Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

Christian Lin · Follow

6 min read · Oct 13, 2023

Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

Ze Liu[†*] Yutong Lin[†*] Yue Cao[*] Han Hu[*‡] Yixuan Wei[†]
Zheng Zhang Stephen Lin Baining Guo
Microsoft Research Asia
{v-zeliu1,v-yutlin,yuecao,hanhu,v-yixwe,zhez,stevelin,bainguo}@microsoft.com

Abstract

This paper presents a new vision Transformer, called Swin Transformer, that capably serves as a general-purpose backbone for computer vision. Challenges in adapting Transformer from language to vision arise from differences between the two domains, such as large variations in the scale of visual entities and the high resolution of pixels in images compared to words in text. To address these differences, we propose a hierarchical Transformer whose representation is computed with Shifted windows. The shifted windowing scheme brings greater efficiency by limiting self-attention computation to non-overlapping local windows while also allowing for cross-window connection. This hierarchical architecture has the flexibility to model at various scales and has linear computational complexity with respect to image size. These qualities of Swin Transformer make it compatible with a broad range of vision tasks, including image classification (87.3 top-1 accuracy on ImageNet-1K) and dense prediction tasks such as object detection (58.7 box AP and 51.1 mask AP on COCO test-dev) and semantic segmentation (53.5 mIoU on ADE20K val). Its performance surpasses the previous state-of-the-art by a large margin of +2.7 box AP and +2.6 mask AP on COCO, and +3.2 mIoU on ADE20K, demonstrating the potential of Transformer-based models as vision backbones. The hierarchical design and the shifted window approach also prove beneficial for all-MLP architectures. The code and models are publicly available at https://github.com/microsoft/Swin-Transformer.

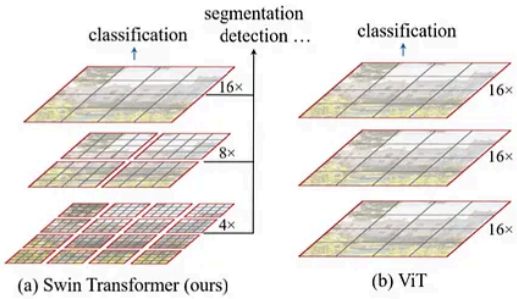


Figure 1. (a) The proposed Swin Transformer builds hierarchical feature maps by merging image patches (shown in gray) in deeper layers and has linear computation complexity to input image size due to computation of self-attention only within each local window (shown in red). It can thus serve as a general-purpose backbone for both image classification and dense recognition tasks. (b) In contrast, previous vision Transformers [20] produce feature maps of a single low resolution and have quadratic computation complexity to input image size due to computation of self-attention globally.

greater scale [30, 76], more extensive connections [34], and more sophisticated forms of convolution [70, 18, 84]. With CNNs serving as backbone networks for a variety of vision tasks, these architectural advances have led to performance improvements that have broadly lifted the entire field.

On the other hand, the evolution of network architectures in natural language processing (NLP) has taken a different path, where the prevalent architecture today is instead the

Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

This study introduces a new tool for computer vision called the Swin Transformer. This tool is built on the Transformer model, which was originally made for language tasks. Adapting it for vision is tricky because images and words are very different, especially in terms of size and detail. Our solution is a multi-level Transformer that uses a special windowing technique. This technique only focuses on certain parts of the image at a time but still connects them all together. This makes our tool both efficient and able to handle different image sizes. The Swin Transformer works well for many vision tasks. For instance, it's really good at classifying images and detecting objects in them. In tests, it even did better than previous top-performing tools. Plus, our special design and windowing approach can also improve other similar tools.

Preliminary

Transformer in NLP:

Transformers first made their appearance in the groundbreaking paper "Attention Is All You Need". Before this, sequence-to-sequence models in NLP relied heavily on recurrent neural networks (RNNs) and their variants. The Transformer architecture discarded recurrence in favor of self-attention mechanisms, providing parallel processing benefits and capturing long-range dependencies in data.

Key Features: The self-attention mechanism, the heart of the Transformer, allows the model to focus differently on various parts of the input data. For instance, in a sentence, some words might be more relevant to understanding the meaning of a particular word. Self-attention weighs this relevance, allowing the model to make contextually informed decisions.

Challenges in Adapting Transformers to Vision:

- **Scale Variations:** Unlike text where words typically have uniform significance, visual data is replete with entities of varying scales. For example, an image might contain a vast landscape and tiny birds in the distance. A model should be adept at recognizing both the landscape and the birds with equal finesse.

- **High Resolution:** Images are grids of pixels. An average image can have thousands to millions of pixels. If each pixel or a small patch is treated as a sequence token (akin to words in NLP), it becomes computationally expensive for the Transformer to process due to its quadratic complexity with sequence length.

Initial Attempts:

Vision Transformers (ViTs) were among the first to adapt the Transformer architecture for vision. They divided images into fixed-size patches, treated each patch as a token by linearly embedding them, and then processed them as a sequence. While ViTs showcased the potential of Transformers in vision, they typically required vast amounts of data and heavy computation to excel, often relying on large-scale pre-training.


shed as a conference paper at ICLR 2021

IMAGE IS WORTH 16X16 WORDS:
ANSFORMERS FOR IMAGE RECOGNITION AT SCA

Alexey Dosovitskiy*,†, Lucas Beyer*, Alexander Kolesnikov*, Dirk Weissenborn*, Xiaohua Zhai*, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby*,†

