How Transformers are Changing the Direction of Deep Learning Architectures

Tom Michiels, System Architect

Synopsys ARC® Processor Summit 2022

CNNs Have Dominated Many Vision Tasks Since 2012

CV
50%

Improved Accuracy:
from 65% to 75%

Improved Efficiency:
over 10X

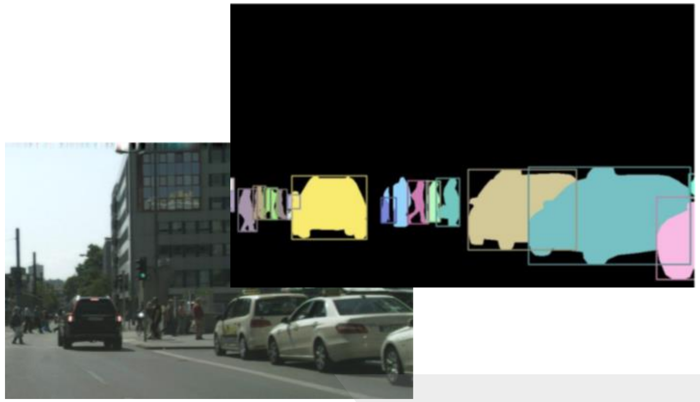
Improved Accuracy:
up to 90%

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

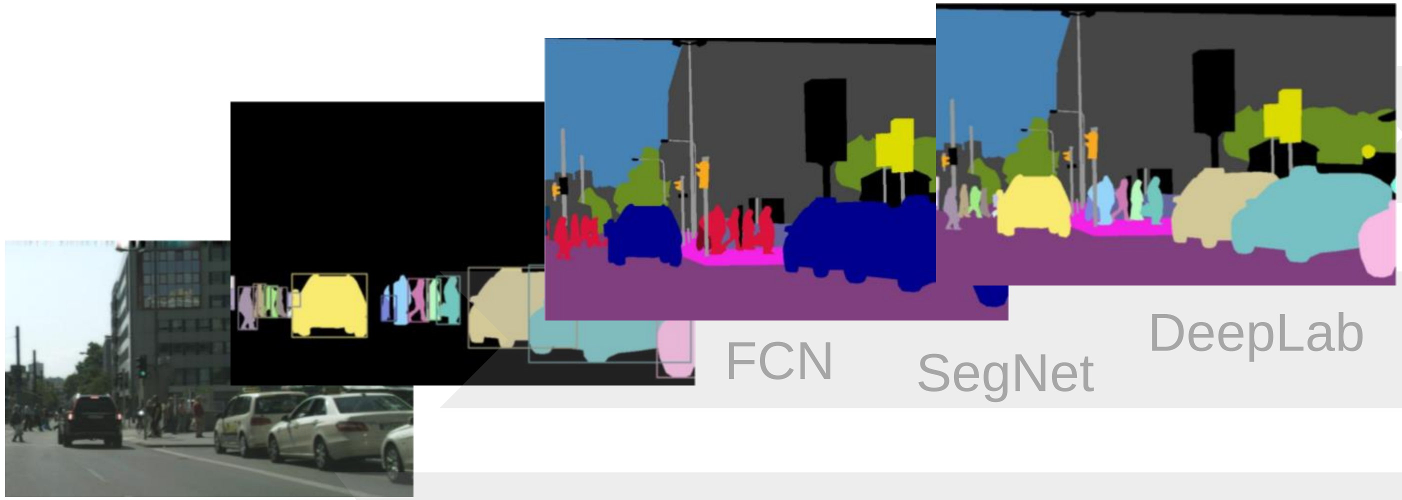
FCN SegNet DeepLab DSNet

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



PFPN
YoloP
DeeperLab
EfficientPS

FCN
SegNet
DeepLab
DSNet

RCNN
FRCNN
SSD
YOLOV2
Mask RCNN
YoloV3
YoloV4/V5
EfficientDet

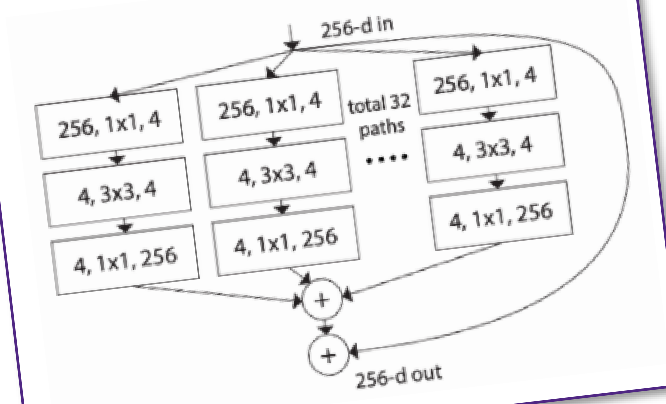
AlexNet
VGG
ResNet
MobileNet V1
MobileNet V2
EfficientNet
MobileNet v3

