GDAŃSK UNIVERSITY OF TECHNOLOGY

WYDZIAŁ ELEKTRONIKI, TELEKOMUNIKACJI I INFORMATYKI

Vision Transformers

Vision Transformers

Magdalena Mazur-Milecka

15 maja 2025

Definition and Concept

Transformer - proposed in 2017 by Google for text analysis: Natural Language Processing (translation, understanding, generation) - "Attention is Aall You Need."

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin. 2017. Attention is all you need. In Proceedings of the 31 International Conference on Neural Information Processing Systems (NIPS17).

Definition and Concept

Vision Transformers (ViT) - proposed in 2020 by Google - "An Image is Worth 16*16 Words: Transformers for Image Recognition at Scale".

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

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

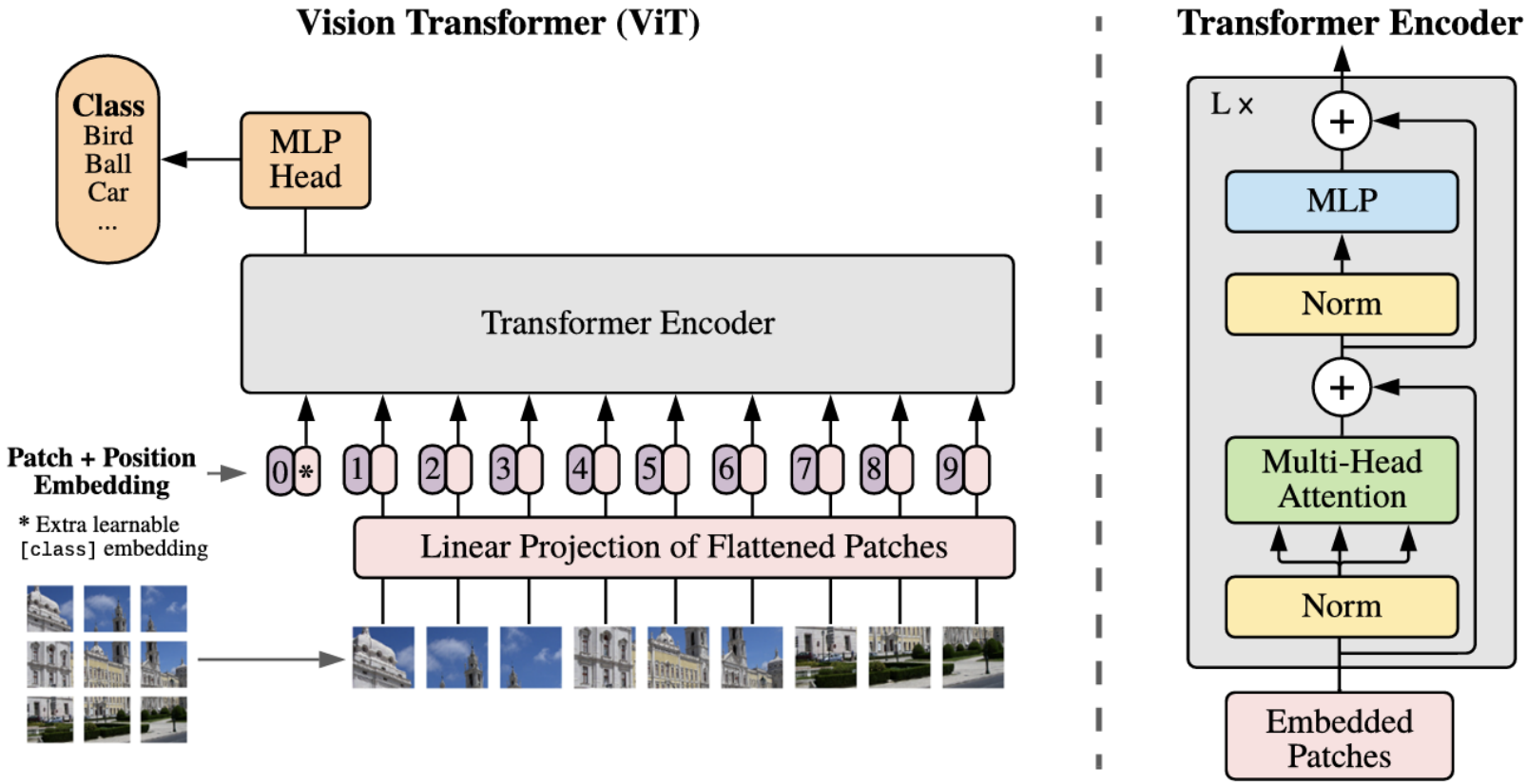
Google Research, Brain Team

{adosovitskiy, neilhoulsby}@google.com

[2] Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

Definition and Concept

Vision Transformers - basic idea



[2] Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

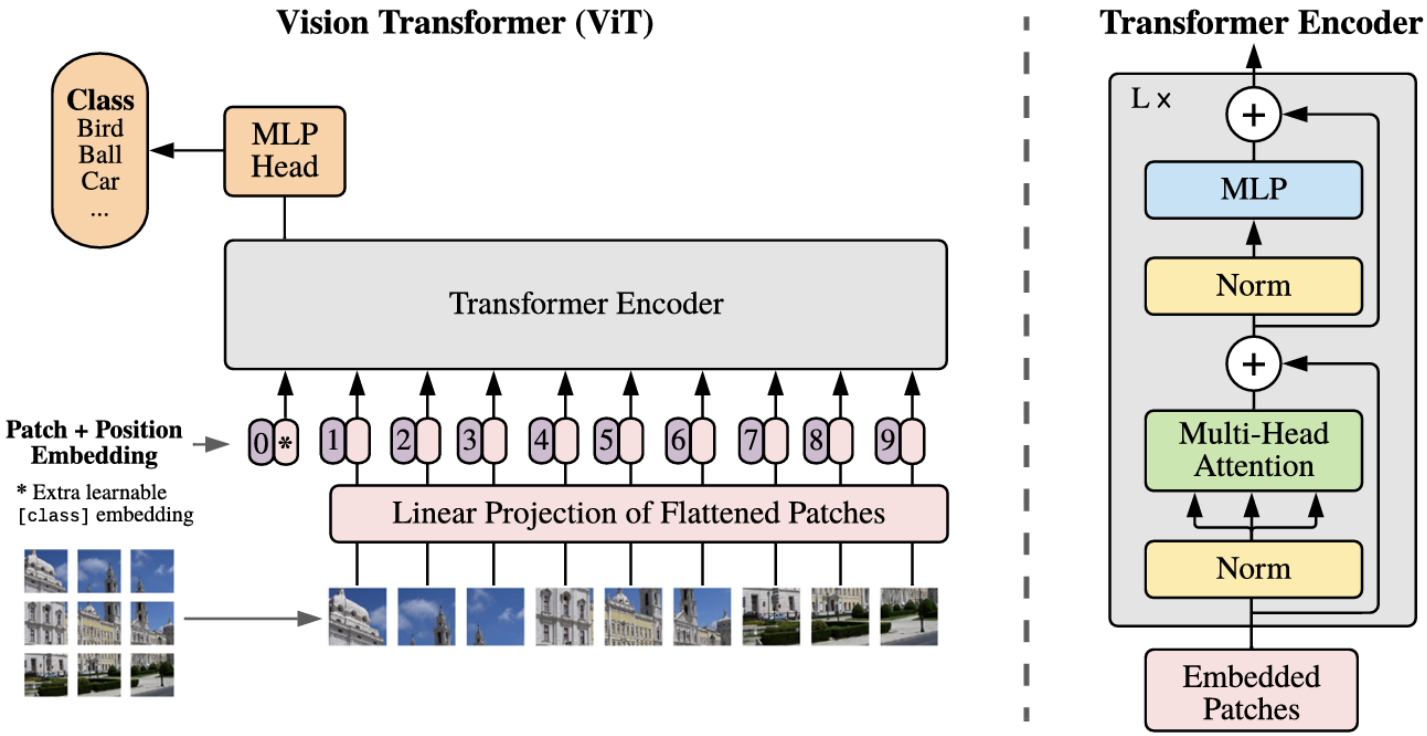
Agenda

3. Vision Transformers: Architecture and Theory

- **ViT Architecture Overview**
  - Patch Embedding
  - Position Embeddings
  - Classification Head
  - Transformer Encoder Layers
- **Significance**
  - Comparison with traditional methods
  - Strengths and Limitations of CNNs
  - Why ViTs are important in modern computer vision

Vision Transformers

- Based on the Transformer "Attention is all you need"
- Huge training set (300 million images)