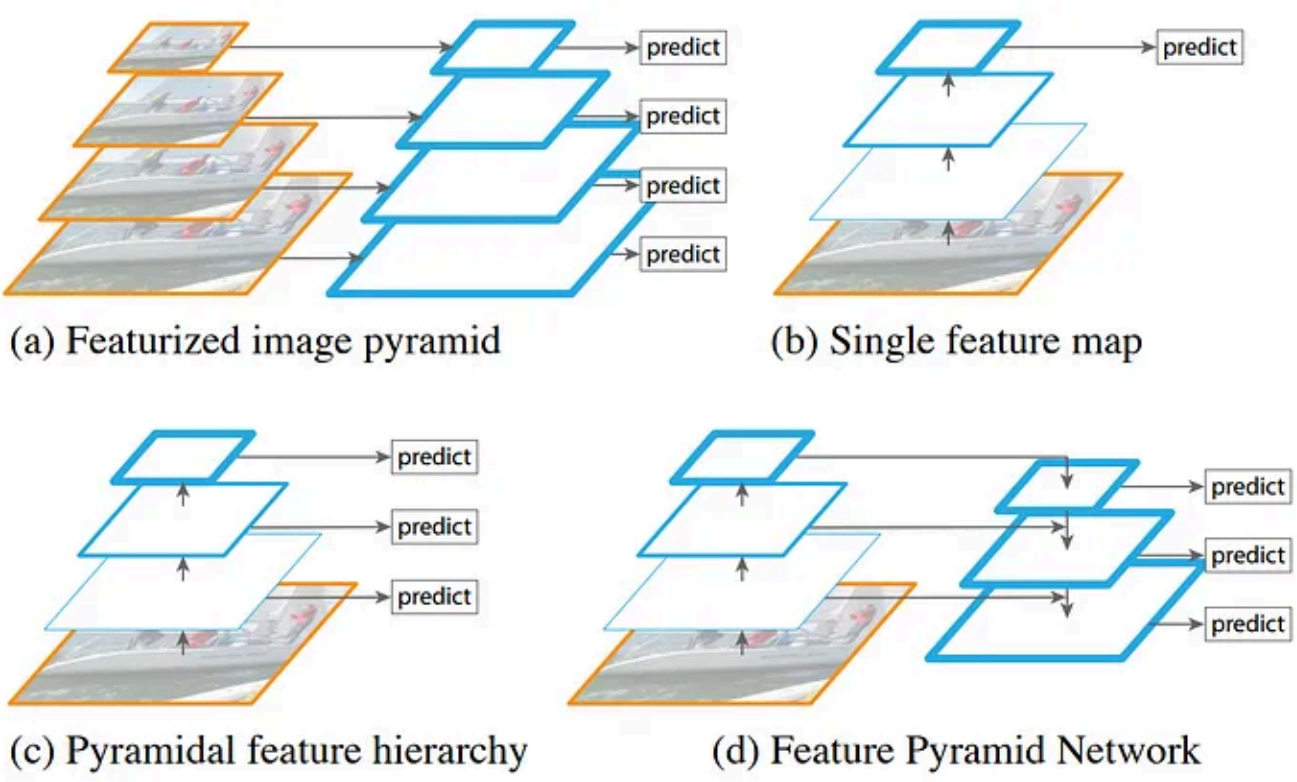
*equal technical contribution, †equal advising
Google Research, Brain Team
{adosovitskiy, neilhoulsby}@google.com

ABSTRACT

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.[1]

The Need for Hierarchical Structures:

Given the multi-scale nature of visual information, a model should ideally process data at multiple resolutions or scales. Hierarchical structures in neural networks facilitate this by having layers dedicated to different scales. This approach is reminiscent of the way convolutional neural networks (CNNs) operate, where initial layers capture fine details and deeper layers capture more abstract, larger-scale features.



Four common hierarchical structures applied in computer vision

Window-Based Mechanisms:

CNNs, the de facto standard for vision tasks before Transformers, use a mechanism where each neuron in a layer looks at a small local region (or receptive field) of the input. This allows for efficient processing of high-resolution images. Adapting a similar local processing mechanism in Transformers can ensure they're more adept at handling visual data. Window-based mechanisms restrict the self-attention to local regions, making computations more manageable and efficient.

Shifted Windowing:

While local processing is efficient, it's also crucial that different parts of an image communicate or share information. The shifted windowing technique addresses this. Imagine dividing an image into non-overlapping windows and processing them. In the next layer, these windows are shifted (like a sliding puzzle) so that they overlap with neighboring windows from the previous layer. This ensures that distant regions of an image can eventually "talk" to each other in deeper layers, providing a global context.

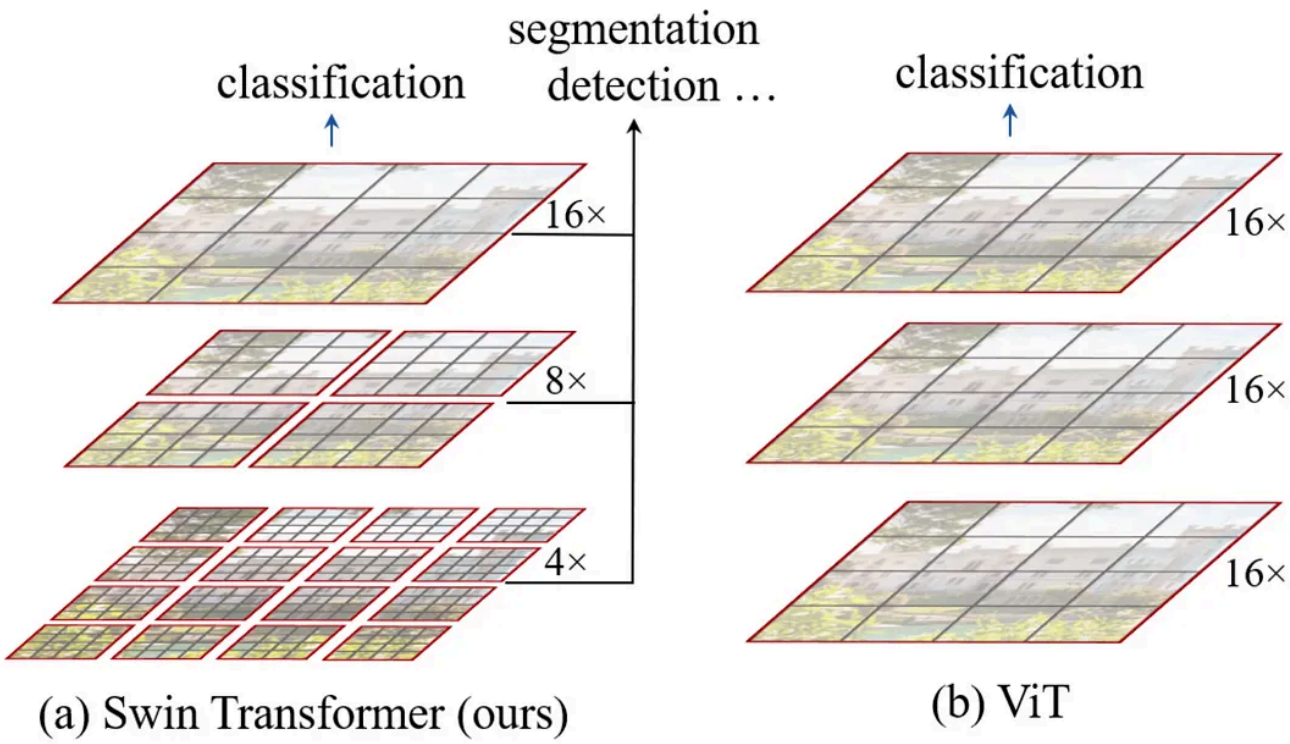
Methodology

Hierarchical Representation with Non-overlapping Patches:

The Swin Transformer initiates its processing by partitioning an input image into fixed-size, non-overlapping patches, analogous to the way Vision Transformers (ViTs) function. Each of these patches captures a localized segment of the image,

ensuring that spatial relationships and visual cues within this segment are maintained. These patches are then linearly embedded, converting two-dimensional visual information into a sequence of tokens suitable for sequential processing.

