Journal Club Presentation

Presenter: Malika Bakhtawar

Research Article Information

Paper Title: Vision Transformer: To discover the “four secrets” of image patches

Paper Authors: Tao Zhou, Yuxia Niu, Huiling Lu, Caiyue Peng, Yujie Guo, Huiyu Zhou

Published Date: 11 January 2024

Journal: ELSEVIER, Information Fusion

Impact Factor: 17.564

Paper Highlights

In Vision Transformers ViT, there are four main steps, which are "four secrets", the existing research on transformer mainly focuses on the these four steps.
Therefore, they are crucial to improve ViT performance.

How to Select Patch?

How to Select Token?

How to select Position encoding?

How to Calculate Self Attention?

Introduction

- Transformer is an encoder-decoder architecture model which is first in the field of natural language processing.
- It utilizes its self attention mechanism to not only model the long dependencies but also supports parallel computation both in inference and training.
- CNN based methods has the problem of limited receptive field, which makes it difficult to model the long dependency.
- Transformer obtains global representation through attention mechanism, because each position in input sequence has access to information from all other positions.
- Through this Transformer can model global features and long-dependencies.

Basic Principle of ViT

To convert the image into the input sequence of Transformer encoder

Step 1:

The 2D image with the dimension
$X \epsilon R^{HxWxC}$
divided into a series of patches
$X_p \epsilon R^{Nx(P^2.C)}$
where,

(H, W) → size of the original image
C → number of channels
(P, P) → size of each patch
$N = \frac{H.W}{p^2}$ → number of patches

