embedded VISION summit

How Transformers are Changing the Direction of Deep Learning Architectures

Tom Michiels
System Architect
Synopsys

Outline

- The Surprising Rise of Transformers in Vision
- The Structure of Attention and Transformer
- Transformers applied to Vision and Other Application Domains
- Why Transformers are Here to Stay for Vision

CNNs Have Dominated Many Vision Tasks Since 2012

AlexNet VGG ResNet MobileNet V1 MobileNet V2 EfficientNet MobileNet v3

2012 ... 2014 2015 2016 2017 2018 2019 2020 2021 2022

Image Classification

CNNs Have Dominated Many Vision Tasks Since 2012

RCNN
FRCNN
SSD
YOLOV2
Mask RCNN
YoloV3
YoloV4/V5
EfficientDet

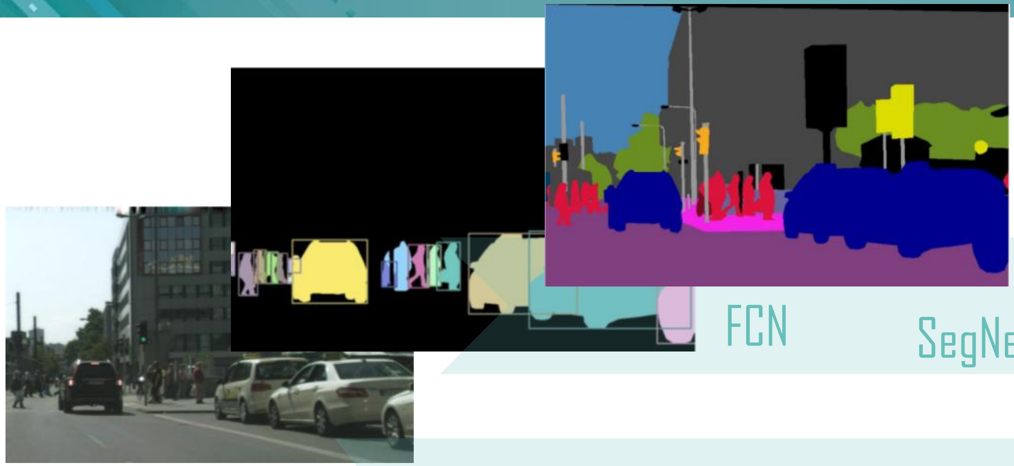
AlexNet
VGG
ResNet
MobileNet V1
MobileNet V2
EfficientNet
MobileNet v3

2012 ... 2014 2015 2016 2017 2018 2019 2020 2021 2022

Object Detection

CNNs Have Dominated Many Vision Tasks Since 2012

embedded VISION summit



FCN SegNet DeepLab DSNet

RCNN FRCNN SSD YOLOV2 Mask RCNN YoloV3 YoloV4/V5 EfficientDet

AlexNet VGG ResNet MobileNet V1 MobileNet V2 EfficientNet MobileNet v3

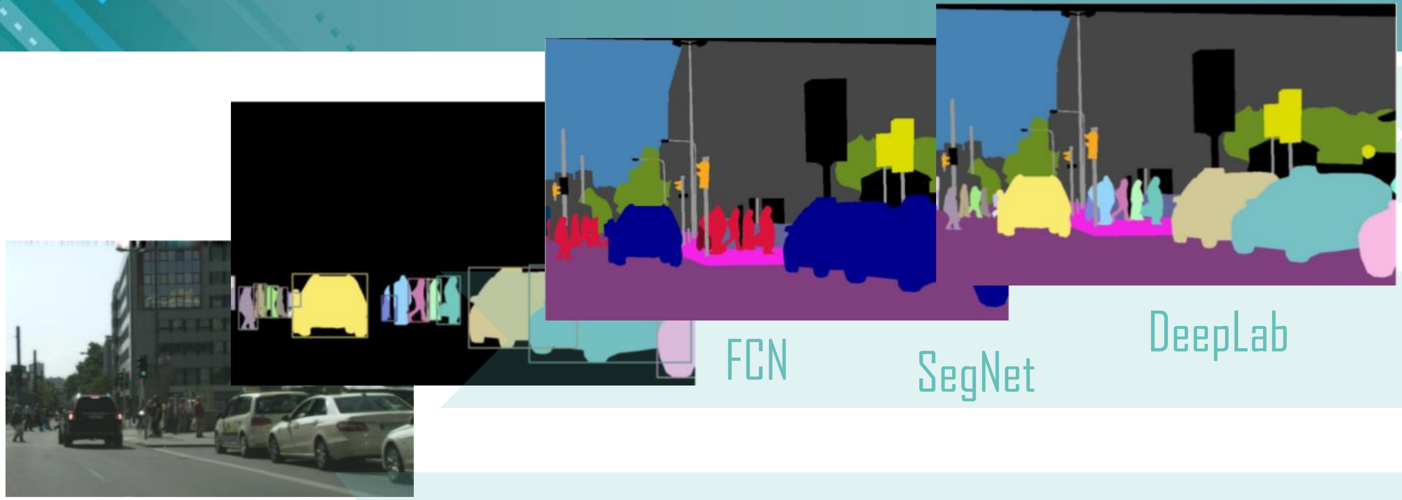
2012 ... 2014 2015 2016 2017 2018 2019 2020 2021 2022

Semantic Segmentation

SYNOPSYS®

CNNs Have Dominated Many Vision Tasks Since 2012

embedded VISION summit



DeeperLab PFPN EfficientPS YoloP

FCN SegNet DeepLab DSNet

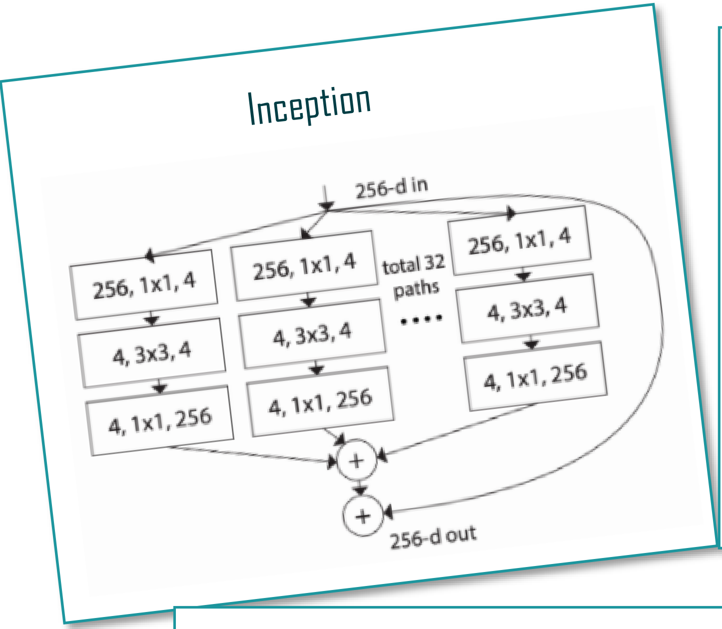
RCNN FRCNN SSD YOLOV2 Mask RCNN YoloV3 YoloV4/V5 EfficientDet

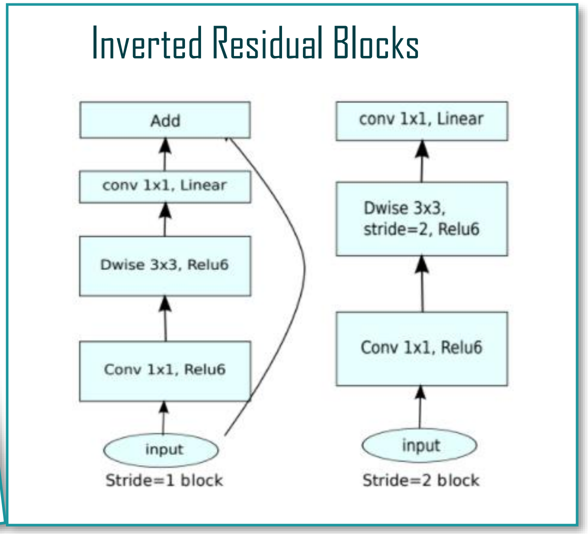
AlexNet VGG ResNet MobileNet V1 MobileNet V2 EfficientNet MobileNet v3

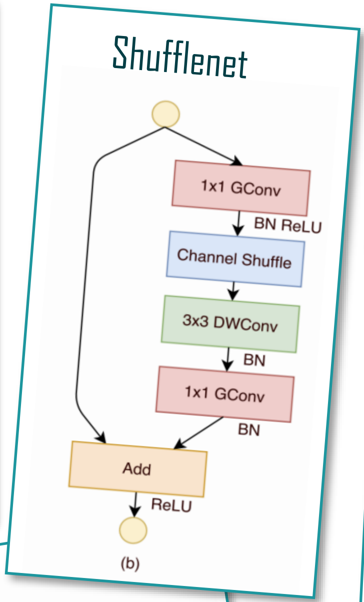
2012 … 2014 2015 2016 2017 2018 2019 2020 2021 2022