The foundational step of translating visual data into token sequences ensures that spatial structures are not lost. Instead, they are reformatted into a representation that allows them to be processed similarly to textual sequences, leveraging the power of Transformer architectures.



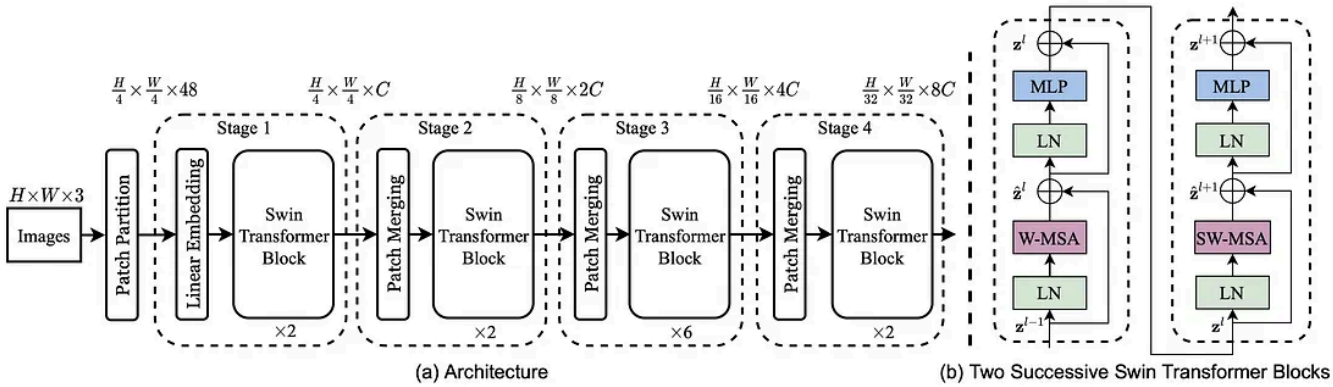
(a) The Swin Transformer, as proposed, constructs multi-level feature maps by integrating image patches (illustrated in gray) in its more advanced layers. Its linear computational complexity relative to the input image size is a result of limiting self-attention computations to individual local windows (depicted in red). Consequently, it has the versatility to act as a foundational backbone for both image categorization and detailed recognition tasks. (b) On the other hand, earlier ViT generates feature maps at a singular, reduced resolution and exhibit quadratic computational complexity relative to image size because they compute self-attention across the entire image.

Shifted Window-based Self-attention:

Within each Transformer layer, self-attention operations are confined to specific non-overlapping windows. Rather than the traditional all-to-all attention mechanism, where every token examines every other token, in this local processing setup, a token's attention span is restricted to its immediate window. This results in a dramatic reduction in computational overhead, as the number of attention computations is curtailed.

To ensure that the model doesn't develop myopia, focusing only on local contexts, the windows undergo a shifting procedure across layers. This strategic shifting enables tokens in subsequent layers to be influenced by and interact with tokens from neighboring windows. Over depth, this creates a network where tokens can accumulate information from increasingly distant parts of the original image.

Through this dual approach of local windowed attention and strategic shifting, the model efficiently processes information while preserving the capacity to integrate broader context.



(a) The architecture of a Swin Transformer (Swin-T); (b) two successive Swin Transformer Blocks. W-MSA and SW-MSA are multi-head self attention modules with regular and shifted windowing configurations, respectively.

Efficient batch computation for shifted configuration:

In the Swin Transformer, self-attention is not computed globally (as with traditional transformers) but within localized windows. This ensures that tokens within a window only attend to other tokens in the same window, reducing computational complexity.

However, to ensure that tokens from different windows can interact and "see" each other at different layers of the model, the Swin Transformer uses a "shifted" windowing scheme. After one layer's self-attention computations are completed using these localized windows, the windows are "shifted" for the next layer. This way, tokens in a window in one layer overlap with tokens from neighboring windows in the subsequent layer.

The challenge here is how to efficiently compute these shifted window-based self-attentions, especially when processing multiple data items in a batch (common in deep learning training for better GPU utilization and convergence).

Swin Transformer uses a special algorithmic arrangement to perform these shifted window self-attention computations across the entire batch of data in a unified and efficient manner. By organizing the data properly and leveraging parallel processing capabilities of modern GPUs, the Swin Transformer can rapidly compute self-attentions for these shifted windows across many data items simultaneously, without excessive memory usage or computational overhead.

This efficient computation ensures that the Swin Transformer maintains its advantages in modeling both local and global contexts without sacrificing speed or efficiency, even when trained with large batches of data.

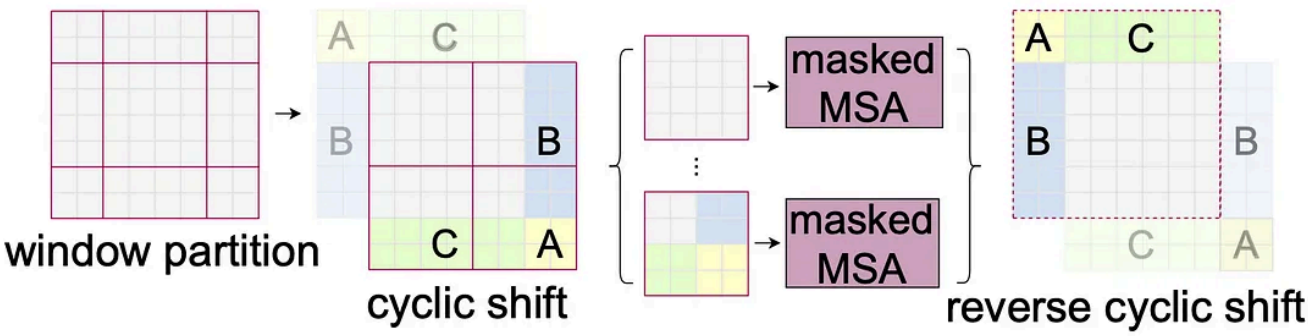


Illustration of an efficient batch computation approach for self-attention in shifted window partitioning.

