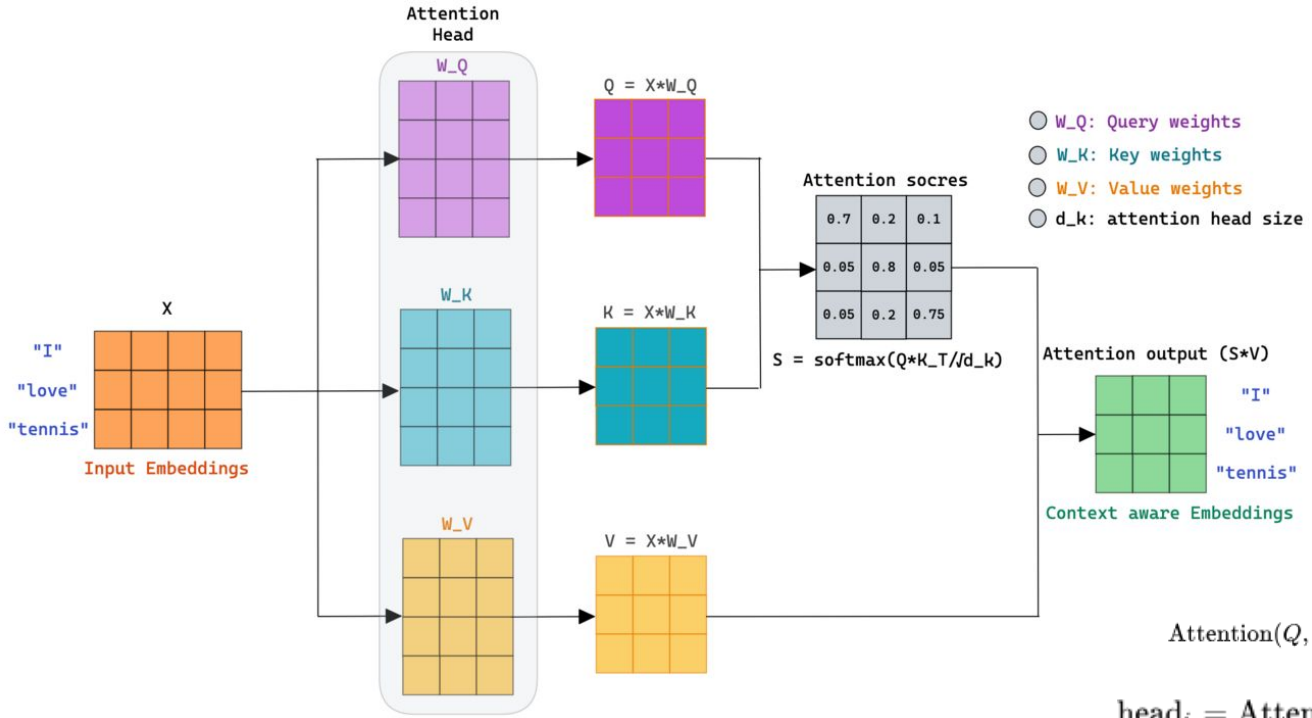
CV Transformers

Attention is All you Need

Attention is All you Need

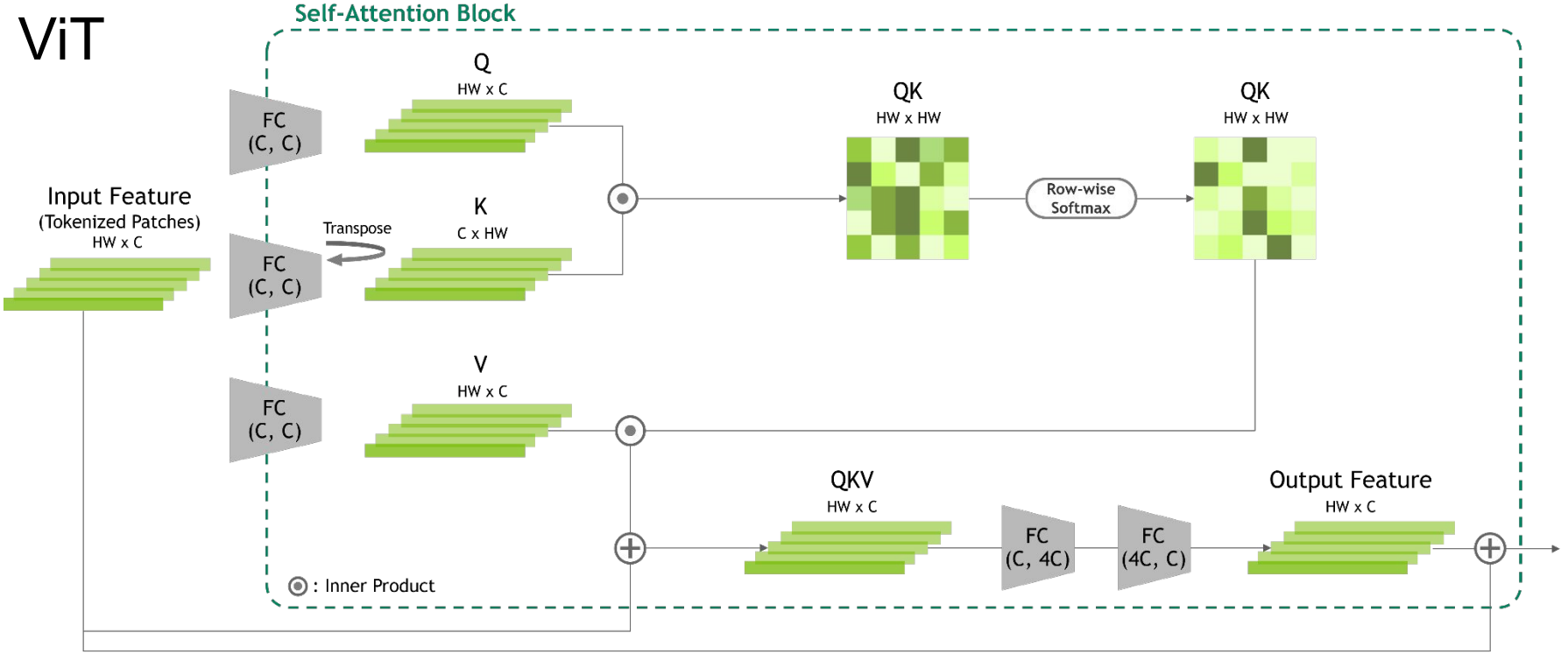


$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\mathsf{T}}{\sqrt{d_k}}\right)V,$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V).$$

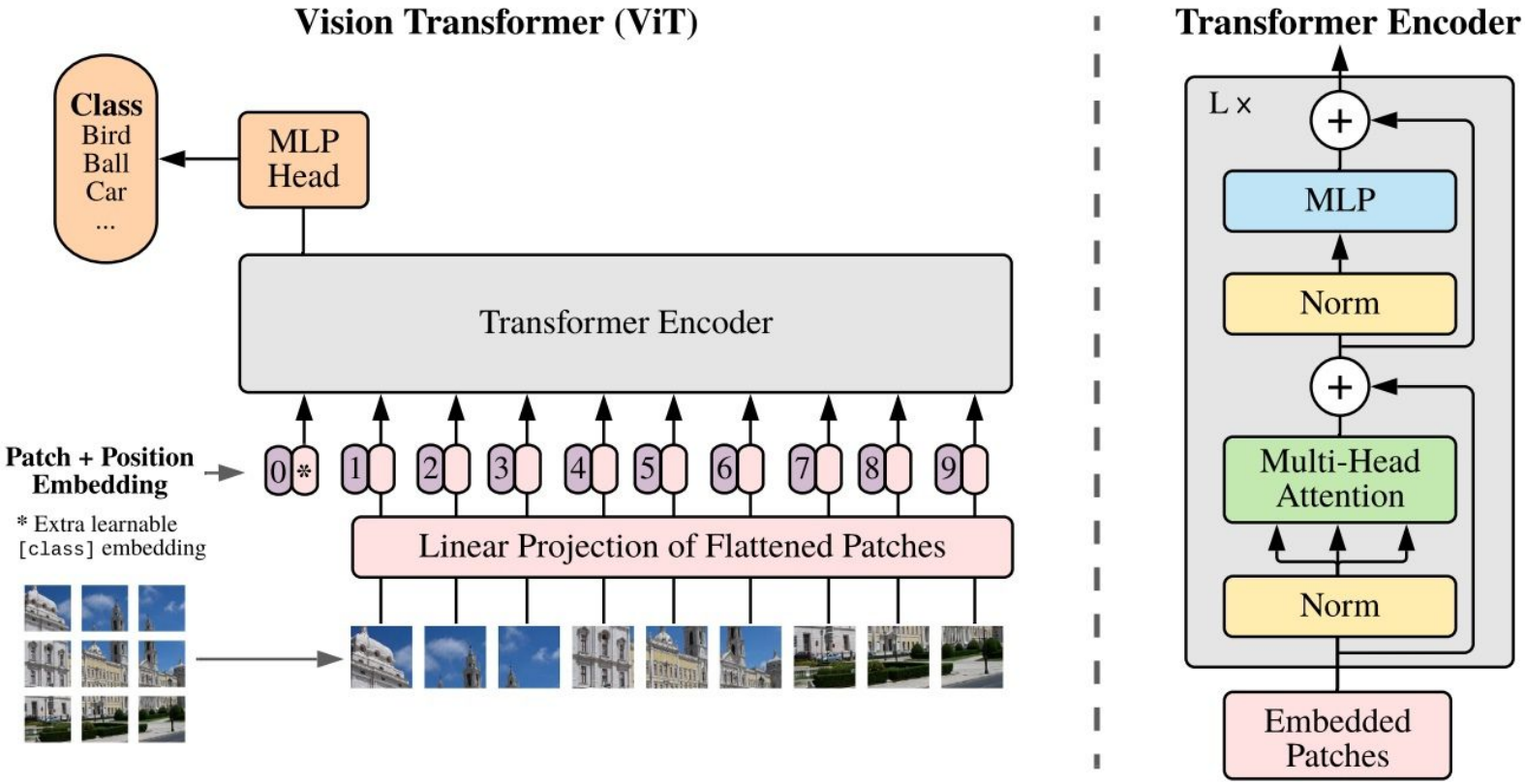
$$\text{MultiHeadAttention}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O,$$