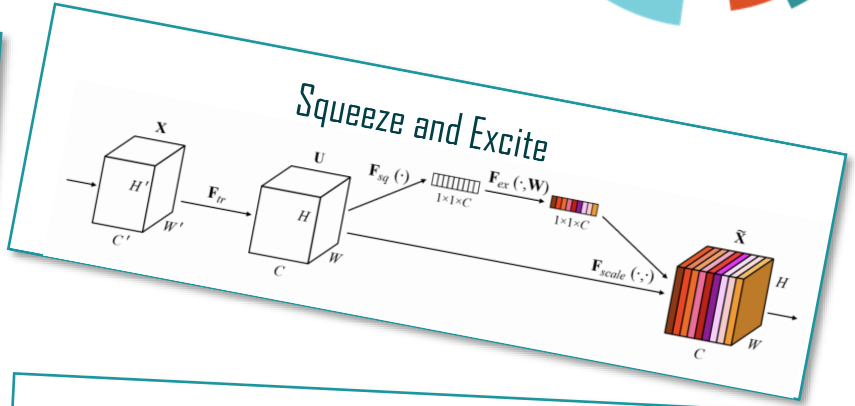
Panoptic Vision

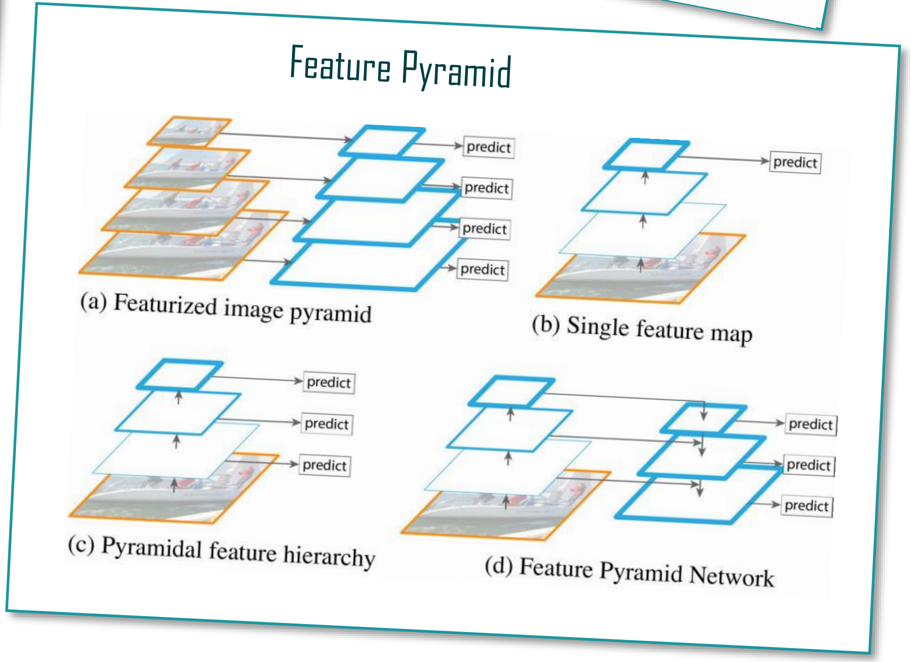
A Decade of CNN Development…

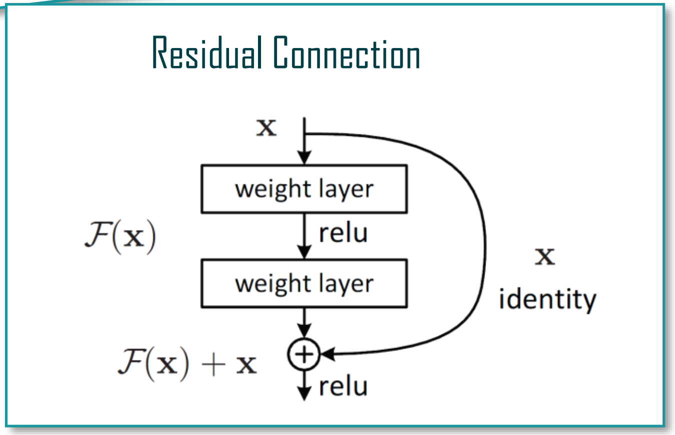


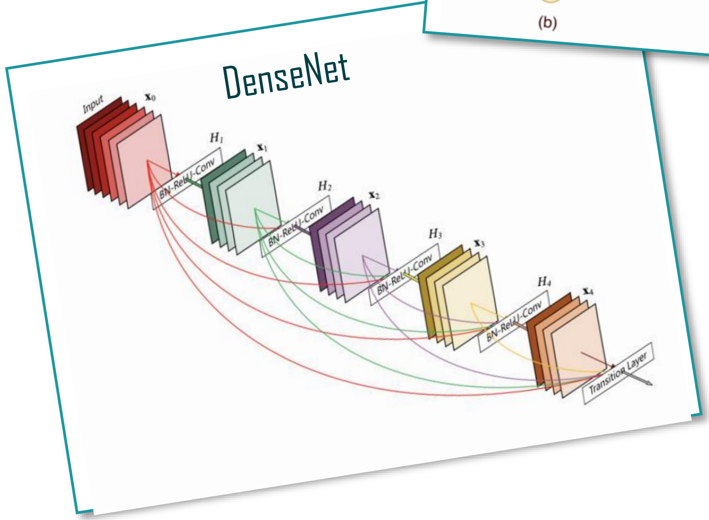






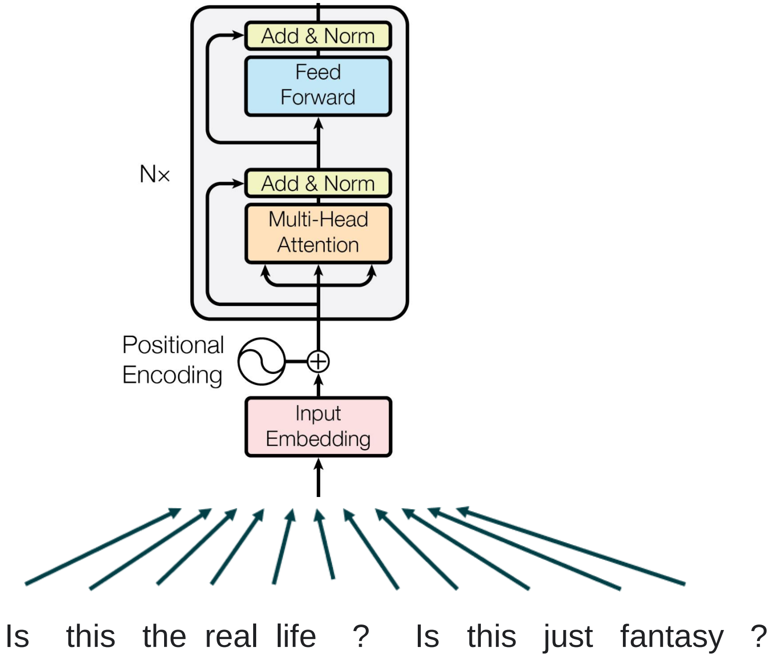




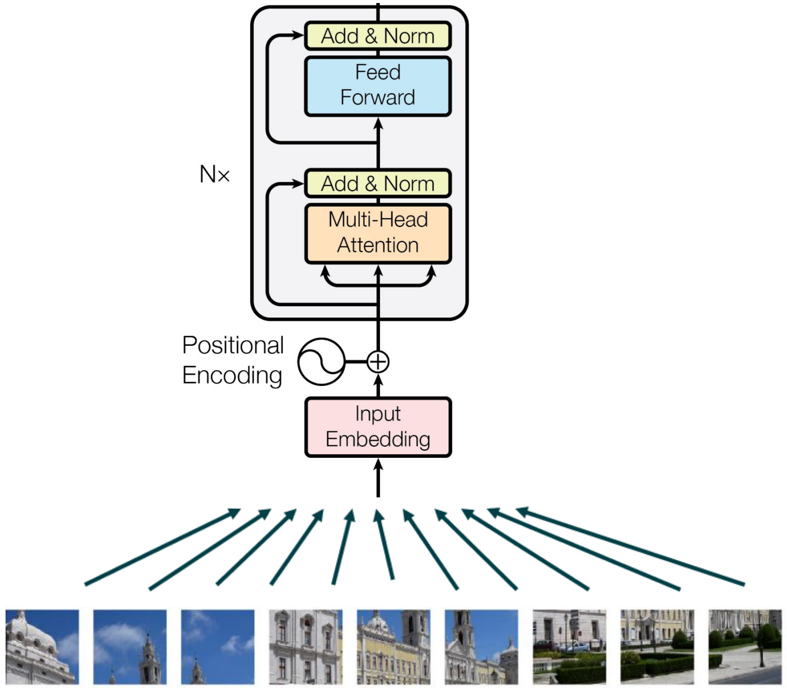


Beaten in Accuracy by Transformers

Transformer, a model designed for natural language processing

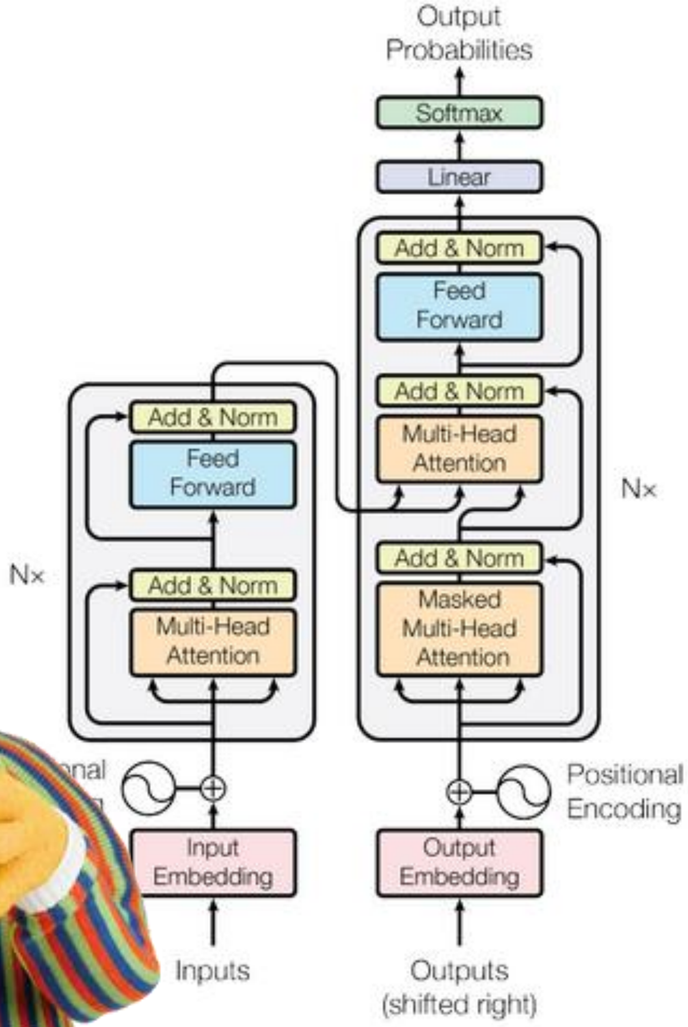


… without any modifications applied to image patches, beats the highly specialized CNNs in accuracy



The Structure of Attention and Transformer

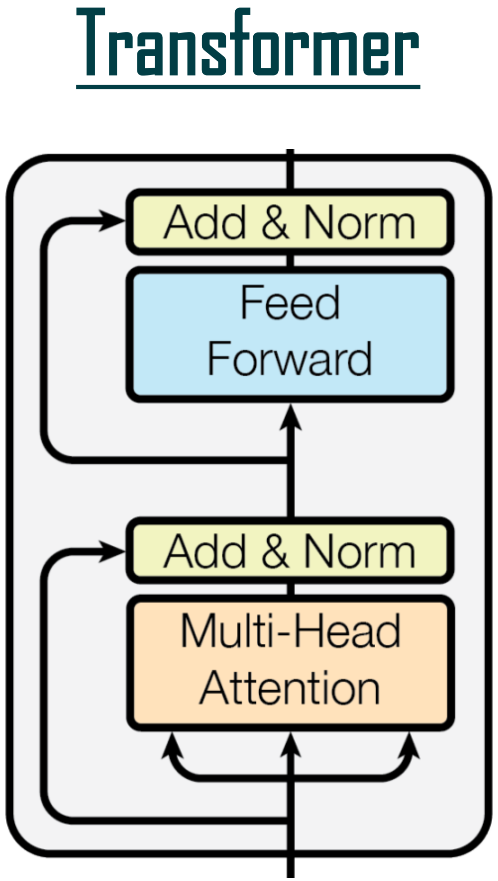
Bert and Transformers

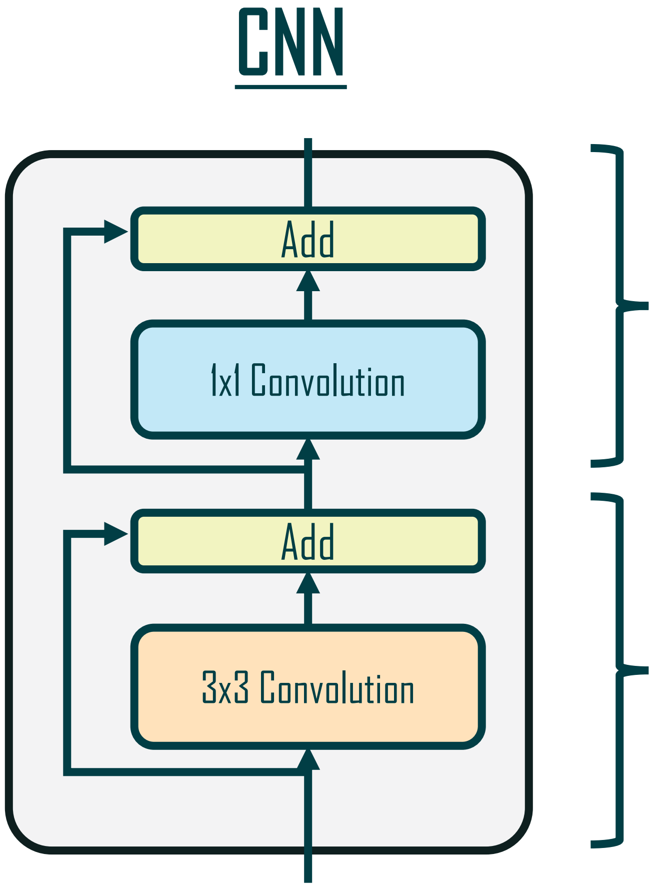


- Attention is all you need!(*)
- **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