In this article, we provide the big picture of Swin Transformer model. If you want to know more detail about this, you can follow the links shown as follow to further understand the related foundation and knowledge.

Swin Transformer: Hierarchical Vision Transformer using Shifted Window...

. . .

In this article, I briefly share my viewpoints on the paper. I hope you can learn more about it after reading it. I also offer the video link about the paper, hope you guys like it!!!!

If you like the article, please give me some 👏 , share the article, and follow me to learn more about the world of multi-agent reinforcement learning. You can also contact me on **LinkedIn**, **Instagram**, **Facebook** and **Github**.

Computer Vision | Deep Learning | Transformers | Machine Learning | Artificial Intelligence

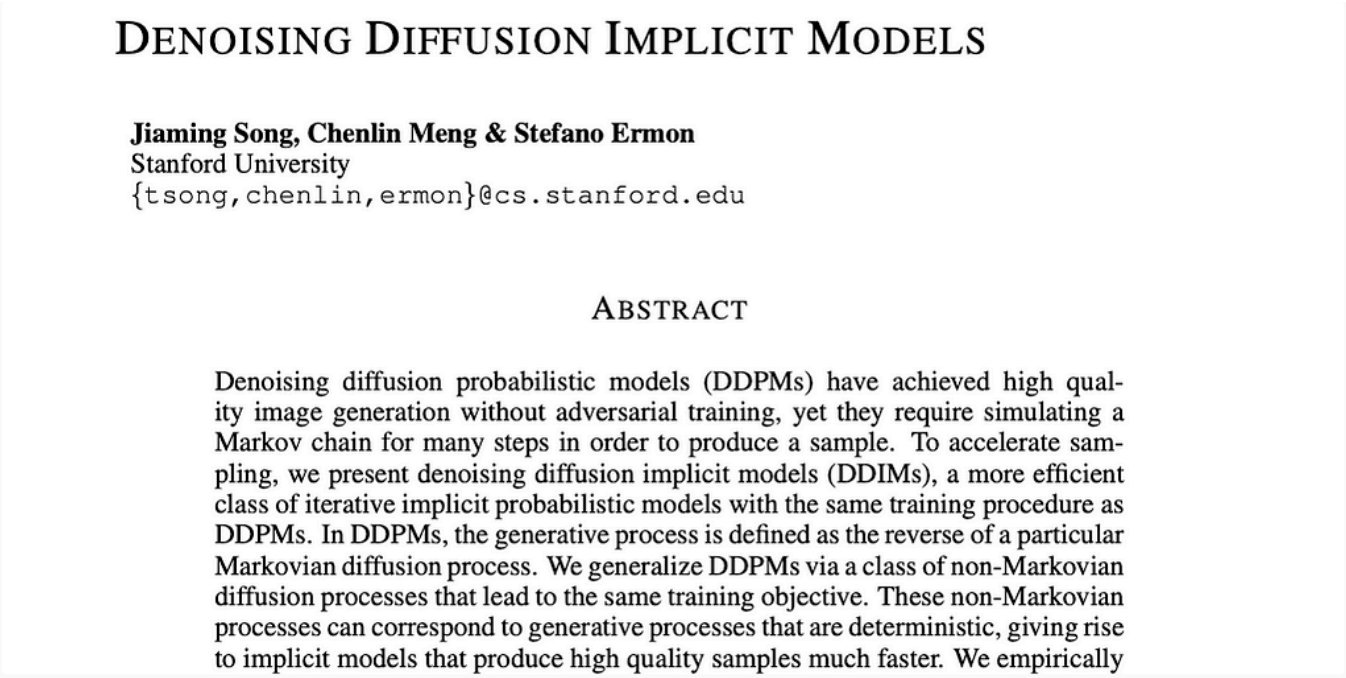
Follow

Written by Christian Lin

73 Followers

A master CS student used to work at ShangShing as an iOS full-end developer. Now, I dive into AI field, especially Multi-agent RL and Bio-inspired intelligence.

More from Christian Lin



DENOISING DIFFUSION IMPLICIT MODELS

Jiaming Song, Chenlin Meng & Stefano Ermon
Stanford University
{tsong,chenlin,ermon}@cs.stanford.edu