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

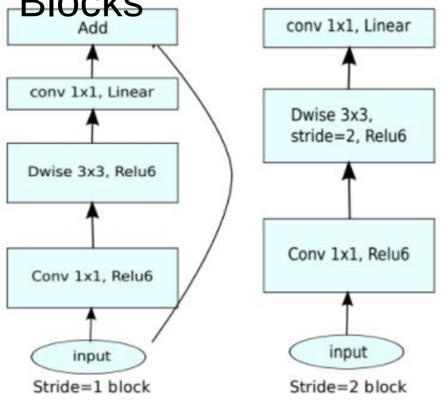
Panoptic Vision

A Decade of CNN Development… ➡ 90.0% Accuracy, 2021

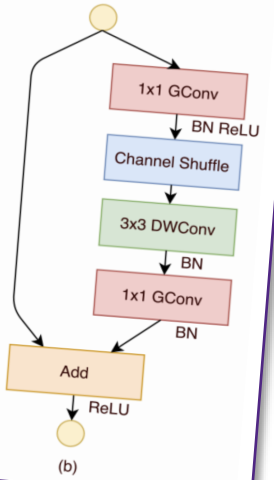
Inception



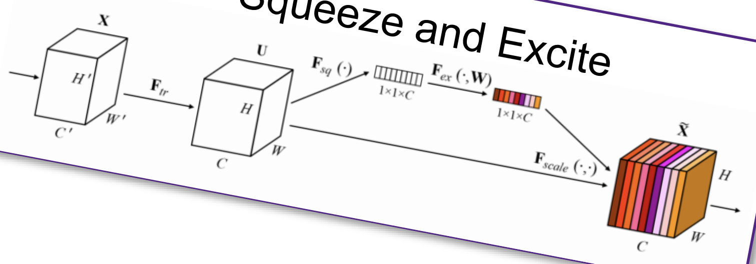
Inverted Residual Blocks



