UE21CS343BB2

Topics in Deep Learning

Dr. Shylaja S S
Director of Cloud Computing & Big Data (CCBD), Centre for Data Sciences & Applied Machine Learning (CDSAML)
Department of Computer Science and Engineering
shylaja.sharath@pes.edu

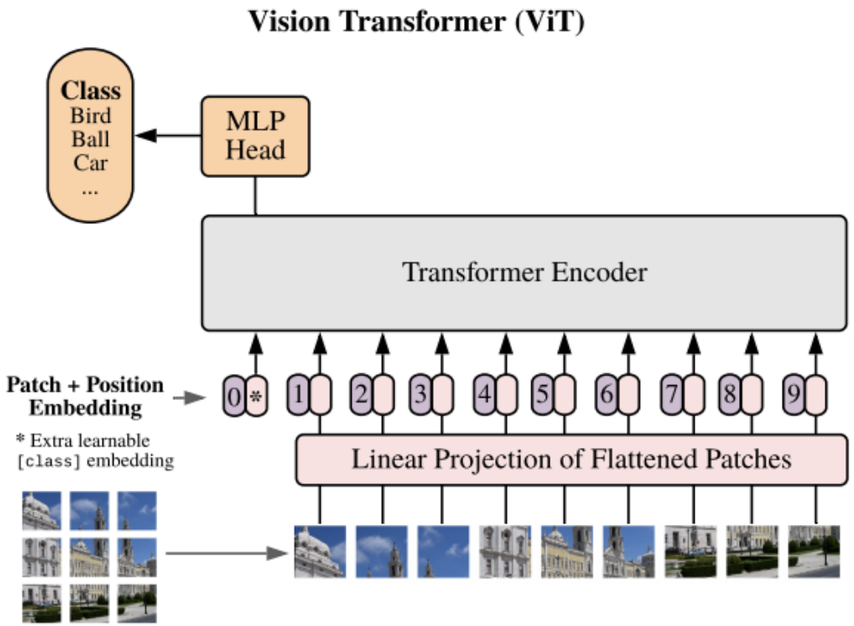
Aryan Sharma,
Teaching Assistant

UE21CS343BB2: Topics in Deep Learning

Introduction: Vision Transformers

What is a Vision Transformer?

A vision transformer, also known as ViT, is a type of artificial neural network architecture designed for image recognition tasks.



We'll explore their inner workings, delve into their advantages and limitations, and discuss their potential future impact.

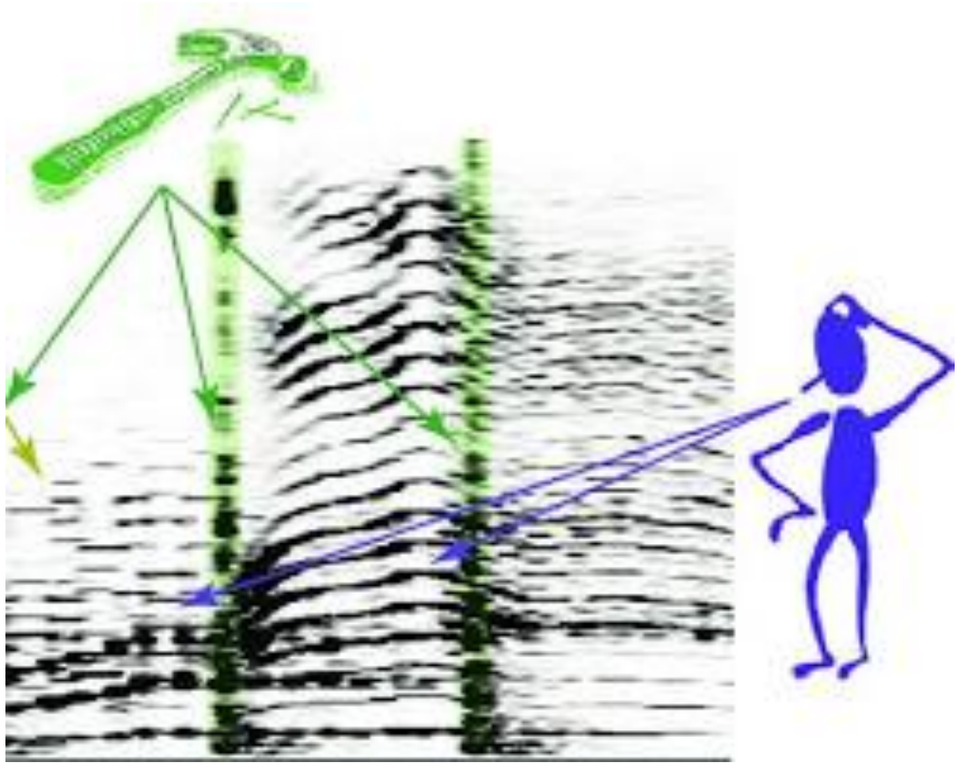
What is a Vision Transformer?

Traditional computer vision tasks aim to extract meaningful information from images.

Tasks include:

- **Object recognition:** Identifying and locating objects within an image.
- **Image classification:** Assigning an image to a specific category (e.g., cat, dog, car).
- **Image segmentation:** Grouping pixels into meaningful regions corresponding to objects.

What is a Vision Transformer?



As we can see here, this image has various objects that require understanding relationships between them for accurate recognition. From self-driving cars to medical diagnosis, the ability to understand visual content is crucial. However, accurately recognizing objects in complex images remains a challenge, especially when understanding the relationships between them becomes important.