ViT

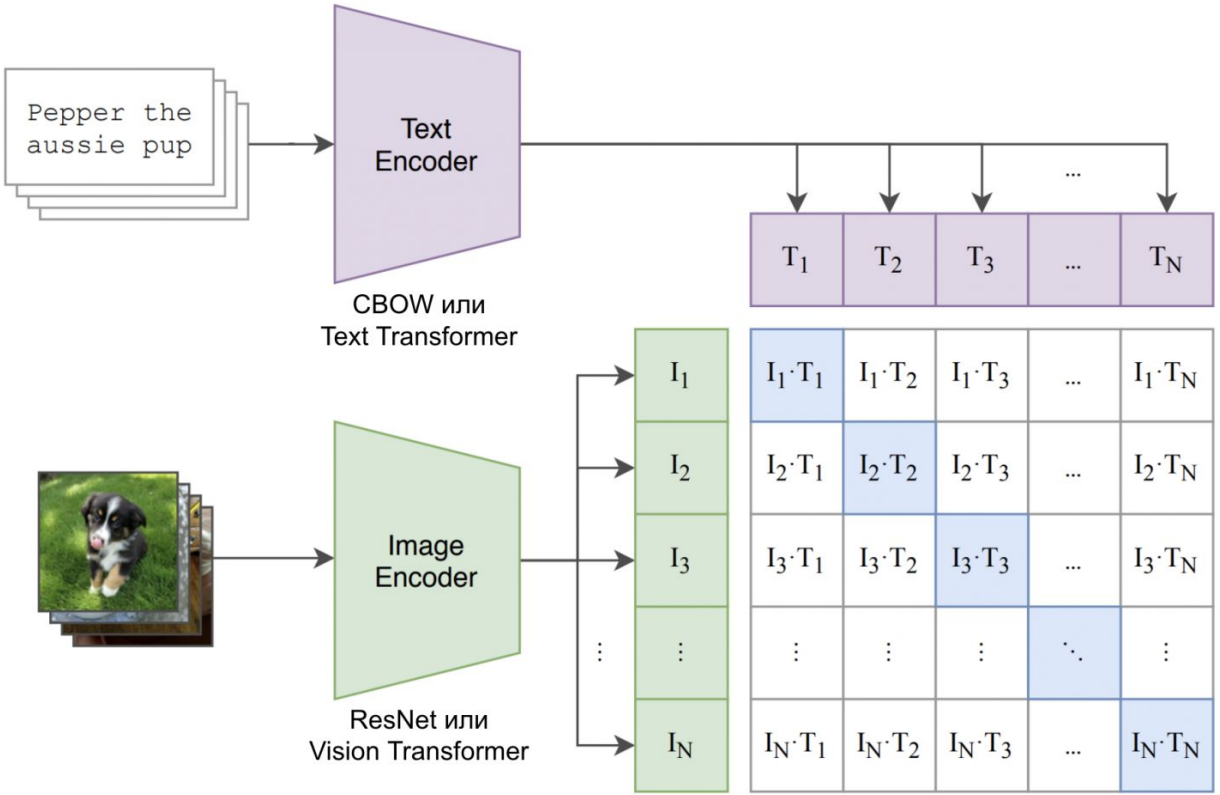

Self-Attention Block
Input Feature
(Tokenized Patches)
HW x C
FC
(C, C)
Q
HW x C
Transpose
K
C x HW
V
HW x C
QK
HW x HW
Row-wise
Softmax
QK
HW x HW
QKV
HW x C
FC
(C, 4C)
FC
(4C, C)
Output Feature
HW x C
: Inner Product