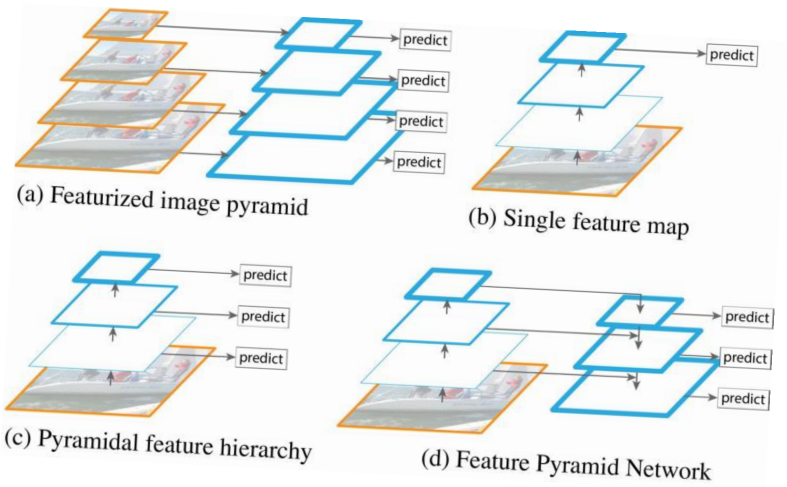
Shufflenet



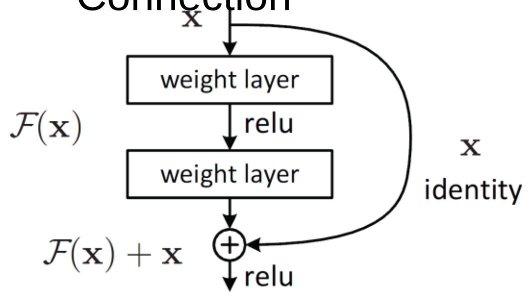
Squeeze and Excite



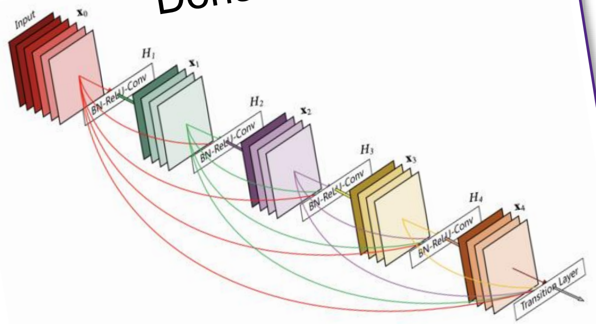
Feature Pyramid



Residual Connection



DenseNet

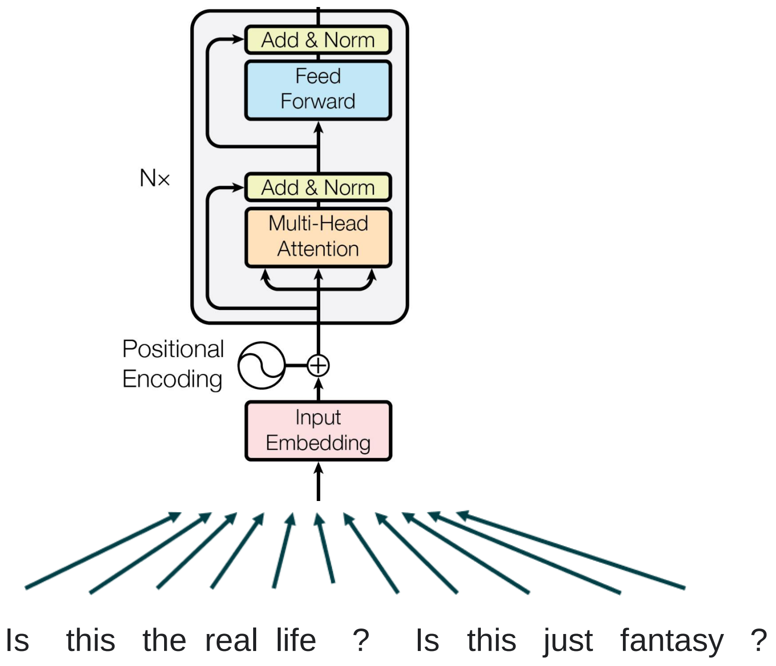


The Appearance of Transformers