- A Transformer is a deep learning model that uses attention mechanism
- Transformers were primarily used for Natural Language Processing
  - Translation
  - Question Answering
  - Conversational AI
- Successful training of huge transformers
  - MTM, GPT-3, T5, ALBERT, RoBERTa, T5, Switch
- Transformers are successfully applied in other application domains with promising results for embedded use

(*) https://arxiv.org/abs/1706.03762

Convolutions, Feed Forward, and Multi-Head Attention





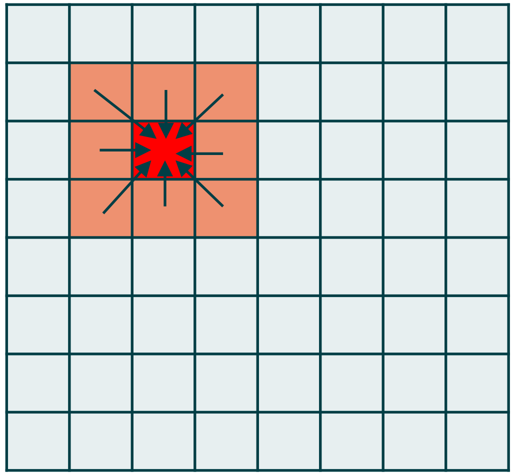
- The Feed Forward layer of the Transformer is identical to a 1x1 Convolution
- In this part of the model, no information is flowing between tokens/pixels

- Multi-Head Attention and 3x3 Convolution layers are the layers responsible for mixing information between tokens/pixels

