DEEP NETWORK DEVELOPMENT

Imre Molnár
PhD student, ELTE, AI Department
imremolnar@inf.elte.hu
curiouspercibal.github.io

Tamás Takács
PhD student, ELTE, AI Department
tamastheactual@inf.elte.hu
tamastheactual.github.io

Deep Network Development

Lecture 10.

Vision Transformers

Budapest, 25th April 2025

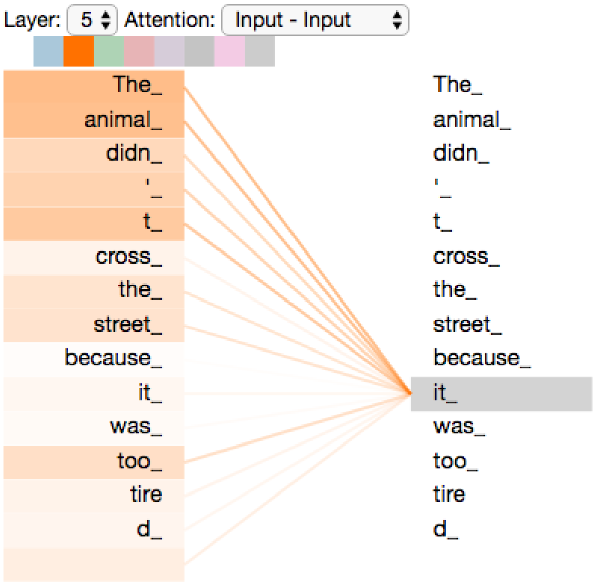
1 Transformer Network

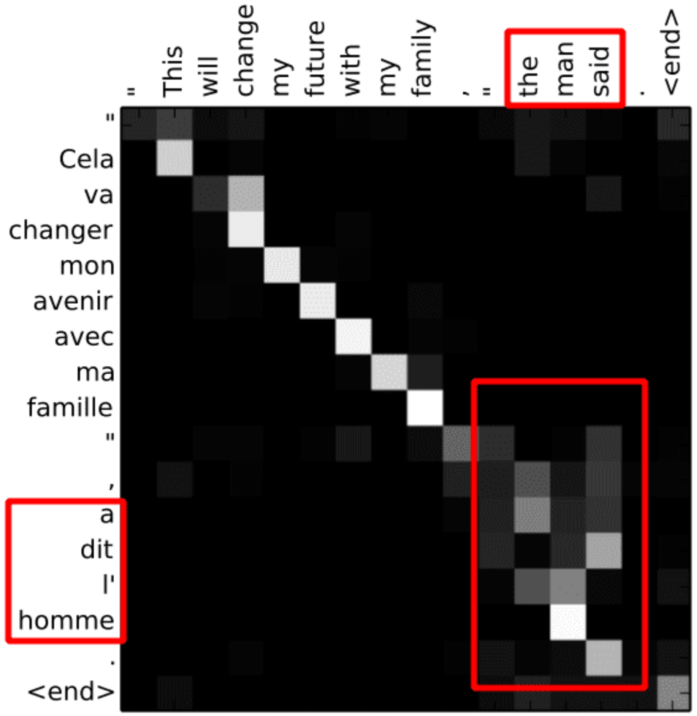
2 Vision Transformers

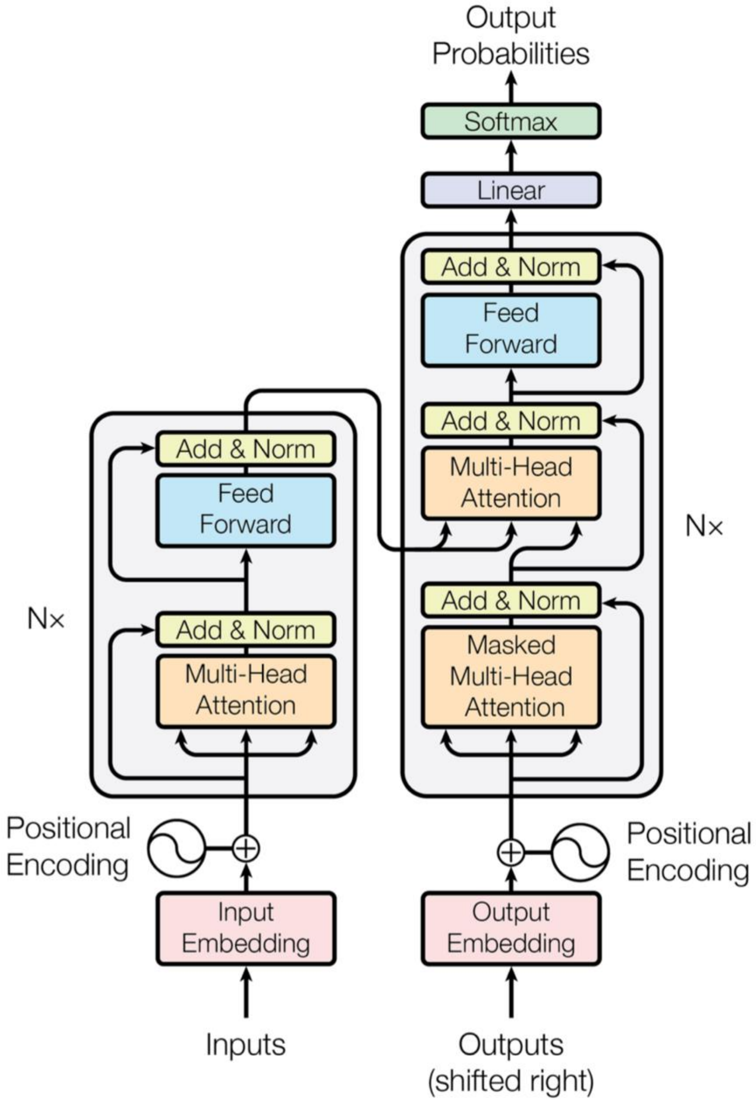
3 State of the Art

Previously on Lecture 9

- Attention, Self-Attention, Multi-Head Attention
- Transformers







Deep Network Development

Lecture 10.