link

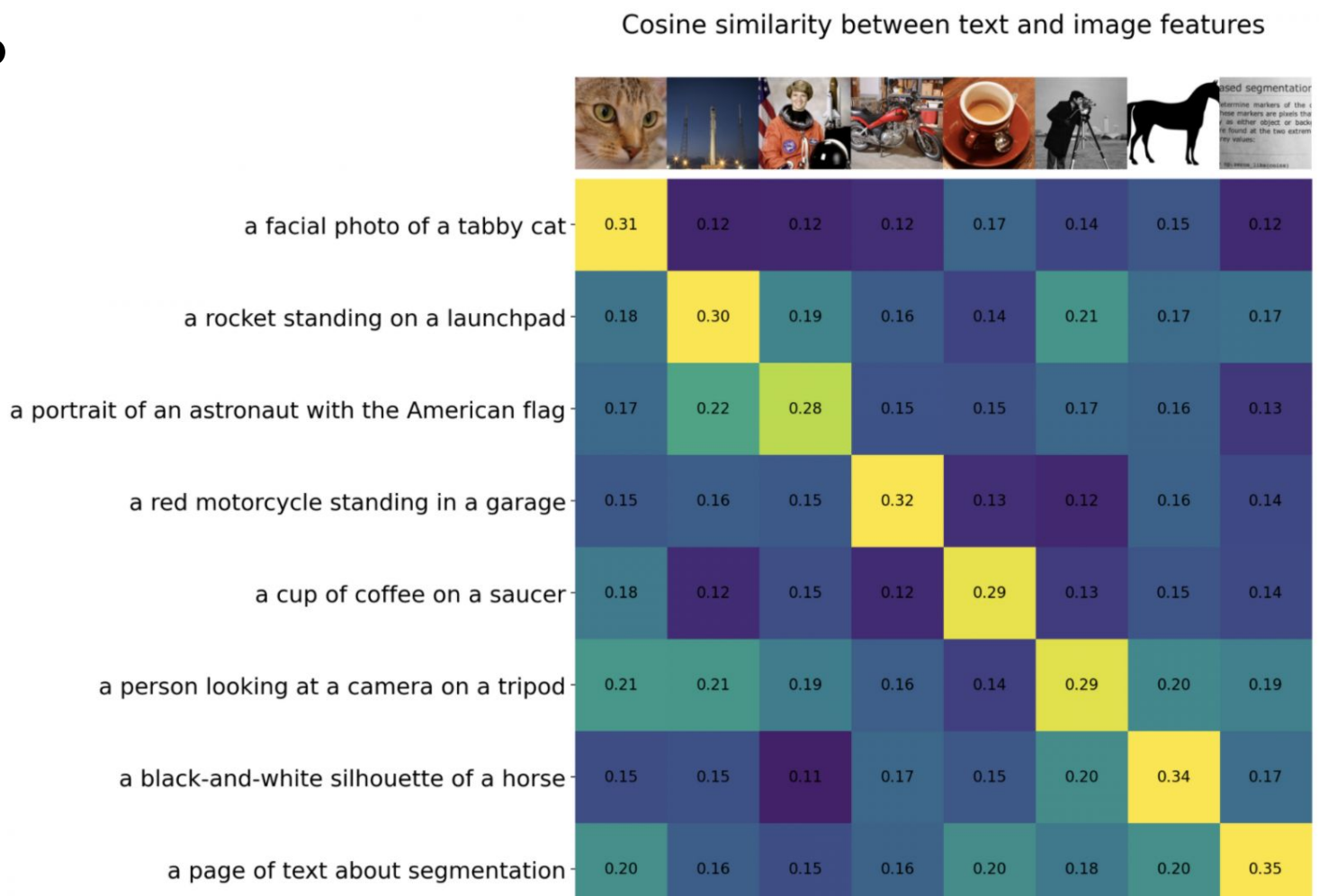
ViT


Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0 * 1 2 3 4 5 6 7 8 9
Linear Projection of Flattened Patches
Transformer Encoder
L x
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches