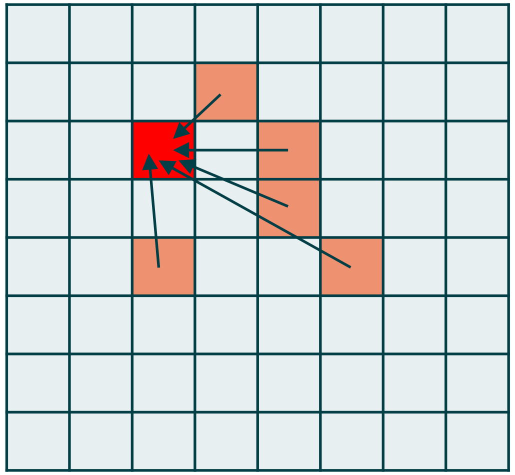
Convolutions as Hard-Coded Attention

Both Convolution and Attention Networks mix in features of other tokens/pixels

Convolution



Attention

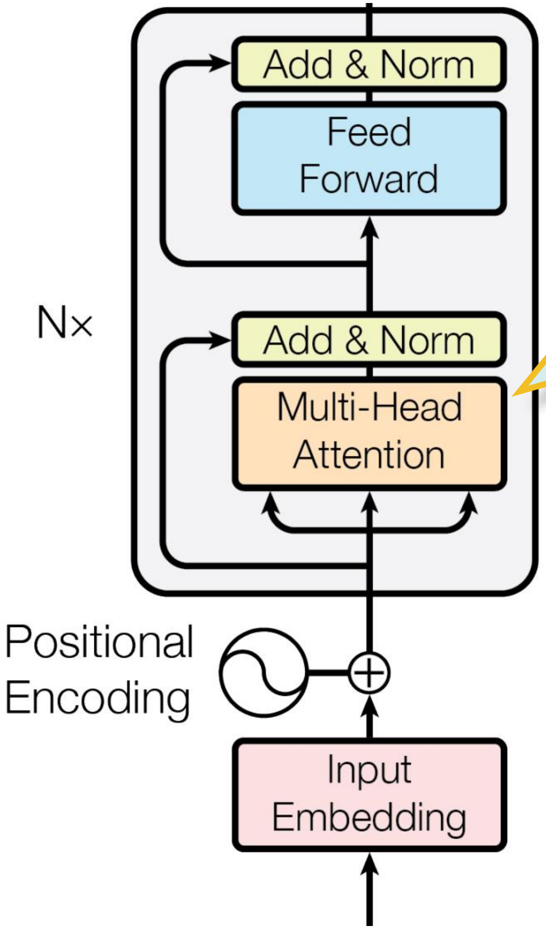


Convolutions mix in features from tokens based on fixed spatial location

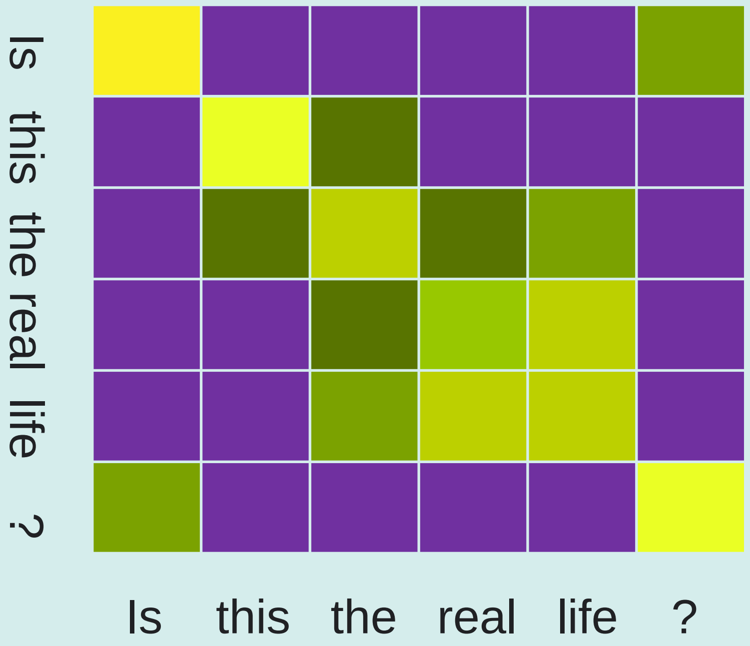
Attention mix in features from tokens based on learned attention

The Structure of a Transformer: Attention

Multi-Head Attention

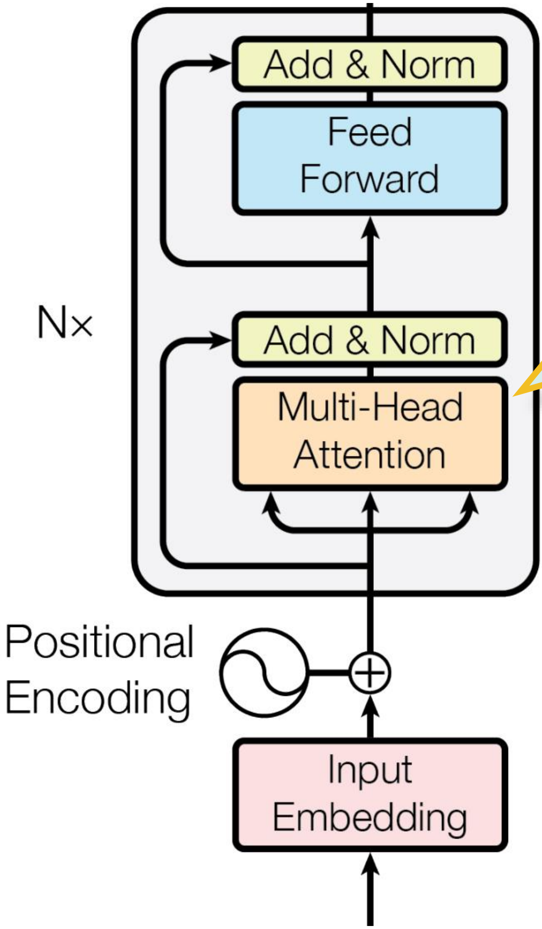


Attention: Mix in Features of Other Tokens

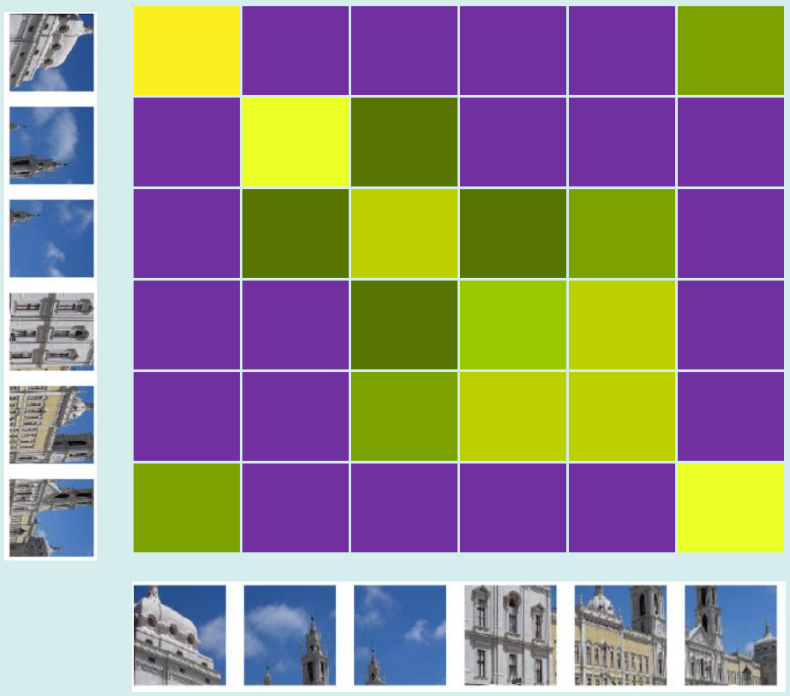


The Structure of a Transformer: Attention

Multi-Head Attention

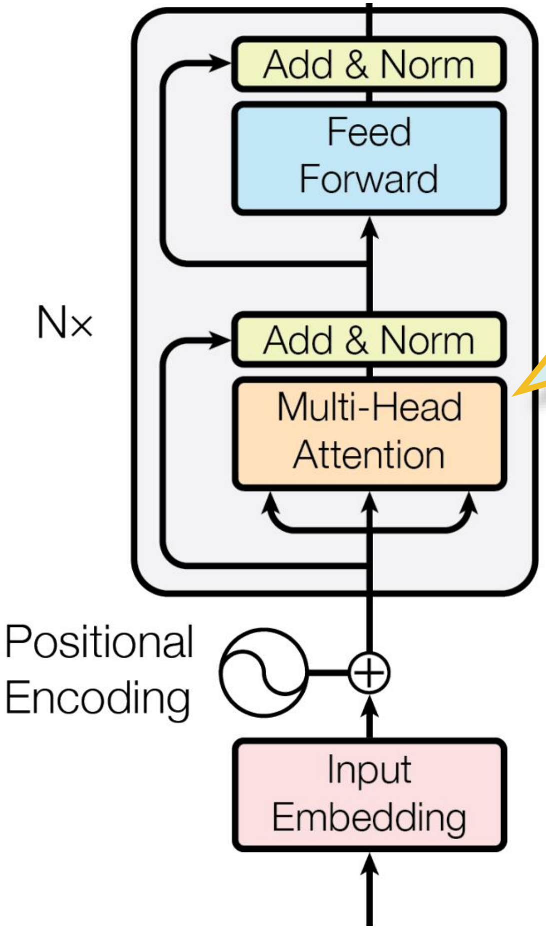


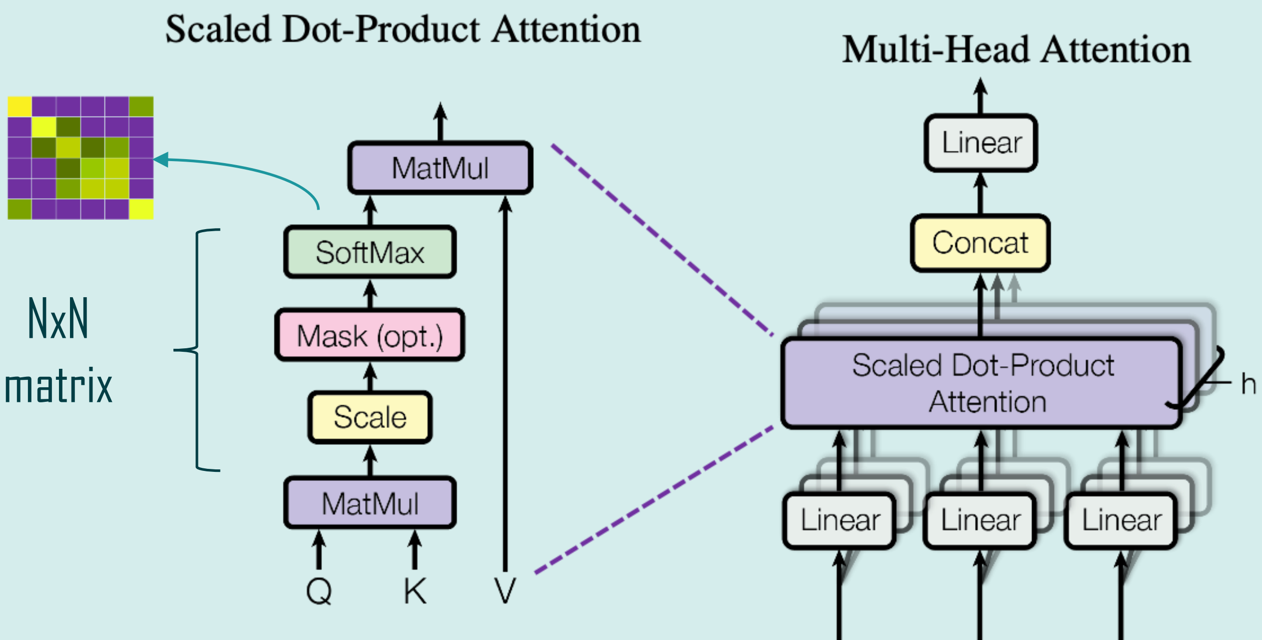
Attention: Mix in Features of Other Tokens