Vision Transformers

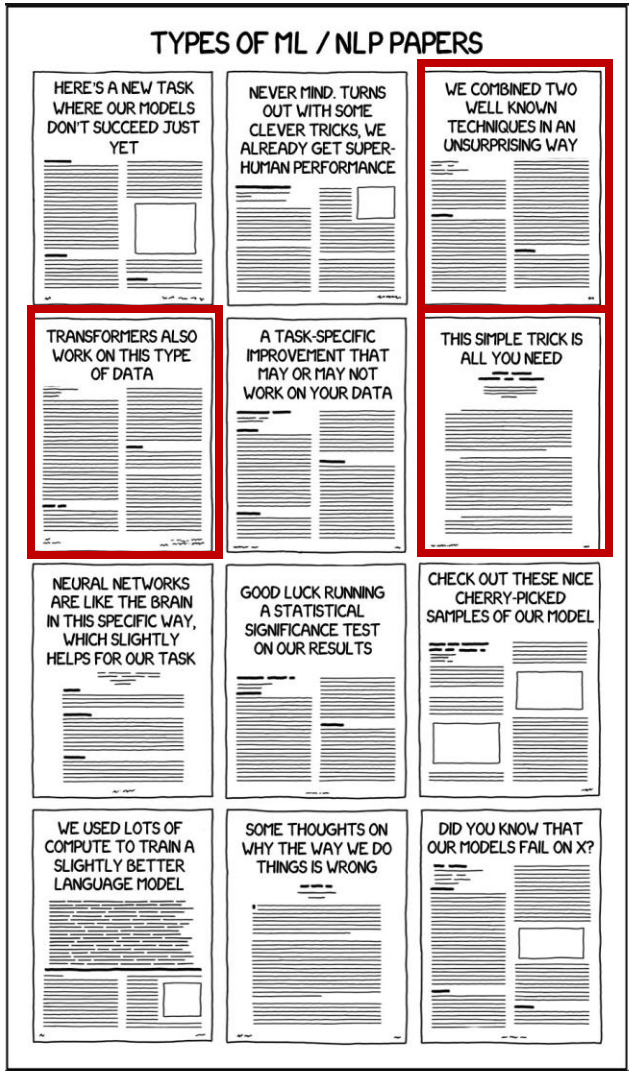
Budapest, 25th April 2025

1 Transformer Network

2 Vision Transformers

3 State of the Art

Transformer Networks



Generative Pre-trained Transformer (GPT) – June 2018

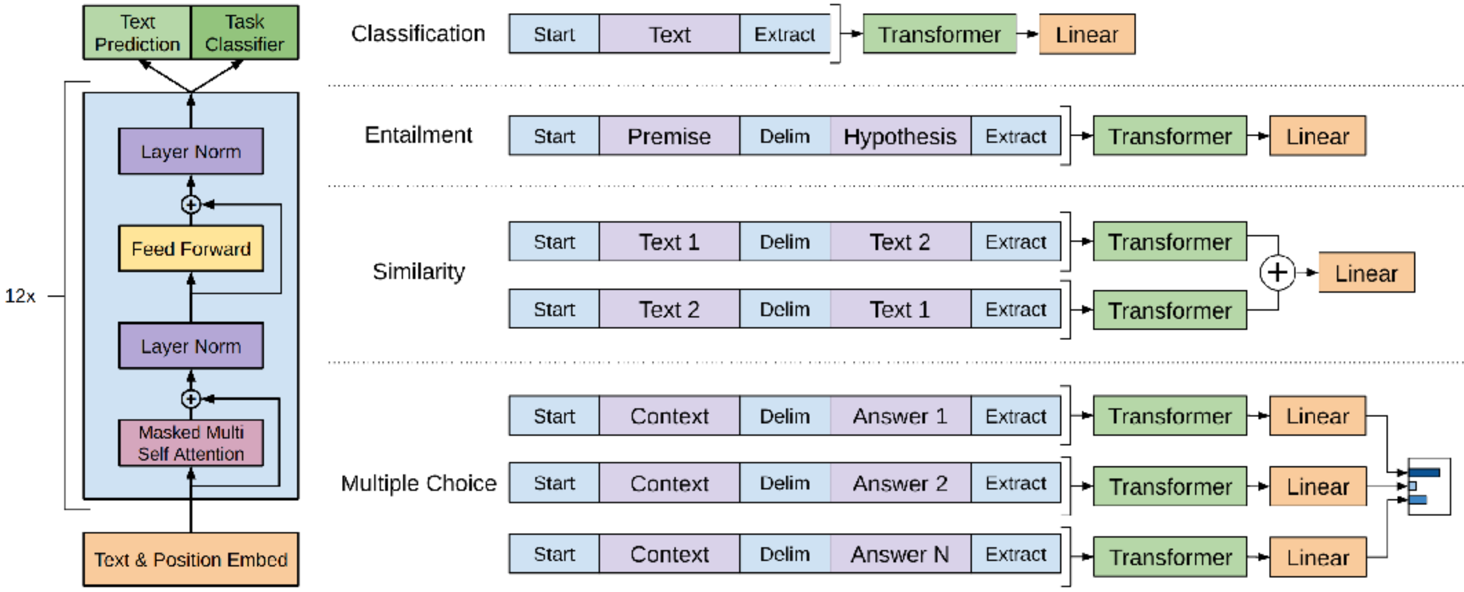
The original goal is Language Modelling (LM)

Uses Masked Self-Attention to limit the attention to the previous tokens only (left-to-right)

Two stage training:

1. Unsupervised pre-training:
   - The goal is to predict the next token based on the previous tokens.

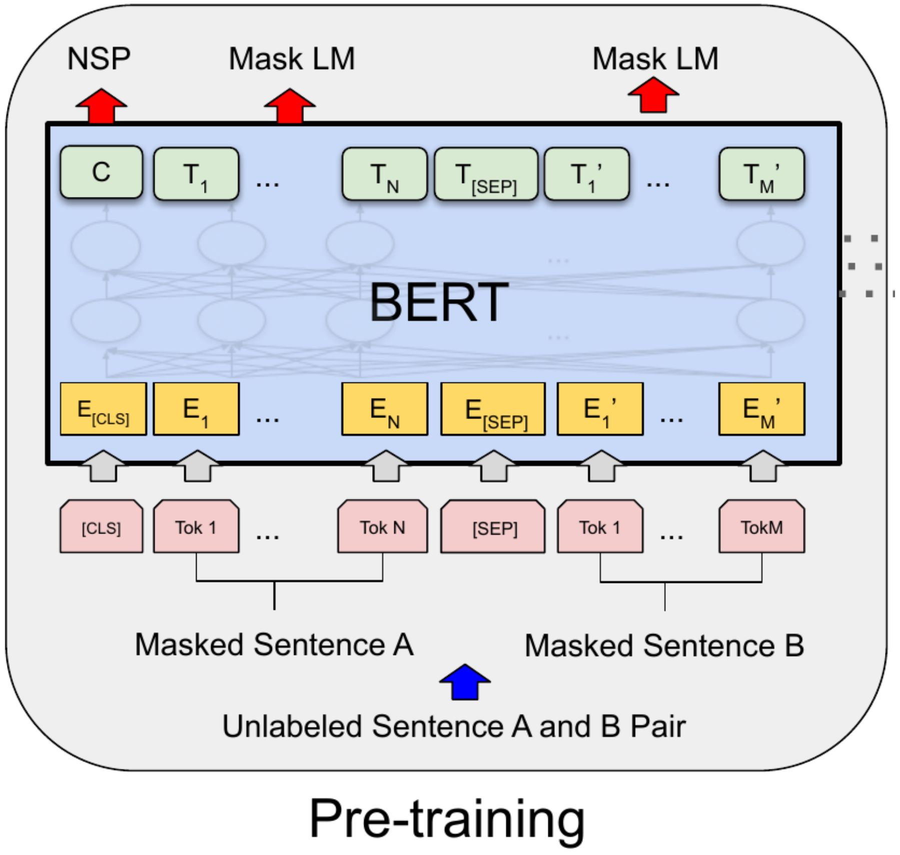
2. Supervised fine-tuning:
   - Predict the label (y) based on the input tokens