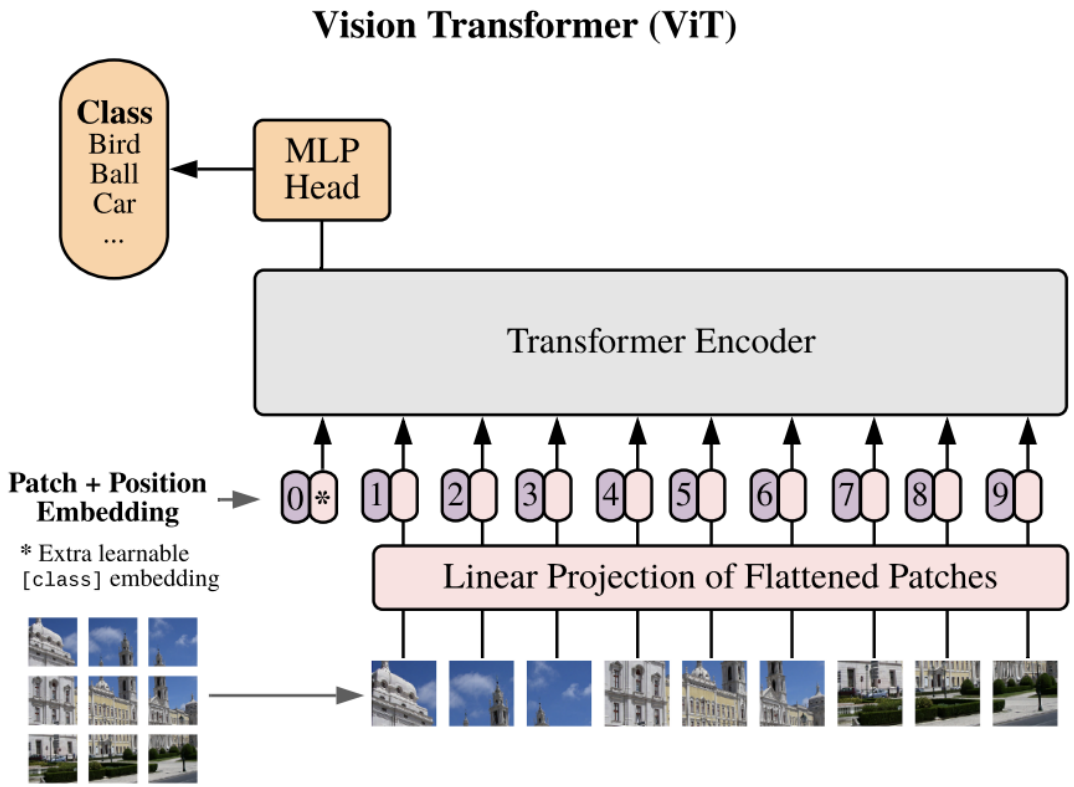


Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

Vision Transformers

Pipeline:

1. Dividing the image into patches,
2. Vectorizing patches (flattening),
3. Creating patch embeddings,
4. Adding positional embedding,
5. Feeding this data to the Transformer encoder,
6. Final classification using Multi-Layer Perceptron



Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

Vision Transformers

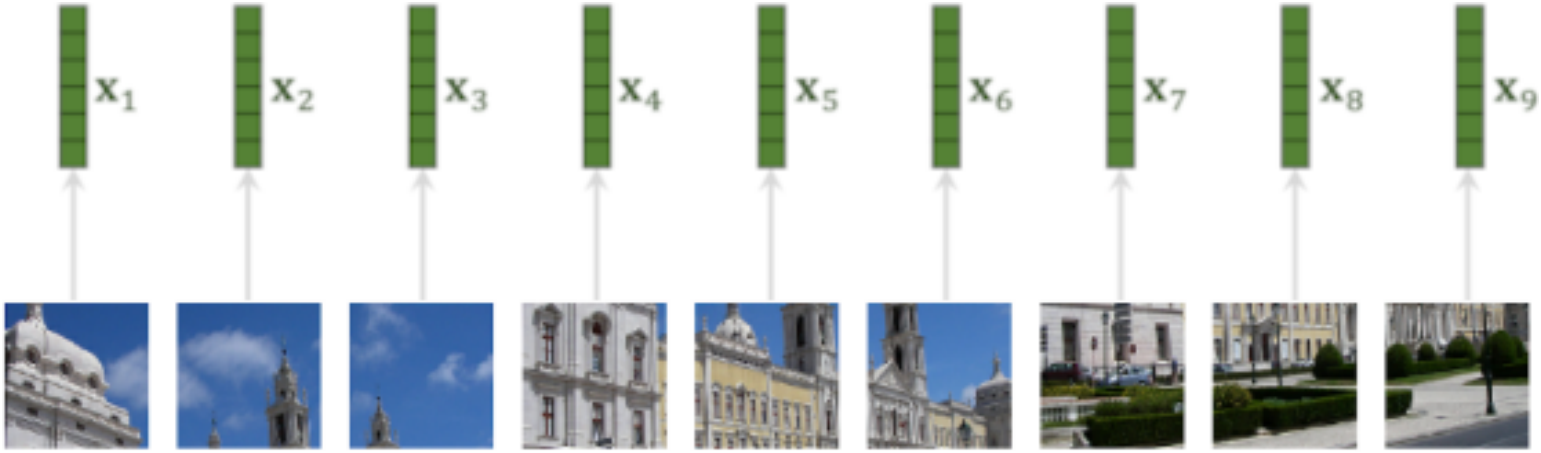
1. Dividing the image into patches

ViT divides images into visual tokens - unroll non-overlapping patches of fixed size (e.g., 16x16) into sequence,

Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

Vision Transformers

2. Vectorizing patches (flattening)

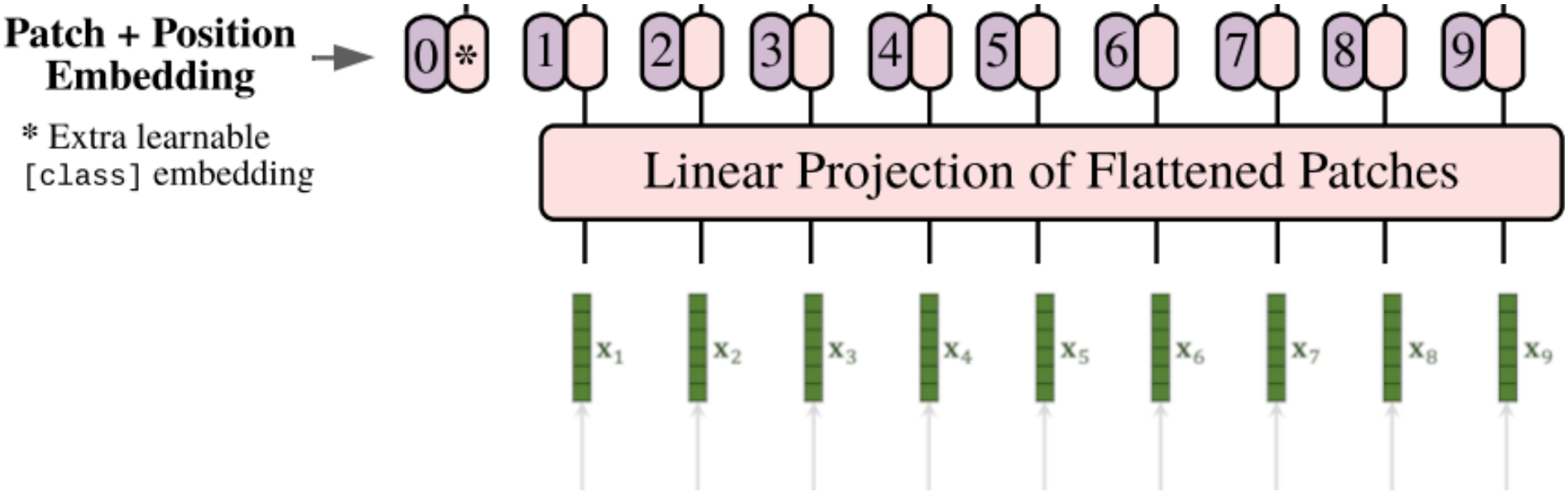


Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

Vision Transformers

3. Creating patch embeddings

The each patch vector is then linearly projected into a higher-dimensional embedding space using a trainable linear layer



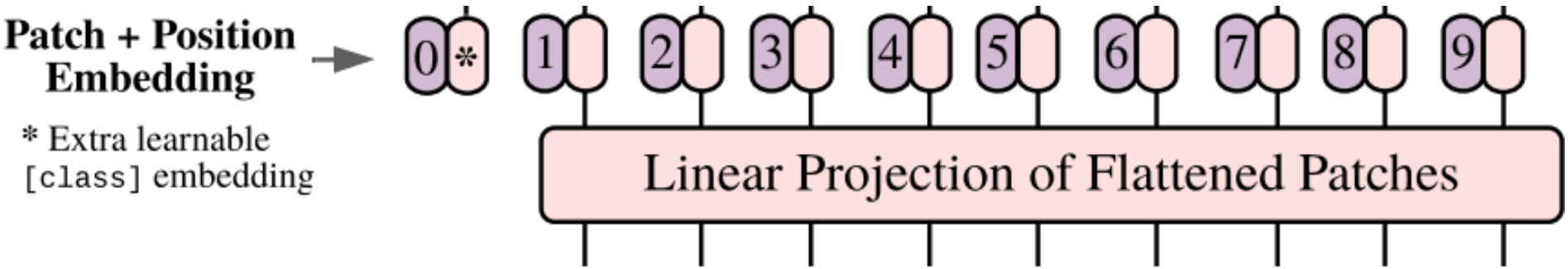
Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

Vision Transformers

4. Adding position embedding

To each patch embedding (pink), positional embedding (purple) is added - also a trainable parameter

0∗ - "classification token" (CLS token)