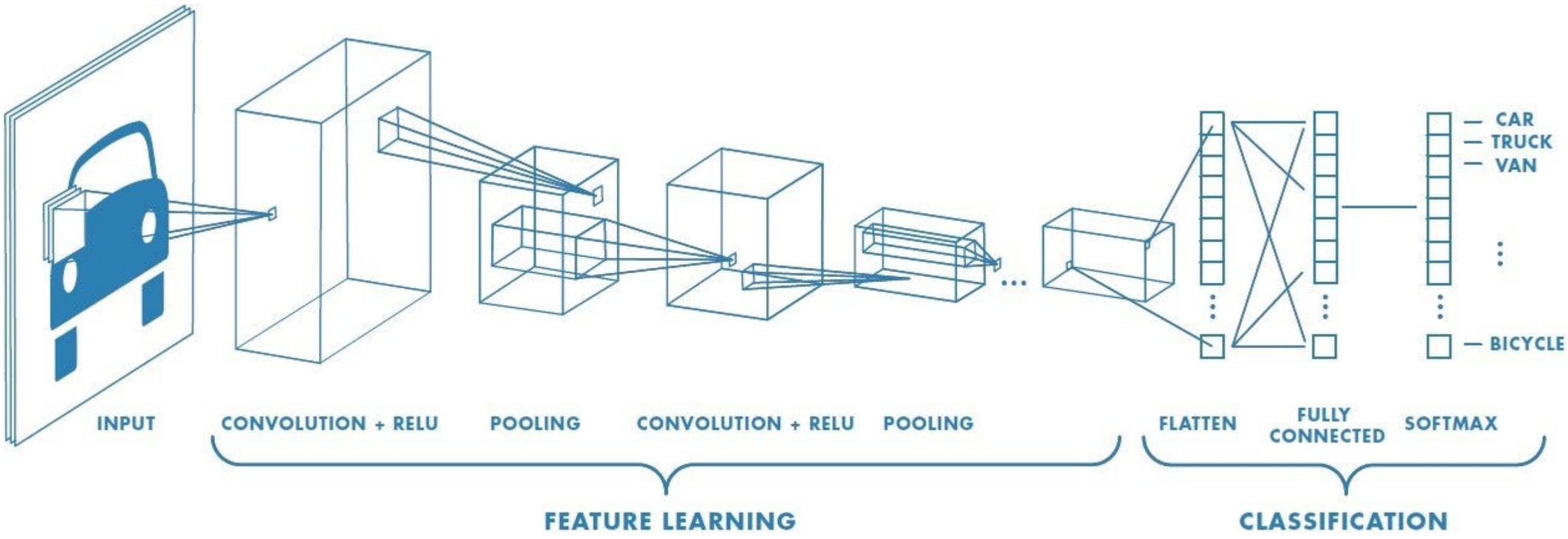
Limitations of CNNs

- CNNs have been the dominant architecture for image recognition for decades.

- <u>Strengths:</u>
  Excellent at capturing local features like edges and textures.
  Effective in learning spatial hierarchies of features.

- <u>Weaknesses:</u>
  Struggles with long-range dependencies: Understanding relationships between distant parts of an image.

Limitations of CNNs



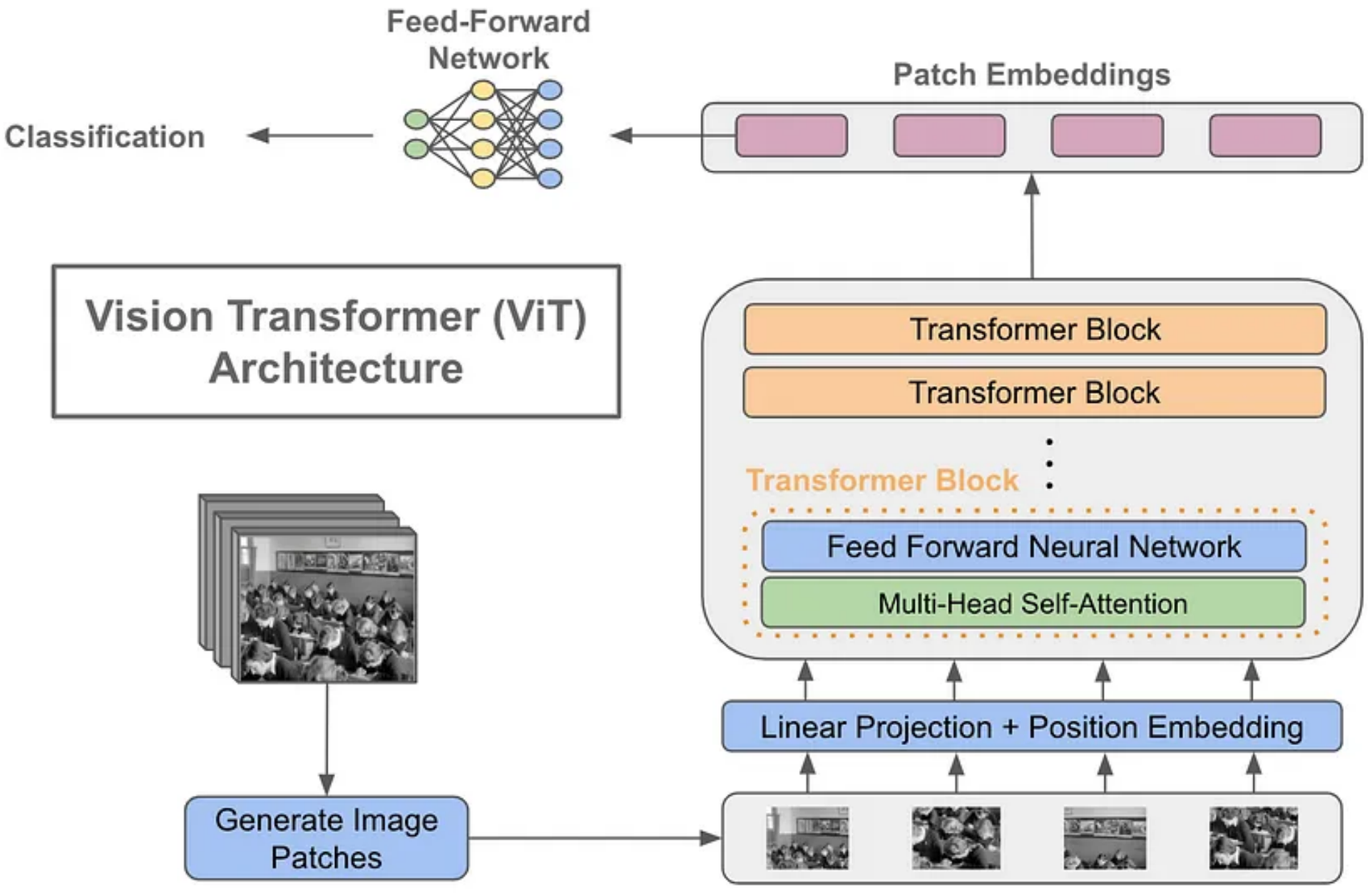
CNNs excel at building a hierarchy of features, going from basic edges to more complex shapes. However, CNNs can struggle with understanding the relationships between objects in distant parts of an image, which is crucial for accurate recognition in complex scenes.

Introducing Vision Transformers (ViTs)

- A novel deep learning architecture specifically designed for image recognition.
- Leverages the power of transformer networks, commonly used for natural language processing.

- Processes images by dividing them into patches and encoding them using vector representations.
- **Key idea:** Analyze relationships between image patches through attention mechanisms.