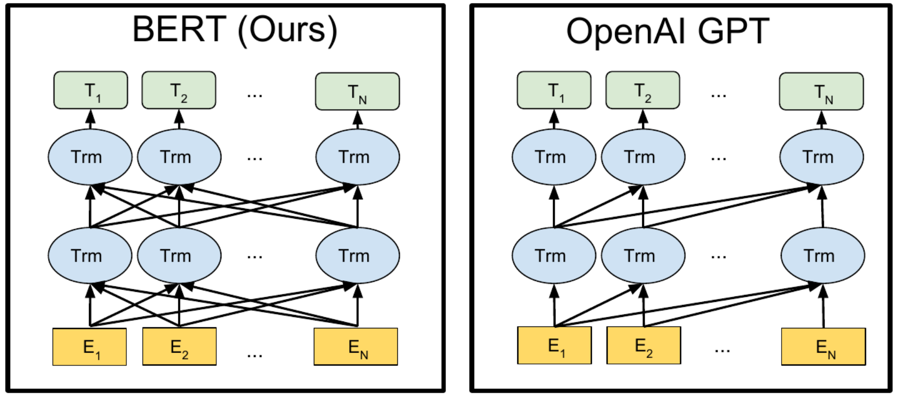


Bidirectional Encoder Representations from Transformers (BERT) – May 2019

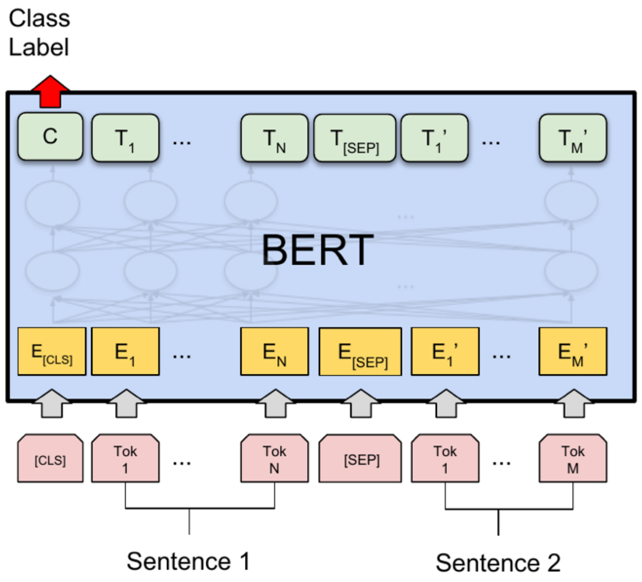
- Compared to OpenAI GPT it uses a bidirectional self-attention
- Trained on 2 tasks at the same time during pre-training
  1. Masked LM (15% of the tokens are masked)
  2. Next Sentence Prediction

- A special [CLS] token is introduced at the beginning of each sequence.

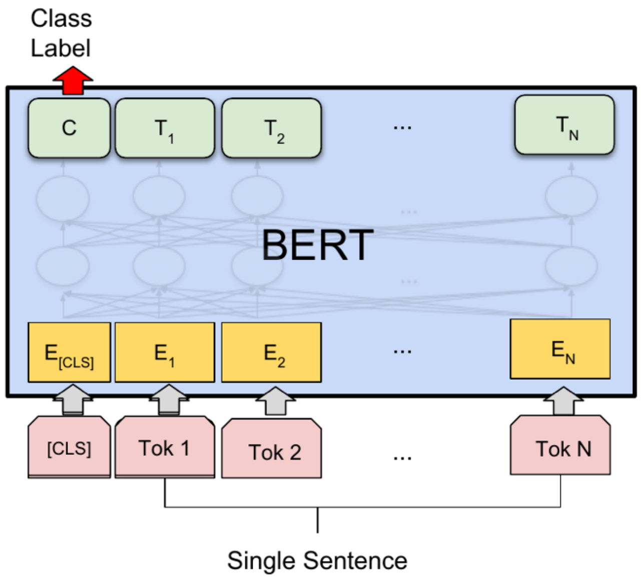




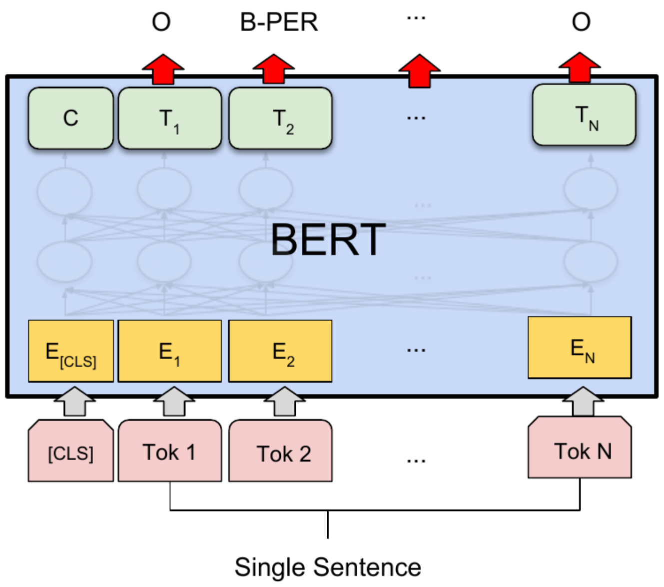
Bidirectional Encoder Representations from Transformers (BERT) – May 2019



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Deep Network Development

Lecture 10.

Vision Transformers

Budapest, 25th April 2025

Introduction

- Jun 2021 – *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* - *arxiv.org/abs/2010.11929*

- 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.
- 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