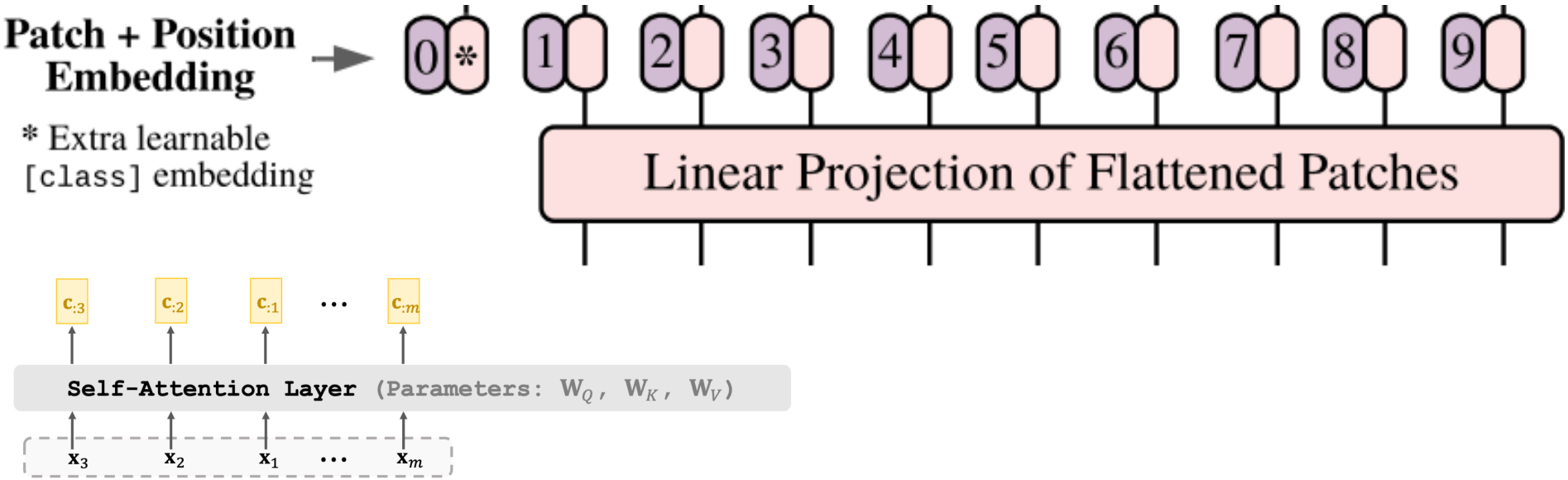
Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

Vision Transformers

4. Adding position embedding

To each patch embedding (pink), positional embedding (purple) is added - also a trainable parameter

0∗ - "classification token" (CLS token)

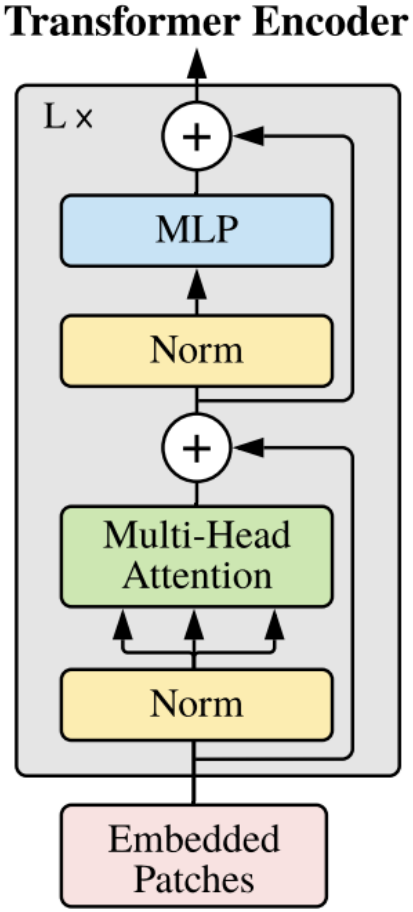


Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

Vision Transformers

5. Feeding this data to the Transformer encoder,

The Transformer encoder contains L blocks:



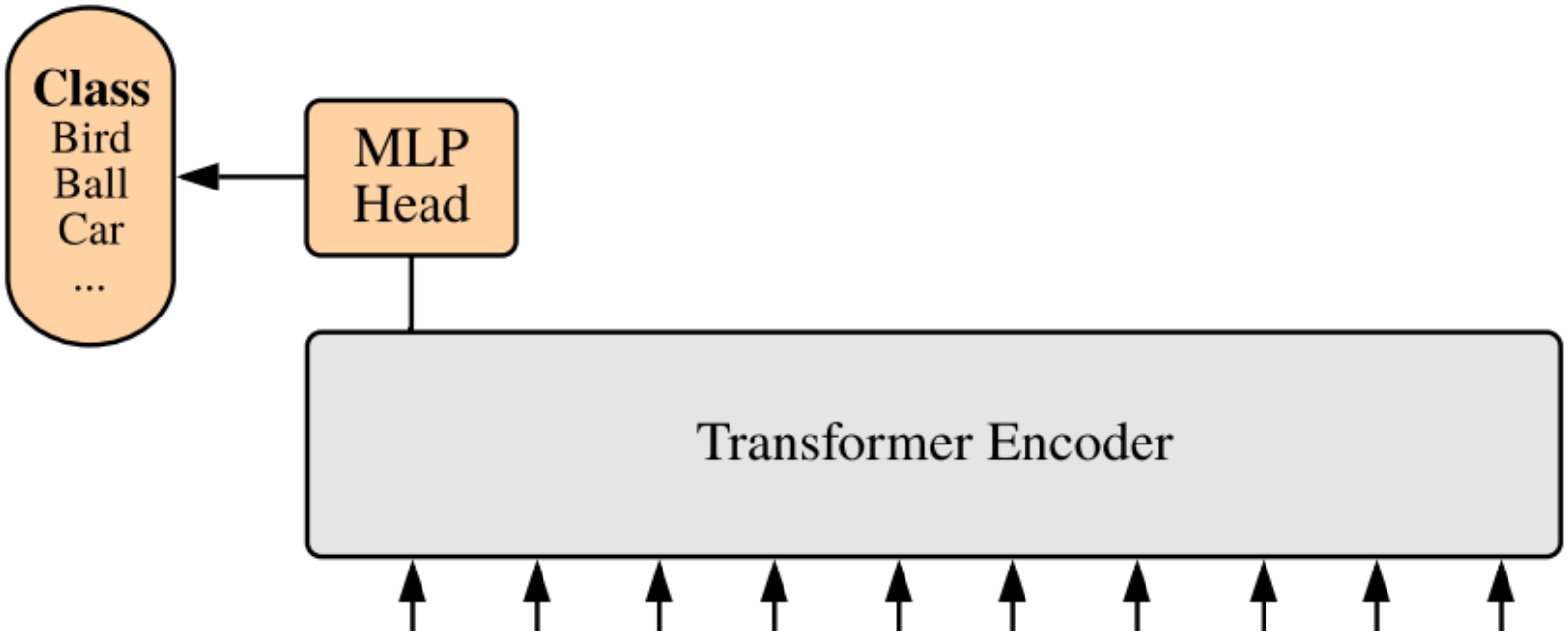
1. Multi-Head Self Attention Layer (MSP)
2. Multi-Layer Perceptrons (MLP):
3. Normalization Layer (LN):
4. Skip connections

Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

Vision Transformers

6. Multi-Layer Perceptron Head

The size of the standard Transformer encoder (Self-Attention) output matches the input. For classification, we only consider first element that corresponds to the CLS token, which is designed to capture the overall representation of the input image after passing through the transformer.

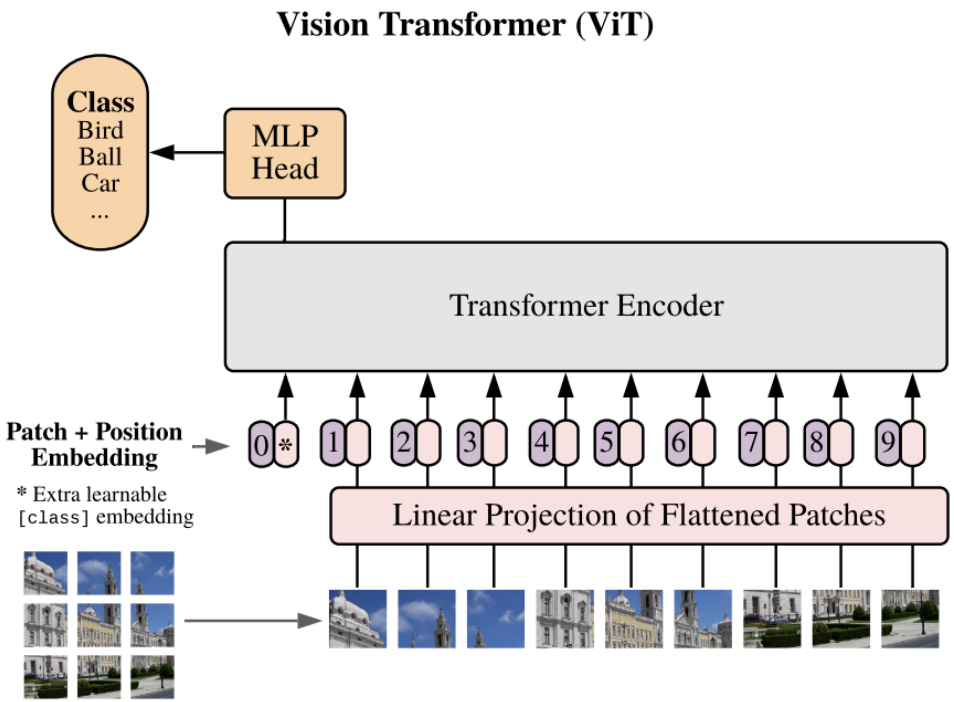


Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

Vision Transformers

Key Takeaways:

- Utilizes Transformer Encoder and Self-Attention
- Transformer inputs are flattened from multiple patches
- Patch and position embeddings are learnable parameters
- In the final transformer encoder layer we only consider the first output element



Demo

myViT_v3.ipynb

Implementation Walkthrough

- Loading and Preprocessing Image Data
- Creating Patch Embeddings
- Building the ViT Model in Code
- Training the ViT Model

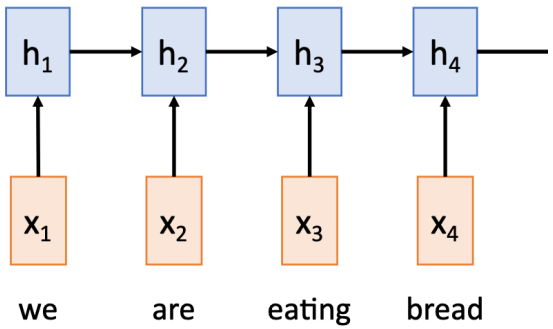
Transformers vs. RNN / LSTM

RNN / LSTM:

- Order of input matters
- Must compute sequentially
- Difficult to train due to complex dependency types
- Words from the beginning are forgotten after some time
- Sequence - inductive prior
- **Constrained by the structure**

Transformer:

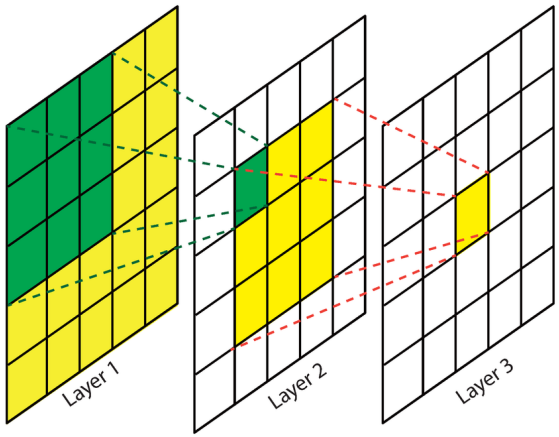
- Input order is not important - no imposed way of analyzing it
- Can compute in parallel
- Can analyze larger parts of text (remembers longer)
- Enables modeling longer dependencies



ViT vs. CNN

CNN:

- Filter receptive field is initially local, only in subsequent layers it extends to global,
- Imposed way of analyzing the image through filters - inductive prior



Transformers:

- Analyzes the image at different scales from the beginning,
- No imposed way of analyzing the image,
- Requires more training data - needs to learn image structure analysis,

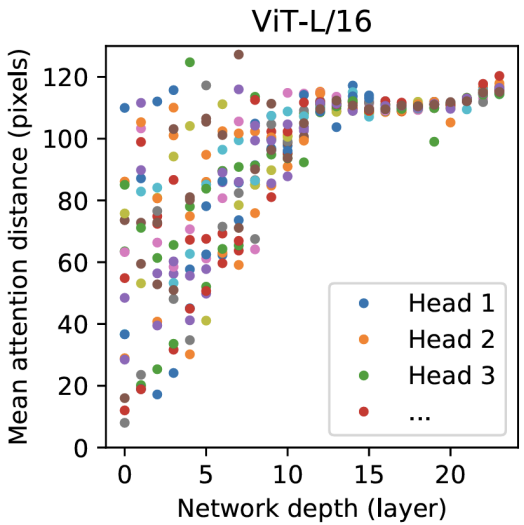
---

L, Haoning, S, Zhenwei, Z. Zhengxia. (2017). Maritime Semantic Labeling of Optical Remote Sensing Images with Multi-Scale Fully Convolutional Network. Remote Sensing

ViT vs. CNN

CNN:

- Filter receptive field is initially local, only in subsequent layers it extends to global,
- Imposed way of analyzing the image through filters - inductive prior



Transformers:

- Analyzes the image at different scales from the beginning,
- No imposed way of analyzing the image,
- Requires more training data - needs to learn image structure analysis,

Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

ViT vs. CNN

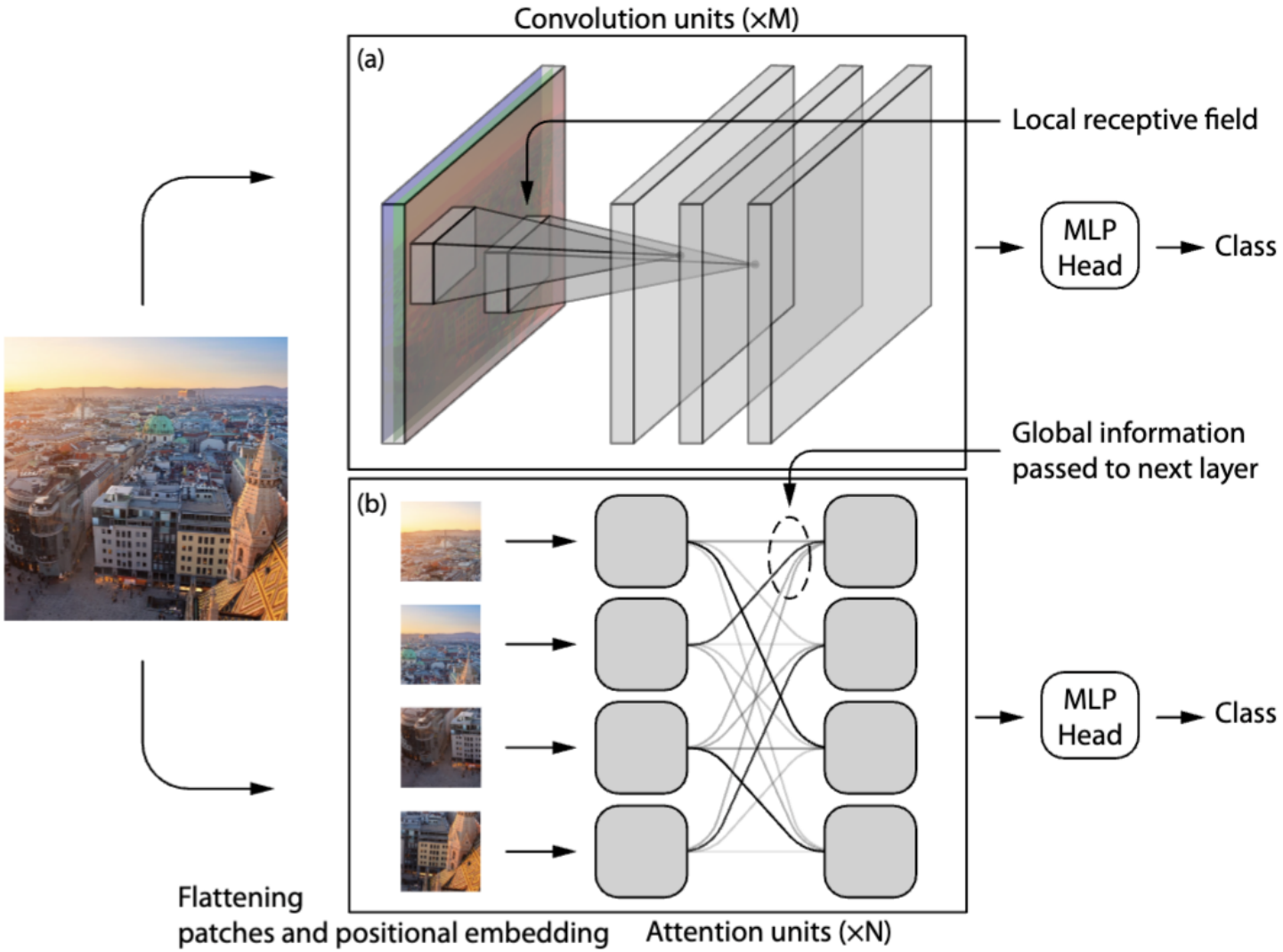
CNN:

- Filter receptive field is initially local, only in subsequent layers it extends to global,
- Imposed way of analyzing the image through filters - inductive prior
- **Constrained by the structure**

Transformers:

- Analyzes the image at different scales from the beginning,
- No imposed way of analyzing the image,
- Requires more training data - needs to learn image structure analysis,

ViT vs. CNN



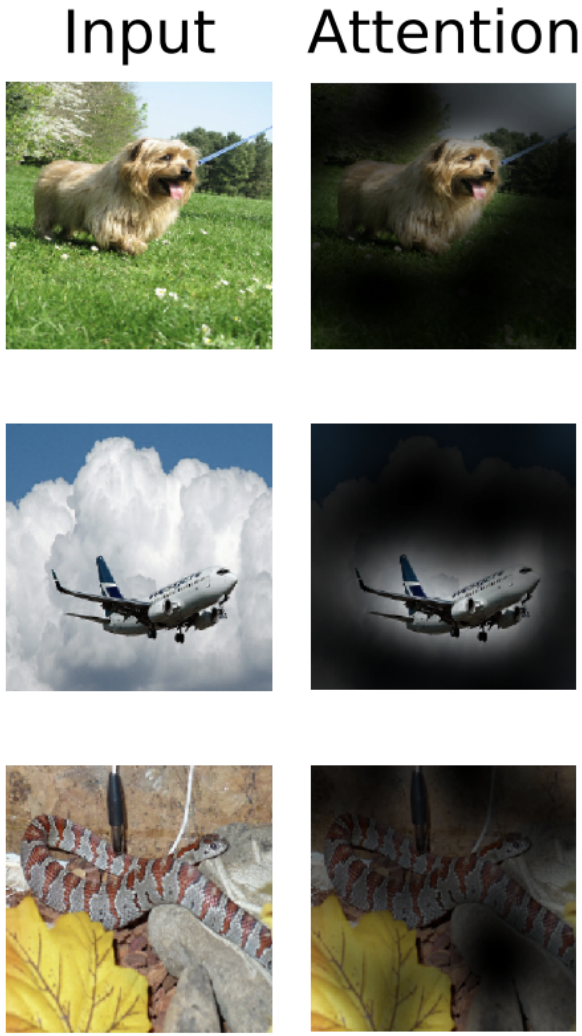
Tuli, S., Dasgupta, I., Grant, E., Griffiths, T.L. (2021). Are Convolutional Neural Networks or Transformers more like human vision? ArXiv, abs/2105.07197.