The Structure of a Transformer: Attention

Multi-Head Attention



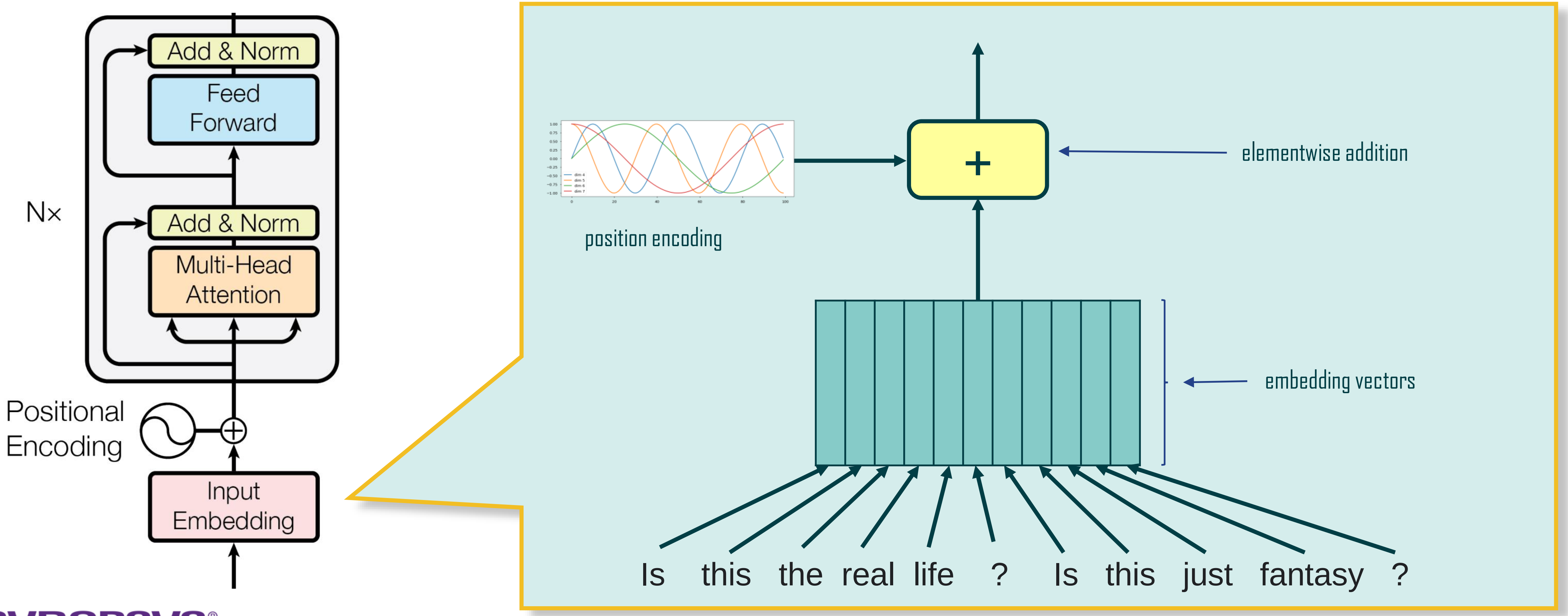


$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

N: number of tokens

The Structure of a Transformer: Embedding

Embedding of input tokens and the positional encoding



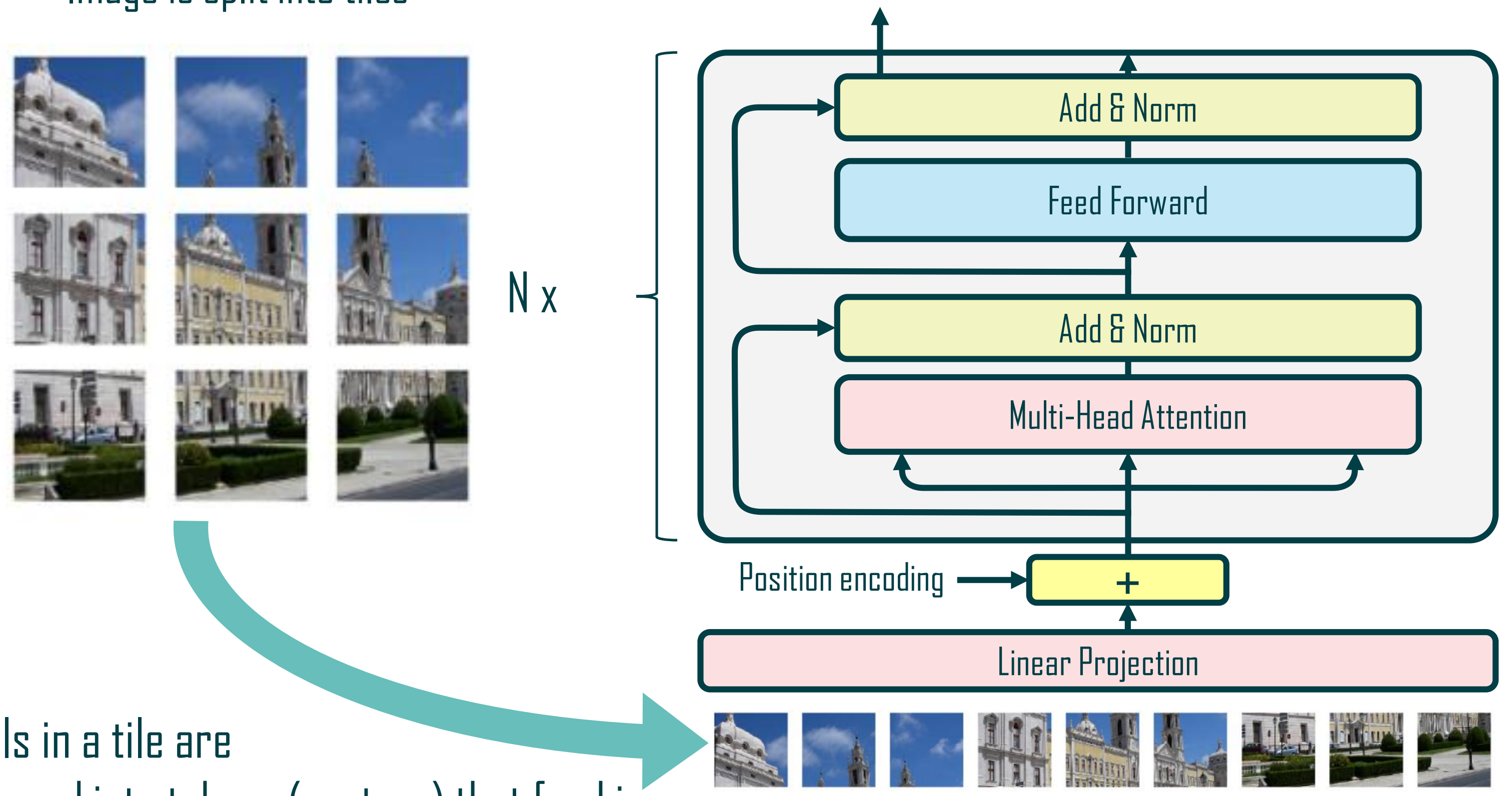
Other Application Domains:
Vision, Action Recognition, Speech Recognition

SYNOPSYS®

Vision Transformers (ViT/L16 or ViT-G/14)

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale(*)