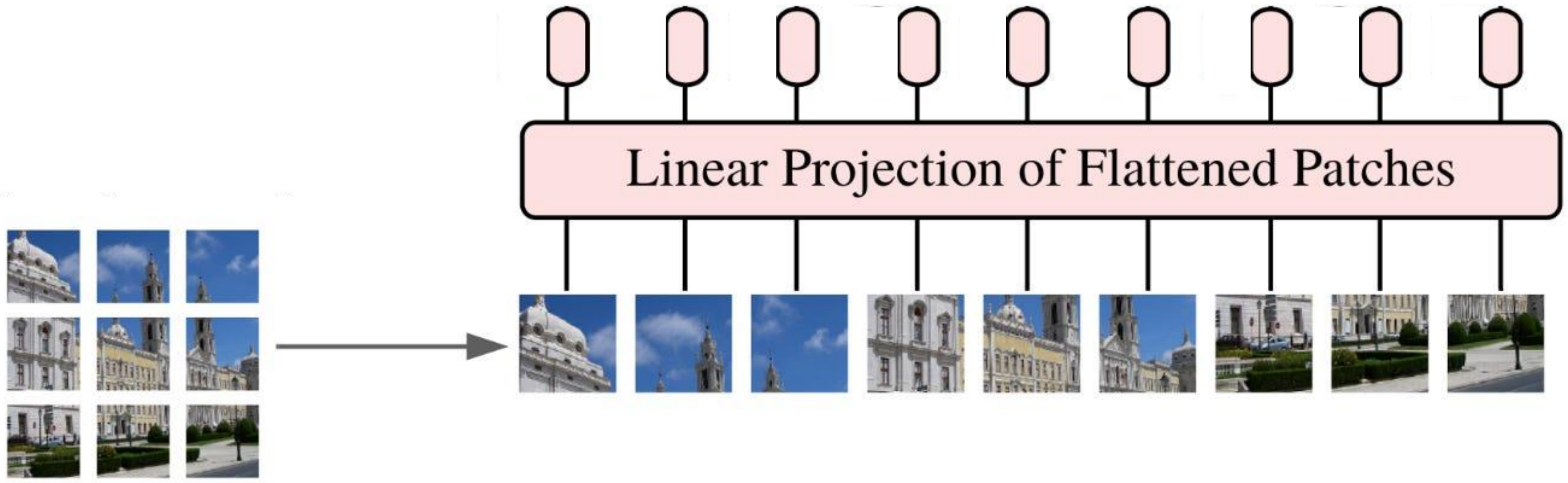
- Transformers operate on a sequence of tokens
- How do we transform an image into tokens?

Image Tokenization

1. Reshape images of $x \in \mathbb{R}^{\mathrm{H} \times W \times \mathrm{C}}$ into $N = \frac{HW}{P^2}$ patches
   - (H, W) – image resolution
   - C – number of channels
   - (P, P) – patch resolution

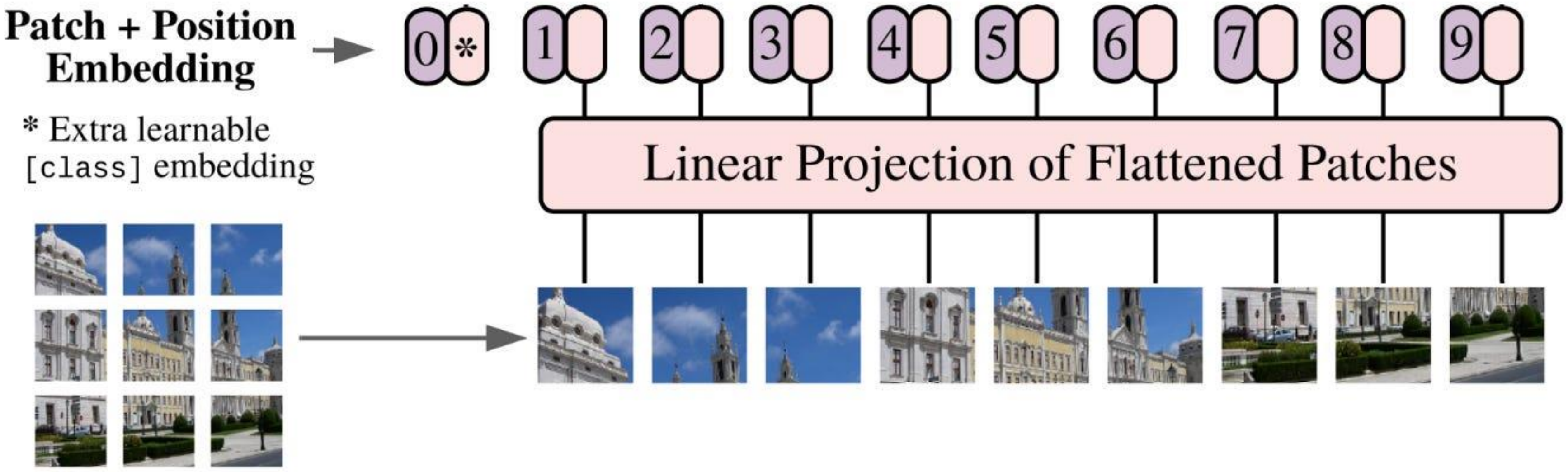
Image Tokenization

2. The Transformer uses constant latent vector size D through all of its layers
   - Flatten all the patches and apply a learnable linear projection (*patch embeddings*)