Why ViTs are so important

- Global Context Awareness
- Flexibility and Adaptability - general architecture
- Scalability - fine-tuning
- Performance - SwinTransformer, DeiT
- Efficiency - high-resolutions
- Explainability

Why ViTs are so important

- Global Context Awareness
- Flexibility and Adaptability - general architecture
- Scalability - fine-tuning
- Performance - SwinTransformer, DeiT
- Efficiency - high-resolutions
- Explainability



Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

Agenda

5. Examples of ViTs

- Review of Successful ViTs
- The Hybrid: ViT + CNN

Vision Transformer Problems

1. Dividing the image into patches

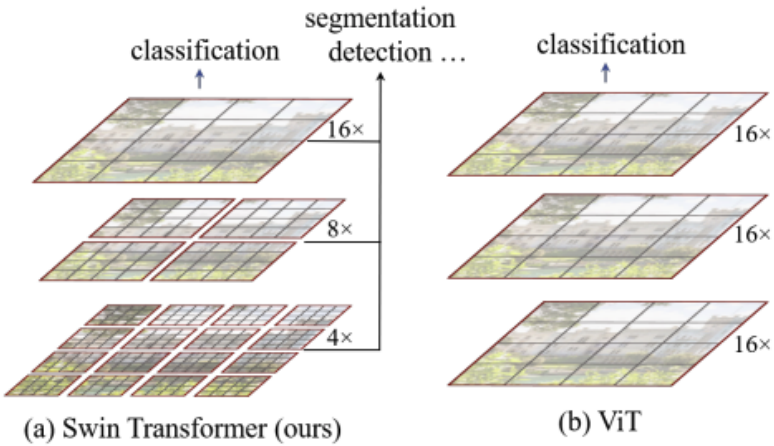
ViT divides images into visual tokens non-overlapping patches of fixed size (e.g., 16x16),

Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

SwinTransformer

Hierarchical Vision Transformer using **S**hifted **Win**dows

- Adapting Transformers to the image structure
- Suitable for detail detection where 16 x 16 pixels are too large (e.g., segmentation)
- Patches are smaller (4 x 4), then their size is increased using **Patch Merging**
- Different tasks use information from different patch sizes - **Hierarchical Architecture**

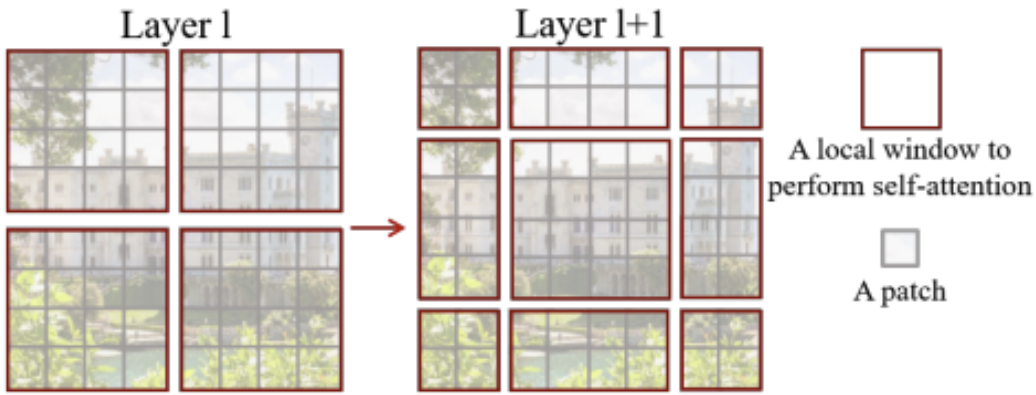


Liu, Ze et al. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows." 2021 IEEE/CVF International Conference on Computer Vision (ICCV)

SwinTransformer

Shifted **Win**dows

- Introducing **Shifted Window based Self-Attention** (only M patches from the neighborhood are considered)

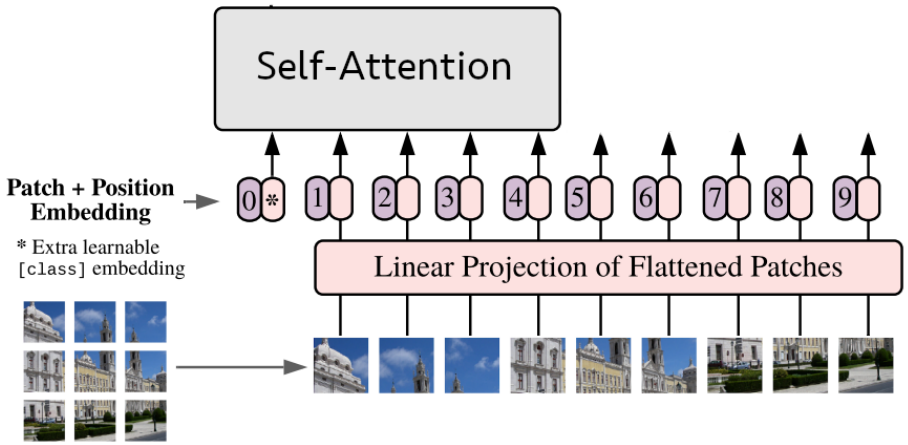


Liu, Ze et al. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows." 2021 IEEE/CVF International Conference on Computer Vision (ICCV)

SwinTransformer

Shifted **Win**dows

- Introducing **Shifted Window based Self-Attention** (only M patches from the neighborhood are considered)

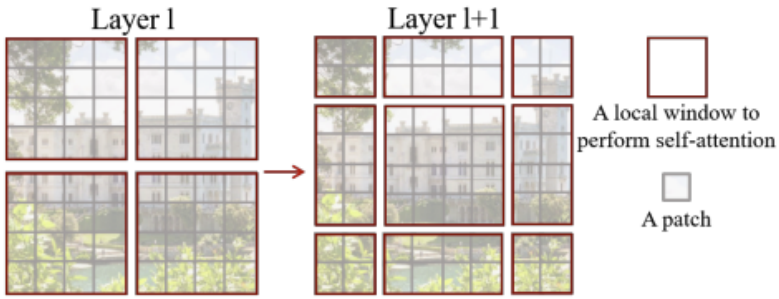


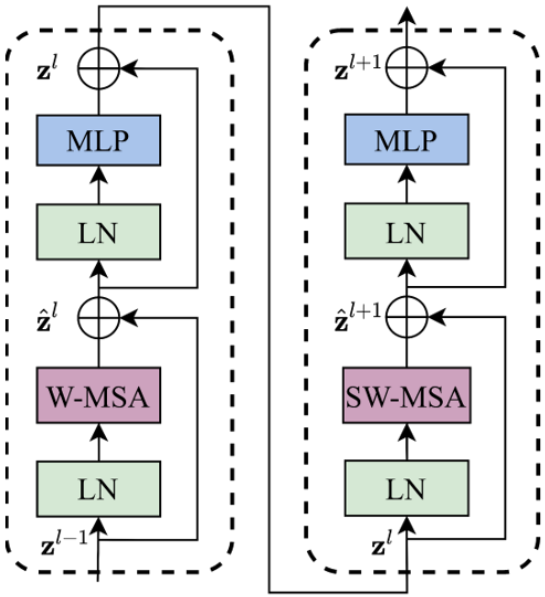
Liu, Ze et al. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows." 2021 IEEE/CVF International Conference on Computer Vision (ICCV)

SwinTransformer

Shifted **Win**dows

- Introducing **Shifted Window based Self-Attention** (only M patches from the neighborhood are considered)
- Windows are shifted between consecutive layers to capture cross-window connections and improve the receptive field





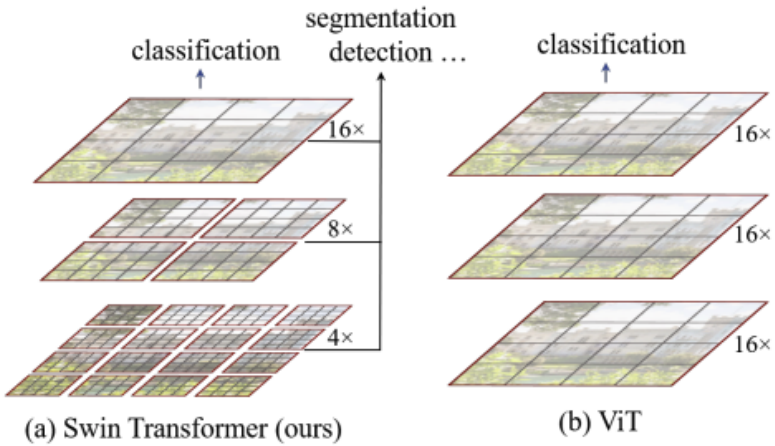
(c) Two Successive Swin Transformer Blocks