Introducing Vision Transformers (ViTs)



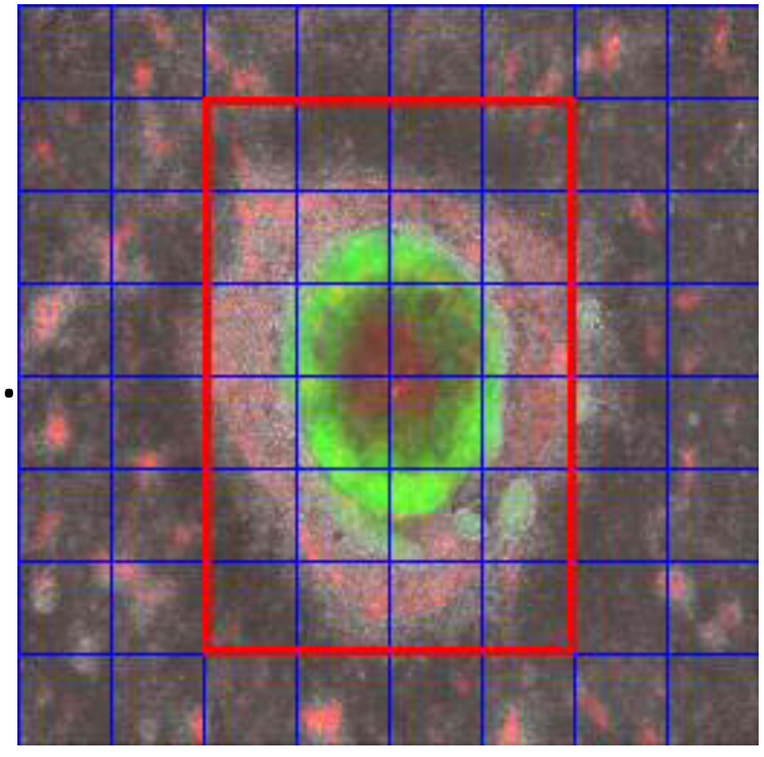
Introducing Vision Transformers (ViTs)

- Transformers excel at capturing relationships within sequences.
- For the ViT, the decoder is not as important in comparison to architectures like GPT where sequential data is involved.
- ViTs process images by dividing them into fixed-size patches, converting these patches into vectors, and then feeding them into the transformer for analysis.
- The core idea lies in using the transformer's attention mechanism to analyze the relationships between these image patches, allowing the model to understand how different parts of the image relate to each other.

Patch Processing in ViTs

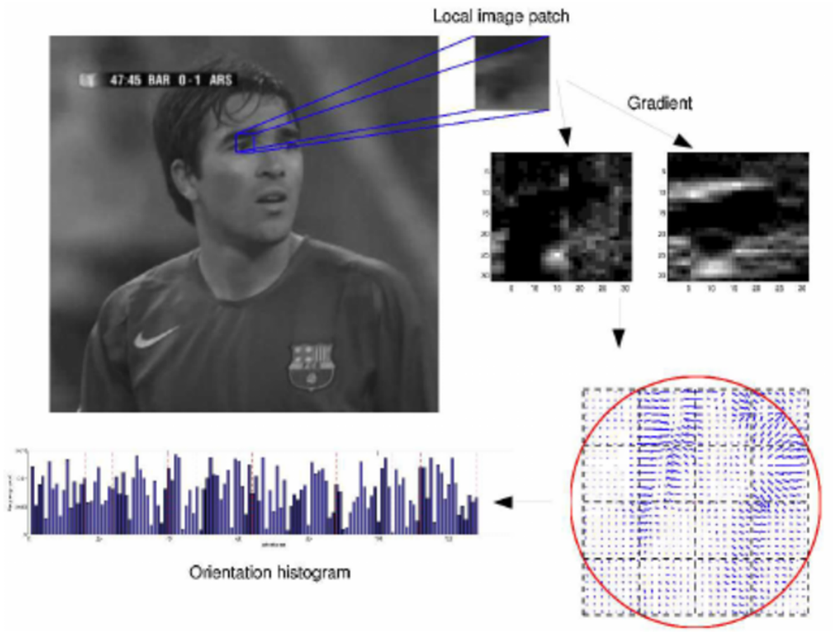
- The image is divided into a grid of fixed-size patches.
- Each patch is flattened into a vector representation.
- This vectorization process allows ViTs to treat image data similarly to sequences of words.

- The first step in ViT processing involves dividing the image into a grid of overlapping or non-overlapping patches.
- Each patch essentially captures a small local region of the image.



Patch Processing in ViTs

- These patches are then flattened into vectors, transforming the spatial information into a format suitable for the transformer architecture.
- This allows ViTs to leverage the strengths of transformers, which were originally designed to analyze sequences of words.