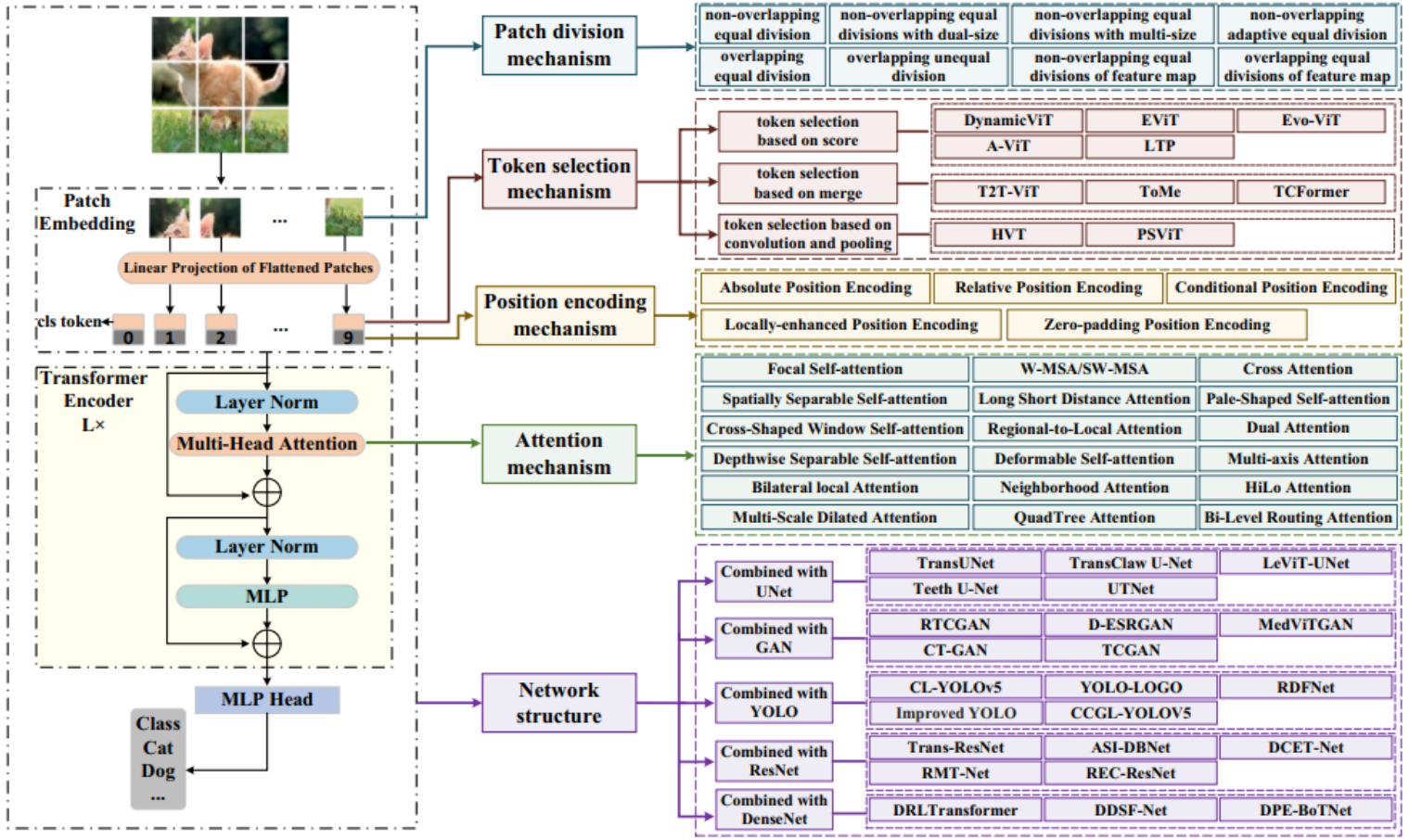


Fig. 1. The overall framework of ViT.

Step 2:

Each patch is then converted from 2D grid of pixels to one dimensional vector x_p^i, while keeping the information from all the channels.

Where $x_p^i \epsilon R^{Nx(P^2.C)}$ and i ϵ $(1, \ldots . N)$

Then by linear mapping through full connect network E, its corresponding token is obtained.

Mathematically,

$$z = [x_{class}, x_p^1 E, x_p^2 E, \ldots . . x_p^N] + E_{pos}$$

Where $E_{pos} \epsilon R\hat{}((N+1)xD)$

The resultant z is the sequence which inputs to the transformer encoder.

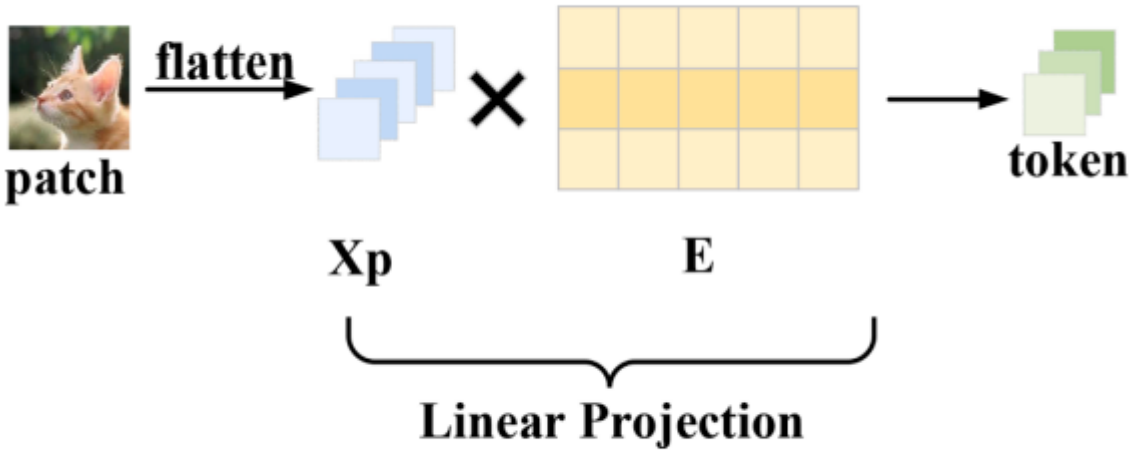


Fig. 2. Linear Projection of Flattened Patches.

Step 3:

Transformer encoder

The encoder consists of L identical Transformer modules stacked on top of each other.