ABSTRACT

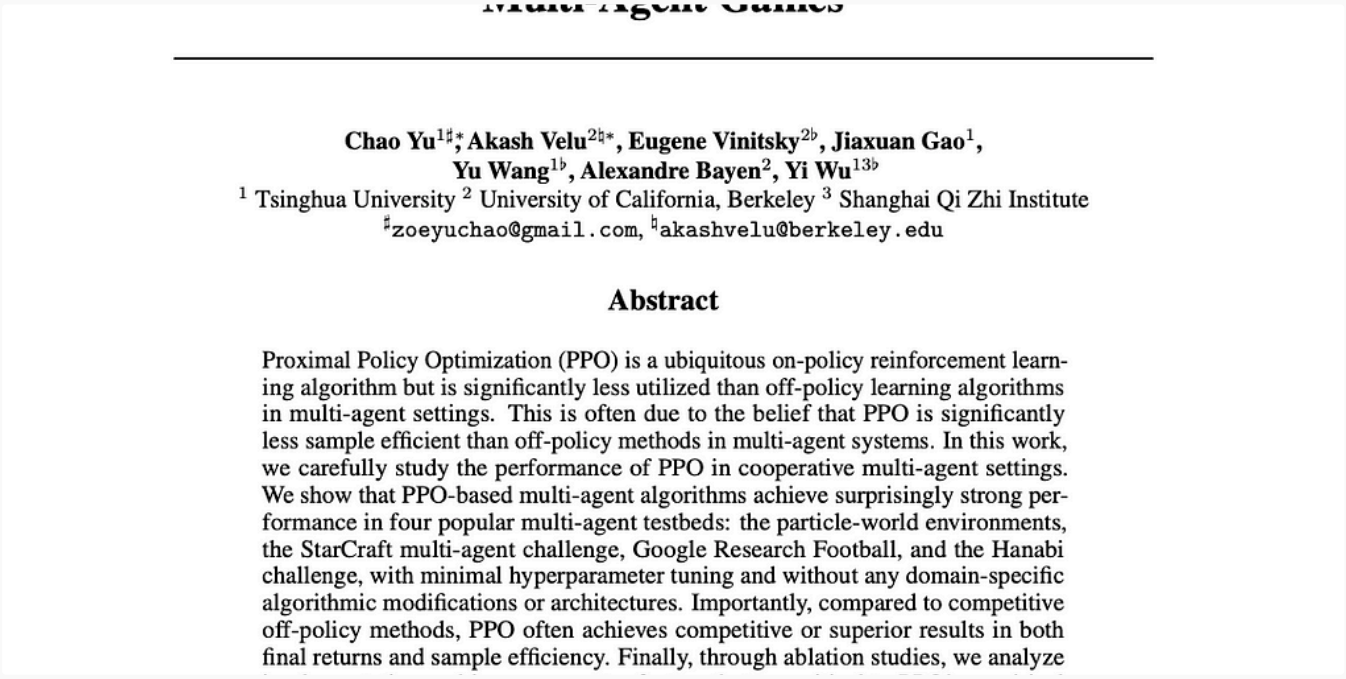
Denoising diffusion probabilistic models (DDPMs) have achieved high quality image generation without adversarial training, yet they require simulating a Markov chain for many steps in order to produce a sample. To accelerate sampling, we present denoising diffusion implicit models (DDIMs), a more efficient class of iterative implicit probabilistic models with the same training procedure as DDPMs. In DDPMs, the generative process is defined as the reverse of a particular Markovian diffusion process. We generalize DDPMs via a class of non-Markovian diffusion processes that lead to the same training objective. These non-Markovian processes can correspond to generative processes that are deterministic, giving rise to implicit models that produce high quality samples much faster. We empirically

Christian Lin

Computer Vision Paper ~ Denoising Diffusion Implicit Models

Denoising diffusion probabilistic models (DDPMs) can generate high-quality images without adversarial training, but it takes many steps to...

Apr 8, 2023 5



Chao Yu[1♯]*, Akash Velu[2♮]*, Eugene Vinitsky[2♭], Jiaxuan Gao[1],
Yu Wang[1♭], Alexandre Bayen[2], Yi Wu[13♭]
[1] Tsinghua University [2] University of California, Berkeley [3] Shanghai Qi Zhi Institute
♯zoeyuchao@gmail.com, ♮akashvelu@berkeley.edu

Abstract

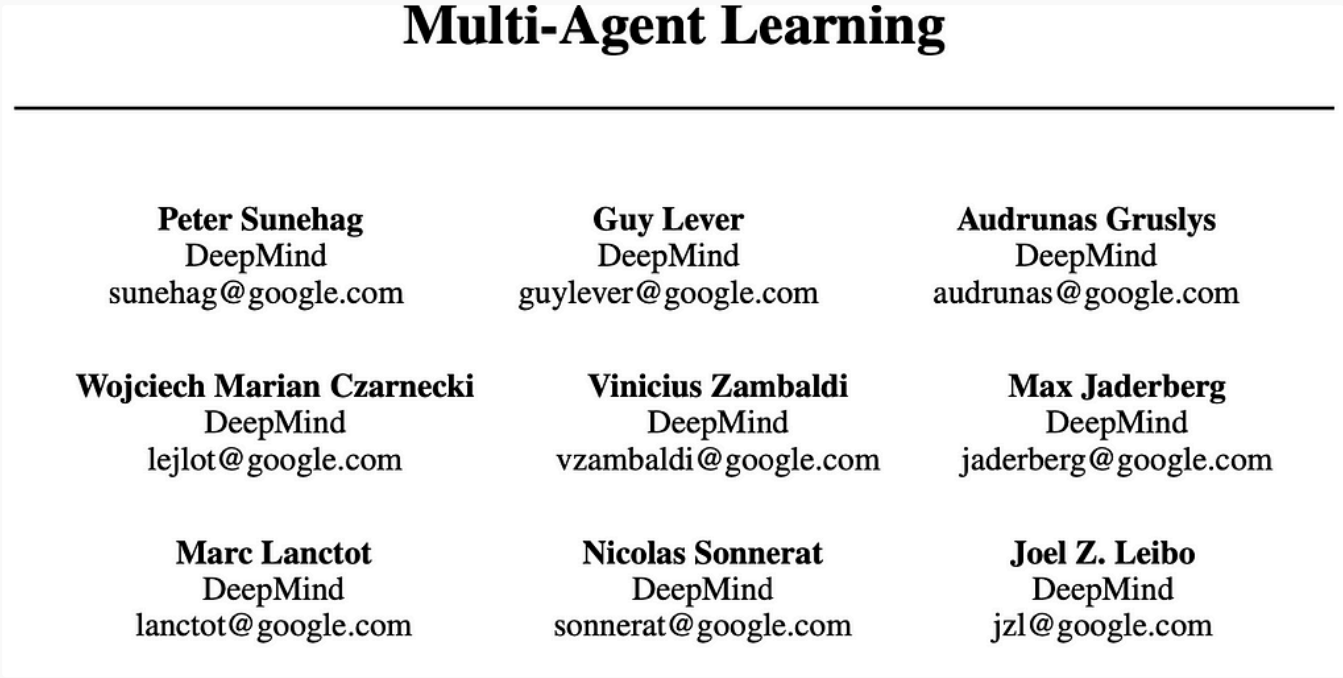
Proximal Policy Optimization (PPO) is a ubiquitous on-policy reinforcement learning algorithm but is significantly less utilized than off-policy learning algorithms in multi-agent settings. This is often due to the belief that PPO is significantly less sample efficient than off-policy methods in multi-agent systems. In this work, we carefully study the performance of PPO in cooperative multi-agent settings. We show that PPO-based multi-agent algorithms achieve surprisingly strong performance in four popular multi-agent testbeds: the particle-world environments, the StarCraft multi-agent challenge, Google Research Football, and the Hanabi challenge, with minimal hyperparameter tuning and without any domain-specific algorithmic modifications or architectures. Importantly, compared to competitive off-policy methods, PPO often achieves competitive or superior results in both final returns and sample efficiency. Finally, through ablation studies, we analyze

Christian Lin

Multi-agent Reinforcement Learning Paper Reading~ The Surprising Effectiveness of PPO in...