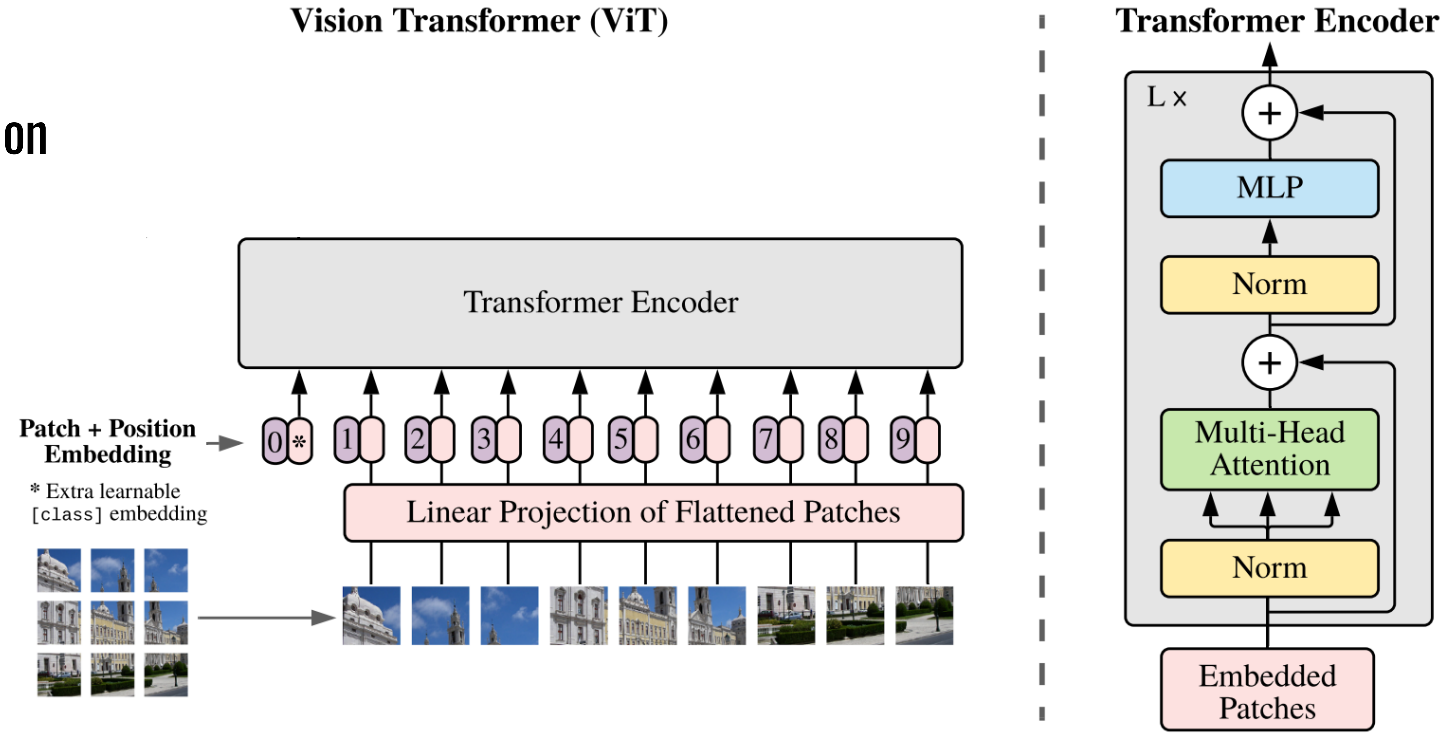
Token Processing

3. Processing the tokens
    - Prepend a learnable embedding to the patch embeddings
    - Apply position embedding



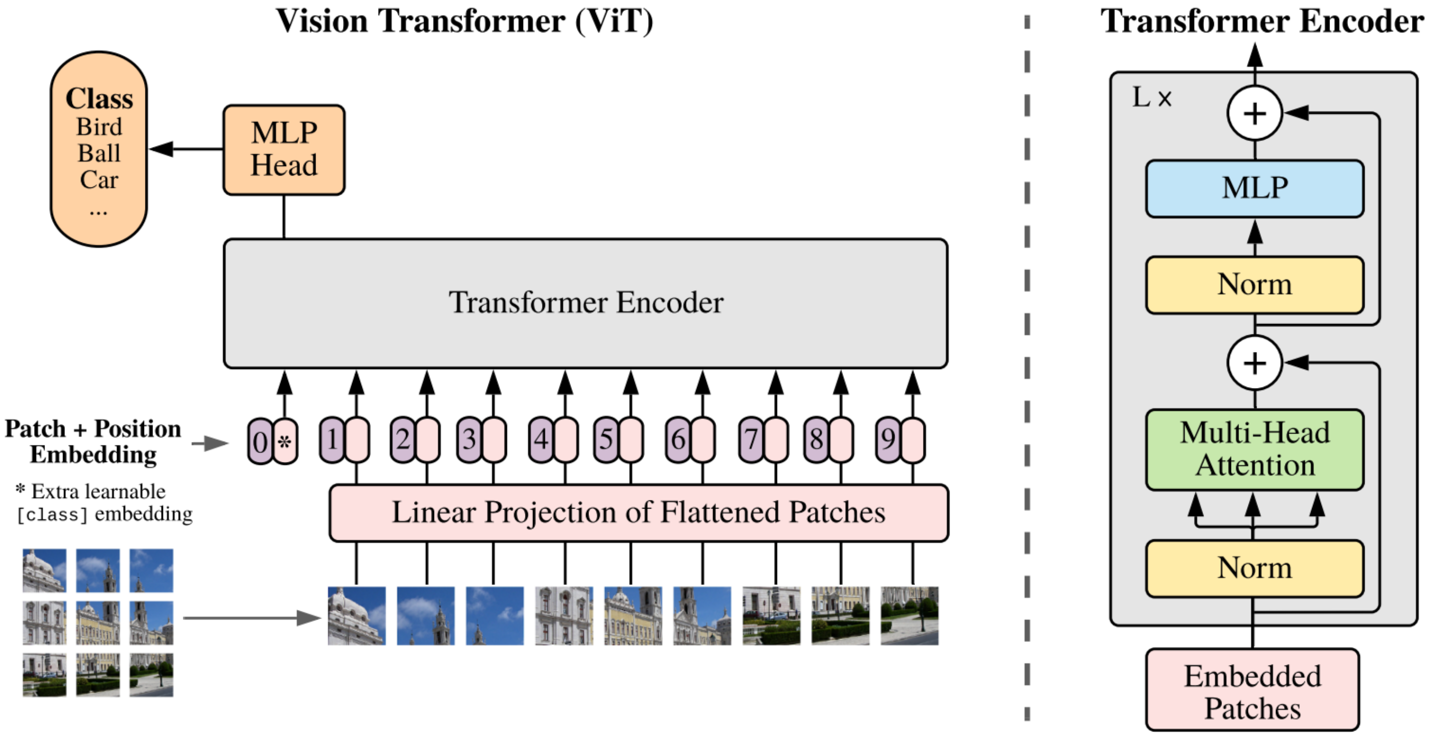
Encoder Block

4. Feeding the embedded patches to the Transformer Encoder
   - Input the sequence of embedded patches ($[\boldsymbol{z}_0^0, z_0^1, \ldots, z_0^N]$)
   - At the end we get the image representation ($[\boldsymbol{z}_L^0, z_L^1, \ldots, z_L^N]$)