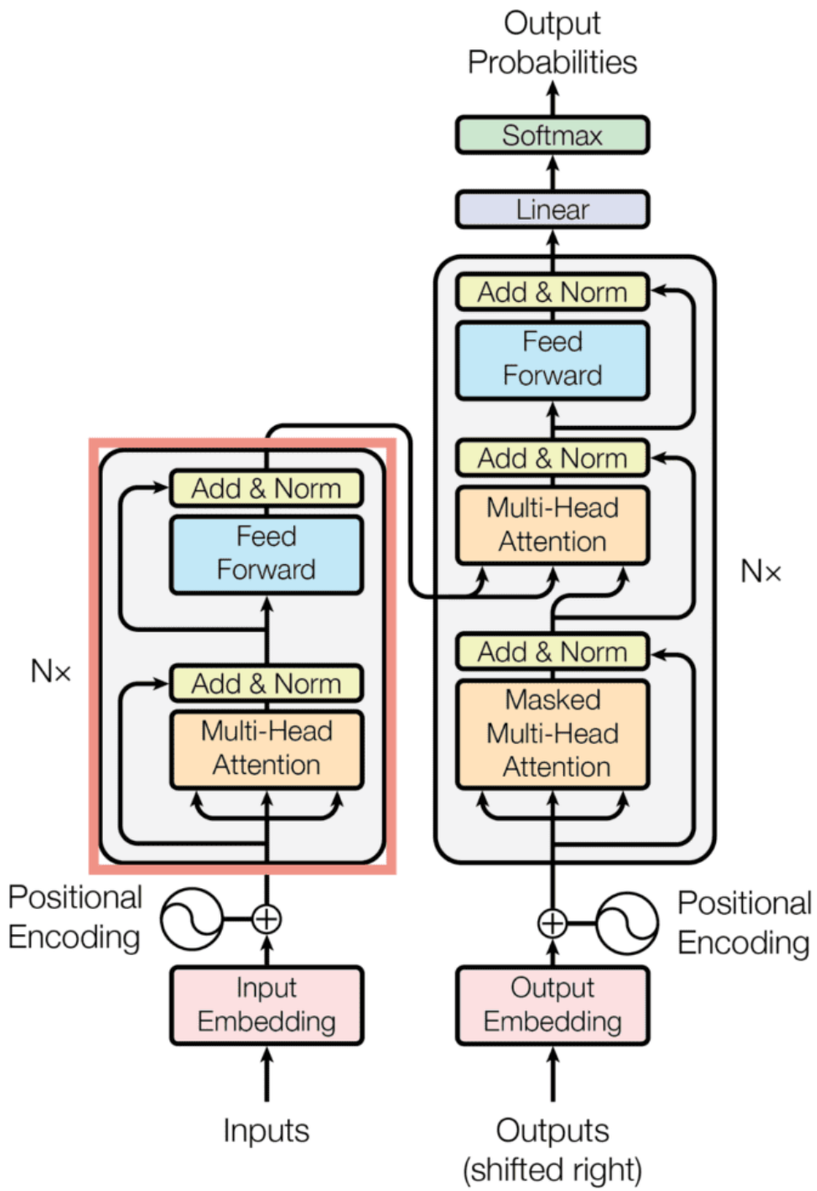


Transformer Encoder - The Core of ViTs

The transformer encoder is the powerhouse of ViTs. It consists of multiple stacked encoder layers. Each layer employs two crucial sub-layers:

1. **Multi-head attention:** This mechanism allows the model to focus on specific patches and analyze their relationships with other patches in the image. This is analogous to how we selectively focus on specific parts of a scene to understand the bigger picture.
2. **Feed-forward network:** This sub-layer introduces non-linearity into the model, allowing it to learn more complex relationships between patches.

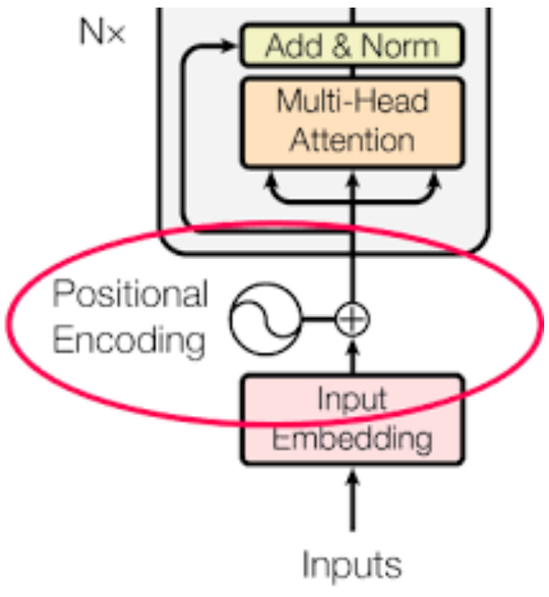
Transformer Encoder - The Core of ViTs



Transformer Encoder - The Core of ViTs

Positional Encoding:

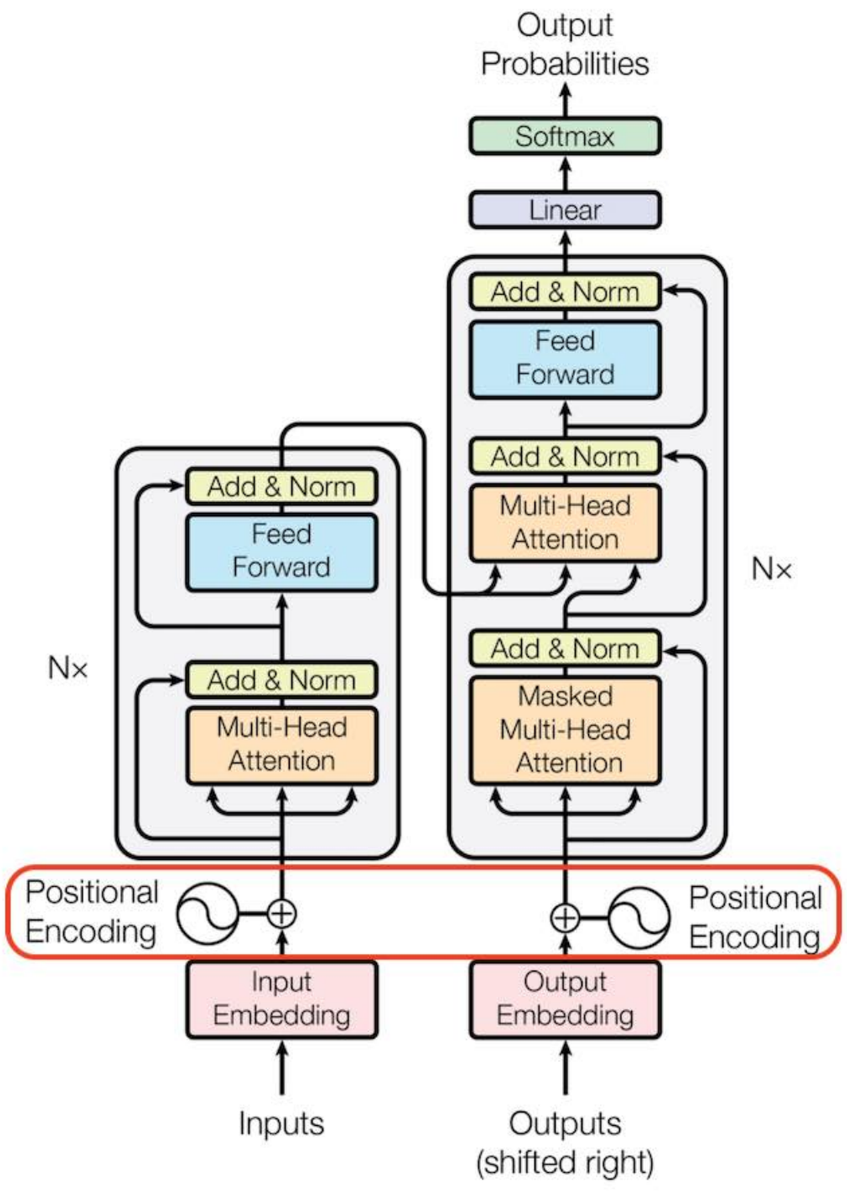
While convolutional neural networks inherently capture spatial information through their filters, ViTs, which process image data as patches, lack this built-in mechanism. To address this, positional encoding is often applied to the patch embeddings before feeding them into the transformer encoder.



Transformer Encoder - The Core of ViTs

- Positional encoding injects information about the relative or absolute position of each patch within the image. This allows the model to understand the spatial relationships between different parts of the image, even though the image is processed in patches.
- There are different techniques for positional encoding, with common approaches including sine and cosine functions.

