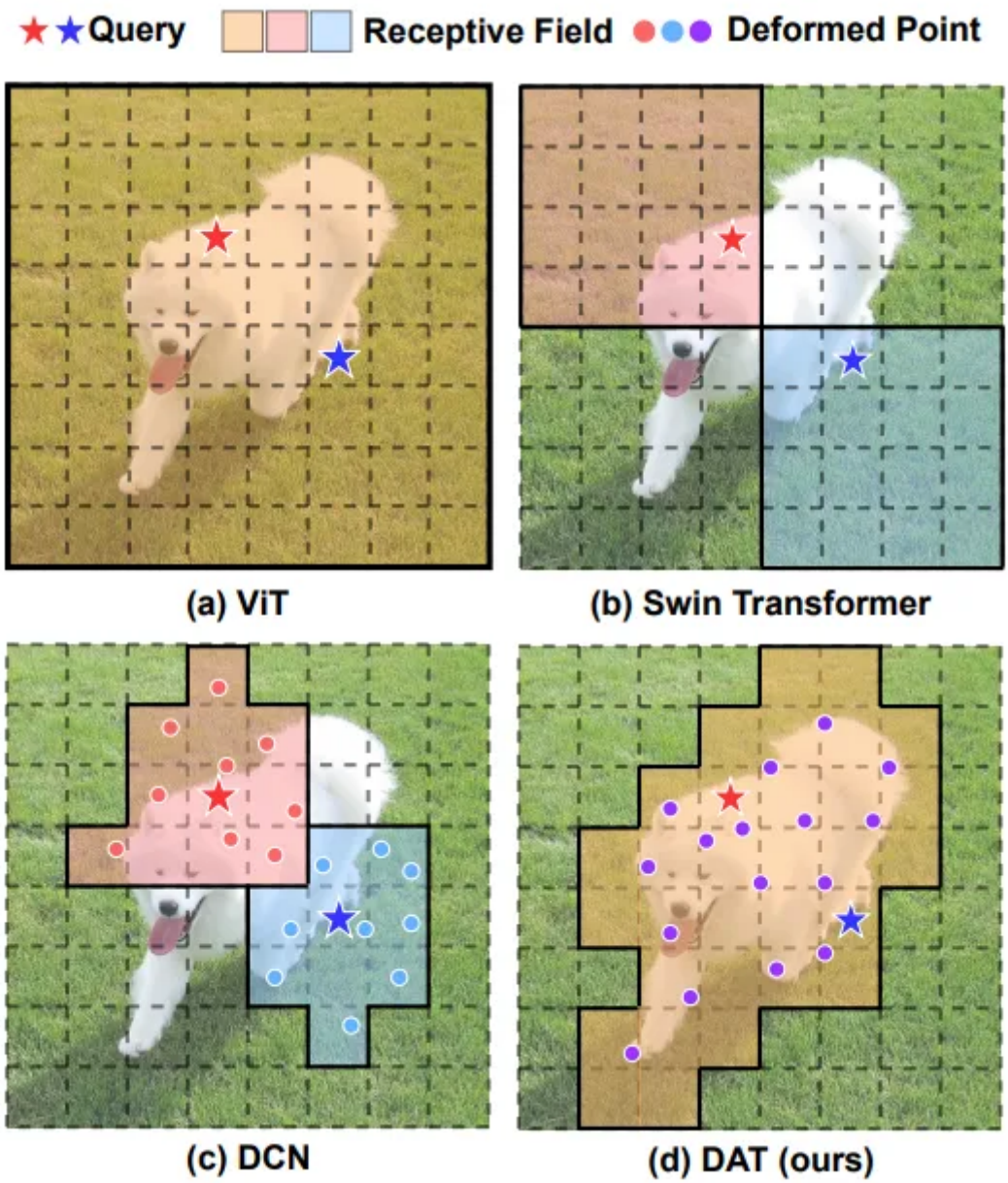
Joe El khoury · Follow

13 min read · Jun 21, 2023

Listen Share More

This post is based on findings made in this paper



Introduction

In recent years, Transformer, which uses the Attention mechanism, has shown remarkable performance in the field of natural language processing, and has become the de facto standard in the field of natural language. In the field of image processing, CNN using the convolution mechanism was the de facto standard, but since then attempts have been made to incorporate Transformer into the field of image processing.

Initially, models combined with CNN were devised, but after the announcement of Vision Transformer (ViT), which eliminates CNN and is built only with Transformer, models based only on this Transformer have also been used in the field of image processing and recognition. Moreover, it is known for treating images as “sequence data” consisting of a series of image patches.

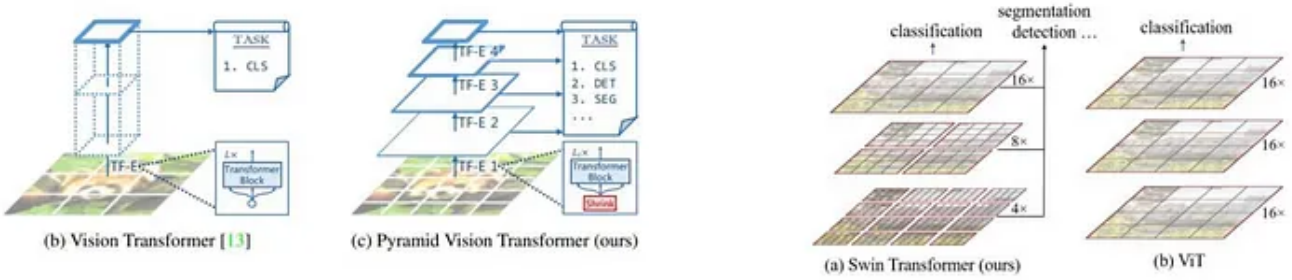
The advantage of using Transformer for image processing like ViT is its wide

receptive field. By suppressing a wider area than CNN, etc., it is possible to obtain better feature values.

On the other hand, ViT, which is simply a repurposed Transformer, is also called Dense Transformer, and has the major disadvantage of increased memory requirements, high computational cost, delayed learning convergence and dangers of overlearning. since the relationship of all images is obtained, the calculation cost is enormous, and there is no inductive bias such as close pixels having a deep relationship, and it can surpass CNN only by learning with a large amount of data. There is a problem that accuracy cannot be achieved. Another problem is that irrelevant locations can affect the features.

Therefore, the Pyramid Vision Transformer (PVT), also known as the Sparse Transformer, and the SwinTransformer were created as improvements to ViT. These are models that improve memory efficiency and computational efficiency by paying attention to areas in the image that have been narrowed down to some extent. Compared to ViT, performance and improved, but on the other hand since the area in the image is narrowed down, it is possible that the extensive relationship information obtained from the original area is lost.

On the other hand, Swin does not calculate Self-Attention for the entire image, but divides the image into small regions and performs Self-Attention within the small regions. Compared to ViT, Swin has improved accuracy in ImageNet-1k, but it can only acquire relationships within a small area and may lose information on global relationships. In the receptive field of self-attention determined manually like Swin, important information may be missed.



However, an Attention range constructed manually, such as the Swin Transformer, may not be optimized in terms of efficiency. It is also possible that important Key/Value relationships were dropped while unnecessary ones were used. Ideally,

the Attention range should be freely transformable for each input image, while allowing only important regions to be used.

Deformable Attention Transformer (DAT) was proposed to solve the problem. it is a general backbone model with deformable attention for both image classification and dense prediction tasks.

The Transformer model revolutionized the implementation of attention by dispensing with recurrence and convolutions and, alternatively, relying on self-attention. This allows the model to adapt to the input data and improve its performance. The Deformable Attention Transformer (DAT) proposed this time uses Deformable self-attention, which enables selection of areas with more influence relationships when narrowing down areas like PVT and Swin Transformer.