Prediction Processing

5. Classification MLP head
   - Attaching a classifier head to z_L^0



Overall

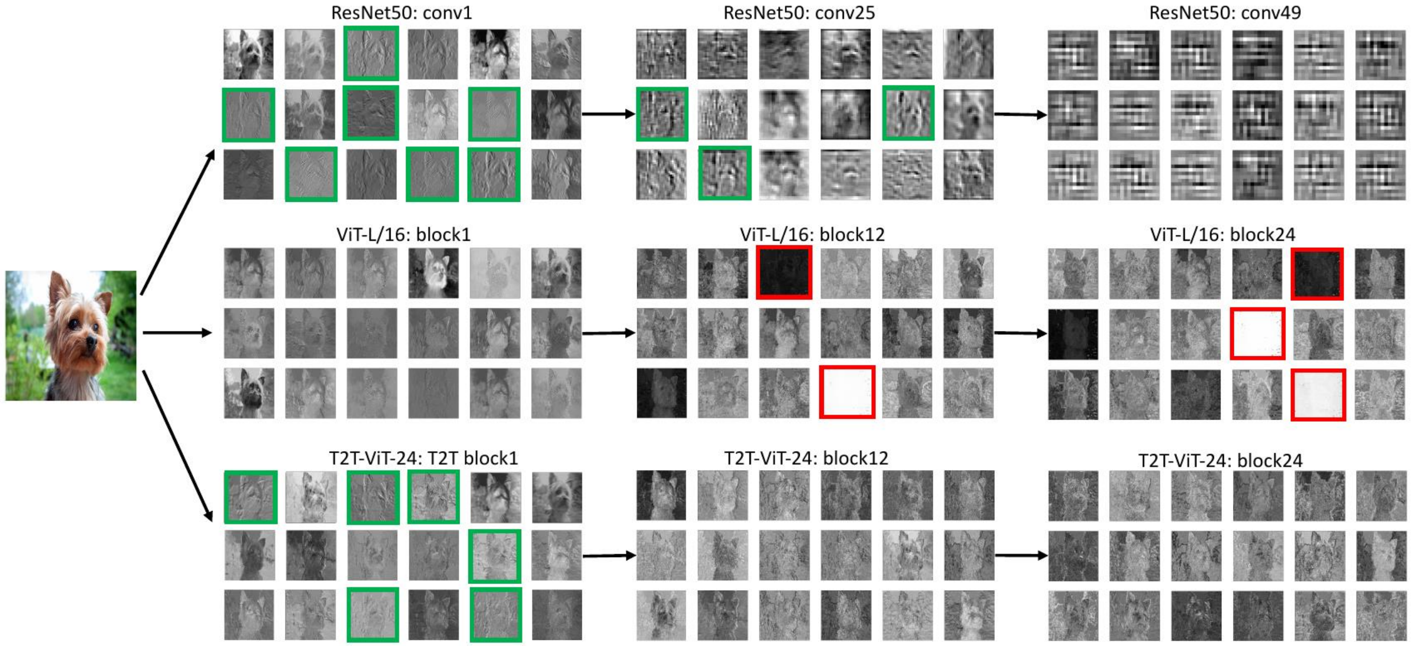
Comparison with Convolutional Networks

- ViT has much less image-specific inductive bias
  - features like **edges**, **textures**, and **patterns** are spatially localized and translationally invariant
- In CNNs, locality, two-dimensional neighbourhood structure, and translation equivariance are baked into the whole model
- Position embedding does not carry information about the 2D position of the patches

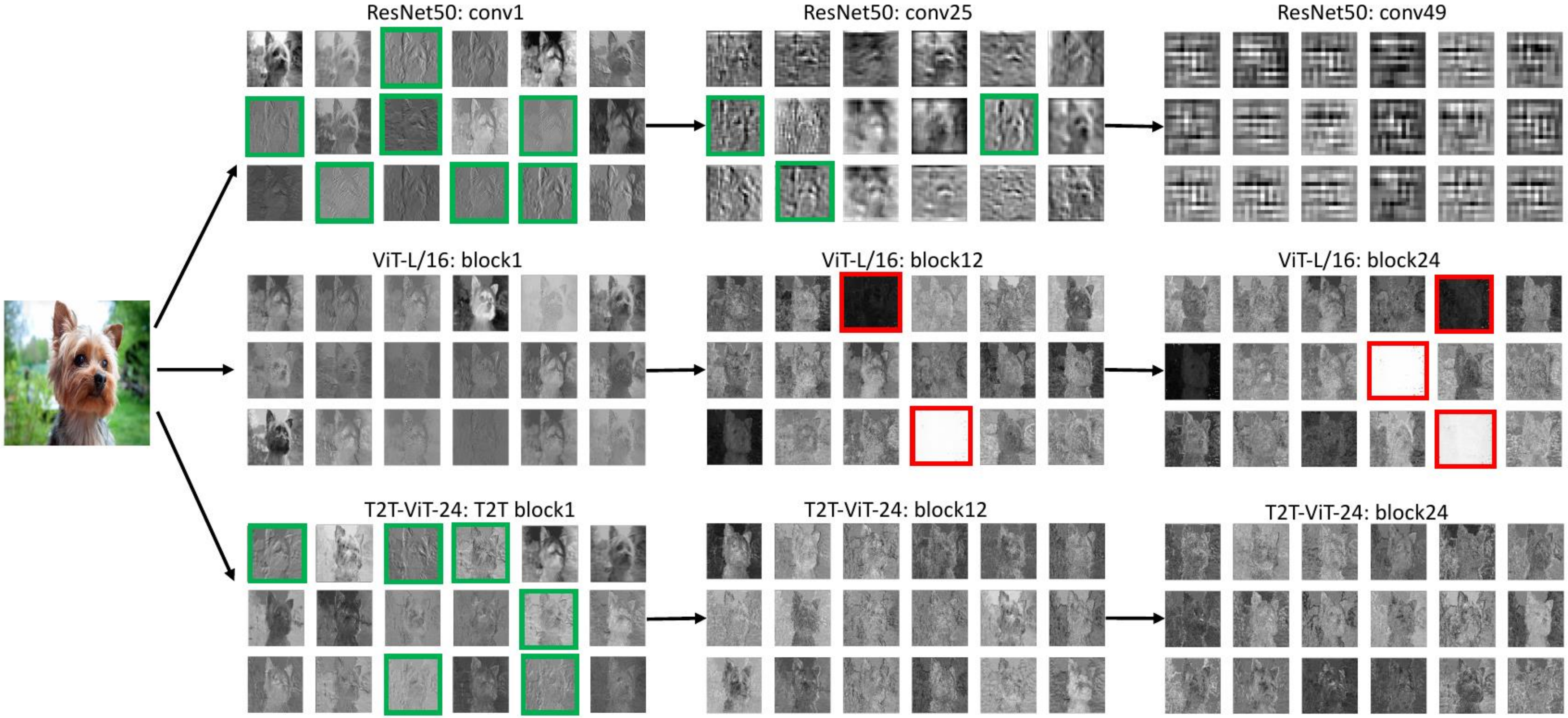
Token-to-Token ViT (T2T-ViT) – Nov 2021

- Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet - arxiv.org/abs/2101.11986

- ViT achieves inferior performance to CNNs when trained on a midsize dataset
  1. The tokenization fails to model the important local structure such as edges, lines, etc.
  2. The redundant attention backbone leads to limited feature richness



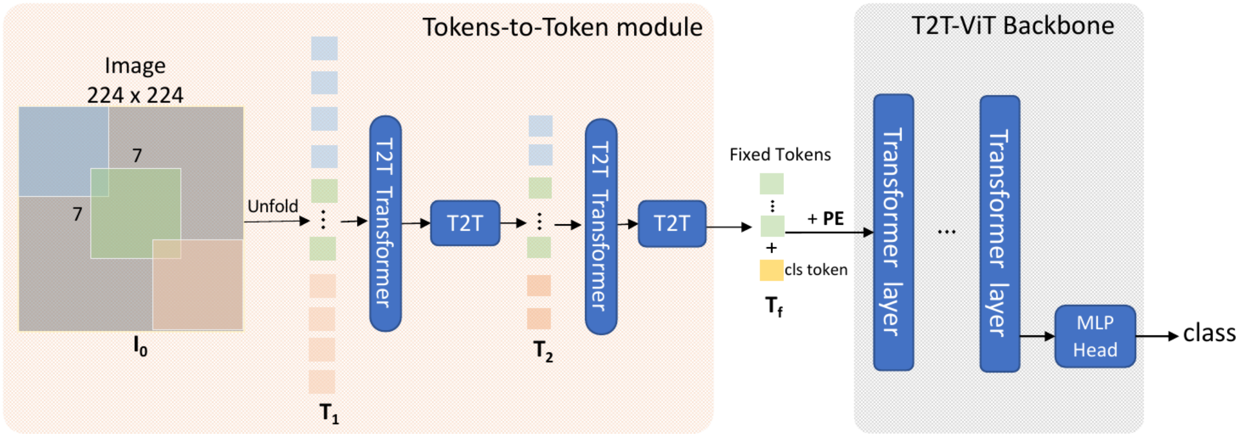
Token-to-Token ViT (T2T-ViT)