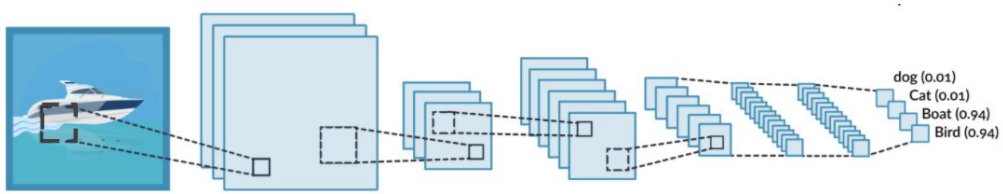
CLIP


Pepper the aussie pup
Text Encoder
CBOW или Text Transformer
T1 T2 T3 ... TN
Image Encoder
ResNet или Vision Transformer
I1 I2 I3 ⋮ IN
I1·T1 I1·T2 I1·T3 ... I1·TN
I2·T1 I2·T2 I2·T3 ... I2·TN
I3·T1 I3·T2 I3·T3 ... I3·TN
⋮ ⋮ ⋮ ⋱ ⋮
IN·T1 IN·T2 IN·T3 ... IN·TN

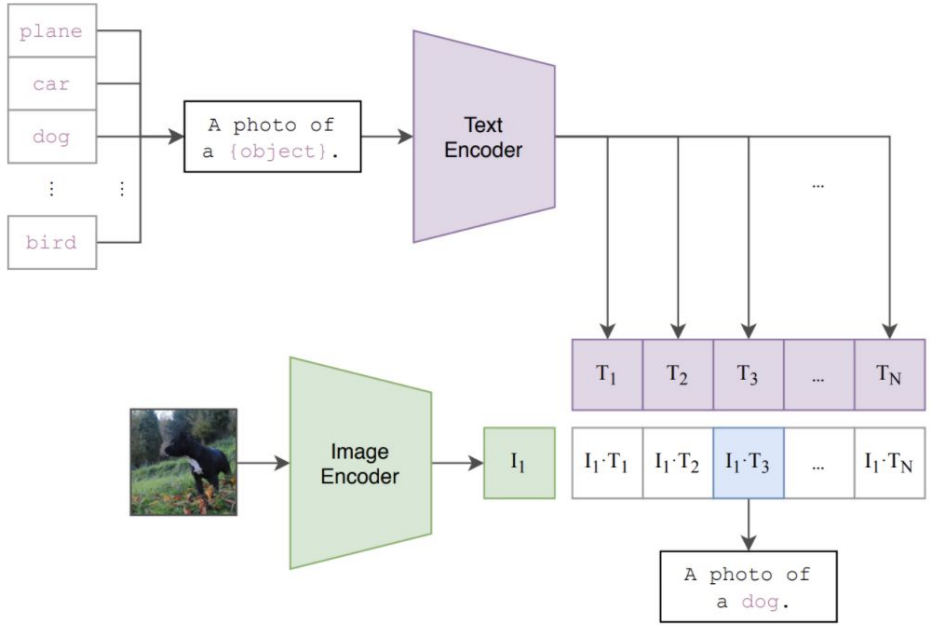
CLIP


Cosine similarity between text and image features
a facial photo of a tabby cat
a rocket standing on a launchpad
a portrait of an astronaut with the American flag
a red motorcycle standing in a garage
a cup of coffee on a saucer
a person looking at a camera on a tripod
a black-and-white silhouette of a horse
a page of text about segmentation
0.31 0.12 0.12 0.12 0.17 0.14 0.15 0.12
0.18 0.30 0.19 0.16 0.14 0.21 0.17 0.17
0.17 0.22 0.28 0.15 0.15 0.17 0.16 0.13
0.15 0.16 0.15 0.32 0.13 0.12 0.16 0.14
0.18 0.12 0.15 0.12 0.29 0.13 0.15 0.14
0.21 0.21 0.19 0.16 0.14 0.29 0.20 0.19
0.15 0.15 0.11 0.17 0.15 0.20 0.34 0.17
0.20 0.16 0.15 0.16 0.20 0.18 0.20 0.35