This is an improved model that utilizes deformable self-attention so that more relevant regions can be selected when limiting the range of self-attention. In other words, it is a model that can control the range of self-attention more flexibly. So it learns different deformed points for each query in a data-dependent way. As a result, it has succeeded in improving efficiency and performance over conventional image processing models. Thus, for this reason this model has been used in various applications, including image classification and dense prediction tasks.
As a result, it achieved SOTA in class classification, object detection, and segmentation tasks by hitting out accuracy that surpasses Swin.

Short overview about Deformable Convolution Network

As previously said , visual recognition tasks, such as object detection or image classification, often grapple with handling object scale, pose, viewpoint, and distortion. Previous methodologies, although effective to some extent, have inherent drawbacks.

Let’s look at the two prevalent approaches:

1. One strategy is to enhance data diversity through augmentation and increase the model capacity. While this may yield more nuanced models, it comes with the demand for large datasets and potentially gargantuan models.
2. Another approach involves leveraging deformation-invariant features and algorithms, like SIFT and max-pooling. However, these are typically hand-crafted, limiting their generality.

Moreover, traditional Convolutional Neural Networks (CNNs), due to their geometrical structure, can be susceptible to geometric deformations. They perform operations on fixed positions, and all layers share the same receptive field shape and size, which might not always be suitable for tasks like object semantic recognition.

To overcome these challenges, Deformable convolution network (DCN) was proposed as a novel solution involving deformable convolutions, Region of Interest (RoI) pooling.

To understand better the deformable attention transformer, let's talk first about Deformable Convolution Network. The latter is based on the idea that "the receptive field should adapt according to the scale and shape of the object".

Deformable Convolutions and RoI Pooling

that solution introduces adjustable 2D offsets into standard grid-like convolutions and a learnable offset to bin positions in RoI pooling. This flexibility allows the model to handle a wider variety of shapes and sizes of objects in images.

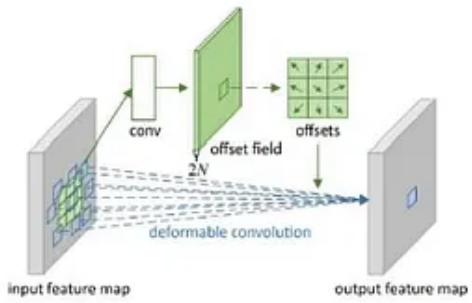


Figure 2: Illustration of 3 × 3 deformable convolution.

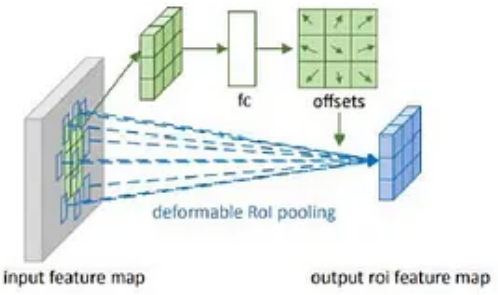


Figure 3: Illustration of 3 × 3 deformable RoI pooling.

while ordinary convolution tries to calculate y (p0) as:

$$\mathbf{y}(\mathbf{p}_0) = \sum_{\mathbf{p}_n \in \mathcal{R}} \mathbf{w}(\mathbf{p}_n) \cdot \mathbf{x}(\mathbf{p}_0 + \mathbf{p}_n)$$

In a deformable convolution an offset is added as

$$y(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n + \mathrm{D}p_n)$$

p_0: pixels on the output map obtained by Convolution
$\mathcal{R}, p_n$: pixel set and pixels on the input map used in normal Convolution
$\mathrm{D}p_n$: displacement, offset
$x()$: Pixel value calculated by bilinear interpolation
$w()$: weight

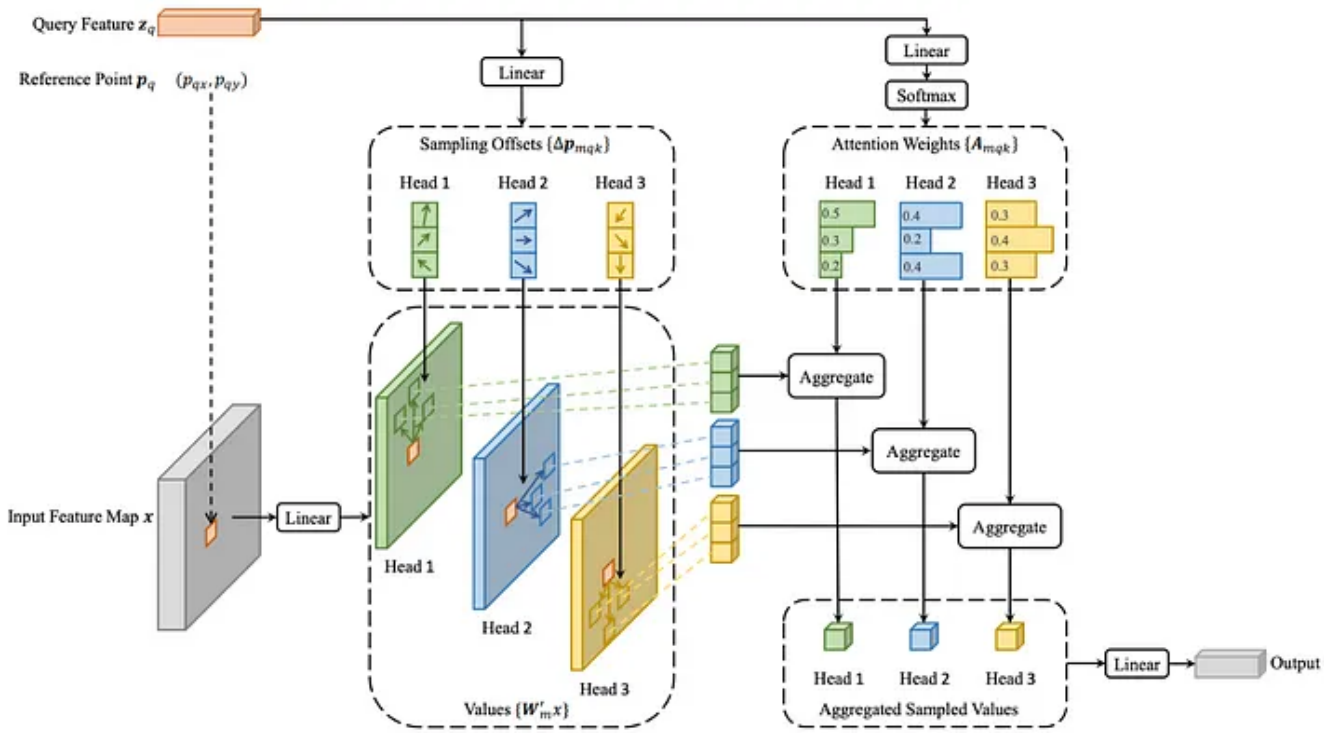
A simple application of what is done in the DCN to the Transformer would require high memory and computational costs, making it impractical.

DAT and its surroundings

Deformable attention transformer (DAT), which can be a core network for image classification, object detection, and segmentation tasks, brings flexibility and efficiency to the image recognition domain.

The key component of DAT is Deformable Attention (DA). It efficiently models relationships among tokens by focusing on important regions in the feature map. The attention area is obtained using deformable sampling points learned from queries with offset networks.