Proximal Policy Optimization (PPO, following link provides the original PPO paper for people!!!) is a ubiquitous on-policy reinforcement...

Feb 19, 2023 13



Multi-Agent Learning

Peter Sunehag
DeepMind
sunehag@google.com

Guy Lever
DeepMind
guylever@google.com

Audrunas Gruslys
DeepMind
audrunas@google.com

Wojciech Marian Czarnecki
DeepMind
lejlot@google.com

Vinicius Zambaldi
DeepMind
vzambaldi@google.com

Max Jaderberg
DeepMind
jaderberg@google.com

Marc Lanctot
DeepMind
lanctot@google.com

Nicolas Sonnerat
DeepMind
sonnerat@google.com

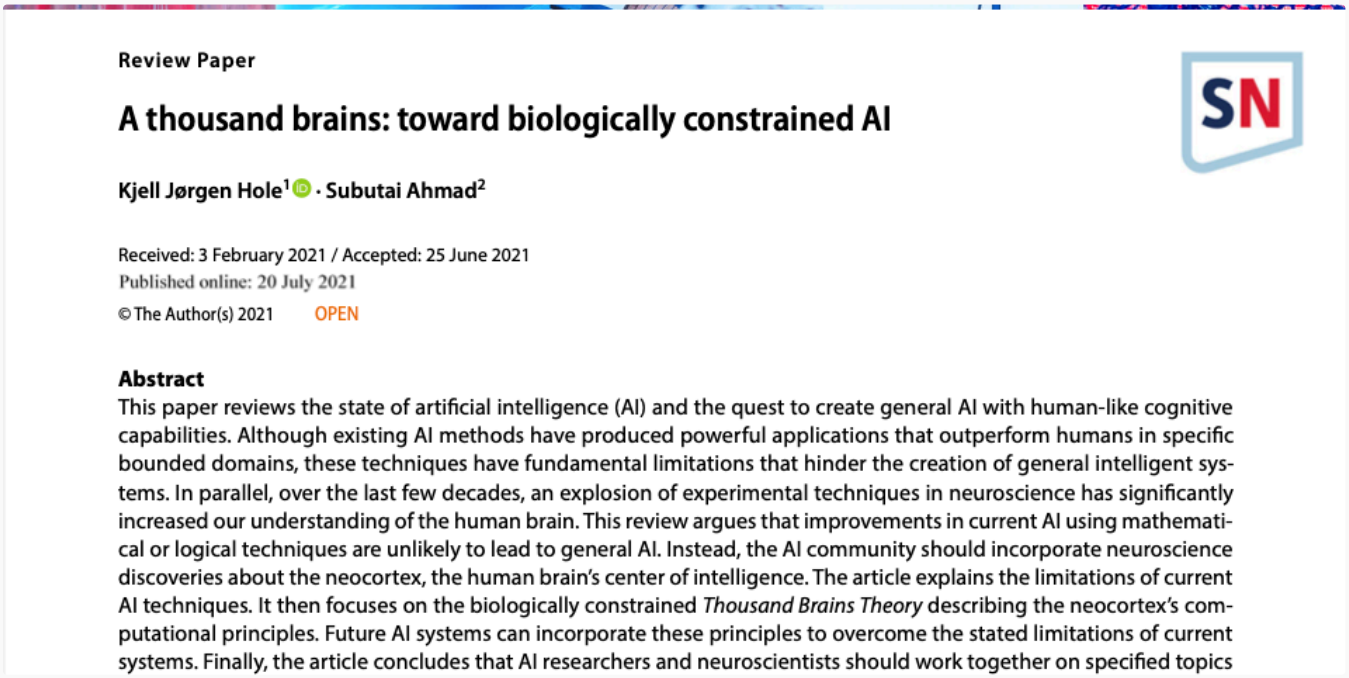
Joel Z. Leibo
DeepMind
jzl@google.com

Christian Lin

Multi-agent Reinforcement Learning Paper Reading ~ VDN

Compared with the single-agent environment task, agents in the MARL task usually face a more inconsistent environment due to the attendance...

Oct 9, 2022 13



Review Paper

A thousand brains: toward biologically constrained AI

Kjell Jørgen Hole[1] · Subutai Ahmad[2]

Received: 3 February 2021 / Accepted: 25 June 2021
Published online: 20 July 2021
© The Author(s) 2021 OPEN

Abstract
This paper reviews the state of artificial intelligence (AI) and the quest to create general AI with human-like cognitive capabilities. Although existing AI methods have produced powerful applications that outperform humans in specific bounded domains, these techniques have fundamental limitations that hinder the creation of general intelligent systems. In parallel, over the last few decades, an explosion of experimental techniques in neuroscience has significantly increased our understanding of the human brain. This review argues that improvements in current AI using mathematical or logical techniques are unlikely to lead to general AI. Instead, the AI community should incorporate neuroscience discoveries about the neocortex, the human brain's center of intelligence. The article explains the limitations of current AI techniques. It then focuses on the biologically constrained *Thousand Brains Theory* describing the neocortex's computational principles. Future AI systems can incorporate these principles to overcome the stated limitations of current systems. Finally, the article concludes that AI researchers and neuroscientists should work together on specified topics