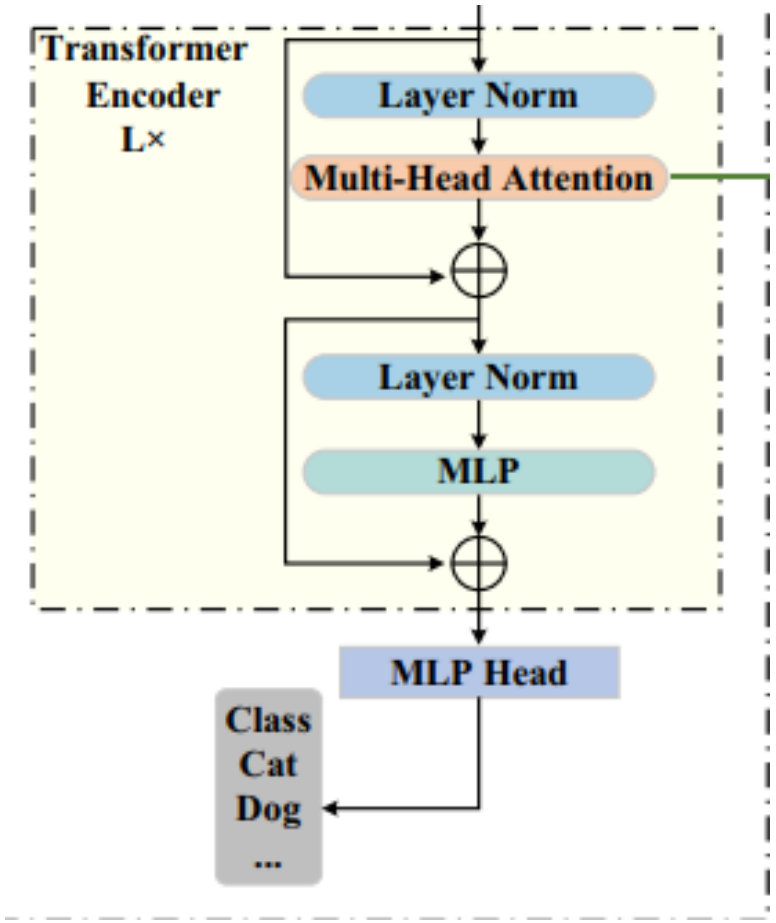
The output of one module becomes the input to the next.

Each module is composed of Layer Norm (LN), Multi-Head Self-Attention (MHSA), Multi-Layer Perceptron (MLP), and a residual connection.

Mathematically,

$$Z_i = MHSA\big(LN(Z_{i-1})\big) + Z_{i-1}$$

$$Z_i = MLP\big(LN(Z_i)\big) + Z_i$$



Patch Division Mechanism

- By dividing the image into multiple patches, local features in the image can be better extracted, thus the efficiency and accuracy of image processing are improved.

Several methods for Patch Division are developed such as

- Non-Overlapping Equal Division (Dual-size, multi-size)
- Non-Overlapping Adaptive Equal Division
- Non-Overlapping Equal Divisions of feature map
- Overlapping Equal Division
- Overlapping Unequal Division
- Overlapping Equal Division of feature map