Unlike Deformable Convolutional Networks (DCN) that learn different regions for different pixels in the feature map, DAT learns query-agnostic region groups. Recent studies show that global attention results in almost identical attention patterns for different queries, which allows for focused key/values in important areas and more efficient computations.



Overcoming High Calculation Costs

However, despite these benefits, the calculation costs can be potentially high. To address this, a strategy was proposed that involves generating reference points from the input feature map, normalizing these reference points, generating offsets using a subnetwork, and then performing bilinear completion on the deformed reference point.

Promoting Diversity in Deformed Points

The feature channel are divided into several groups to encourage diversity in Deformed Points. Similar to the Multi-Head Self-Attention (MHSA) method, features based on each group utilize a shared subnetwork to generate reasonable offsets.

Enhancing Spatial Information with Deformable Relative Position Bias

Incorporating Deformable Relative Position Bias into DAT also enhances spatial information in Attention operations.

DAT's Model Architecture

In terms of computational cost, Deformable Multi-Head Attention (DMHA) compares favorably with models like the PVT and Swin Transformer. The difference lies mainly in the computational complexity of the offset network.

DAT uses a multi-scale feature map and employs Deformable Attention in later stages to model relationships in broader regions. For classification tasks, we use a linear classifier with pooled features. For object detection and segmentation tasks, DAT serves as the backbone of the model, extracting multi-scale features.

(a) **ViT (Vision Transformer)**: ViT applies self-attention to the entire image, leveraging a global receptive field to capture overarching features.

(b) **Swin Transformer**: Unlike the ViT, the Swin Transformer limits its receptive field, performing self-attention within these defined boundaries.

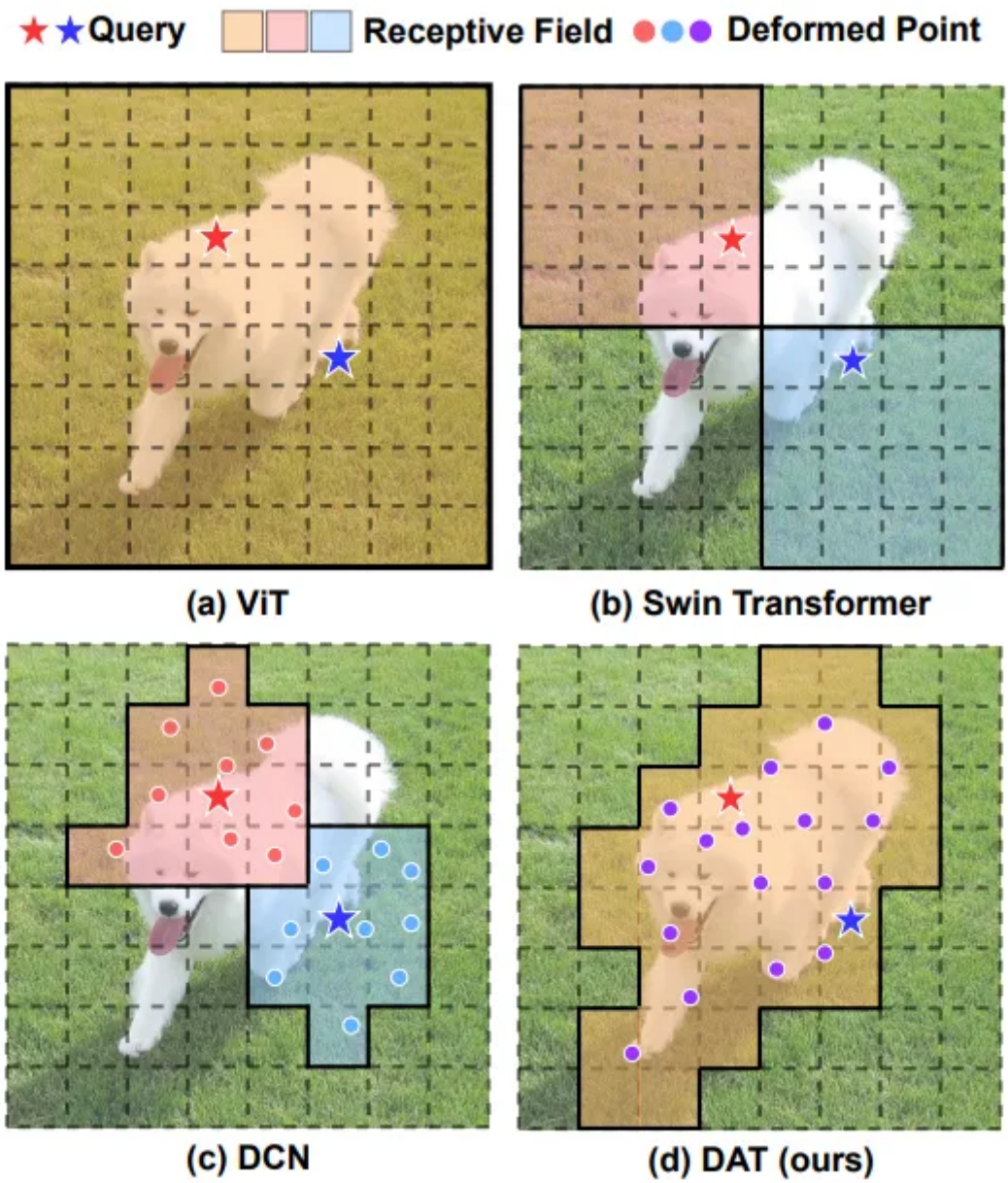
(c) **DCN (Deformable Convolution Networks)**: DCN is a model based on Convolutional Neural Networks (CNN), and it performs operations on deformable receptive fields.

DAT Structure Overview

The overall structure of DAT follows a four-stage hierarchical design similar to ResNet. With each stage progression, the spatial size of the feature map is halved while the number of channels is doubled, leveraging convolutional layers for downsampling between stages.

To reduce computational effort, the first 4x4 convolution downsamples to 1/4 of the image size. Stages 1 and 2 implement local Attention and Shift-Window Attention — restricted receptive field self-attention recognition methods from the Swin Transformer. Meanwhile, stages 3 and 4 employ local attention and deformable attention, performing alternate local and global recognition for accuracy enhancement.

Interestingly, DAT only adopts deformable attention in the latter half of the process. This is due to the ViT model's tendency to prefer local recognition in early recognition stages and an effort to reduce computation.



Model structure

The overall model structure adopts a four-stage hierarchy like ResNet. As the stages progress, the spatial size of the feature map is halved and the number of channels is doubled. This downsampling between stages uses convolutional layers. (k=kernel size, s=stride).
The first 4x4 convolution is downsampled to 1/4 the image size to reduce computational effort.

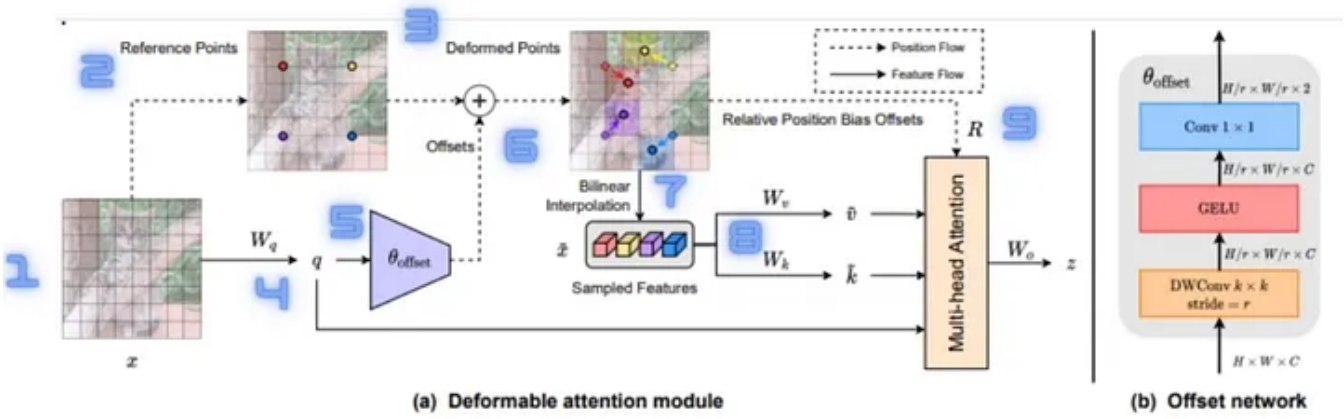
Stages 1 and 2 adopt local Attention and Shift-Window Attention. These are self-attention recognition methods with restricted receptive fields used in Swin Transformer.