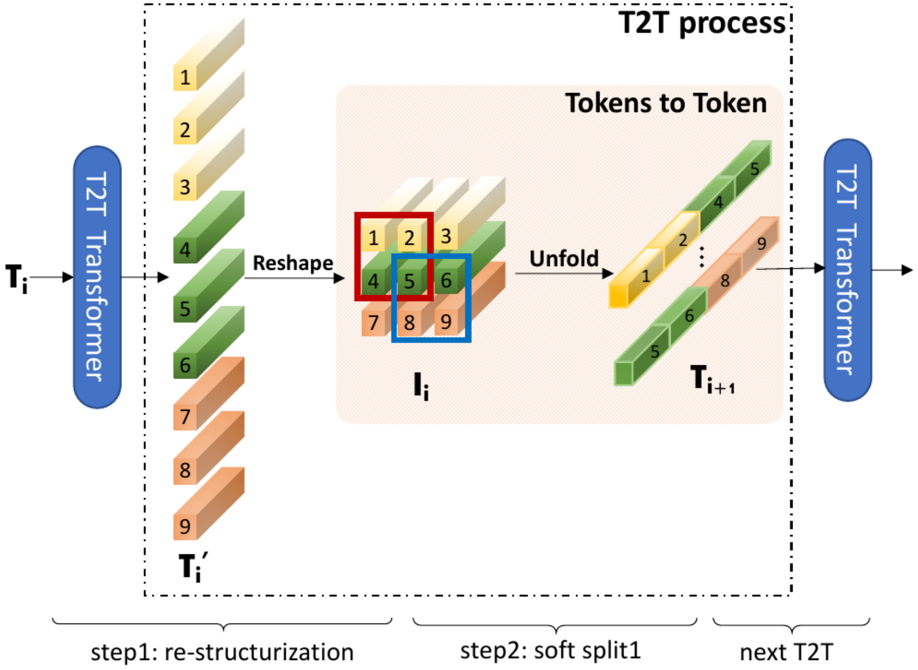


Token-to-Token ViT (T2T-ViT)

1. A layer-wise "Token-to-token module"
2. An efficient "T2T-ViT backbone"

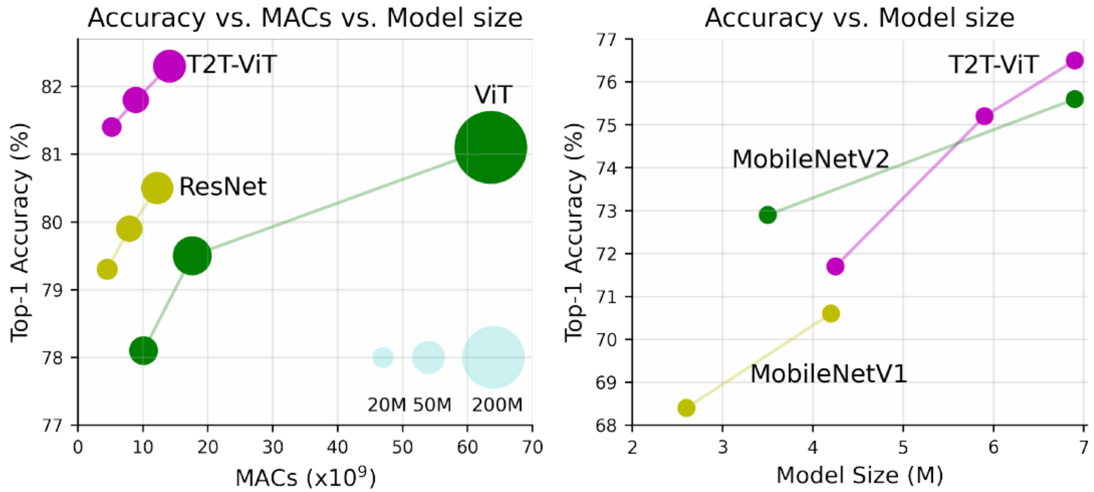
- The generated tokens are reordered like an "image"
- Then areas closer together are grouped together into a new token





Token-to-Token ViT (T2T-ViT)

- While “vanilla” ViT requires a large dataset and more tuneable parameters to beat the “state-of-the-art” (**JFT-300M**) CNN models
- T2T-ViT requires smaller datasets and less tuneable parameter



Deep Network Development

Lecture 10.

Vision Transformers

Budapest, 25th April 2025

1 Transformer Network

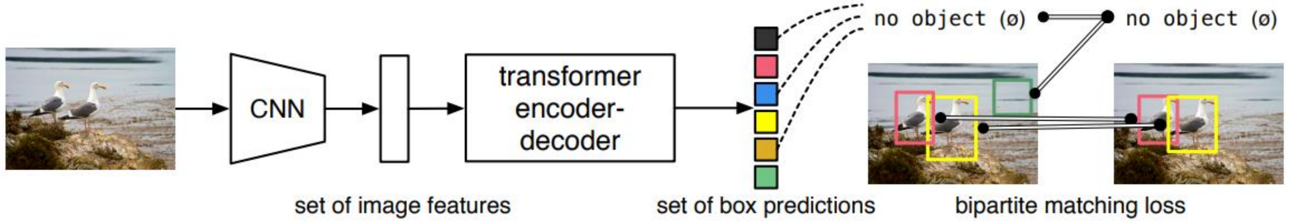
2 Vision Transformers

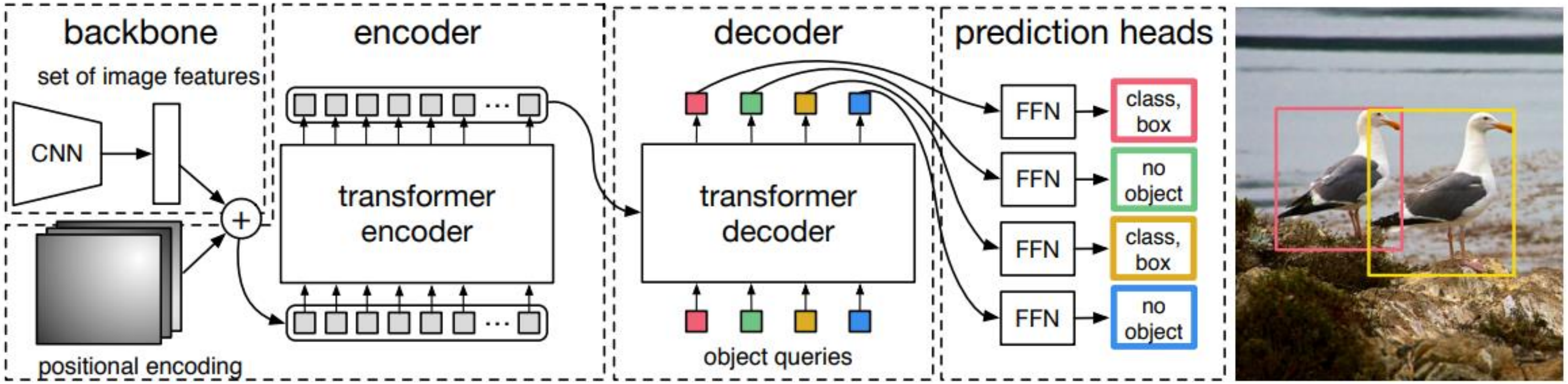
3 State of the Art

DEtection TRansformer (DETR) – May 2020

End-to-End Object Detection with Transformers - arxiv.org/abs/2005.12872

Simple architecture:
1. CNN backbone
2. Encoder-Decoder Transformer
3. Feed Forward Network





