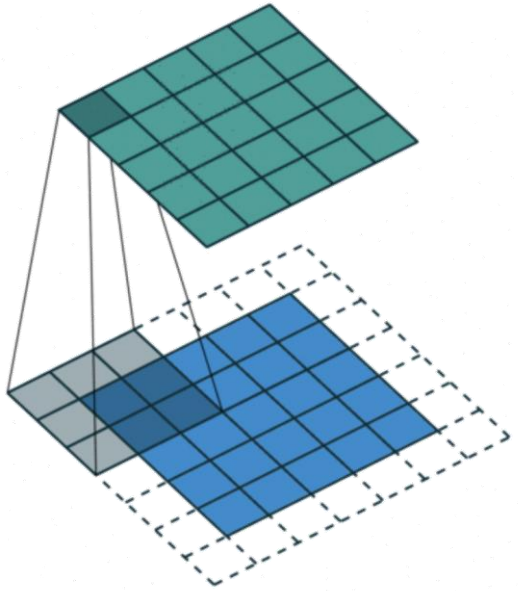


DL4CV: NN Building blocks

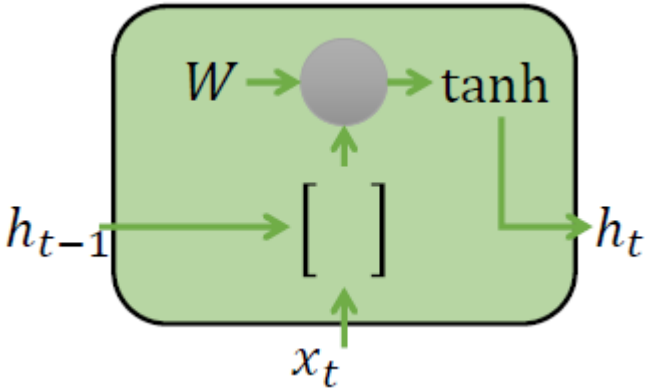
Liner Layer



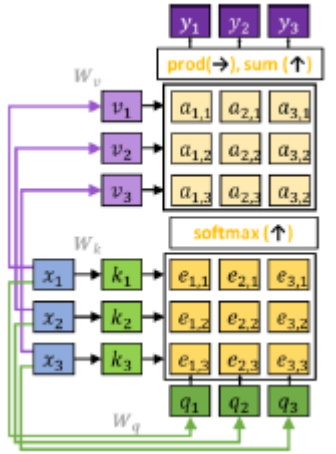
Convolution Layer



Recurrent Layer



Attention Layer



What's next?

Next lecture:

Detection and Segmentation II (Niv)