Transformer Encoder - The Core of ViTs



Transformer Encoder - The Core of ViTs

The "An Image is Worth 16x16 Words" paper by Vaswani et al. (2022) played a significant role in demonstrating the effectiveness of transformers for image recognition tasks. ViTs leverage this concept to capture global dependencies within images, overcoming a limitation of CNNs.
https://arxiv.org/abs/2010.11929

Multi-Head Attention in ViTs

Multi-head attention is a core component of the transformer encoder.

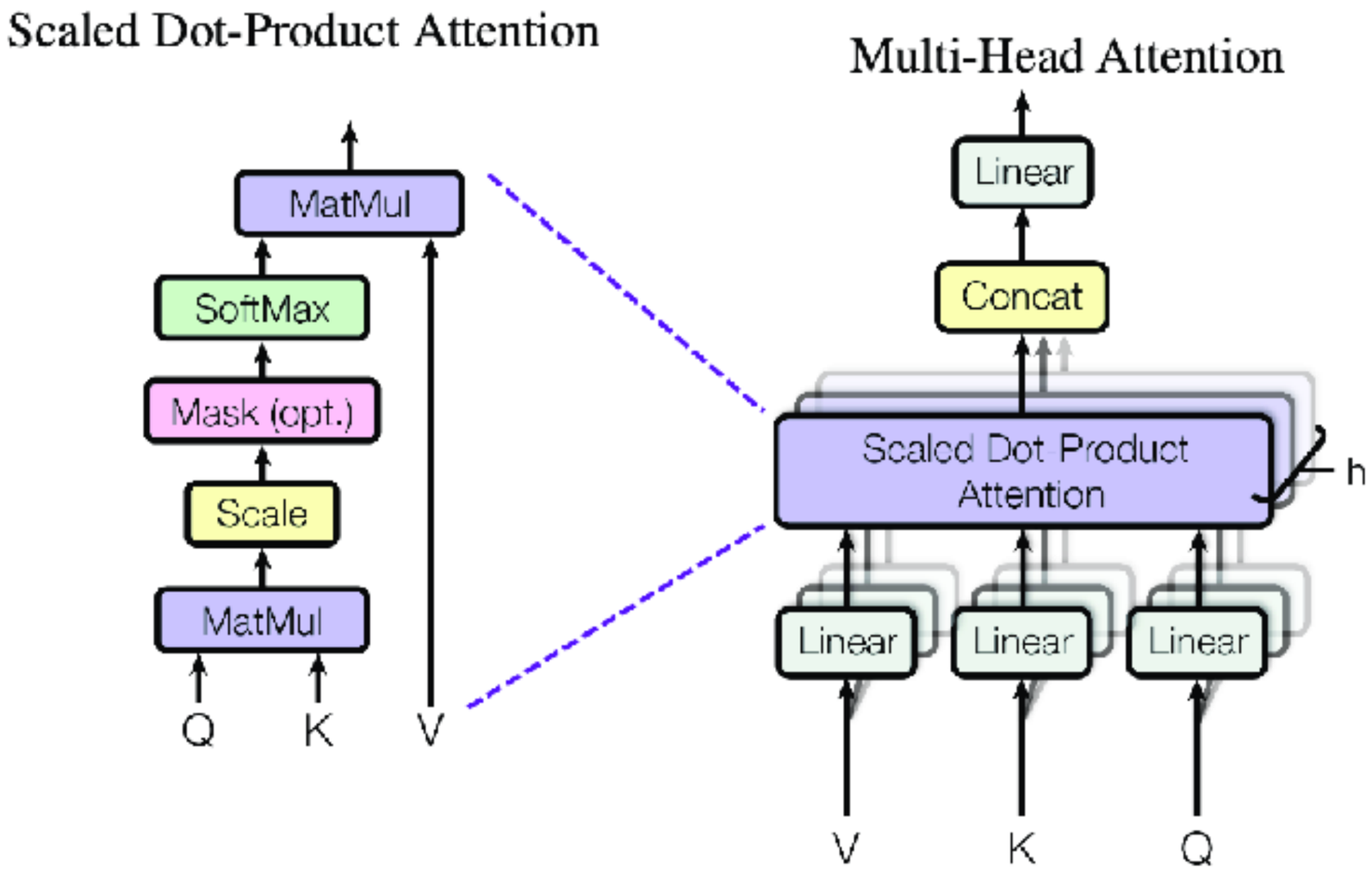
It allows the model to attend to different aspects of the input patches simultaneously.

Each head in multi-head attention learns a different way to relate patches.

Mathematically, it involves calculating:

- **Query**, **Key**, and **Value** vectors for each patch.
- A score is computed between each query and key vector.
- Attention weights are obtained by normalizing the scores.
- The weighted sum of value vectors is calculated, resulting in the output for each head.