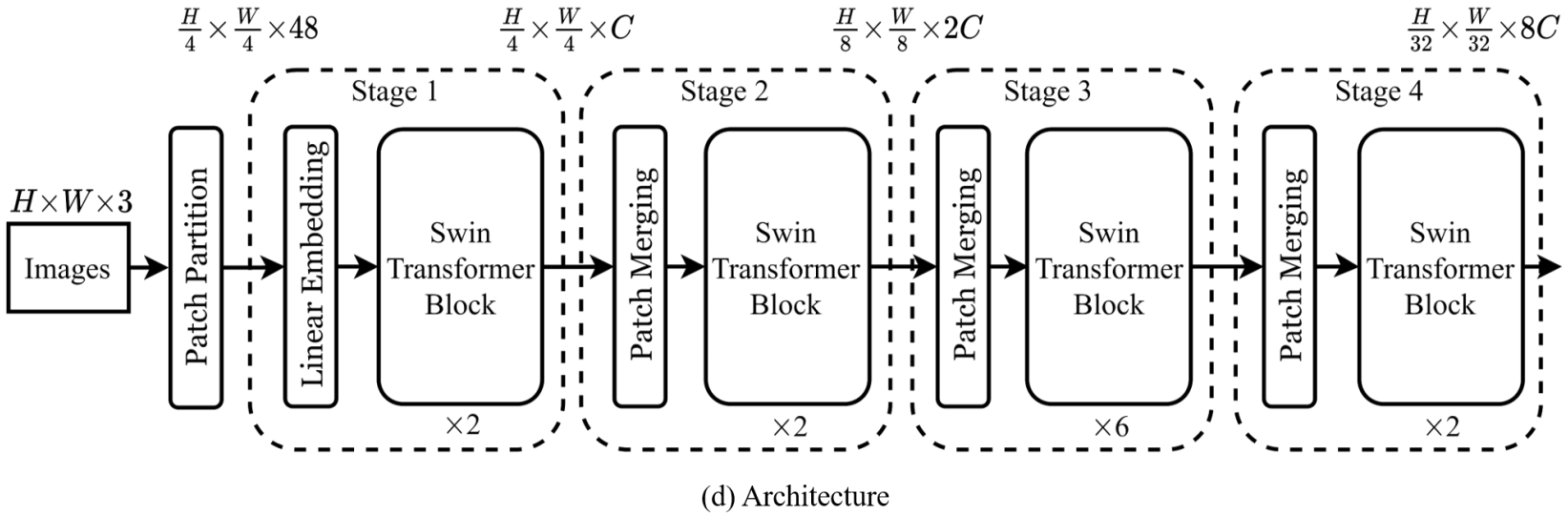
Liu, Ze et al. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows." 2021 IEEE/CVF International Conference on Computer Vision (ICCV)

Swin Transformer

Hierarchical Architecture

- Hierarchical structure where the feature maps are generated at multiple scales
- merging patches in a manner similar to pooling layers in CNNs





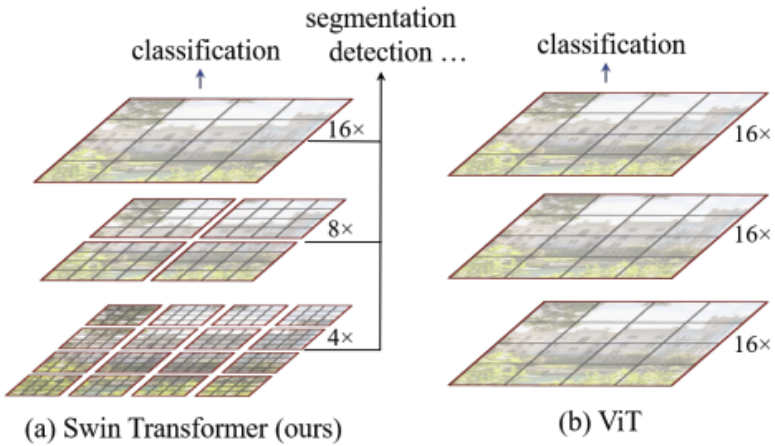
(d) Architecture

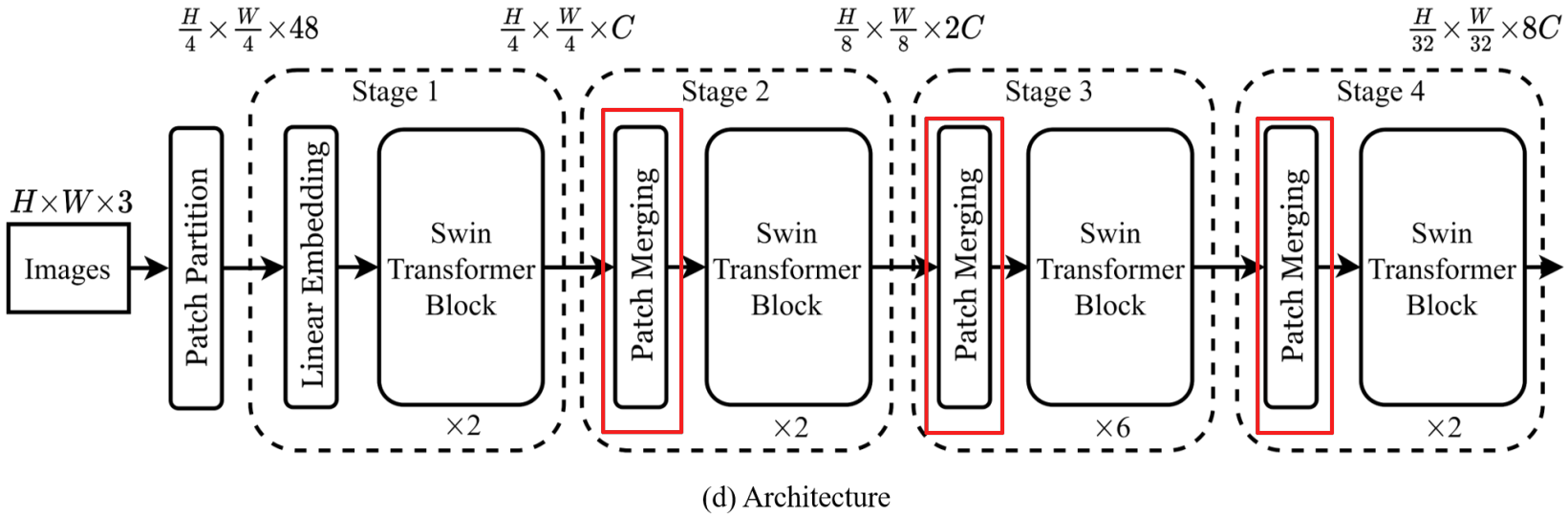
Liu, Ze et al. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows." 2021 IEEE/CVF International Conference on Computer Vision (ICCV)

Swin Transformer

Patch Merging

- merging layers concatenate neighboring patches, reducing the spatial dimensions while increasing the depth of feature channels.





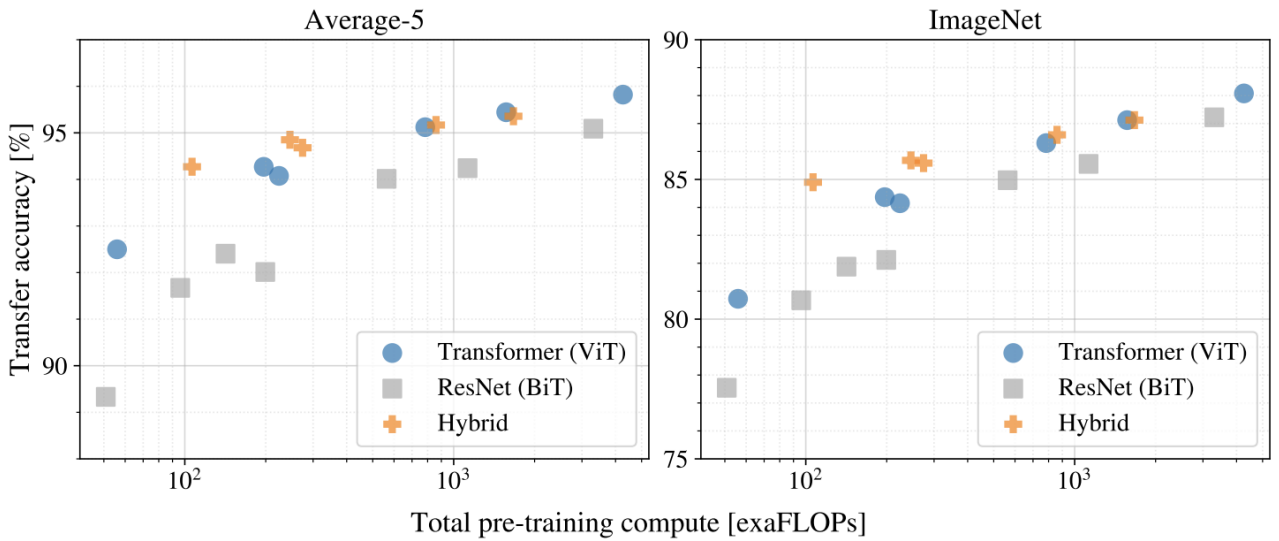
(d) Architecture

Liu, Ze et al. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows." 2021 IEEE/CVF International Conference on Computer Vision (ICCV)