Image is split into tiles



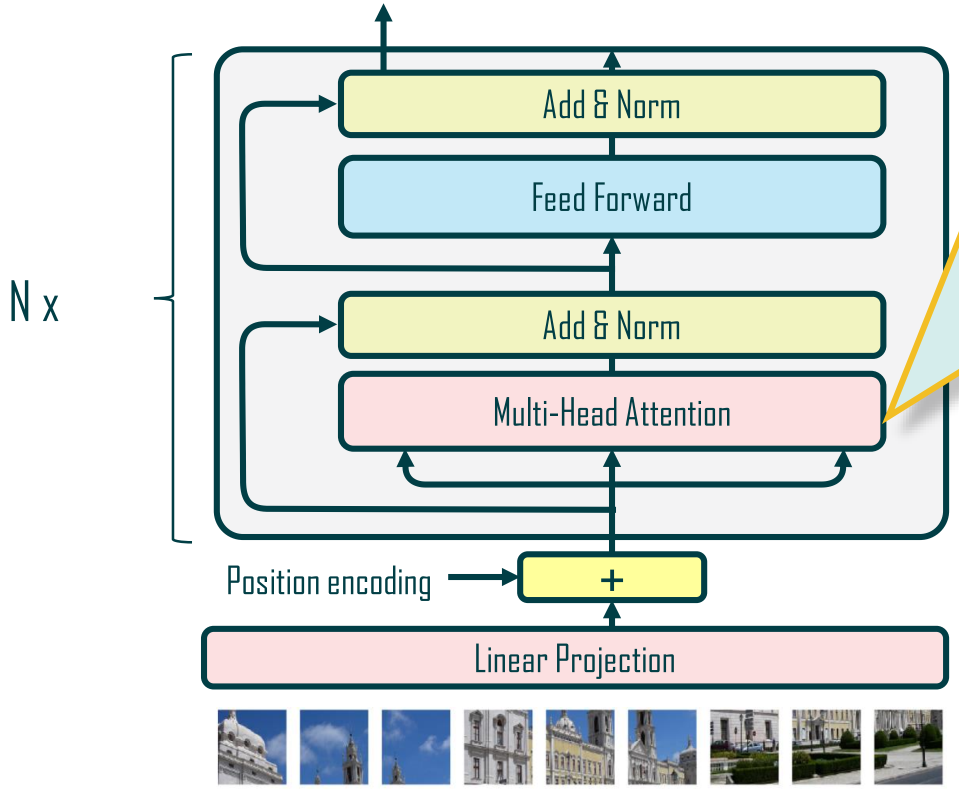
Pixels in a tile are flattened into tokens (vectors) that feed in the transformer

Vision Transformers are at the time of publication **best-known method for image classification**

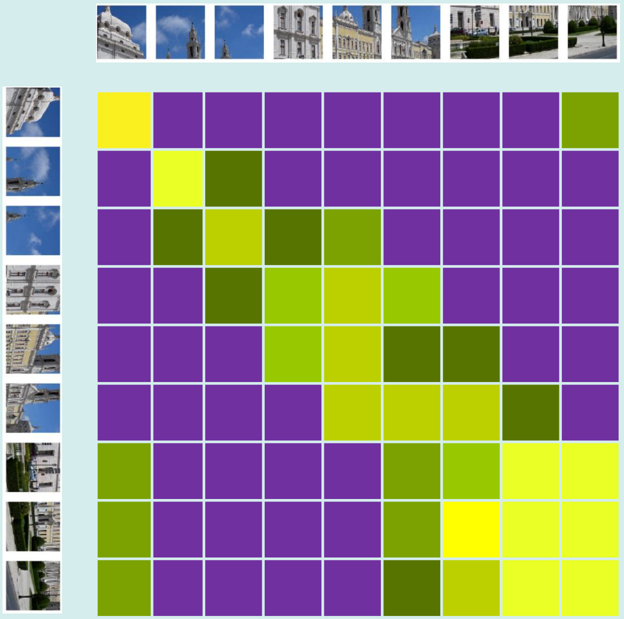
They are beating convolutional neural networks in **accuracy** and **training time,** but **not in inference time.**

(*) https://arxiv.org/abs/2010.11929

Vision Transformer → Increasing Resolution



Attention matrix scales quadratically with the number of patches



N x N matrix

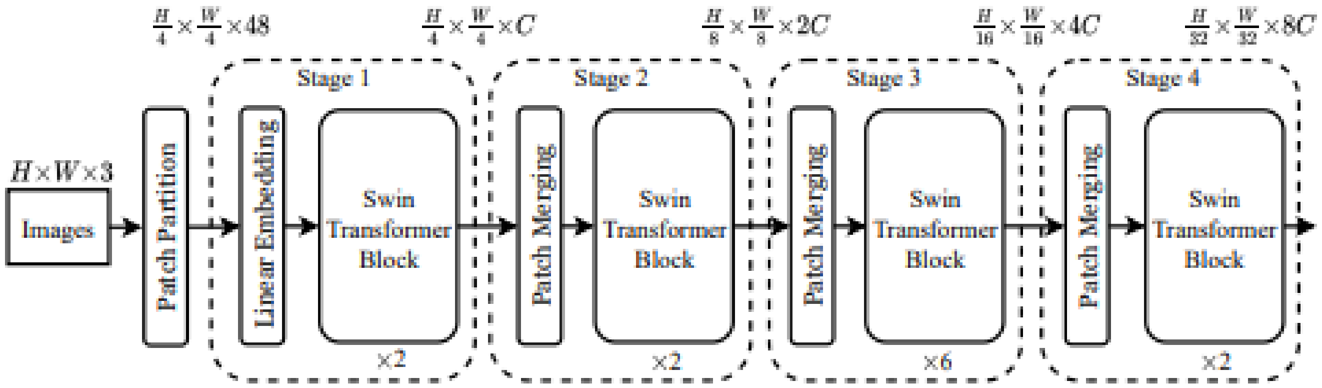
Where N = the number of tokens/patches

Swin Transformers

Hierarchical Vision Transformer using Shifted Windows (*)

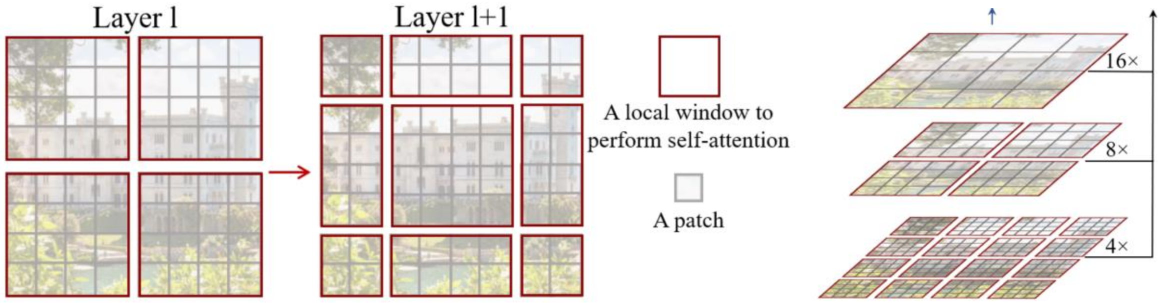
Adaptation makes Transformers scale for larger images:

1. Shifted Window Attention
2. Patch-Merging



State of the Art for

- Object Detection (COCO)
- Semantic Segmentation (ADE20K)