Stages 3 and 4 adopt local attention and deformable attention. Contributes to accuracy improvement by alternately performing local recognition and global

recognition.

The reason why deformable attention is adopted only in the latter half of the process is that the ViT model tends to prefer local recognition in the early stages of recognition, and in order to reduce the amount of computation, deformable attention is adopted only in the second half.

Deformable attention module



The input image x (H×W×C), with reference points (Hg×Wg= HxW/r^2), where r is manually determined, is processed as follows where Hg=H/r and Wg=W/r:

(i) The input is a feature map x with dimensions $H \times W \times C$.

(ii) We select pixels p as reference points from a uniform grid of dimensions $HG \times WG \times 2$ (where $HG = H/r$, and $WG = W/r$, effectively downsampling the original grid).

(iii) These reference points are linearly projected onto a two-dimensional coordinate system with points ranging from (0,0) to ($HG - 1$, $WG - 1$), and normalized between [-1, +1], where the top-left corresponds to (-1,-1).

(iv) To acquire an offset from each reference point, the feature map is linearly projected, resulting in the query token $q = xWq$.

(v) The query token q is then passed into the subnetwork θ *offset* to generate the offset.

$$\Delta p = \theta_{\text{offset}}(q)$$

and in order to ensure a stable learning process, a predefined value s is employed to prevent Δp from becoming too large, via the transformation $\Delta p \leftarrow s \tanh(\Delta p)$.

(vi) The deformed reference points are obtained by combining the reference point and the offset information.

(vii) We then conduct a bilinear interpolation on these deformed reference points, sampling the feature $\tilde{x}$.

$$\tilde{x} = \phi(x; p + \Delta p) \qquad \phi(z; (p_x, p_y)) = \sum_{(r_x, r_y)} g(p_x, r_x) g(p_y, r_y) z[r_y, r_x, :]$$

(viii) A linear projection is carried out on the result from step (viii) to secure the key token $k\tilde{} = x\tilde{} \, Wk$ and the value token $v\tilde{} = x\tilde{} \, Wv$.

(ix) Finally, attention is applied in a way that integrates position embedding-like information, culminating in the final output.

Offset network in brief (subnetwork)

In this subnetwork, the offset value is calculated for each reference point using a query. since the input image x undergoes a linear transformation to obtain the query (q), which is then inputted into the offset network. a subnet with two convolution modules with nonlinear activation functions is implemented.
First a kxk (5 × 5 in the paper) depthwise convolution is used to acquire local features. Then the offset network utilizes the **GelU** function between two convolutions. The convolution's kernel in the DW convolution convolves spatial information.
Then a **1x1 convolution** that convolves in the channel direction compresses to 2 channels (horizontal, vertical). The feature map stores the vertical and horizontal distance values corresponding to each reference point.

Keys and Values

Translate the reference point using the values determined by the offset network. Determine
the value of the reference point to which it is moved by bilinear interpolation (to deal with floating numbers).
Feature map using reference point determined values x （Hg×Wg×C) and create x
Then linearly transform from to key and value.

In order to encourage diversity in the Deformed Points, the feature channel is split into G groups, a strategy reminiscent of the Multi-Head Self-Attention (MHSA) technique. Feature subsets within each group exploit a shared subnetwork to produce correlated offsets. Practically, the Multi-Head Attention units' count is made G times the number of offset groups, ensuring that every transformed group of key and value tokens is assigned multiple Attention Heads.

Additionally, the relative position bias (between 7 and 9) encapsulates the relative position between all possible query-key pairs, enhancing the conventional Attention mechanism with spatial data. Lastly, within the framework of DAT, the normalization value serves as a position embedding, accounting for a continuous relative displacement to cover all potential offset values.

Thus, multi-head attention applies, where the input query, key, and value are derived via:

$$q = xW_q, \tilde{k} = \tilde{x}W_k, \tilde{v} = \tilde{x}W_v$$

Self-attention applies the following equation, where B indicates Deformable relative position bias:

$$z^{(m)} = \text{Softmax}\left(q^{(m)}\tilde{k}^{(m)^T}/\sqrt{d} + \phi\left(\tilde{B}; R\right)\right)\tilde{v}^{(m)}$$

Deformable multi-head atten-tion (DMHA) has similar computational costs, such as PVT and Swin Transformer. The difference is the computational complexity of the offset network.

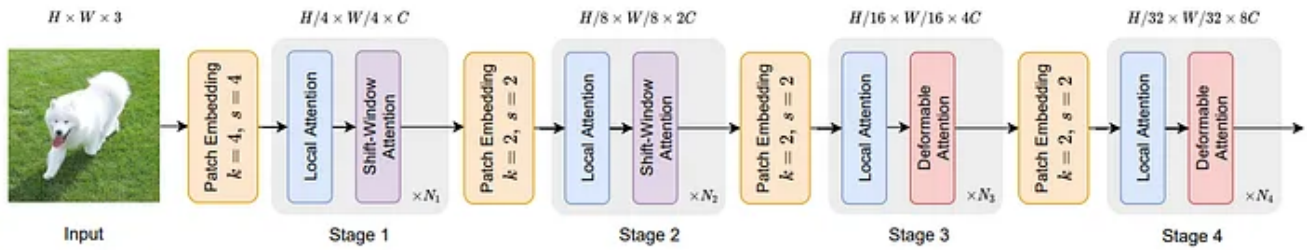
$$\Omega(\text{DMHA}) = \underbrace{2HWN_sC + 2HWC^2 + 2N_sC^2}_{\text{vanilla self-attention module}} + \underbrace{(k^2+2)N_sC}_{\text{offset network}},$$

where Ns= HgxWg

while Swin Transformer has a computation cos of 79.63M Flops, the computational cost incurred by adding subnetworks is approximately 5.08M Flops. Note that the

computational cost can be further reduced by increasing the value of r — the downsampling factor.

Model architecture



The overall architecture of the model follows a four-stage hierarchical structure reminiscent of ResNet. As the stages progress, the spatial dimensions of the feature map halve while the number of channels doubles, enabled by the implementation of convolutional layers. In the initial step, the first 4x4 convolution is reduced to one-fourth the image size to curtail computational effort.

The DAT, recognizing the necessity of multi-scale feature maps for image tasks, employs a similar hierarchical arrangement to form feature pyramids. During the first and second stages, Deformable Attention (DA) isn't implemented as the primary objective is to grasp more local features. DA is avoided here due to the considerable spatial overhead and computational cost. Instead, the model integrates local information with Shift-Window Attention, a window-based local attention mechanism used in the Swin Transformer.

The third and fourth stages bring in Deformable Attention, allowing for the modeling of relationships that transition from more local to broader regions. In classification tasks, a linear classifier is employed alongside pooled features to first normalize the feature map output from the final stage before predicting the logit.