Demo

swin_transformer_flowers.ipynb

Transformers in CV

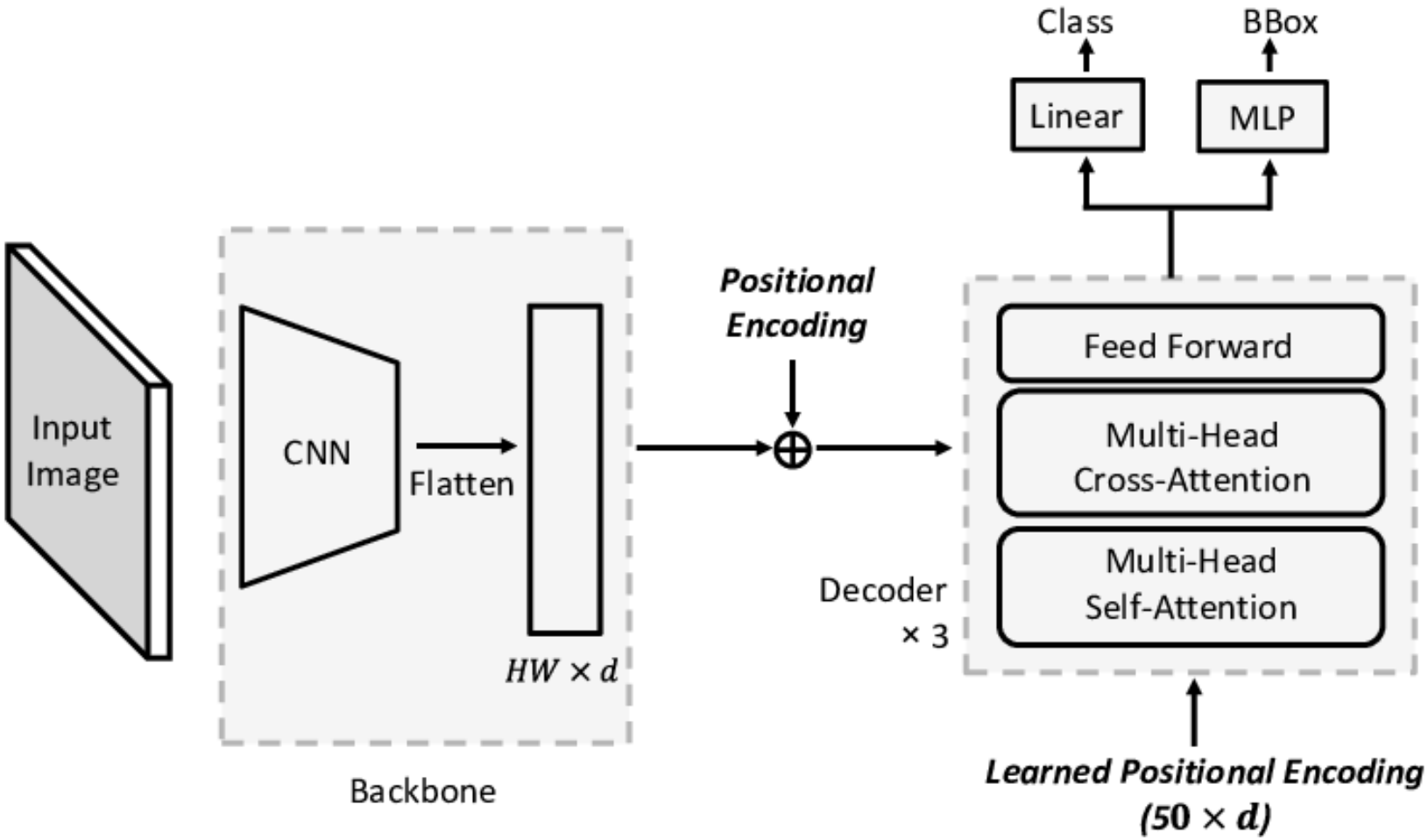


- **Pure Transformer** (ViT) is more efficient and scalable than traditional CNN networks (ResNet BiT) at both smaller and larger computational scales.
- **Hybrid** architecture (CNN + Transformer) performs better than pure Transformer in smaller models and comparably as the model size increases.

Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

Hybrids: Transformer + CNN

Standard approach: CNN creates patches for ViT



Hybrids: Transformer + CNN

Regular ViT uses convolution 16*16 with stride=16. In comparison, CNN convolution 3*3 with stride=2 increases stability and accuracy.
Therefore, in some cases, it is beneficial to combine Transformer and CNN:

- CNN transforms image pixels into a feature map.
- The feature map is translated by the tokenizer into a sequence of tokens, which are then fed to the Transformer.
- The Transformer then applies attention techniques to create the output token sequence.
- Finally, the projector recombines the output tokens with the feature map, allowing key details at the pixel level to be found. This reduces the number of tokens to be checked, significantly lowering costs.

Demo

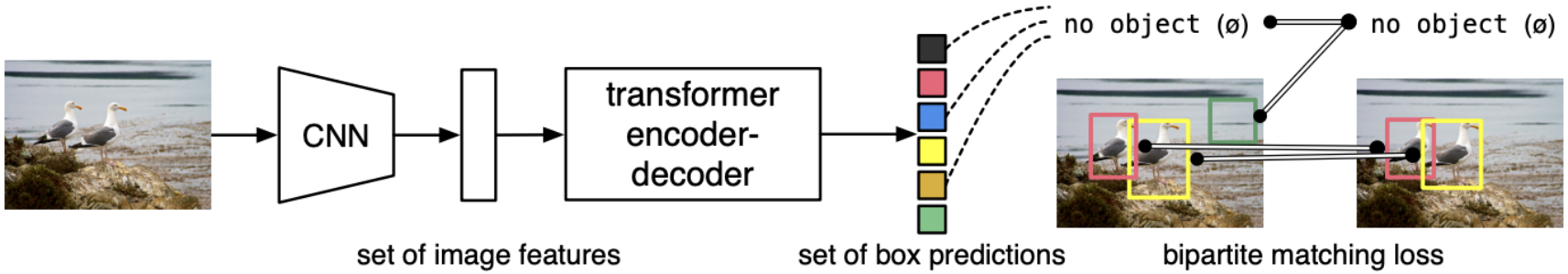
TO DO: Replace the standard ViT with a hybrid version (CNN + ViT) in

myViT_v3.ipynb

Initial Projection Conv2d converts input image into initial patches
- Input: [B, 3, 32, 32] $\rightarrow$ Output: [B, 64, 8, 8]
Try to another Conv2d layer with nn.BatchNorm2d, nn.ReLU(inplace=True) and nn.MaxPool2d to extract features from patches using CNN. The final projection layer: finalproj = nn.Conv2d(dimFromLastConv, embed_dim, kernel_size=1)

DETR

Detection **Tr**ansformer (DETR) is the first object detection architecture that successfully utilized the Transformer as the main component. Combines the efficiency advantages of previous detection methods (Faster R-CNN) with a much simpler architecture.

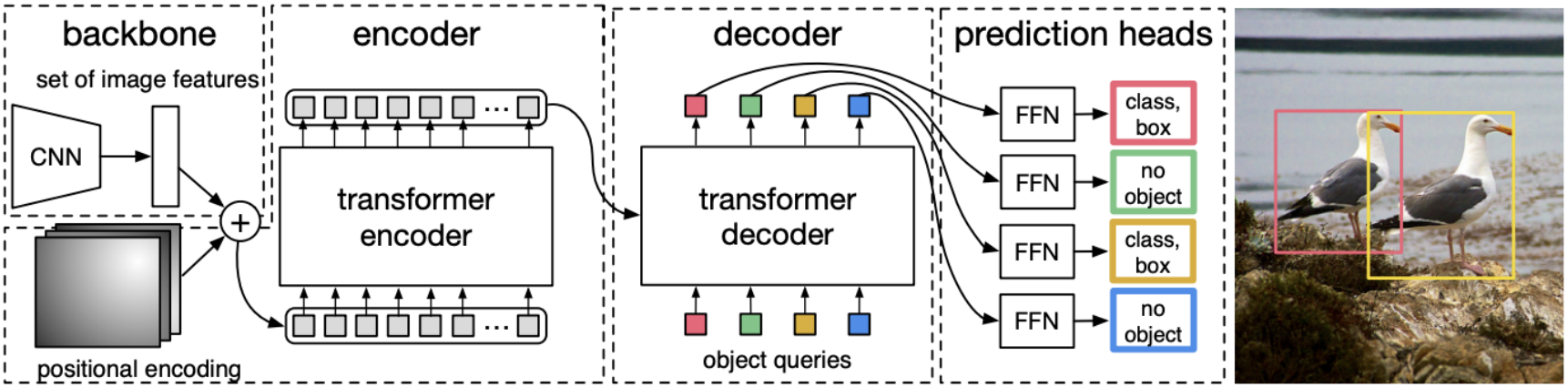


Carion N., Massa F., Synnaeve G., Usunier N., Kirillov A., Zagoruyko S. (2020). End-to-End Object Detection with Transformers.

DETR

Algorithm:

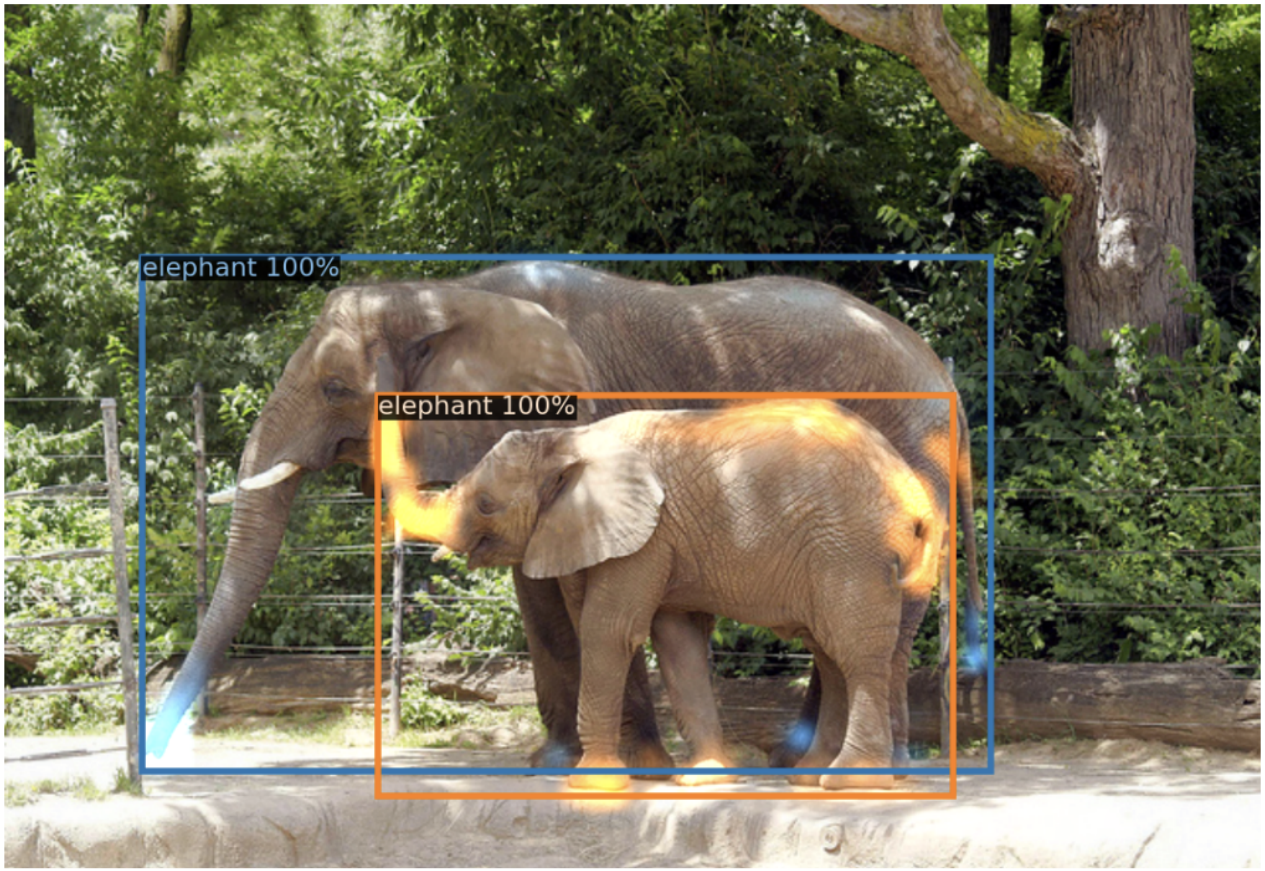
1. CNN is used to learn 2D image representation and extract features.
2. CNN output data is flattened and augmented with positional encodings, which are fed into the standard Transformer encoder.
3. The Transformer decoder passes output proposals (box embeddings) to the feed-forward network (FFN) for classification and bbox detection.