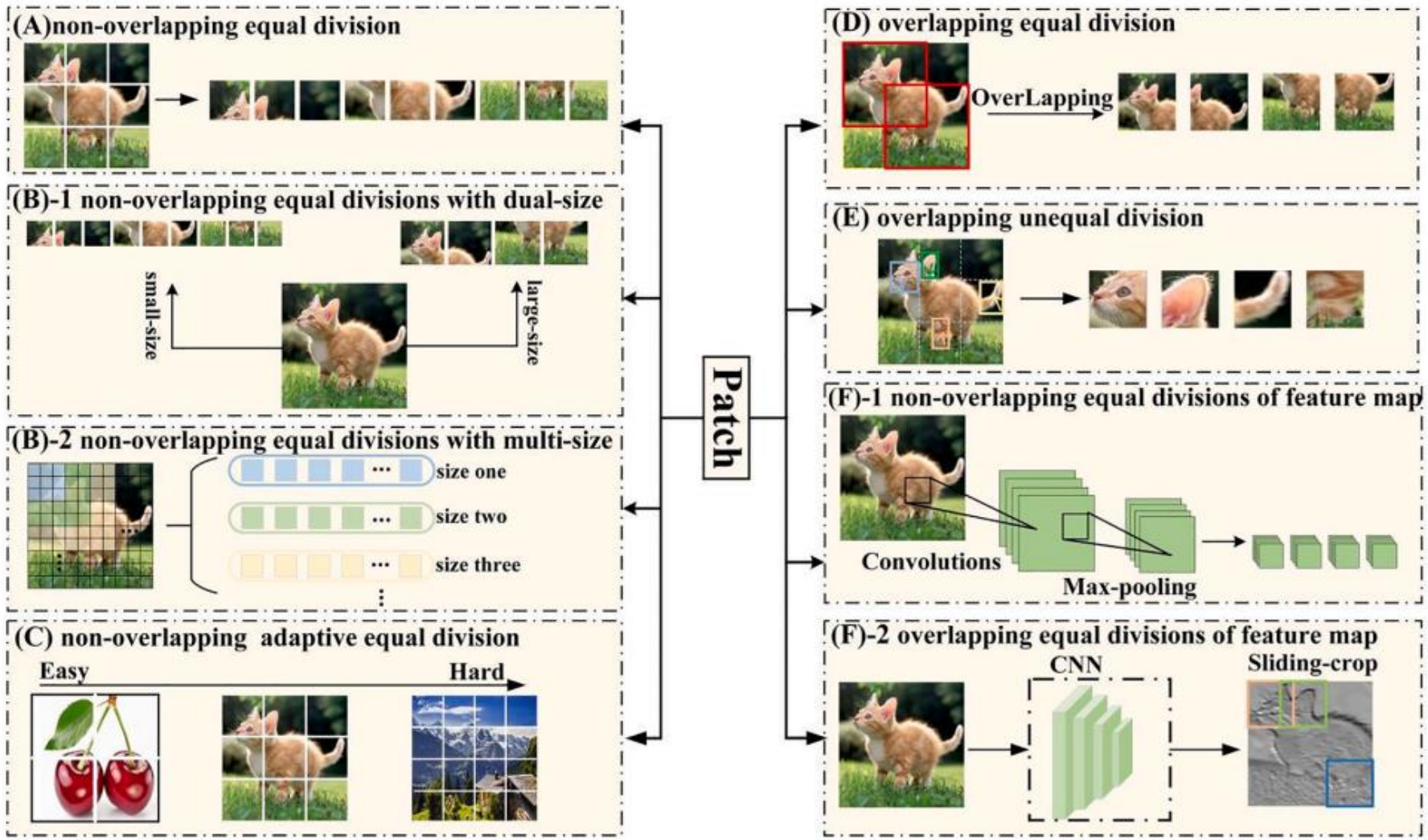


Fig. 4. Patch division mechanism.

From single-size division to multi-size division

Images are divided according to multiple sizes rather than a single size.

In the case of non-overlapping equal division, there are non-overlapping equal divisions with dual-size and multi-size patch division

Why is a dual and multi-size approach needed?

The stronger image features are produced when we combine patches of different sizes.

Uses:

Multi-size approach is effective in computer vision tasks such as detecting and classifying

It ensures that model can handle a wide range of object sizes, improving its overall performance and robustness.

From fixed number division to adaptive number division

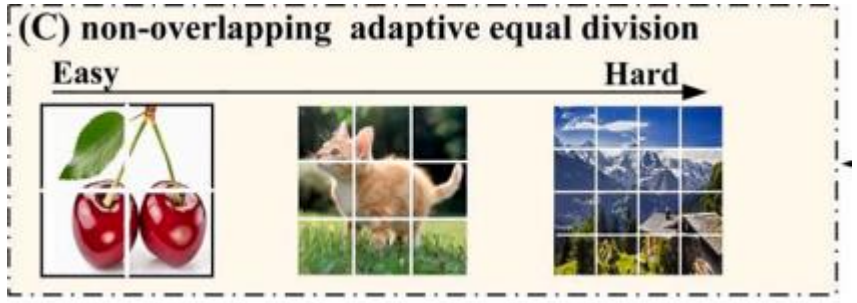
Images are divided into patches in a self-adaptive number rather than a fixed number.

Why is an adaptive number of patches needed?

- More patches contribute to better capturing of the data and improve prediction accuracy however, they also increase computational complexity which may lead to practical limitations, especially in terms of processing time and resource requirements.
- Therefore adaptive number of patches are very important.
- The model is activated sequentially starting from fewer patches,

  once a sufficiently confident prediction is produced,

  the inference is terminated immediately.



For "easy" images, only 2 × 2 patches are satisfied for accurate prediction, while for "hard" images, fine-grained representations are required to reduce information loss and improve computational efficiency.

From non-overlapping division to Overlapping Division

The image is serialized by overlapping and equal division, so that the neighboring patch overlaps half of the area.

Problem with non-overlapping division

Non-overlapping division destructs the image local continuity in some degree.

To solve this problem, overlapping approach is developed to enhance the semantic correlation between neighboring patches.