CLIP

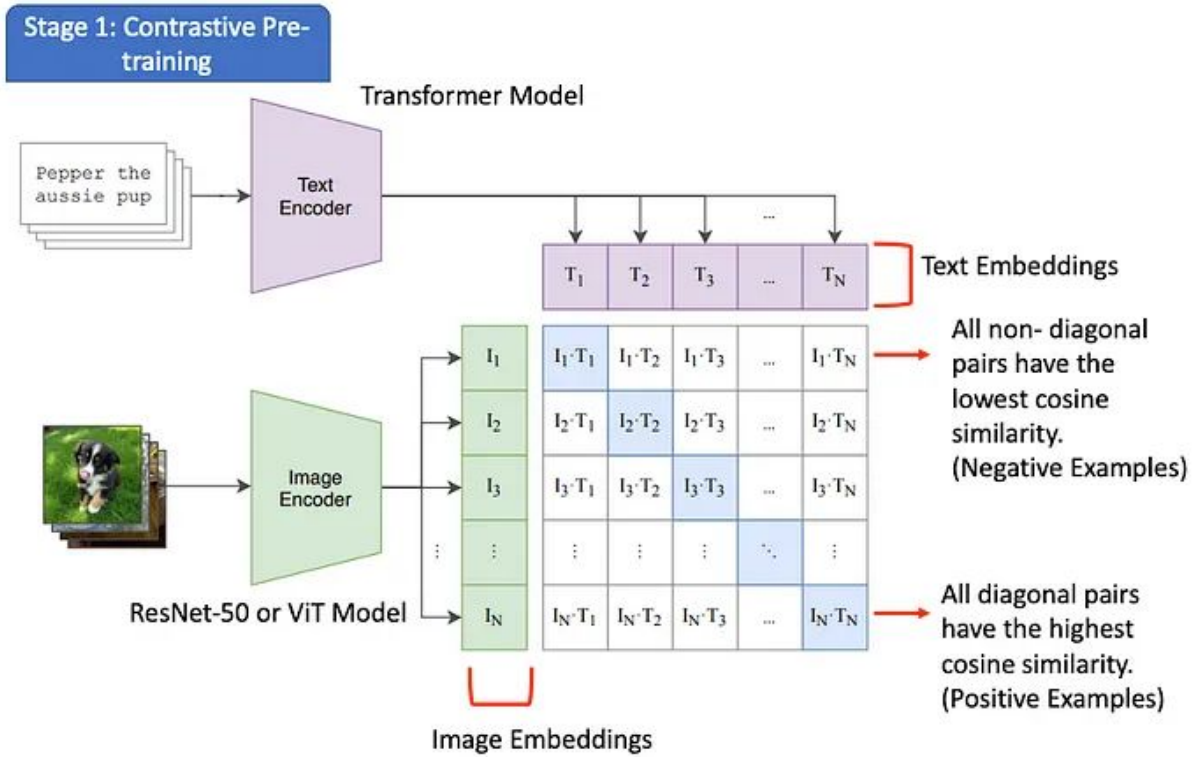


Классификация изображений при помощи сверточной нейросети в рамках “классической” парадигмы