Segment Anything Model (SAM) – Apr 2023

- MAE pre-trained ViT-H/16 as an image encoder
- The mask decoder is a modified transformer
- Prompt encoder from CLIP

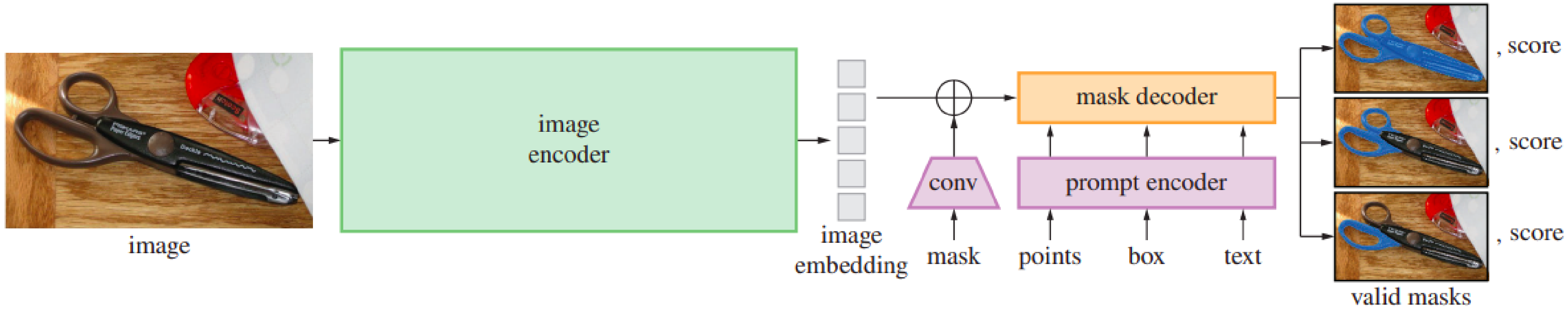


Figure 4: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output multiple valid masks and associated confidence scores.

Masked Autoencoders (MAE) – Dec 2021

The task is to reconstruct the signal given its partial observation

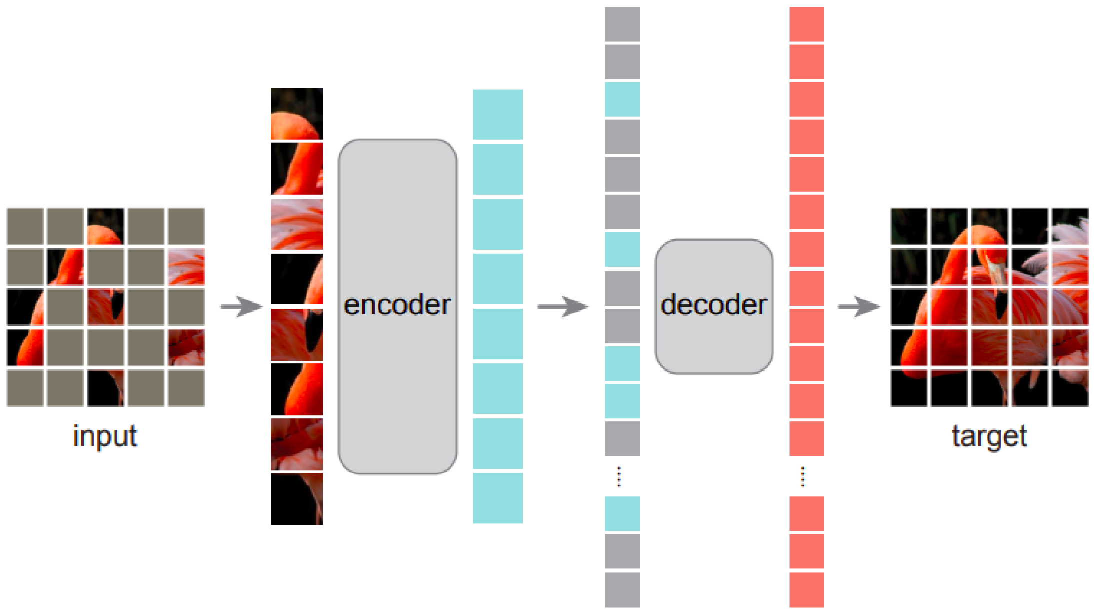
- High masking (75% of the image is masked) eliminates redundancy
- The reconstruction is much harder since the missing part cannot be reconstructed by extrapolation (like in image inpainting)

ViT encoder:

- Only operates on the visible parts (no <MASK> tokens)

MAE decoder:

- Input: Encoded visible tokens and mask tokens
- Each mask token is a shared learned vector
- Positional encoding



And many more

- SAM-2,
- GPT-2, GPT-3, GPT-4
- DALL-E, ViT-VQGAN,
- SORA
- Oasis

Summary

GPT:
- Left-to-right approach,
- Language Modelling and next token prediction

BERT:
- Bidirectional Multihead Self Attention and masking
- Same architecture for all to language modelling tasks

Vision Transformers:
- Image – Patch – Linear Projection – Token
- Fails to capture local structures such as edges, texture and patterns
- Positional embedding does not provide information about locality

Token-to-Token ViT:
- Token reorganization to counter missing locality

DETR:
- Convolutional Feature extraction – Transformer
- Detects N objects at the same time

Segment Anything Model (SAM):
- Segmentation based on user input

Masked Autoencoders (MAE):
- Asymmetric design
- Unique challenge

Resources

Books:

- Courville, Goodfellow, Bengio: Deep Learning
  Freely available: https://www.deeplearningbook.org/
- Zhang, Aston and Lipton, Zachary C. and Li, Mu and Smola, Alexander J.: Dive into Deep Learning
  Freely available: https://d2l.ai/

Courses:

- Deep Learning specialization by Andrew NG
- https://www.coursera.org/specializations/deep-learning

Further Links + Resources

- Attention Is All You Need - *arxiv.org/abs/1706.03762*
- Improving Language Understanding by Generative Pre-Training - *openai.com/index/language-unsupervised/*
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding - *arxiv.org/abs/1810.04805*
- An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale - *arxiv.org/abs/2010.11929*
- Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet - *arxiv.org/abs/2101.11986*
- *medium.com/autonomous-agents/convnets-vs-vision-transformers-mathematical-deep-dive-c7908220e7b3*
- *medium.com/towards-data-science/vision-transformers-explained-a9d07147e4c8*
- End-to-End Object Detection with Transformers - *arxiv.org/abs/2005.12872*
- Segment Anything – *arxiv.org/abs/2304.02643*
- Masked Autoencoders Are Scalable Vision Learners - *arxiv.org/abs/2111.06377*

That’s all for today!