Christian Lin

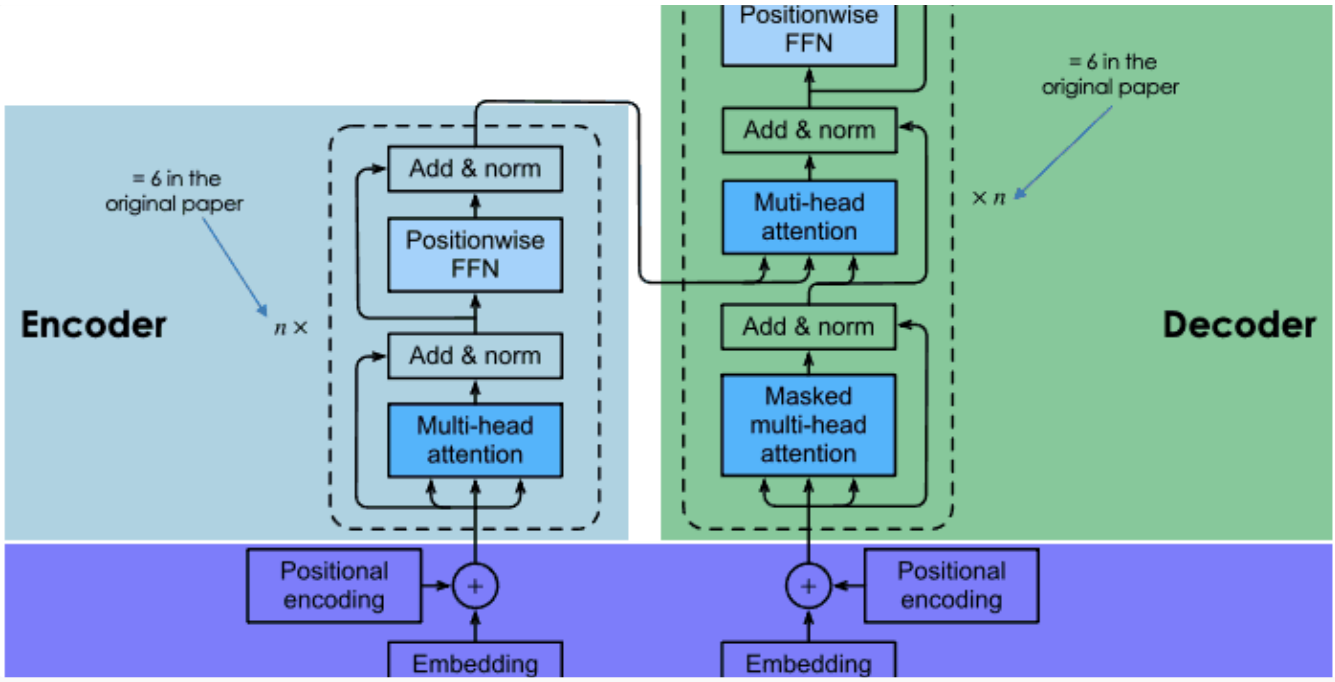
HTM Paper Reading ~A thousand brains: toward biologically constrained AI—PART I

It's exciting to start a topic about artificial intelligence. The reason why I want to explore the realm of AI is about a movie, Iron...

Oct 31, 2022 103

See all from Christian Lin

Recommended from Medium

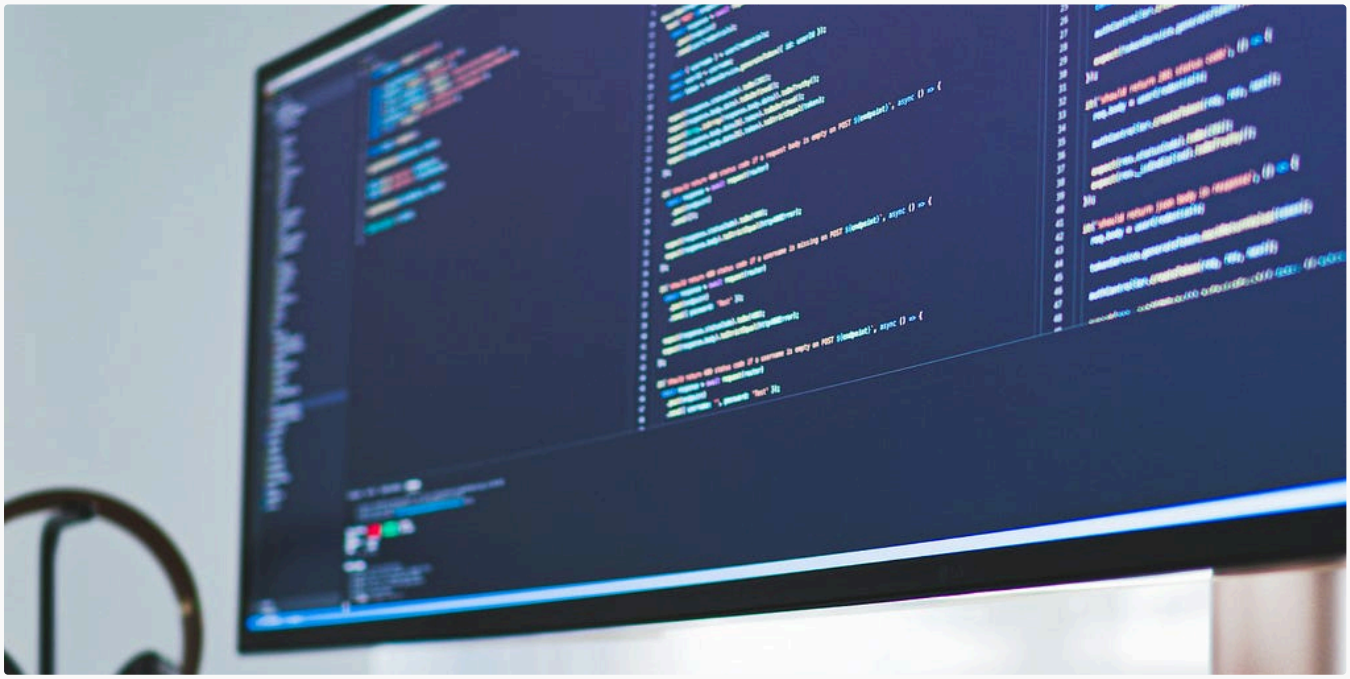


Dr. Ananth G S

Best 3 books to learn “Transformers” in Machine Learning

This post lists the best books to learn about the Transformers network along with the links to purchase the same.