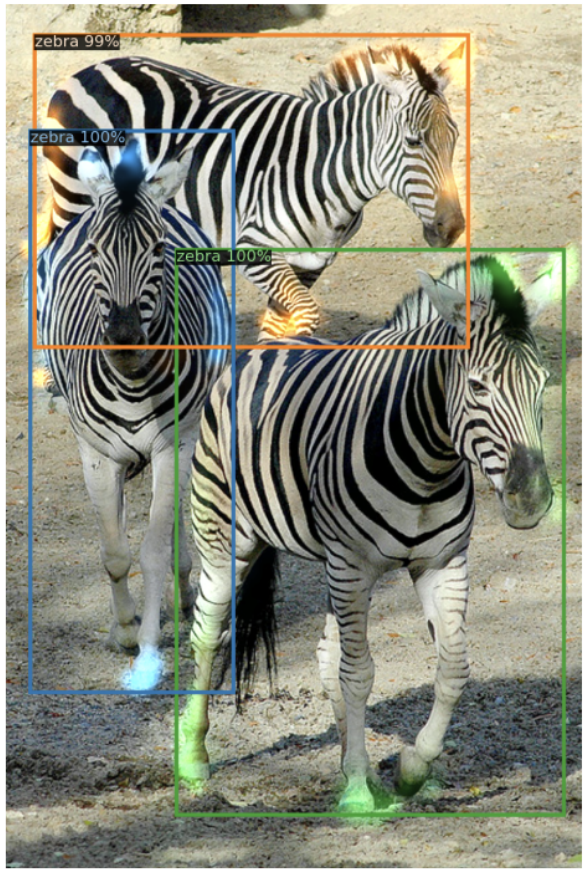


Carion N., Massa F., Synnaeve G., Usunier N., Kirillov A., Zagoruyko S. (2020). End-to-End Object Detection with Transformers.

DETR

Visualization of attention scores presented using different colors for different objects





Carion N., Massa F., Synnaeve G., Usunier N., Kirillov A., Zagoruyko S. (2020). End-to-End Object Detection with Transformers.

DETR

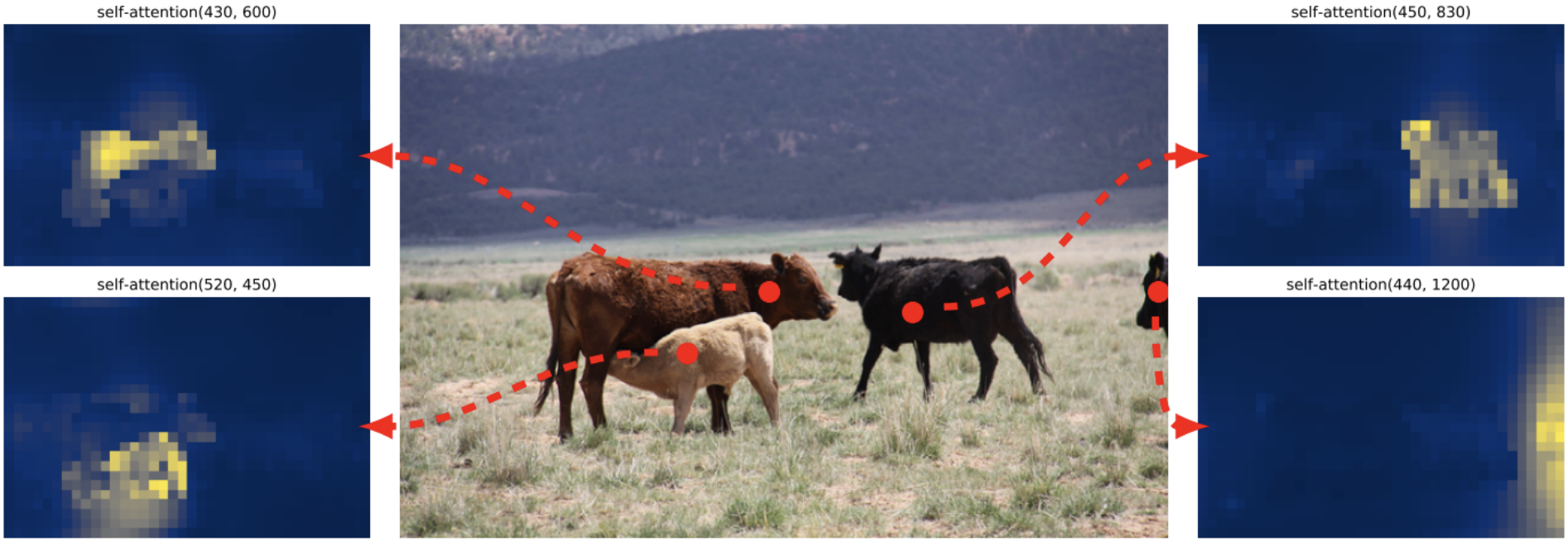
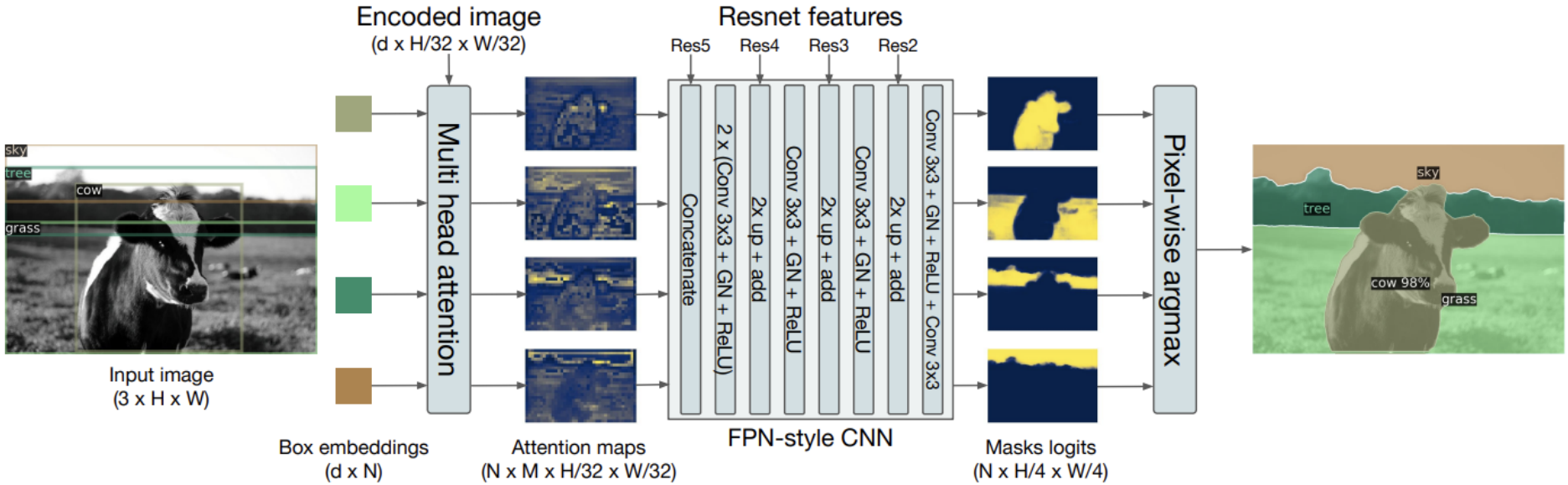


Fig. 3: Encoder self-attention for a set of reference points. The encoder is able to separate individual instances. Predictions are made with baseline DETR model on a validation set image.

Carion N., Massa F., Synnaeve G., Usunier N., Kirillov A., Zagoruyko S. (2020). End-to-End Object Detection with Transformers.

DETR

Segmentation with Attention maps



Carion N., Massa F., Synnaeve G., Usunier N., Kirillov A., Zagoruyko S. (2020). End-to-End Object Detection with Transformers.

The end

Thank you for your **Attention**!!