From semantic segmentation division to semantic aggregation division

The complete local structure of the target object is captured by regular patch is always difficult,

So in this approach the images are adaptively divided into patches of different positions and sizes

Therefore, they effectively capture the complete local structure of the target object.

From original image division to feature map division

Instead of original image, the feature maps are divided into patches.

Feature map generated by adopted pre-trained CNN contained more semantic information.

It includes the non-overlapping equal division of feature maps and the overlapping equal division of feature maps.

Feature map division (using convolutional layer and max pooling layer)

- Overlapping
- Non-overlapping

Token selection Mechanism

Next step after patch division and patch flattening,

One-dimensional vector is obtained by linear mapping is referred to as a "token”

The token selection mechanism is that redundant tokens are dynamically identified and tokens of higher importance are selected in the forward propagation process of the model.

Importance of Token Selection Mechanism

When we use Transformers for visual tasks especially in the context of models like the Vision Transformer (ViT),

The calculated amount increases exponentially as the number of tokens in the input sequence increases.

The final prediction is often influenced more by certain tokens that carry a significant amount of information, while many other tokens may be redundant or contribute less to the final prediction.

Therefore, focusing more on tokens with significant information and considering others as redundant helps improve the efficiency of the model.

Three methods for token selection mechanism

1. Token selection based on score
2. Token selection based on merge
3. Token selection based on convolution and pooling

- The token selection mechanism based on score is that the importance of tokens is scored by the scoring function, high-score tokens are retained through pruning operation, and low-score tokens are deleted.
- The token selection mechanism based on merge is that similar tokens are selected and combined by matching algorithm. The key operations in this method are selection strategy and merge strategy which reduce the information loss and improves the training speed .
- The key operations in this method are convolution strategy and pooling strategy. It is based on convolution and pooling is that the tokens' sequence length is shrunk by convolution and pooling operations, similar to feature maps down sampling in CNN

Table 1
Performance comparison of different token selection mechanisms.

No.	Strategy	Model	Params(M)	FLOPs(G)	Throughput (images/s)	Dataset	Top-1 acc
1	DynamicViT[29]	DynamicViT-LV-S/0.5	26.9	3.7	-	ImageNet-1K	82.0
2	EViT[30]	EViT-LV-S/0.5	-	3.9	3603	ImageNet-1K	82.5
3	Evo-ViT[31]	Evo-LeViT-256/0.5	19.0	-	4277	ImageNet-1K	78.8
4	A-ViT[32]	A-ViT-S	22	3.6	1.1k	ImageNet-1K	78.6
5	LTP[33]	LTP	*				
6	T2T-ViT[34]	T2T-ViT-14	21.5	4.8	-	ImageNet-1K	81.5
7	ToMe[35]	ToMe-DeiT/r_{13}	-	2.7	1552	ImageNet-1K	79.4
8	TCFormer[36]	TCFormer	25.6	5.9	-	ImageNet-1K	82.4
9	HVT[37]	HVT-S	22.09	2.4	-	ImageNet-1K	78.0
10	PSViT[38]	PSViT-1D-Base	-	18.9	-	ImageNet-1K	82.6
		PSViT-2D-Base	-	15.5	-	ImageNet-1K	82.9

Attention Mechanism

- As the key component of ViT, self-attention is a powerful tool to capture long-distance dependency relationships.
- There are 18 attention mechanisms are summarized in this paper and each one is developed to improve performance and reduce computing complexity and storage costs.