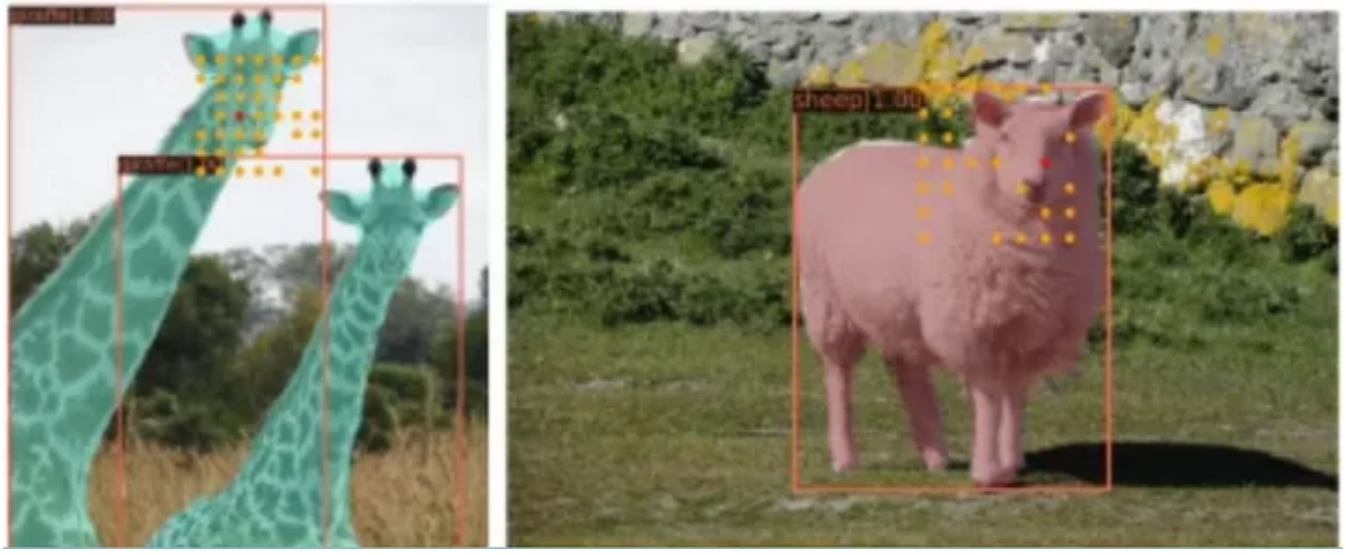
For object detection and segmentation tasks, DAT operates as the backbone of the model, extracting multi-scale features. Moreover, in tasks such as object detection and semantic segmentation decoders, a normalization layer is incorporated into each stage's functionality before feeding it into the subsequent module, a process similar to the FPN approach. This structured approach, which balances local and global recognition and manages computational costs, results in enhanced accuracy and efficiency in various tasks.

Stage adopted by Deformable Attention and Results

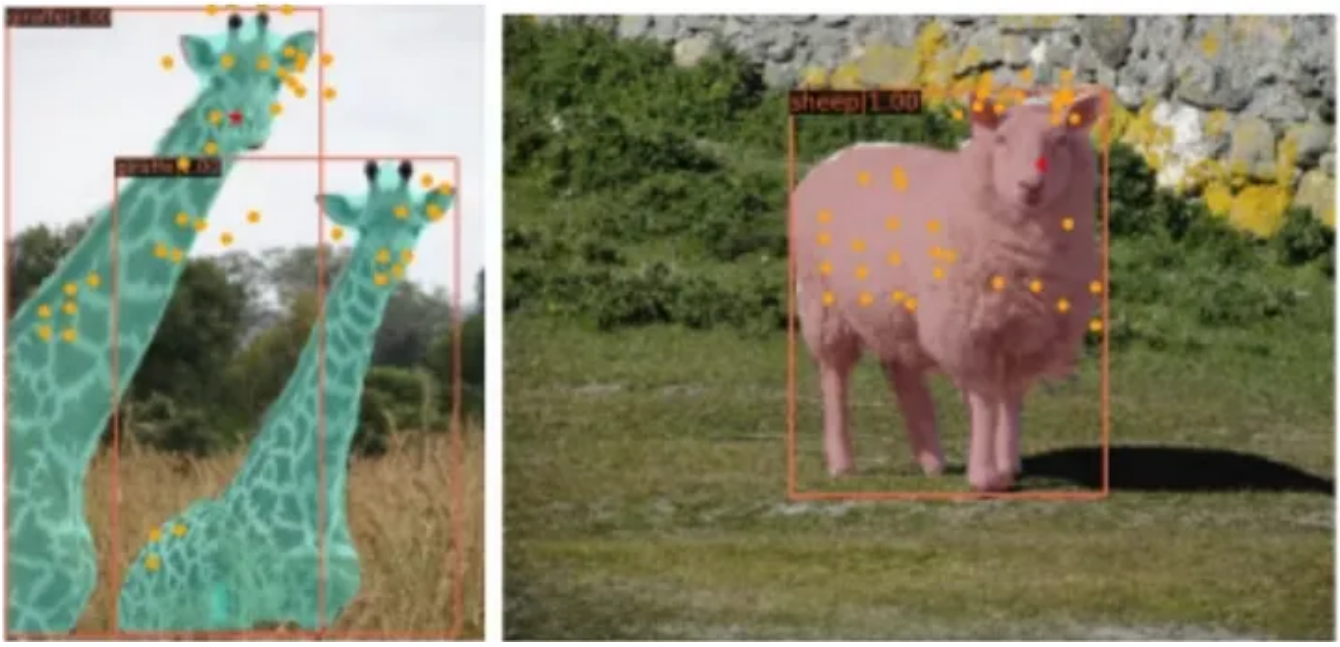
It can be seen that the accuracy is high by adopting it in stages 3 and 4 in the stage adopted by Deformable Attention of the classification task of ImageNet-1k .

DAT outperforms conventional methods in tasks like image classification (ImageNet-1K), object detection (COCO), and segmentation (ADE20K). Particularly in ImageNet-1k classification, stages 3 and 4 of Deformable Attention adoption achieve high accuracy.

Stages w/ Deformable Attention				FLOPs	#Param	Acc.
Stage 1	Stage 2	Stage 3	Stage 4			
✓	✓	✓	✓	4.64G	28.39M	81.7
	✓	✓	✓	4.60G	28.34M	81.9
		✓	✓	4.59G	28.32M	82.0
			✓	4.51G	28.29M	81.4
Swin-T [26]				4.51G	28.29M	81.3



While SWIN on top fails to distinguish between the foreground and background



DAT shifts the reference point to the giraffe, focusing on another giraffe too.

Thus, DAT improves recognition and reduces computational load by moving the reference points closer to the recognition target.

In conclusion, this proposed solution presents a promising way to overcome the existing challenges in visual recognition tasks. By using Deformable Attention Transformer, which aim to bring more flexibility, efficiency, and practicality to the image recognition domain.

References

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In NeurIPS (pp. 5998–6008).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

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

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y. (2017). Deformable convolutional networks. In ICCV (pp. 764–773).

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., & Shao, L. (2021). Pyramid vision transformer: A versatile. In ICCV.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2020). Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159.

Transformer | Artificial Intelligence | Machine Learning | Computer Vision | Deformable Convolution

Follow

Written by Joe El khoury

853 Followers

Doing my MSc in AI, BE Mechanical, MSc in Industrial & Petroleum Eng. Innovating industries through AI. Ever a student, passionate about AI's future.