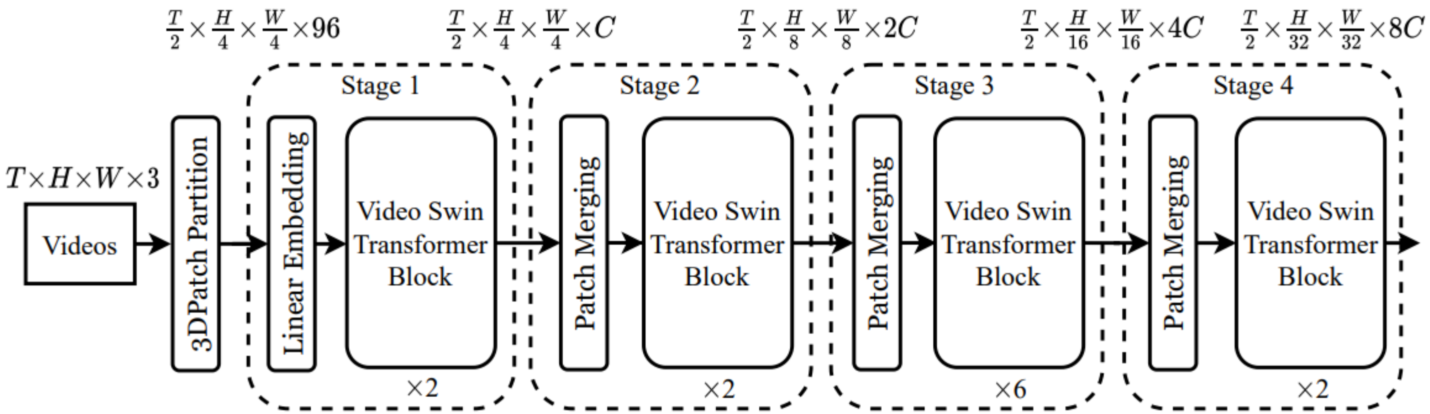
Shifted Window Attention

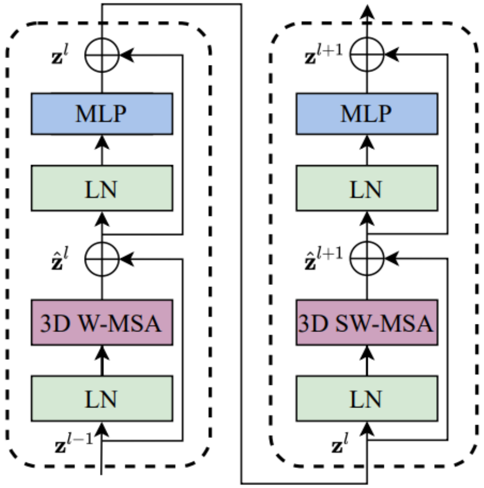
Patch-Merging

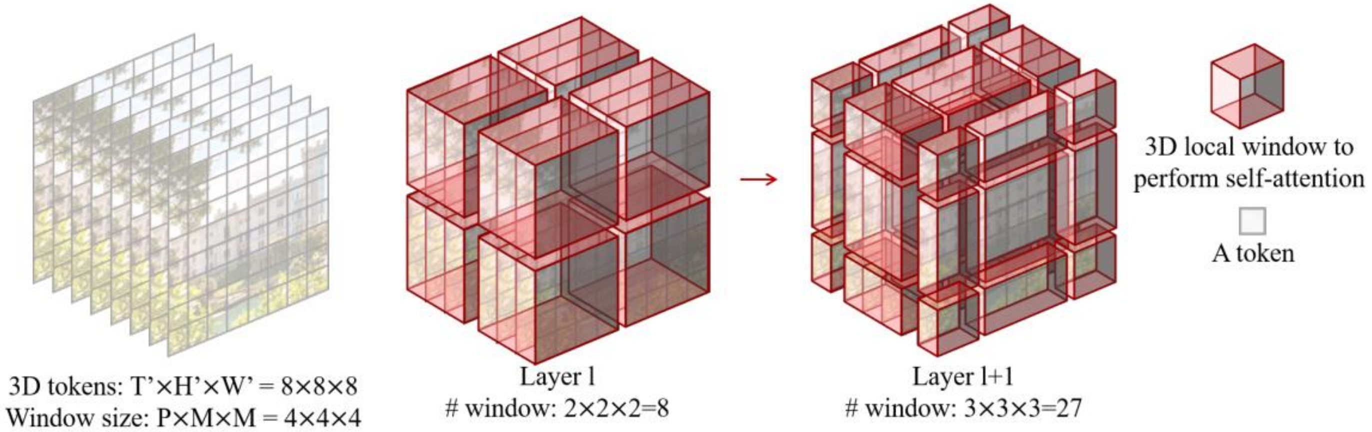
(*) https://arxiv.org/abs/2103.14030

Action Classification with Transformers

Video Swin Transformer







Video Swin Transformers extend the (shifted) window to three dimensions (2D spatial + time)

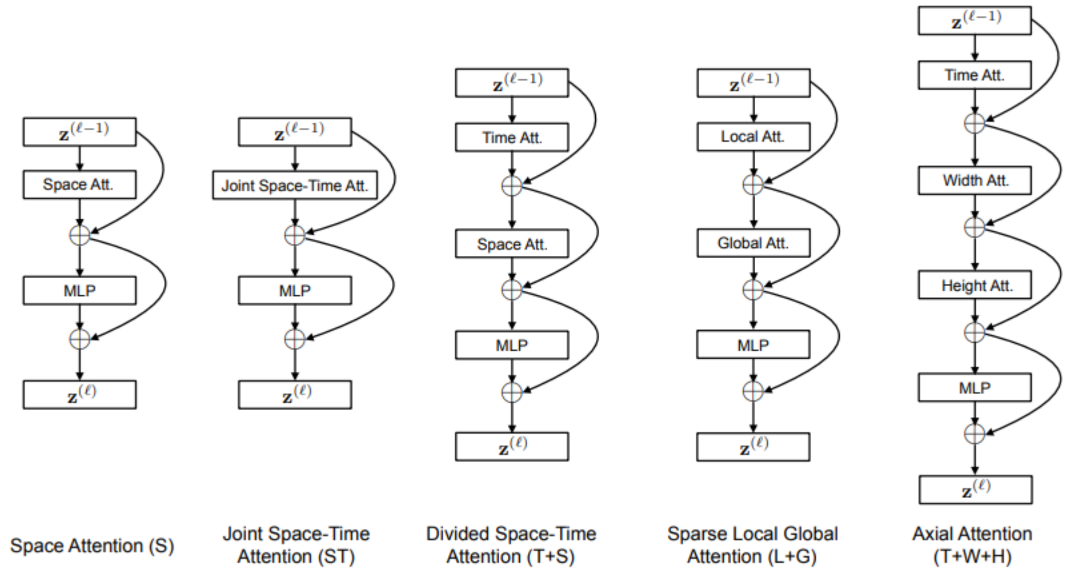
Today's state of the art on Kinetics-400 and Kinetics-600

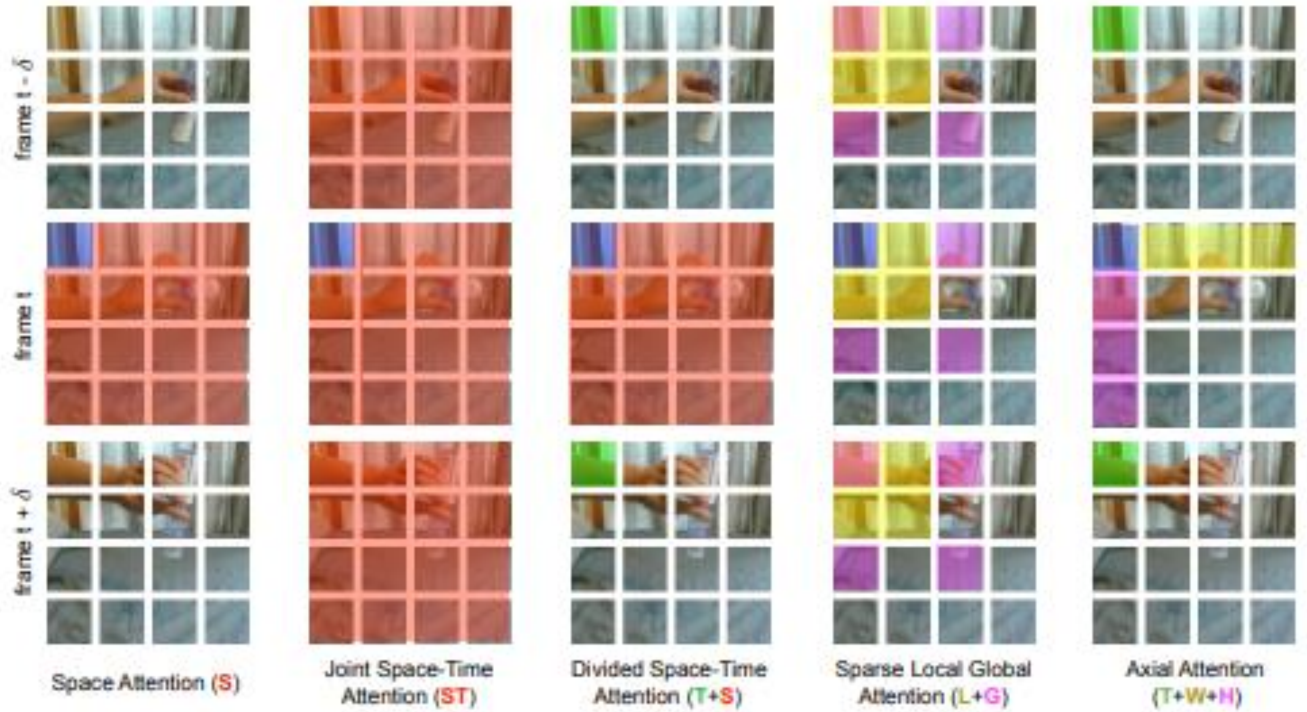
SYNOPSYS®

https://arxiv.org/abs/2106.13230

Action Classification with Transformers

Is Space-Time Attention All You Need for Video Understanding?