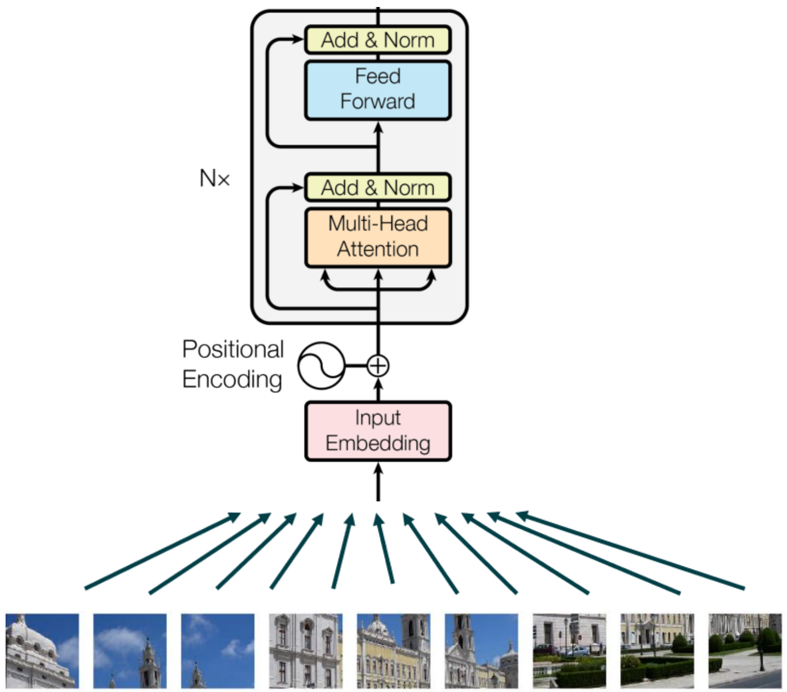
➡ 90.5% Accuracy, 2021

➡ 91.0% Accuracy, 2022

Transformer, a model designed for natural language processing



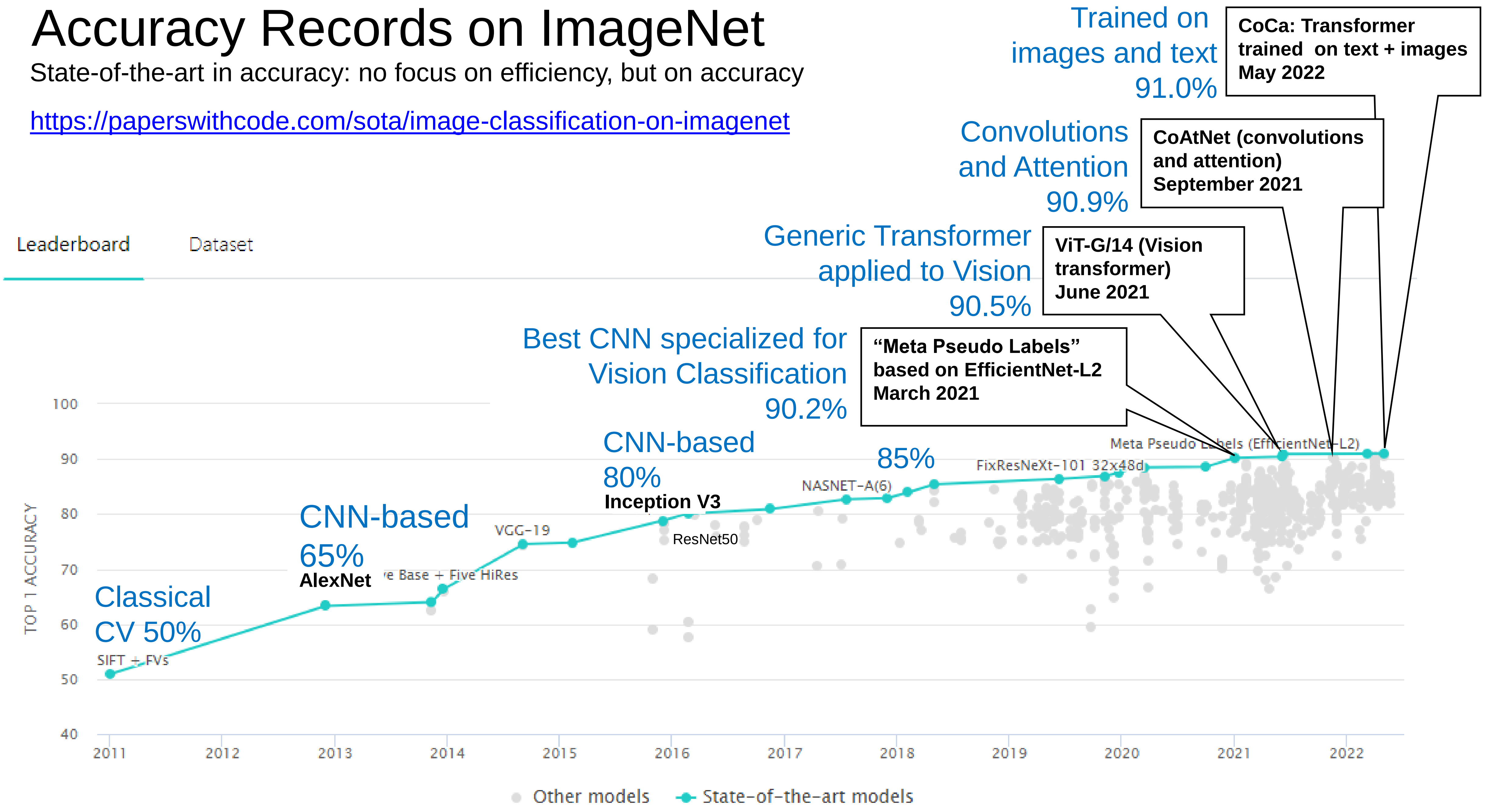
… without any modifications applied to image patches, beats the highly specialized CNNs in accuracy