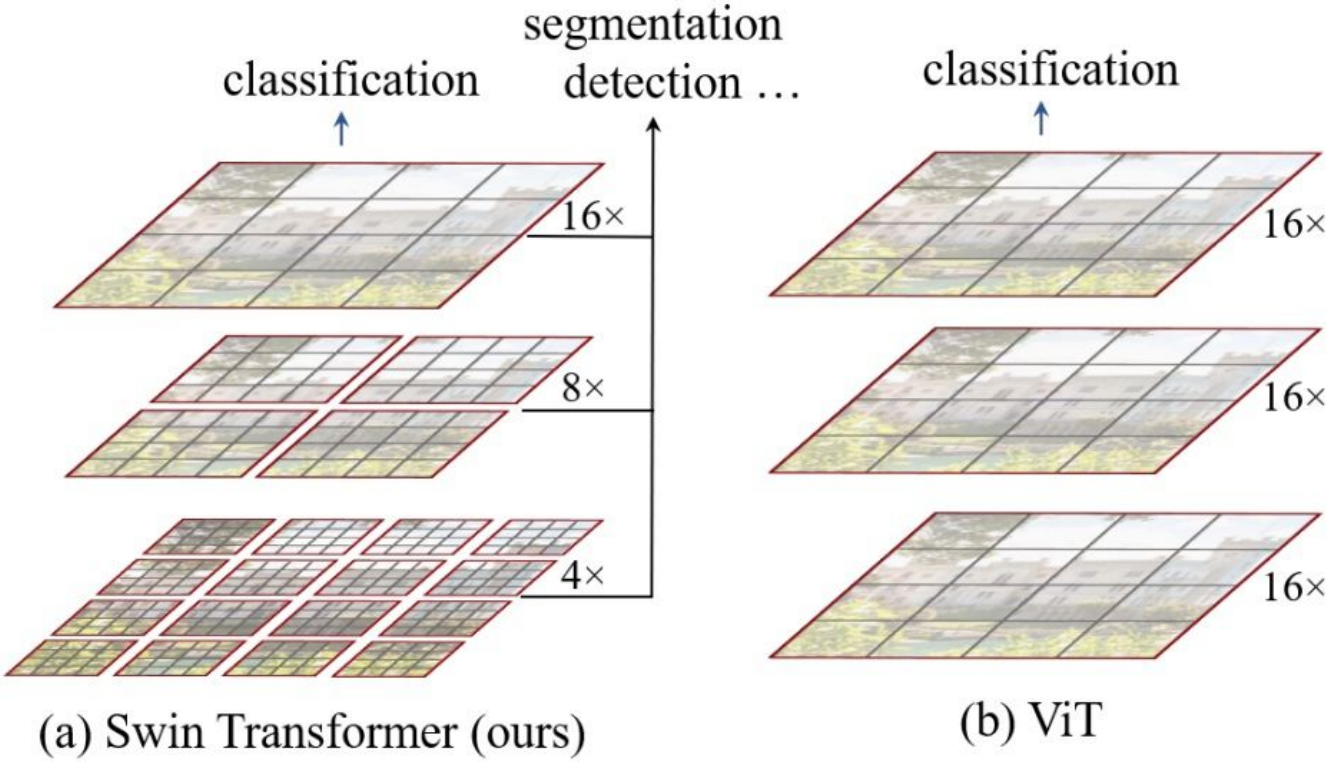
Классификация изображений через схожесть репрезентаций сверточной сети и трансформера в рамках “гибридной” парадигмы