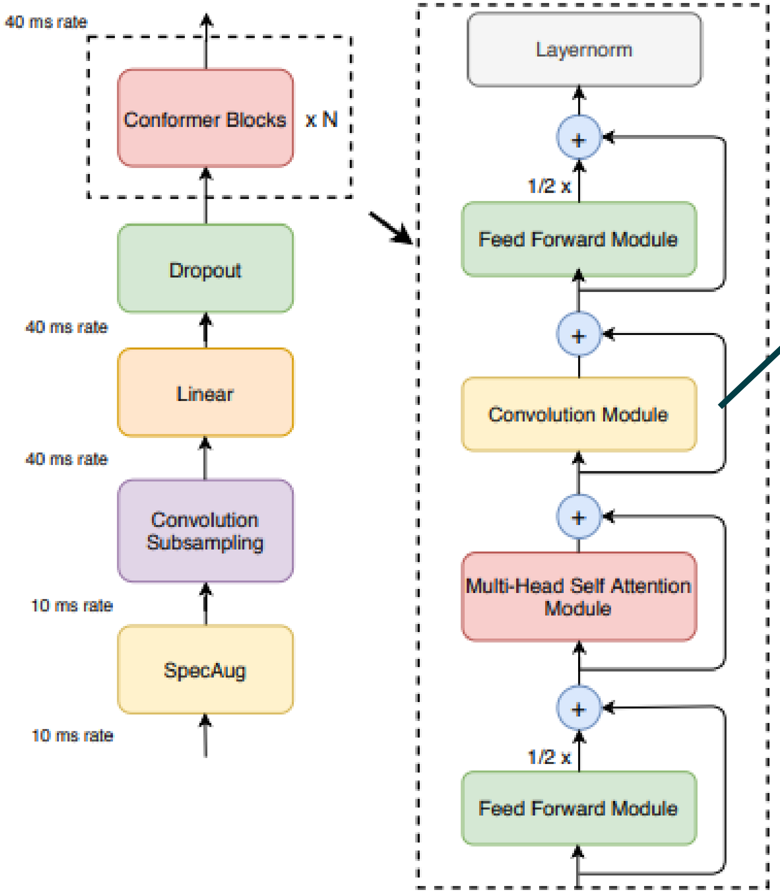




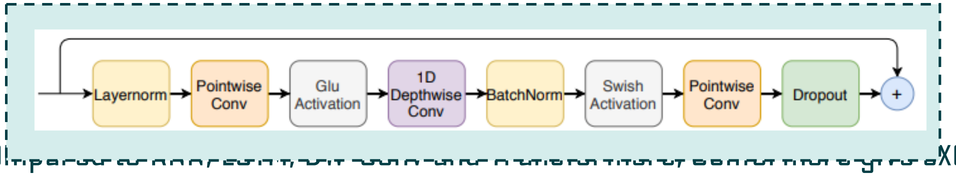
- Transformers can directly be applied to video
- Like for ViT, the video frames are split-up in tiles that feed directly in the Transformer
- Applying attention separately on time and on space "Divided Attention" gives (at time of publication) state of the art results on Kinetics-400 and Kinetics-600 benchmarks

Speech Recognition

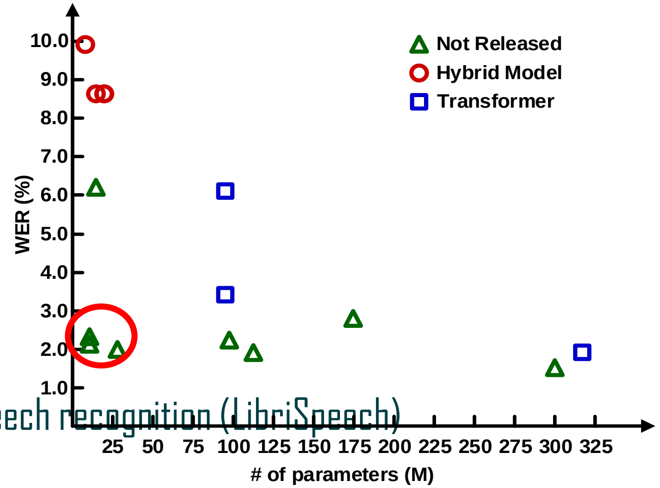
Conformer: Convolution-augmented Transformer for Speech Recognition (*)



- Conformers are Transformer with and additional Convolution Module
- The convolution module contains a pointwise and a depthwise (1D, size=31) convolution:



- Co… xcellent accuracy / size ratio:



- Best known methods for speech recognition (LibriSpeech) are based on Conformers

Why Attention and Transformers are Here to Stay for Vision

SYNOPSYS®

Visual Perception beyond Segmentation & Object Detection

Today

Panoptic Segmentation

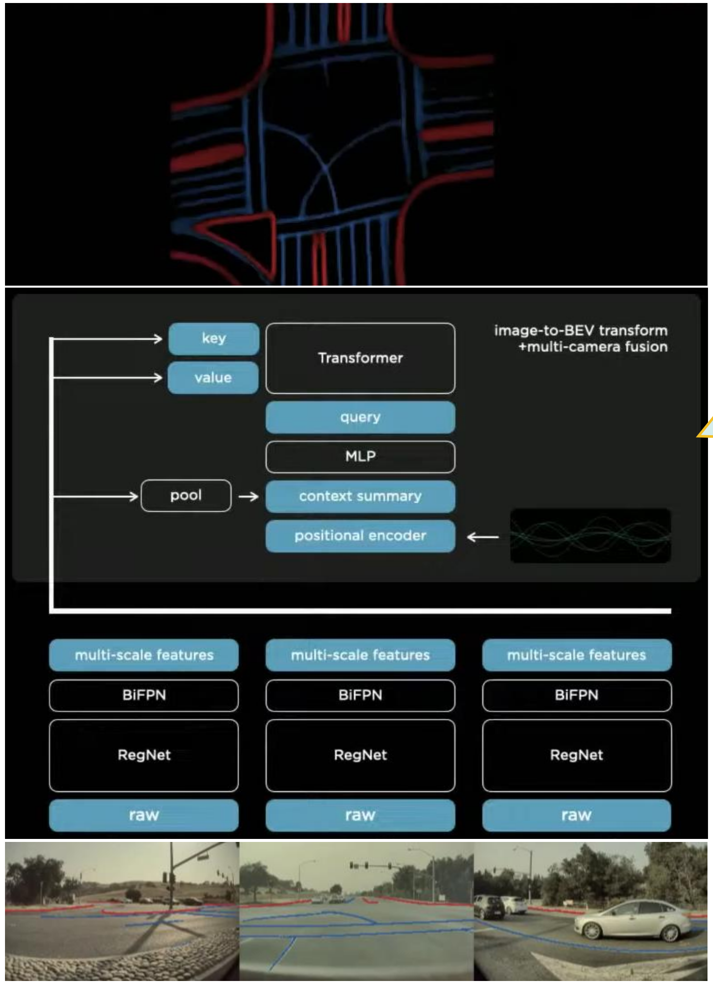
2022-...

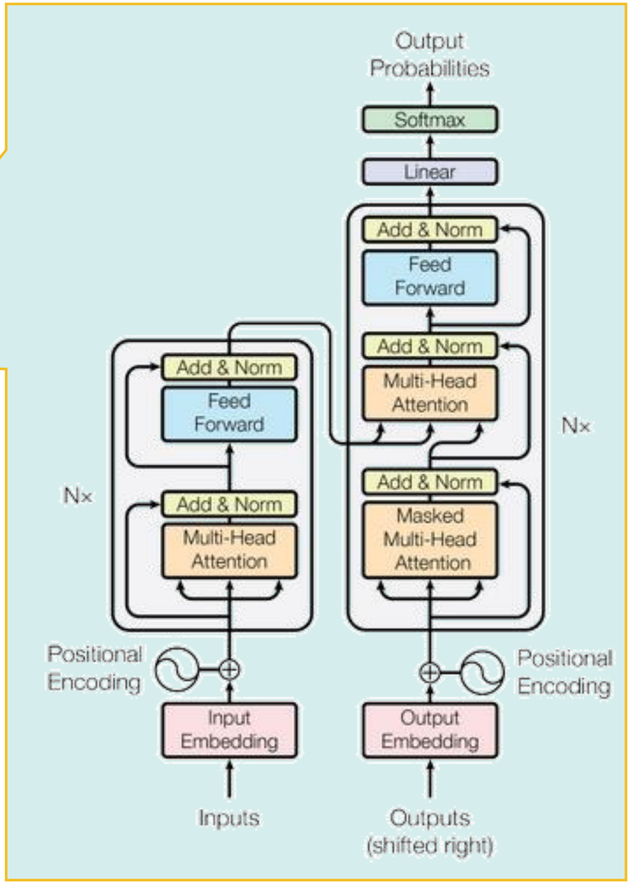
What is happening in this scene?

Future applications like security cameras, personal assistants, storage retrieval,…. require a deeper understanding of the world

→ Merging NLP and Vision using the same knowledge representation backend

Tesla AI Day: Using Transformers Make Predictions in Vector Space





- Convolutional neural network extract features for every camera
- A transformer is used to:
  - Fuse multiple cameras
  - Make predictions directly in bird-eye-view vector space

Why Transformers are Here to Stay in Vision

- Attention based networks outperform CNN-only networks on accuracy
  - Highest accuracy required for high-end applications
- Models that combine Vision Transformers with Convolutions are more efficient at inference
  - Examples: MobileViT(*), CoAtNet(**)
- Full visual perception requires knowledge that may not easily be acquired by vision only
  - Multi-modal learning required for a deeper understanding of visual information
- Application integrating multiple sensors benefit from attention-based networks

(*) https://arxiv.org/abs/2110.02178
(**) https://arxiv.org/abs/2106.04803v2

Summary

- Transformers are deep learning models primarily used in the field of NLP
- Transformers lead to state-of-the-art results in other application domains of deep learning like vision and speech
  - They can be applied to other domains with surprisingly little modifications
  - Models that combine attention and convolutions outperform convolutional neural networks on vision tasks, even for small models
- Transformers and attention for vision applications are here to stay
  - Real world applications require knowledge that is not easily captured with convolutions

Resources

Resources

ARVIX.org

https://arxiv.org/abs/1706.03762

ARC NPX6 NPU IP

www.synopsys.com/arc

Join the Synopsys Deep Dive

Optimize AI Performance & Power for Tomorrow's Neural Network Applications (Thursday, 12-3 PM)

Synopsys Demos in Booth 719

- Executing Transformer Neural Networks in ARC NPX6 NPU IP
- Driver Management System on ARC EV Processor IP with Visidon
- Neural Network-Enhanced Radar Processing on ARC VPX5 DSP with SensorCortek