Multi-Head Attention in ViTs



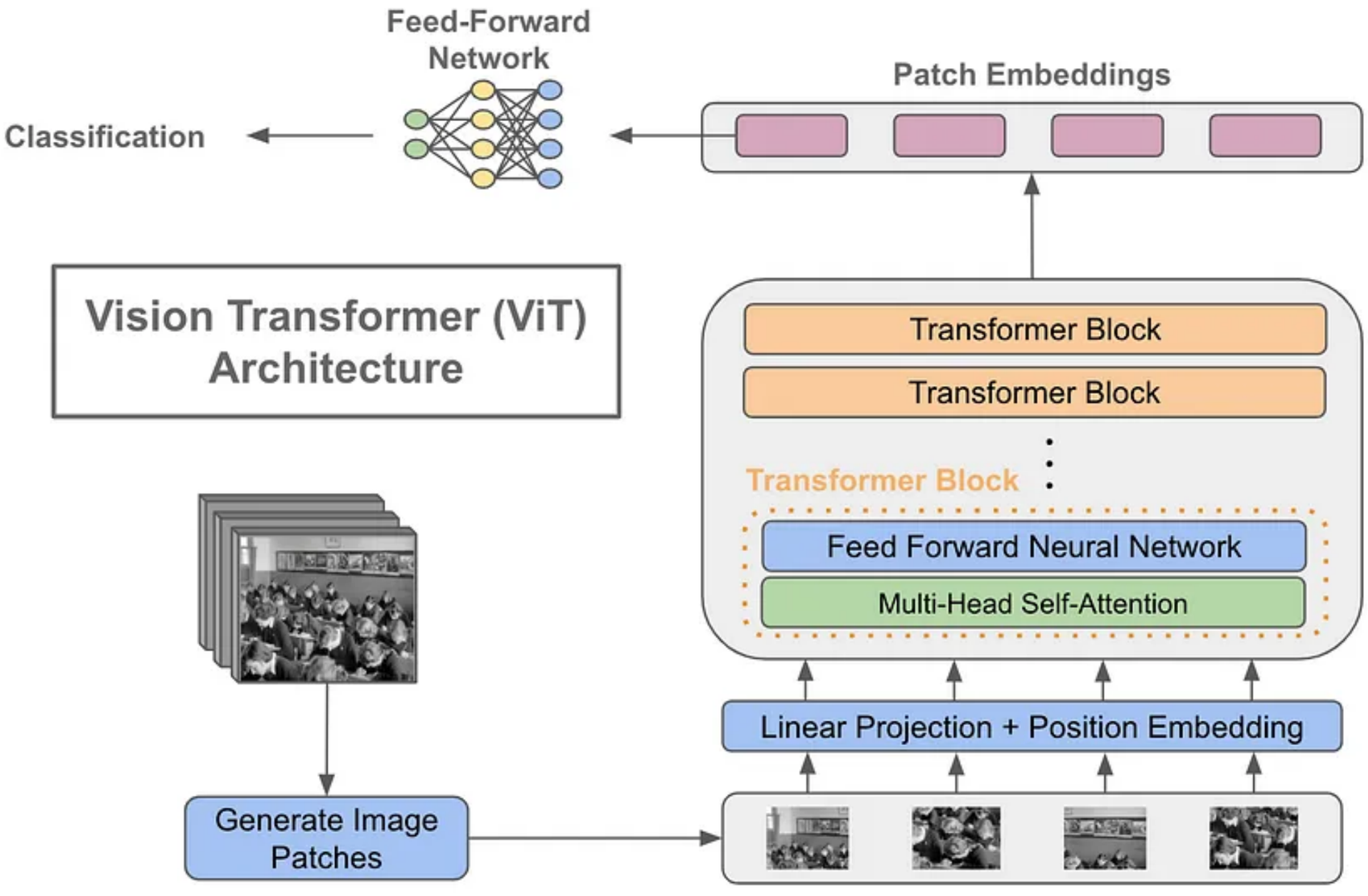
Feed-Forward Network

Feed-Forward Network:

Following the multi-head attention layer within each encoder layer, a feed-forward network is employed. This network is a type of artificial neural network with multiple layers that adds non-linearity to the model.

Function: The feed-forward network helps the model learn more complex relationships between the processed patches. Multi-head attention focuses on relationships, but the feed-forward network allows the model to go beyond simple comparisons and learn more intricate transformations based on the information gathered through attention.

Feed-Forward Network



Layer Normalization

Layer Normalization:

After both the multi-head attention and feed-forward network stages, a layer normalization step is typically applied.

Function: Layer normalization helps stabilize the learning process and improve the model's convergence. It essentially normalizes the outputs of each encoder layer, ensuring the activations remain within a specific range.

Residual Connection and Skip Connection

Residual Connection and Skip Connection:

A residual connection is often implemented around each encoder layer. This involves adding the input of the encoder layer (the processed patches before the multi-head attention) to the output of the layer (after the feed-forward network and layer normalization).

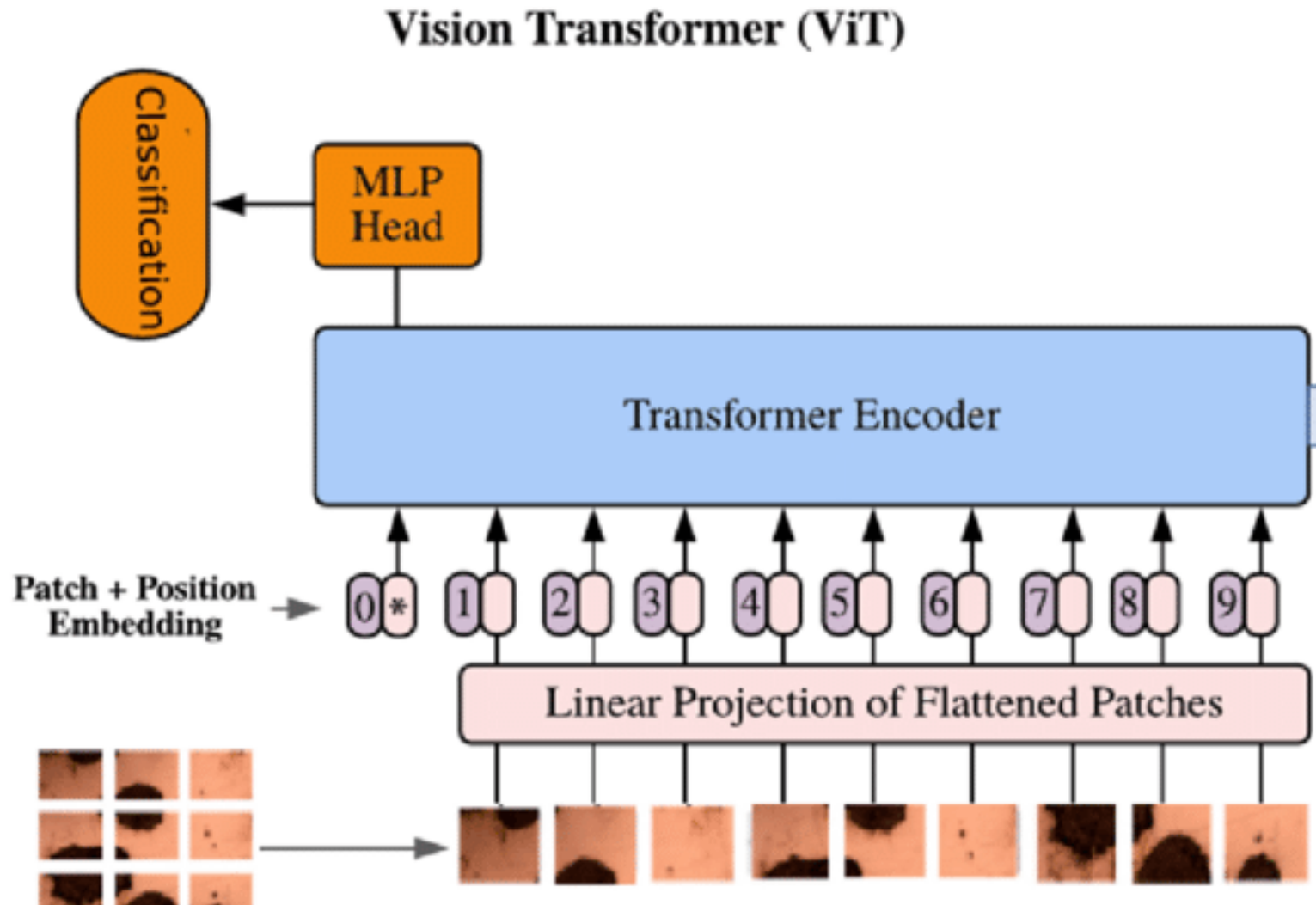
Function: Residual connections help address the **vanishing gradient problem**, which can hinder training in deep neural networks. By adding the original input, the model can learn from the identity mapping (not changing the input) along with the learned transformations.