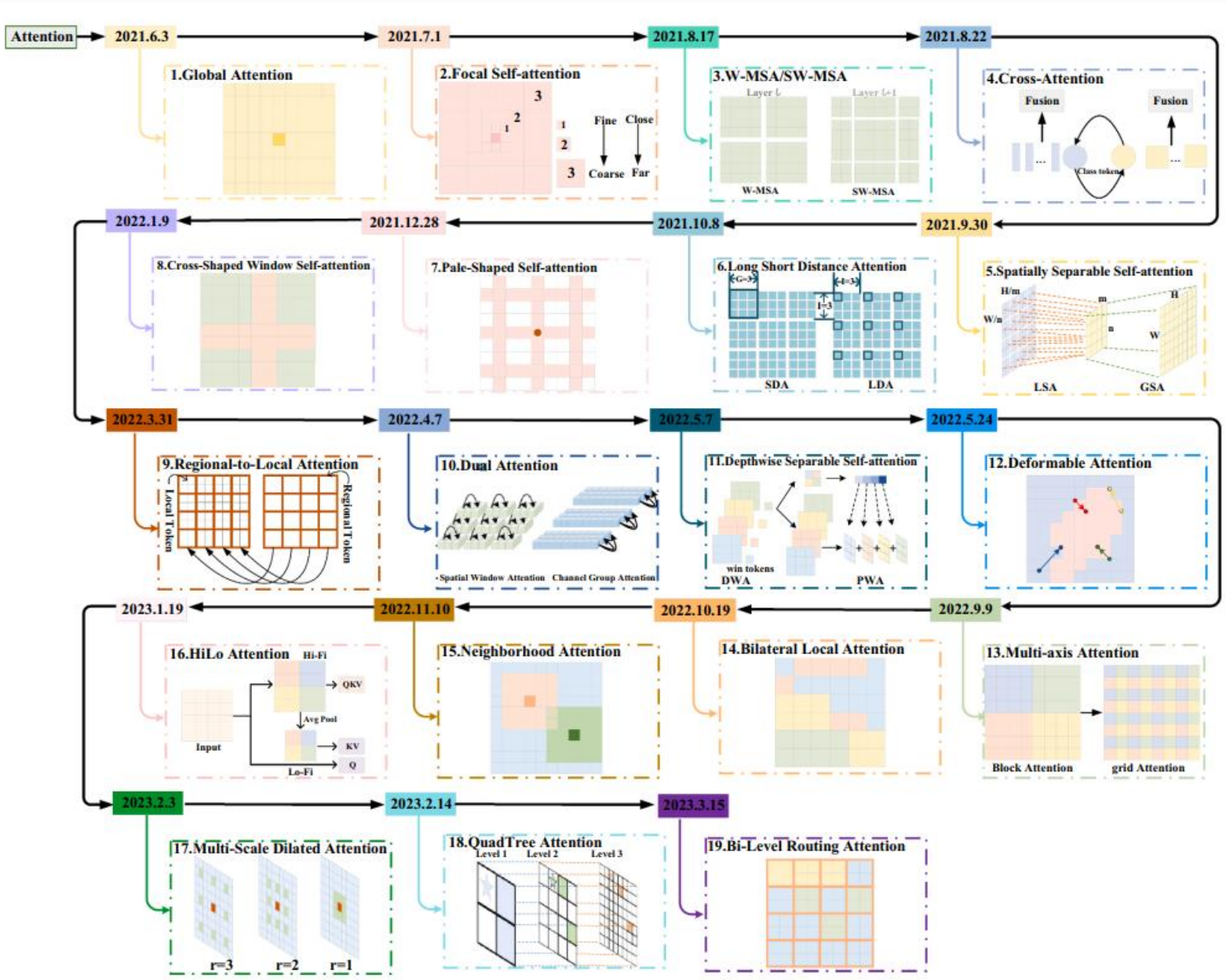


Fig. 6. Attention mechanism.

1: Focal Self-attention(FSA), FSA incorporates both fine-grained local features and coarse-grained global features

Fine-Grained Local Features:

- Each token mainly pays attention to its nearby tokens. This helps the model focus on small details and specific information in close proximity.

Coarse-Grained Global Features:

- Additionally, each token also pays attention to tokens that are far away. This allows the model to capture broader, long-distance relationships and understand the context of the entire input.

2: W-MSA represents Windows Multi-head Self-Attention, in layer l and SW-MSA represents Shifted Windows Multi-Head Self-Attention, in the next layer l + 1. The feature maps are divided into multiple non-overlapping windows, and self-attention is performed in each window.

Both mechanisms optimize attention computation by focusing on specific regions, enhancing efficiency and capturing local and global information in a layered manner.

3: Since the semantic information among all image tokens is already learned by the class token in its own branch, the interaction can fuse information in different scales between the class token at one branch and the image tokens at the other branch. Therefore Cross-Attention(CA) is proposed to realize the information exchange between two branches, and fuse multi-scale features more effectively.

4: Spatially Separable Self-Attention(SSSA), SSSA is composed of locally-grouped self-attention (LSA) and global subsampled attention (GSA). It is proposed to capture fine-grained short-range information and to process global information over long distances.

- In LSA, the feature maps are equally divided into sub-windows, and self-attention calculations are performed within each sub-window
- In GSA, a single representative is used to summarize the important information.