More from Joe El khoury

Joe El khoury

Why Mamba was rejected?

Recently, the International Conference on Learning Representations (ICLR) announced its final decisions for the 2024 conference, drawing...

3 min read · Feb 28, 2024

1.1K 5

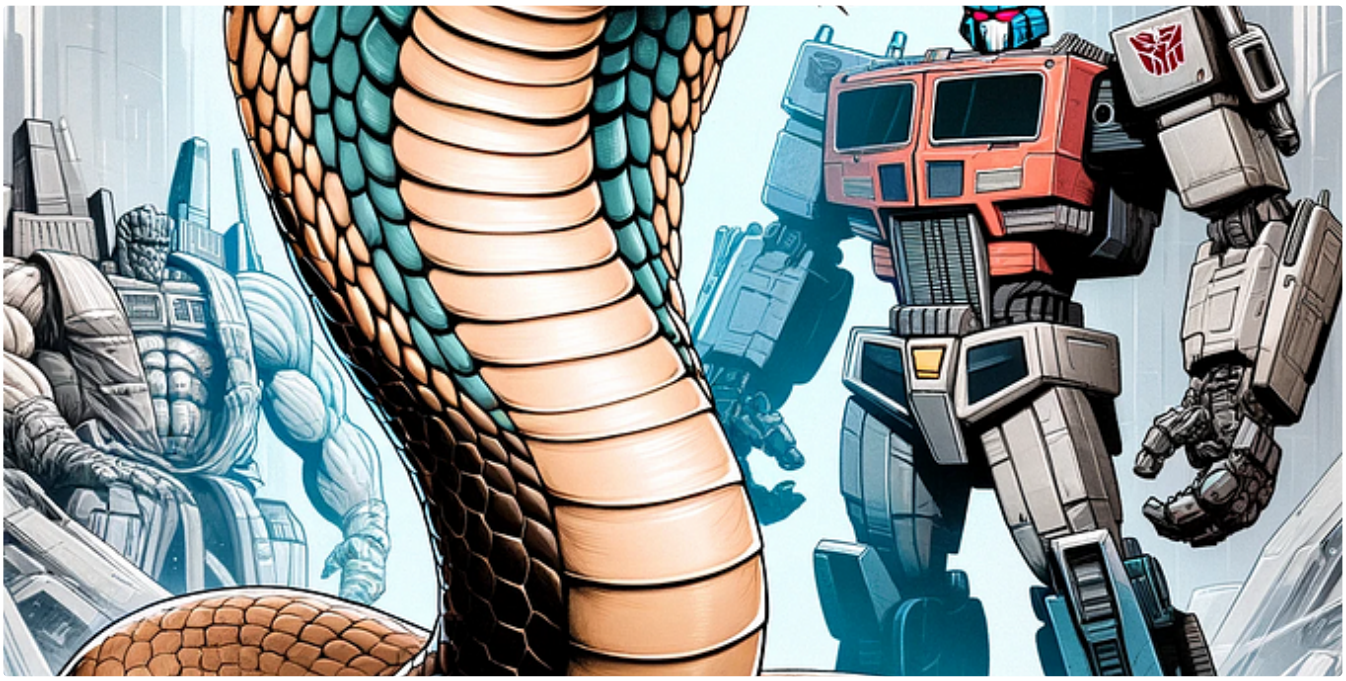
Joe El khoury

What is Mamba?

The following article will just introduce lightly and simply the new Mamba model without diving technically.

6 min read · Dec 12, 2023

878 2



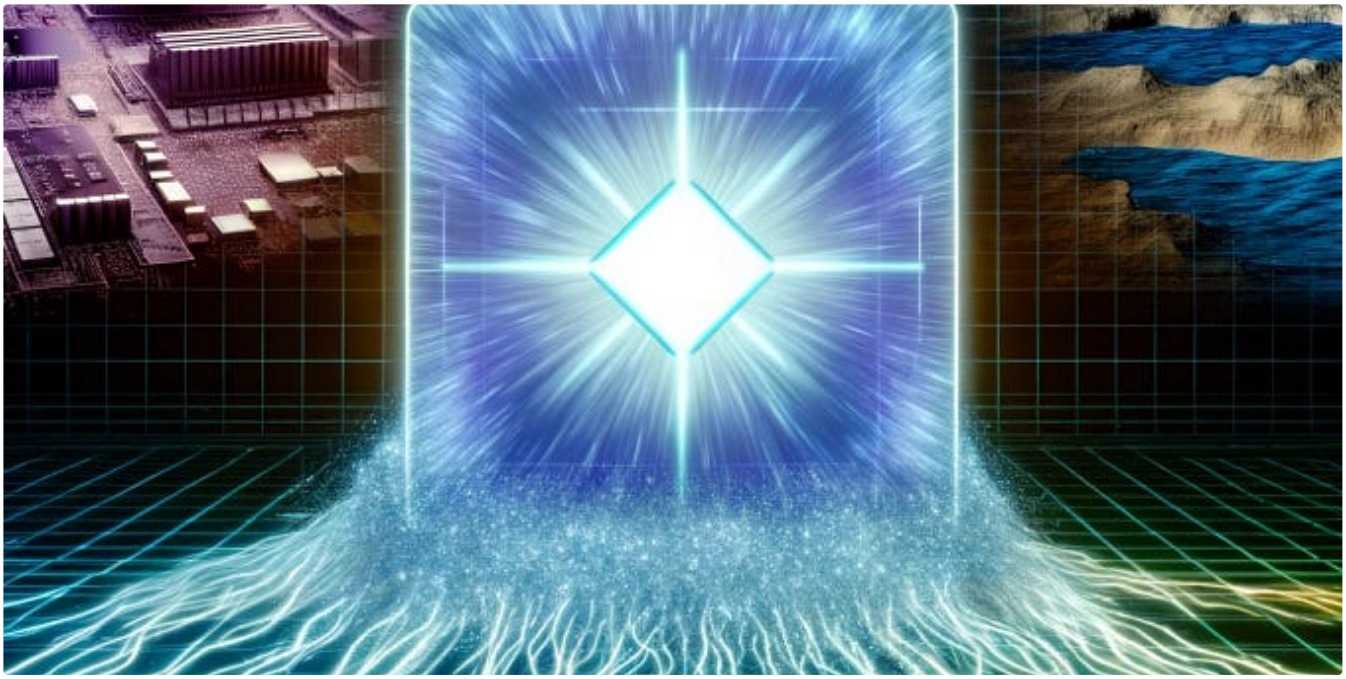
Joe El khoury

Vision Mamba

Mamba is a new design for large language models (LLMs) that deals with long sequences better than older models like Transformers, which l...