Accuracy Records on ImageNet

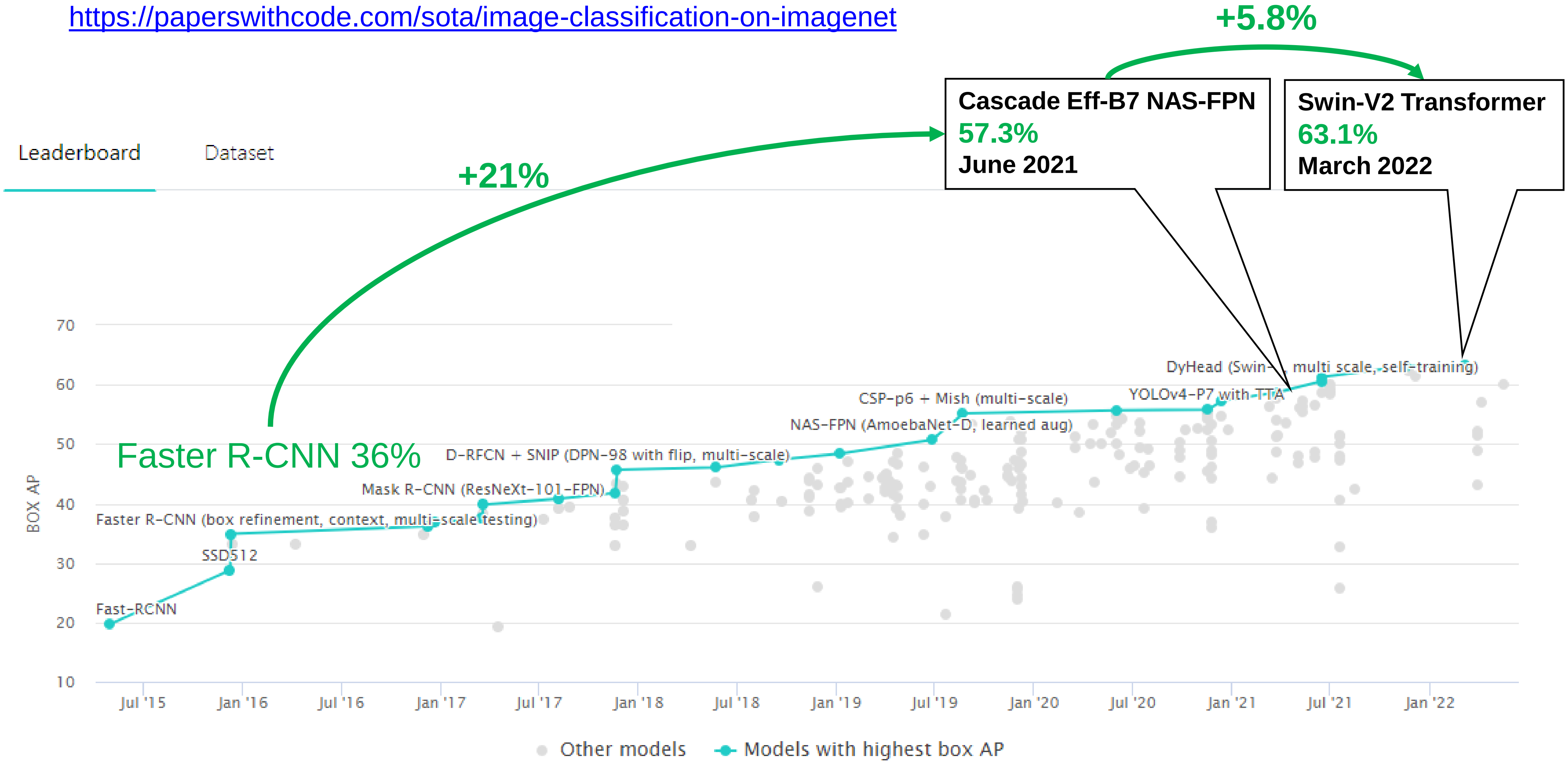
State-of-the-art in accuracy: no focus on efficiency, but on accuracy

https://paperswithcode.com/sota/image-classification-on-imagenet



Object Detection – COCO test-dev, box AP

https://paperswithcode.com/sota/image-classification-on-imagenet



Transformers Compute Requirements and Model Size

- Compute requirements for early Transformer models are much higher
  - Performance comparison (for same NPU configuration)

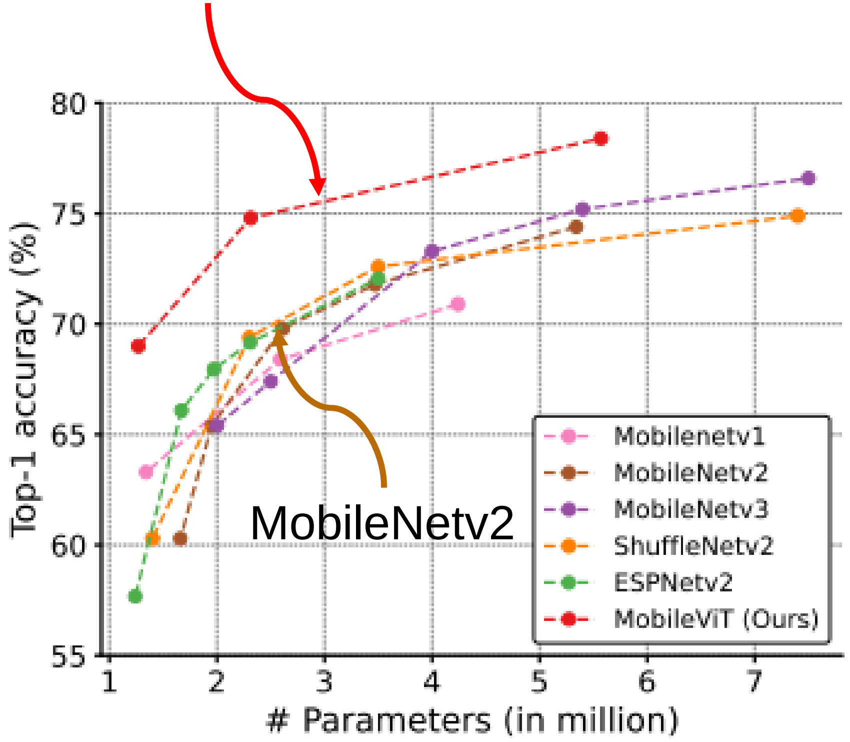
NN Model	Image size	Top 1 Accuracy	Relative GOPS	Relative Frames/sec
MobileNetv2	224x224	69.8%	1X	32X
ViT_B_16	224x224	84.0%	58X	1X