CLIP. Contrastive representation learning

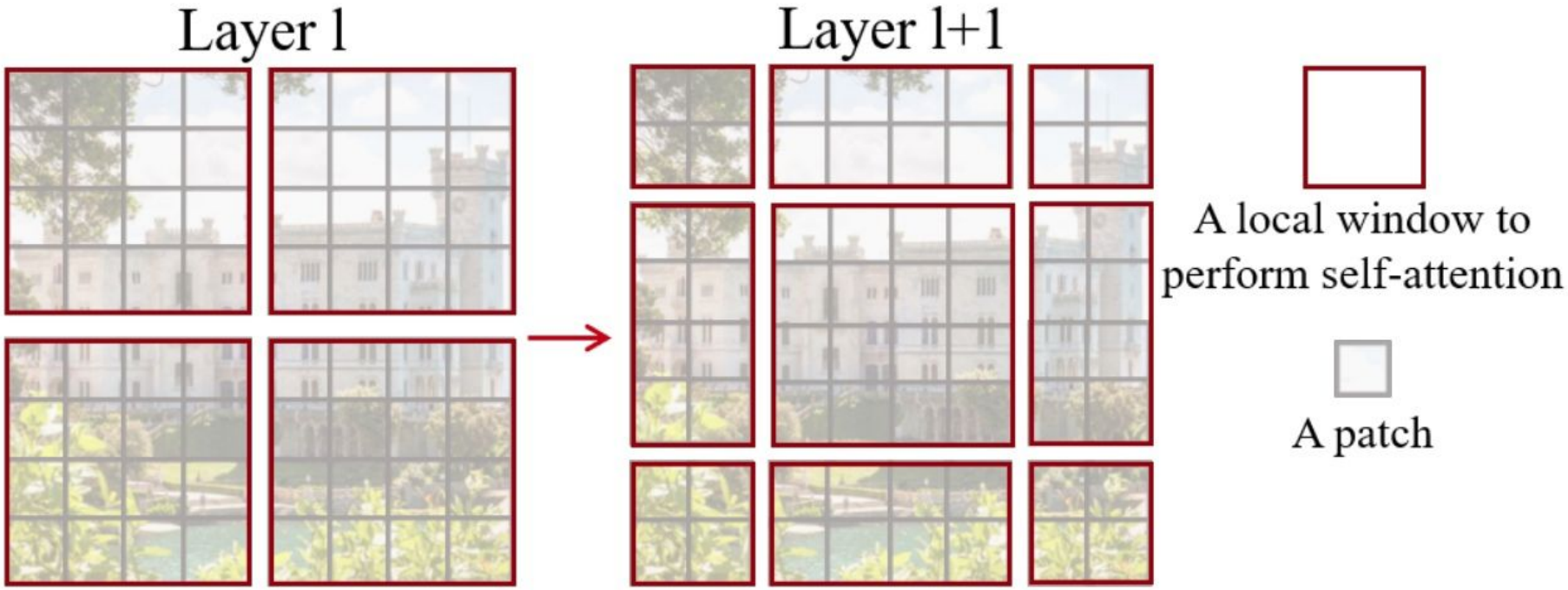


CLIP. Results

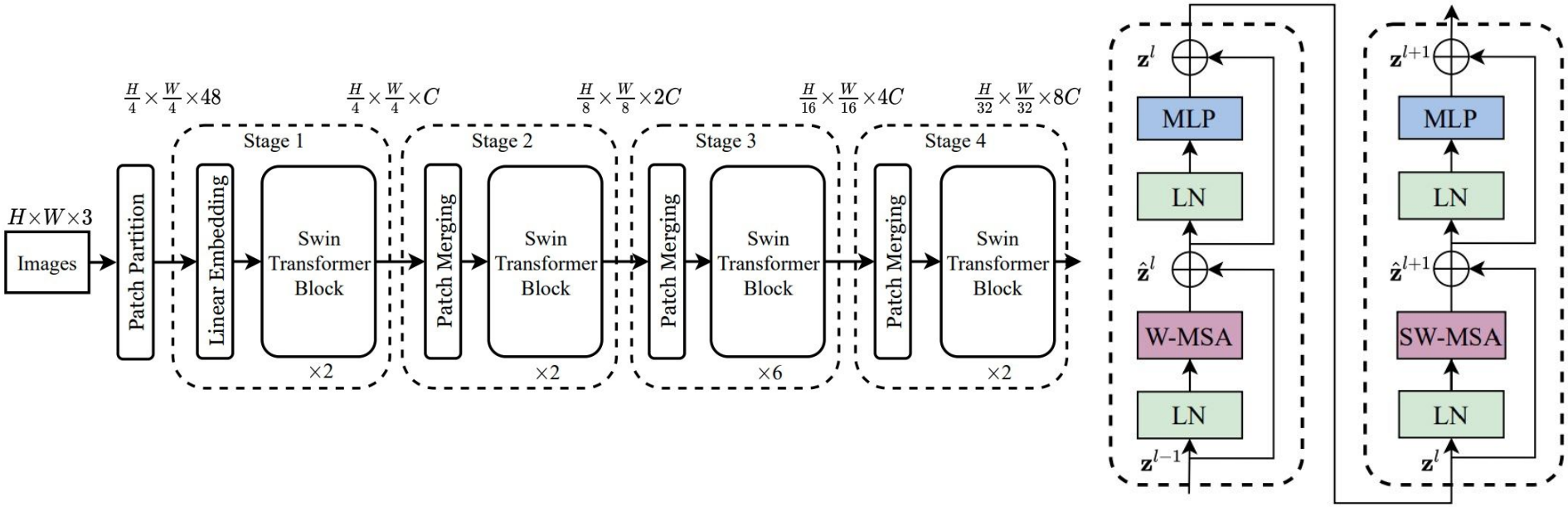
Swin


classification
segmentation
detection …
16×
8×
4×
(a) Swin Transformer (ours)
classification
16×
16×
16×
(b) ViT

Swin


Layer l
Layer l+1
A local window to perform self-attention
A patch

Swin


$H \times W \times 3$
Images
Patch Partition
$\frac{H}{4} \times \frac{W}{4} \times 48$
Stage 1
Linear Embedding
Swin Transformer Block
×2
$\frac{H}{4} \times \frac{W}{4} \times C$
Stage 2
Patch Merging
Swin Transformer Block
×2
$\frac{H}{8} \times \frac{W}{8} \times 2C$
Stage 3
Patch Merging
Swin Transformer Block
×6
$\frac{H}{16} \times \frac{W}{16} \times 4C$
Stage 4
Patch Merging
Swin Transformer Block
×2
$\frac{H}{32} \times \frac{W}{32} \times 8C$
$\mathbf{z}^{l}$
MLP
LN
$\hat{\mathbf{z}}^{l}$
W-MSA
LN
$\mathbf{z}^{l-1}$
$\mathbf{z}^{l+1}$
MLP
LN
$\hat{\mathbf{z}}^{l+1}$
SW-MSA
LN
$\mathbf{z}^{l}$