Apr 16 34



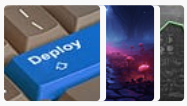
Rahul Beniwal in Level Up Coding

18 Programming Concepts You’ve Never Heard of (But Should!)

Unlock Hidden Programming Gems to Boost Your Coding Superpowers

4d ago 118 3

Lists



Predictive Modeling w/ Python

20 stories · 1549 saves

Natural Language Processing

1715 stories · 1287 saves

AI Regulation

6 stories · 571 saves

Practical Guides to Machine Learning

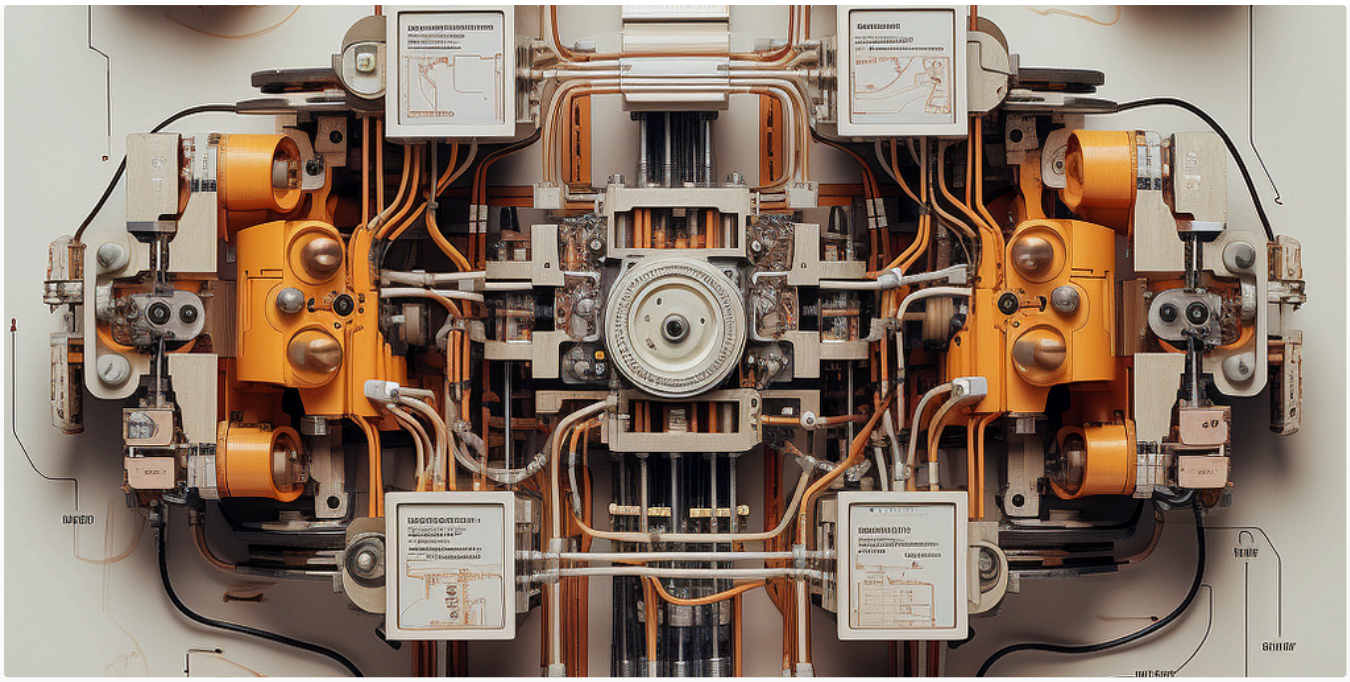
10 stories · 1878 saves

Daniel Warfield in Towards Data Science

CLIP, Intuitively and Exhaustively Explained

Creating strong image and language representations for general machine learning tasks.

Oct 21, 2023 722 7

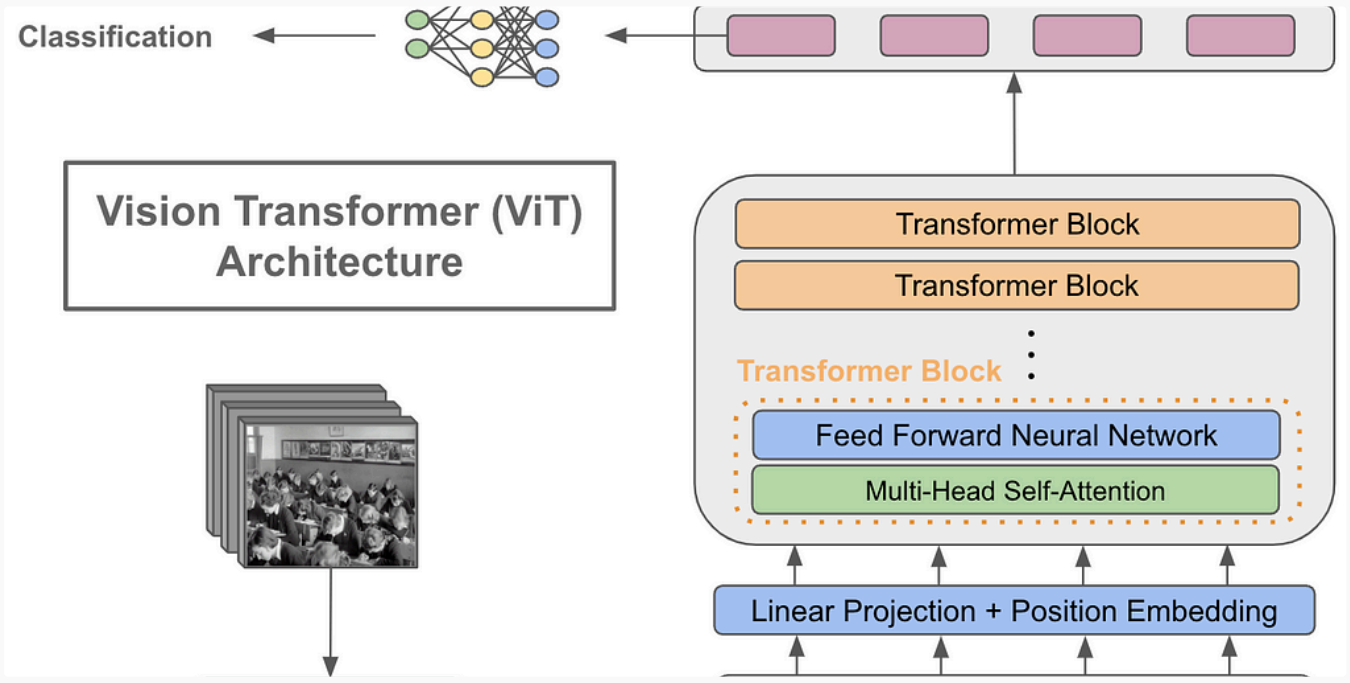


Daniel Warfield in Towards Data Science

Transformers—Intuitively and Exhaustively Explained

Exploring the modern wave of machine learning: taking apart the transformer step by step

Sep 21, 2023 2.6K 20



Cameron R. Wolfe, Ph.D. in Towards Data Science

Using Transformers for Computer Vision

Are Vision Transformers actually useful?

Oct 5, 2022 314 6