- All the state-of-the-art models (CNN and Transformers) are huge
  - Approx. 2G parameters
  - Impractical for use in embedded applications

Mobile ViT: Small Mobile (Paper by Apple, March 2022)

https://arxiv.org/pdf/2110.02178.pdf

MobileVit is a small Transformer + Convolution model that beats convolutions of similar size in accuracy



Model	# Params. ⇓	Top-1 ⇑
MobileNetv1	2.6 M	68.4
MobileNetv2	2.6 M	69.8
MobileNetv3	2.5 M	67.4
ShuffleNetv2	2.3 M	69.4
ESPNetv2	2.3 M	69.2
MobileViT-XS (Ours)	2.3 M	**74.8**

+ 5% Accuracy

Comparison with light-weight CNNs with similar model-size

Mobile ViT: Small Mobile (Paper by Apple, March 2022)

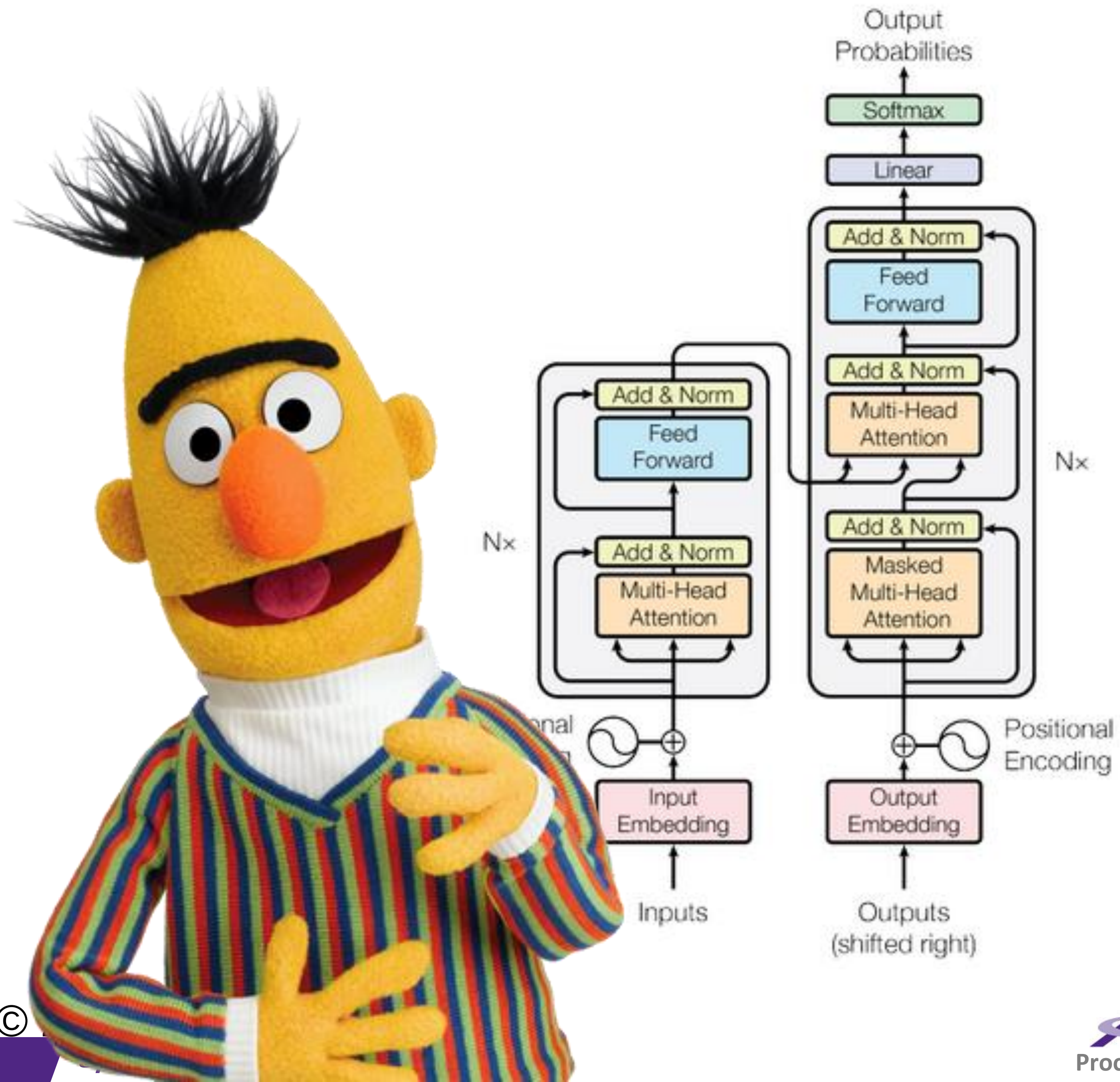
https://arxiv.org/pdf/2110.02178.pdf

Model	# Params ⇓	FLOPs ⇓	Top-1 ⇑	Inference Time (ms)		
				iPhone12 - CPU	iPhone12 - Neural Engine	
MobileNetv2	3.5 M	**0.3 G**	73.3	**7.50 ms**	**0.92 ms**	CPU/NNE = 8.1X
DeIT	5.7 M	1.3 G	72.2	28.15 ms	10.99 ms	
PiT	4.9 M	0.7 G	73.0	24.03 ms	10.56 ms	
MobileViT (Ours)	**2.3 M**	0.7 G	**74.8**	17.86 ms	7.28 ms	CPU/NNE = 2.5X
	0.7X Model Size	2.3X FLOPs	+1.5% Accuracy	2.4X Time	7.9X Time	