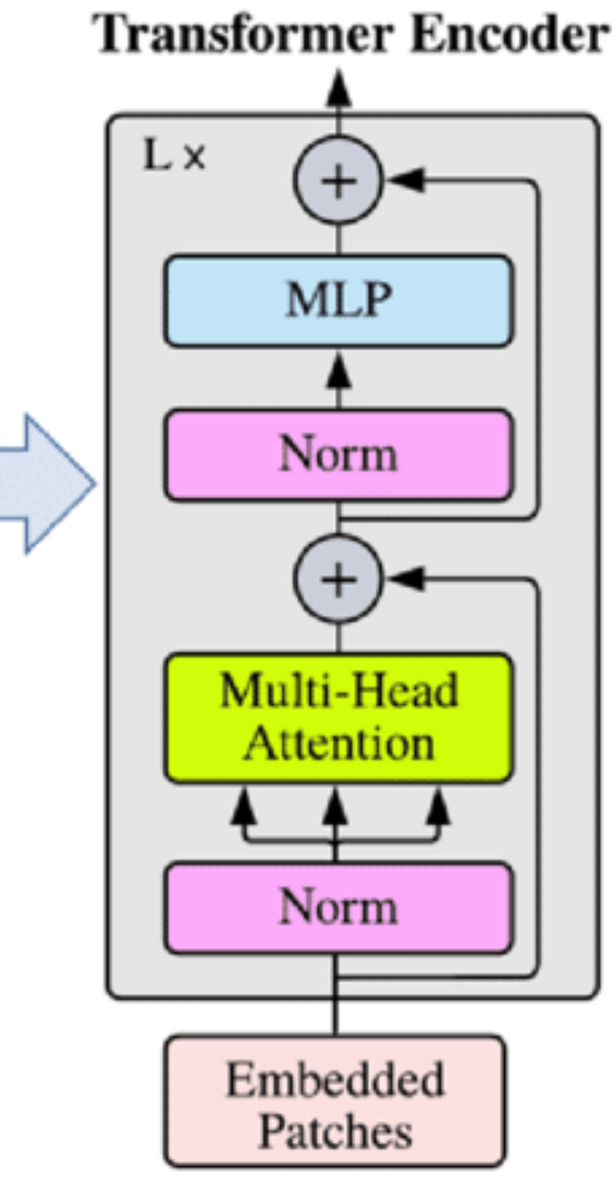
Classification or Regression Head

Depending on the specific task the ViT is being used for, there might be additional steps after the final encoder layer. Here's what could follow:

- **For Image Classification:**
  After the final encoder layer, a global pooling operation (e.g., averaging) might be applied to summarize the encoded feature representation of the entire image.
  This is followed by a fully-connected neural network with a single output neuron (for binary classification) or multiple output neurons (for multi-class classification). The final layer uses a softmax activation function to output class probabilities for the image.

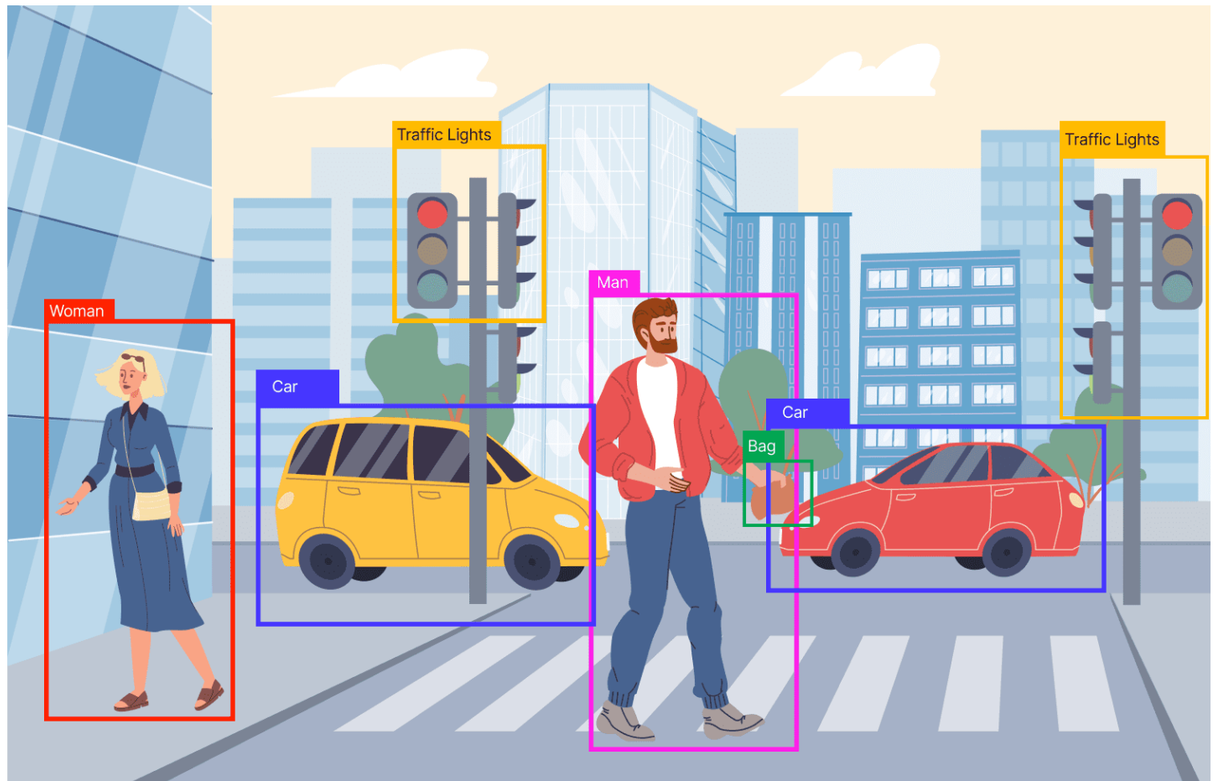
Classification or Regression Head





Classification or Regression Head

- **For Object Detection:**
  The final encoder output might be fed into a separate head for predicting bounding boxes and class labels for detected objects. This head could involve additional convolutional layers or specialized decoder architectures to generate the required outputs.