Why we need improvement in attention mechanism ?

Quadratic Complexity of Self-Attention

The self-attention mechanism has a quadratic time complexity with respect to the sequence length. For each position in the sequence, the model computes attention weights with all other positions. This results in an $O(N^2)$complexity, where N is the number of tokens.

Exponential Increase with More Tokens

As the number of tokens in the input sequence increases, the computational cost of self-attention grows exponentially. For example, doubling the number of tokens quadruples the computational load because each token needs to attend to all other tokens, leading to a quadratic increase in pairwise interactions.

This quadratic nature of the attention mechanism leads to challenges in terms of memory and processing requirements, and make it impractical for very long sequences.

Therefore the series of work is proposed to improve its performance .

Second highest accuracy →Pale-Shaped Self-attention(PSA), In PSA, the input feature maps are split into multiple pale-shaped regions, and the attention calculations are performed within each region.

Top accuracy → Multi-axis Attention(MaxA), MaxA is composed of Block Attention and Grid Attention.

In Block Attention, the feature maps are divided into non-overlapping local Windows, and self-attention is performed within each window.

In Grid Attention, the feature maps are divided into fixed grid blocks, and self-attention is performed at corresponding positions in each grid block.

Table 3
Performance comparison of 18 attention mechanisms.

No.	Model	Attention	Params(M)	FLOPs(G)	Throughput (images/s)	Dataset	Top-1 acc
1	ViT(Base)[9]	Global Attention	86	55.4	85.9	ImageNet-1K	77.9
2	Focal Transformer (Base)[44]	Focal Self-attention	89.8	16.0	-	ImageNet-1K	83.8
3	Swin Transformer (Base)[39]	W-MSA,SW-MSA	88	15.4	278.1	ImageNet-1K	83.5
4	CrossViT (Base)[22]	Cross-Attention	104.7	21.2	239	ImageNet-1K	82.2
5	Twins (Base)[45]	Spatially Separable Self-attention	56	8.6	469	ImageNet-1K	83.2
6	CrossFormer (Base)[46]	Long Short Distance Attention	52.0	9.2	-	ImageNet-1K	83.4
7	Pale Transformer (Base)[47]	Pale-Shaped self-Attention	85	15.6	-	ImageNet-1K	84.9
8	CSWin Transformer (Base)[41]	Cross-Shaped Window Self-attention	78	15.0	250	ImageNet-1K	84.2
9	RegionVIT (Base)[48]	Regional-to-Local Attention	72.7	13.0	-	ImageNet-1K	83.2
10	DaViT (Base)[49]	Dual Attention	87.9	15.5	-	ImageNet-1K	84.6
11	SepViT (Base)[50]	Depthwise Separable Self-attention	82.3	13.1	308	ImageNet-1K	84.0
12	DAT (Base)[51]	Deformable Attention	88	15.8	-	ImageNet-1K	84.0
13	MaxViT (Base)[52]	Multi-axis Attention	120	23.4	133.6	ImageNet-1K	84.95
14	BOAT (Base)[53]	Bilateral Local Attention	90	17.5	-	ImageNet-1K	84.7
15	NAT (Base)[54]	Neighborhood Attention	90	13.7	783	ImageNet-1K	84.3
16	LITv2 (Base)[55]	HiLo Attention	87	13.2	602	ImageNet-1K	83.6
17	DilateFormer (Base)[56]	Multi-Scale Dilated Attention	47	10.0	-	ImageNet-1K	84.4
18	QuadTree (Base)[57]	QuadTree Attention	64.2	11.5	-	ImageNet-1K	84.0
19	BiFormer (Base)[58]	Bi-Level Routing Attention	57	9.8	-	ImageNet-1K	84.3

Conclusion

- In conclusion, the attention mechanism, token selection, and patch division components of the Vision Transformer (ViT) are interconnected in a detailed and intricate manner and form a unified framework for image processing tasks.
- The attention mechanism enables tokens to capture relationships across the entire input sequence, allowing the model to understand both local and global features.
- Patch division introduces spatial organization by breaking down the image into smaller patches, enabling the model to focus on localized information.
- Together, these components cooperatively enhance the ViT's ability to efficiently process images, balancing the extraction of fine-grained details and capturing broader contextual information for robust visual understanding.

Thank You!

Any questions?