- Observations in Paper
  - On embedded devices (iPhone) MobileViT is slower than CNN based methods
  - Because the AI accelerator on iPhone is not as optimized for Transformers as it is for CNN’s
  - The authors expect that future AI accelerators will better support Transformers

The Structure of Attention and Transformers

Bert and Transformers

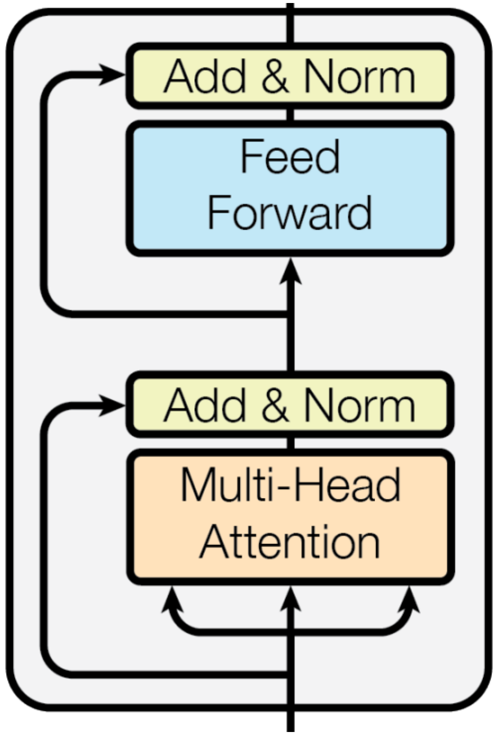


- Attention is all you need!(*)
- A Transformer is a deep learning model that uses Attention mechanism
- Transformers were primarily used for Natural Language Processing
  - Translation
  - Question Answering
  - Conversational AI
- Successful training of huge transformers
  - MTM, GPT-3, T5, ALBERT,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

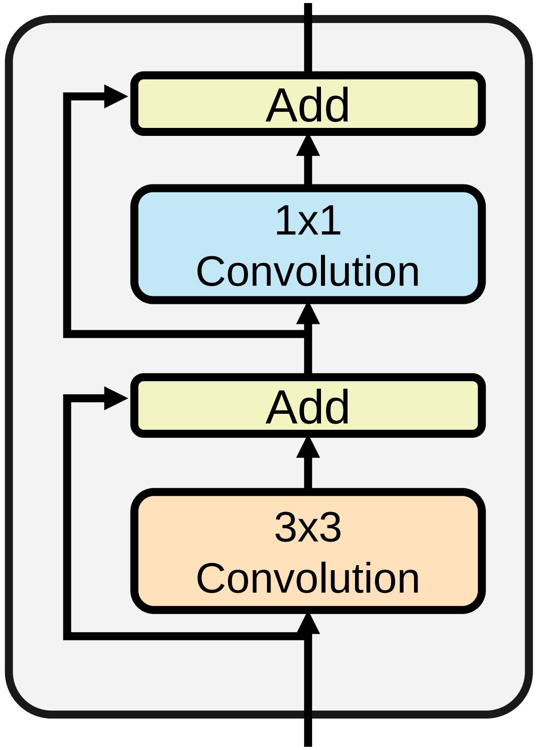
(*) https://arxiv.org/abs/1706.03762

Convolutions, Feed Forward, and Multi-Head Attention

Transformer



CNN



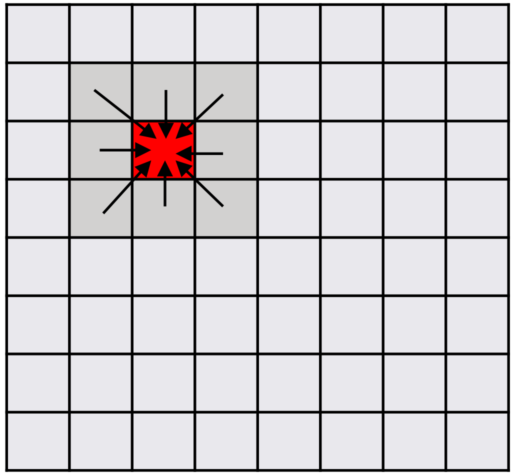
- The Feed Forward layer of the Transformer is identical to a 1x1 Convolution
- In this part of the model, no information is flowing between tokens/pixels