5 min read · Jan 23, 2024

416 2

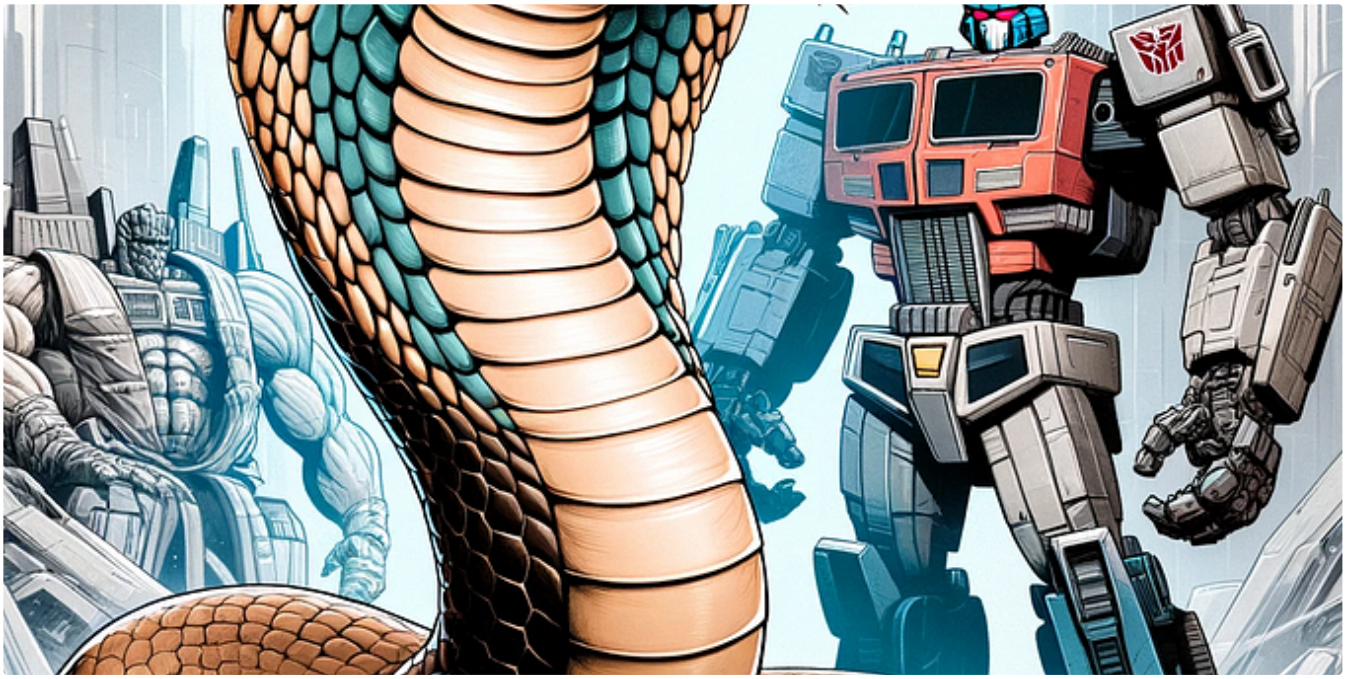


Joe El khoury

BitNet 1.58 Bits

Introduction

9 min read · Mar 1, 2024

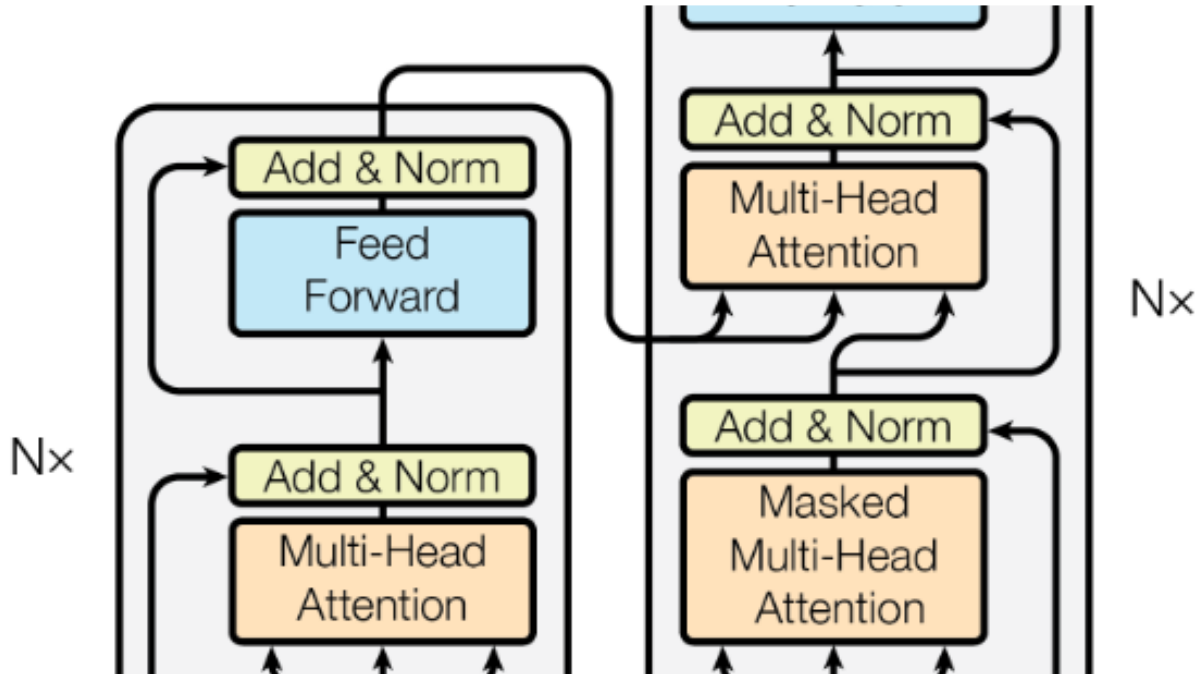


Joe El khoury

Vision Mamba

Mamba is a new design for large language models (LLMs) that deals with long sequences better than older models like Transformers, which l...

5 min read · Jan 23, 2024



Stefan

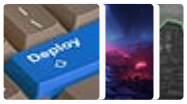
Understanding Attention and Transformers

My notes for understanding the attention mechanism and transformer architecture used by GPT-4 and other LLMs.