Classification or Regression Head

- **For Segmentation Tasks:**
  A decoder network might be employed after the encoder to project the encoded features back into a spatial representation corresponding to the original image resolution.
  This allows pixel-wise classification for tasks like semantic segmentation, where each pixel is assigned a label corresponding to the object it belongs to.
  Basically, every pixel in the image is assigned a label.

Classification or Regression Head

predict

Person
Bicycle
Background

An example for Segmentation Tasks.

Vision Transformers vs. Convolutional Neural Networks (CNNs)

Both ViTs and CNNs are powerful architectures for image recognition, but they have distinct strengths and weaknesses. This slide can present a comparison table highlighting factors like:

Long-range dependency capture: ViTs generally excel in this area.
Computational efficiency: CNNs might be more efficient for very high-resolution images.
Pre-training requirements: ViTs might require more pre-training data for optimal performance.
Performance on specific tasks: Depending on the task, either ViTs or CNNs might be more suitable.

Pre-training for Vision Transformers

- ViTs often require pre-training on large image datasets like ImageNet. Pre-training helps the model learn general visual representations.

IMAGENET

- This pre-training stage neIps the model learn general visual representations and feature extraction capabilities. Once pre-trained, ViTs can be fine-tuned for specific image recognition tasks using a smaller amount of task-specific data. This fine-tuning process allows the model to specialize in recognizing specific objects or categories within an image.
- Fine-tuning on specific tasks can then be performed with less data.

The DeiT (Data-efficient Image Transformer) Model

- The DeiT (Data-efficient Image Transformer) model is a specific ViT architecture designed for efficiency.
- It achieves competitive results with a lower number of pre-training parameters compared to other ViT models.
- This translates to lower computational cost, making DeiT suitable for deployment on devices with limited resources.
- Well-suited for deployment on resource-constrained devices.
- Feel free to explore!
  https://arxiv.org/abs/2012.12877

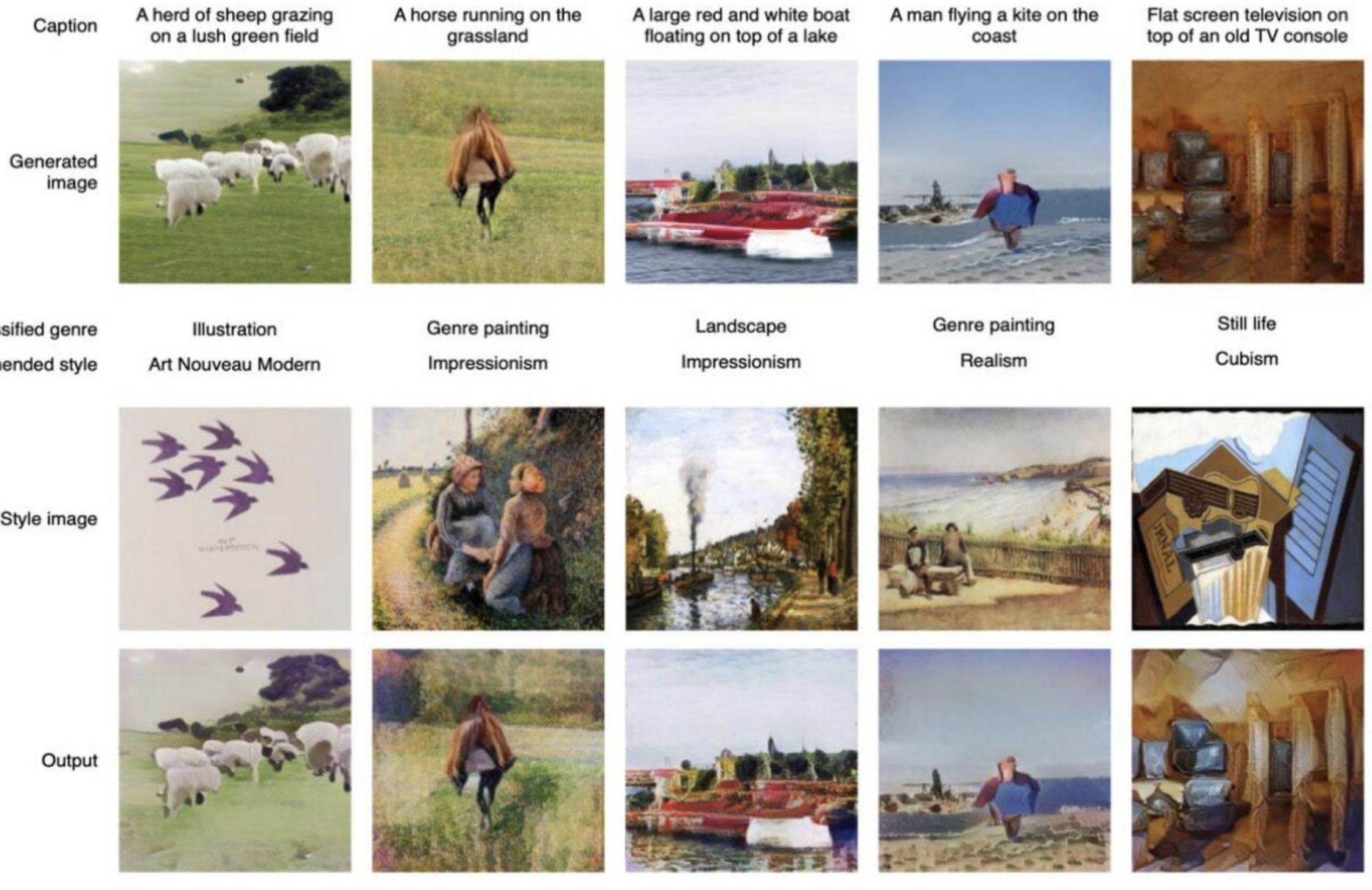
ViT-Based Image Generation

An exciting area of research involves using ViTs for image generation. Here, the model takes a textual description as input and generates a corresponding image.

This technology is still under development but holds vast potential for various applications. For instance, it could be used to generate product mockups.

This is an emerging area of research with vast potential applications.

ViT-Based Image Generation



Acknowledgements & References

- https://towardsdatascience.com/using-transformers-for-computer-vision-6f764c5a078b
- https://www.researchgate.net/figure/a-Image-divided-into-8x8-grid-of-small-patches-b-Highlighted-in-red-is-the-ROI_fig1_377445769
- https://machinelearningmastery.com/the-transformer-model/

THANK YOU