7 min read · Nov 29, 2023

133 1

Lists



Predictive Modeling w/ Python

20 stories · 1195 saves

Natural Language Processing

1455 stories · 963 saves

AI Regulation

6 stories · 454 saves

Practical Guides to Machine Learning

10 stories · 1445 saves

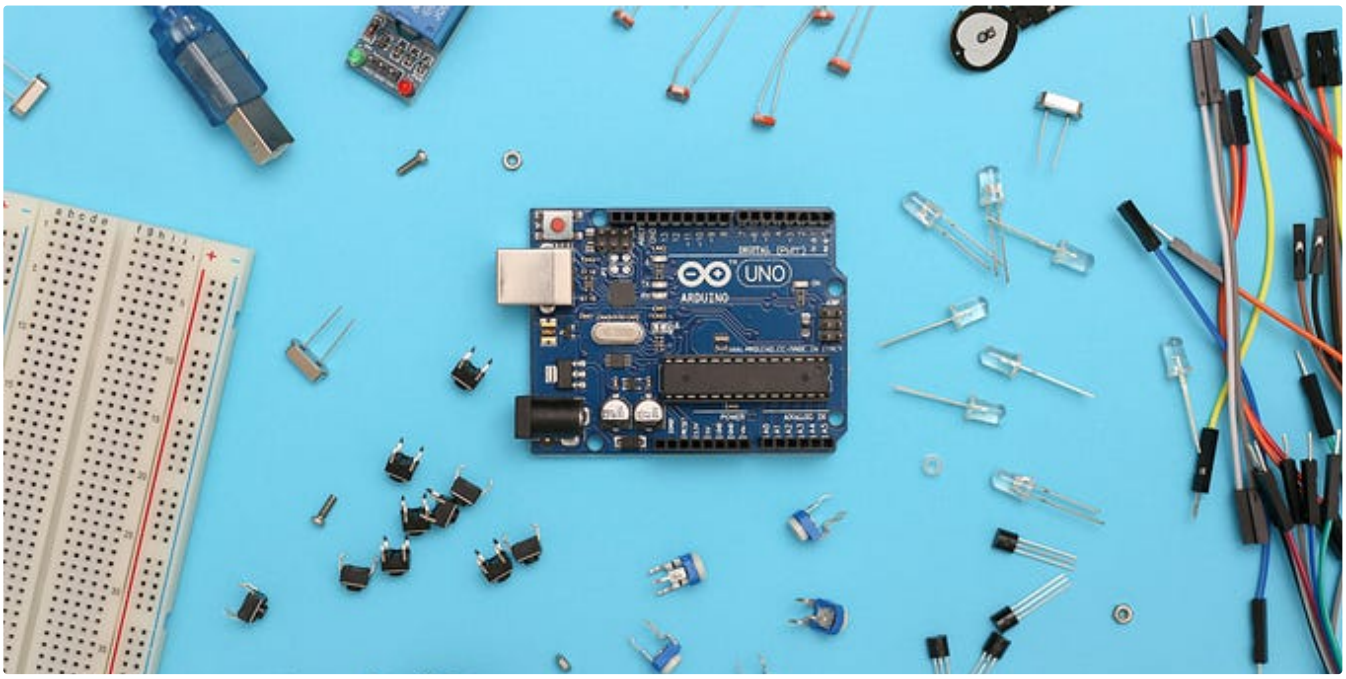
joel varun

Mamba: Revolutionizing Sequence Modeling with Selective State Spaces

Introduction

7 min read · Jan 22, 2024

50



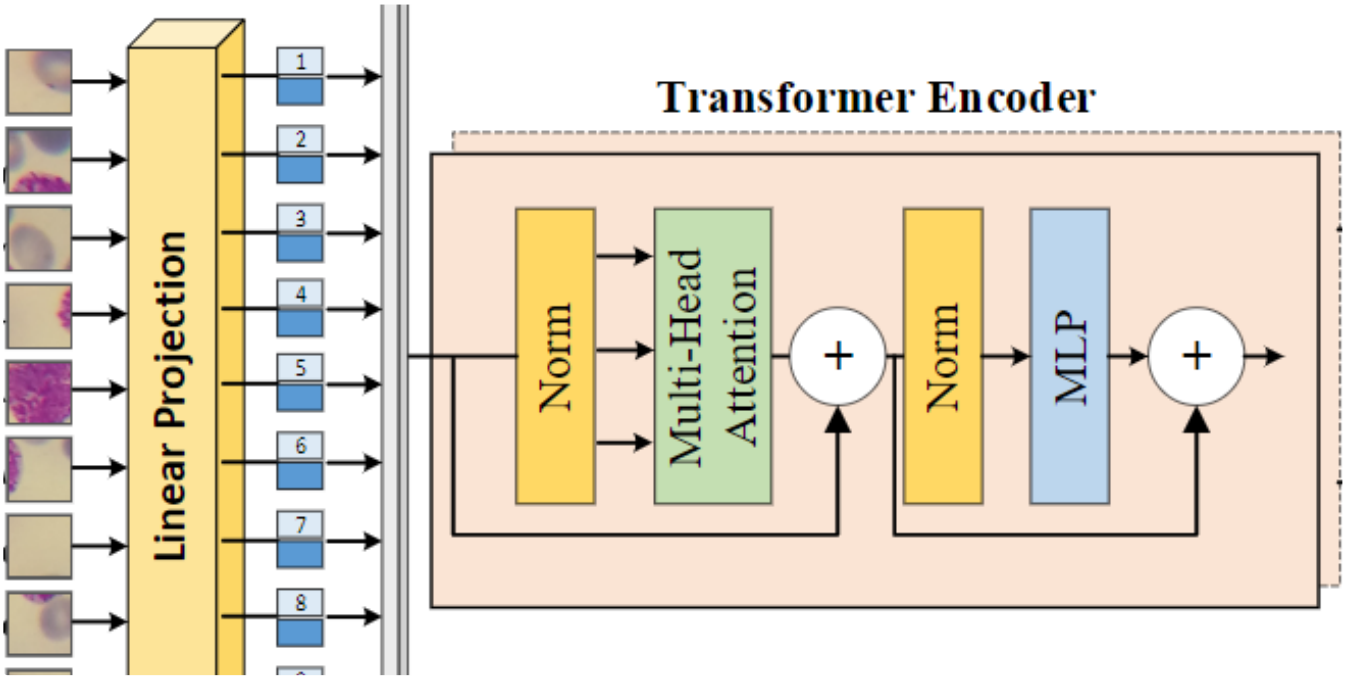
Skylar Jean Callis in Towards Data Science

Vision Transformers, Explained

A Full Walk-Through of Vision Transformers in PyTorch

18 min read · Feb 27, 2024

1K 9



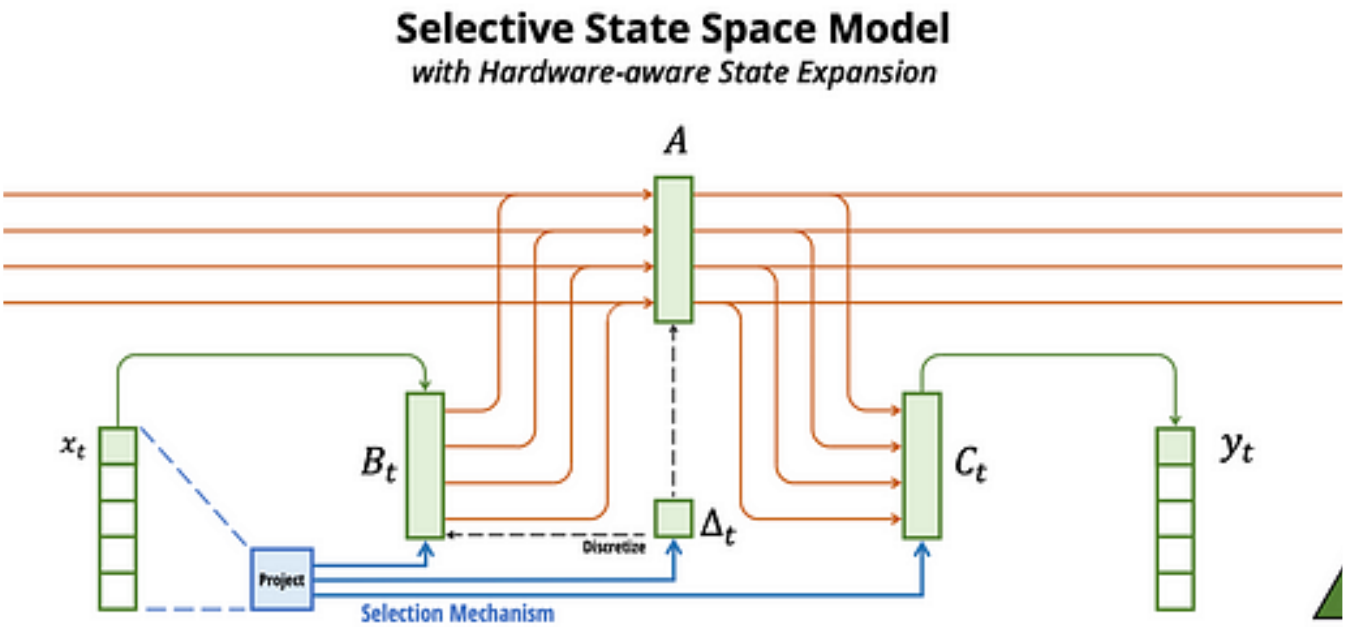
Elven Kim

Vision Transformer : Braand Future

Vision transformer is a recent breakthrough in the area of computer vision. While transformer-based models have dominated the field of...

2 min read · Dec 18, 2023

2



Vishal Rajput in AIGuys

Mamba: Can it replace Transformers?

Solving the quadratic scaling problem of Self-Attention.

· 12 min read · Jan 8, 2024

1K 5

See more recommendations