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

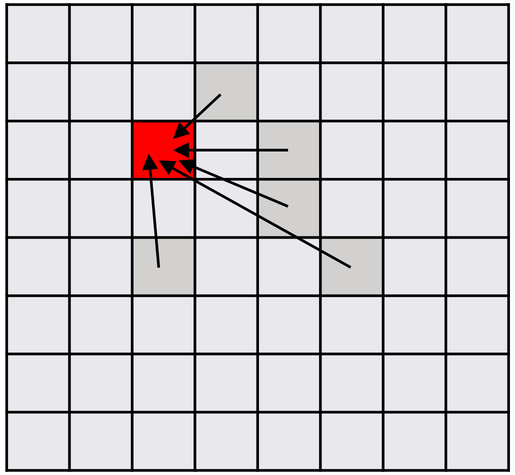
Convolutions as Hard-Coded Attention

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

Convolution



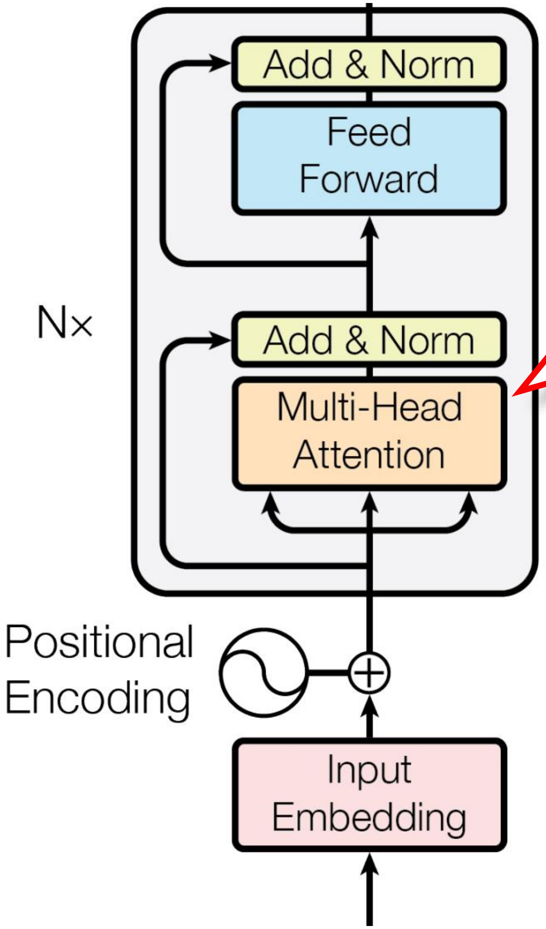
Attention



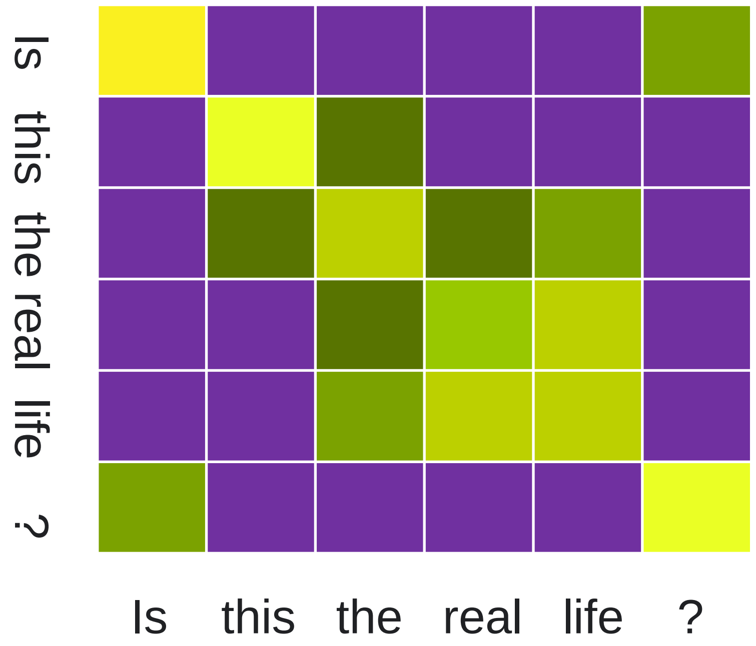
Convolutions mix in features from tokens based on fixed spatial location
Attention mix in features from tokens based on learned attention

The Structure of a Transformer: Attention

Multi-Head Attention

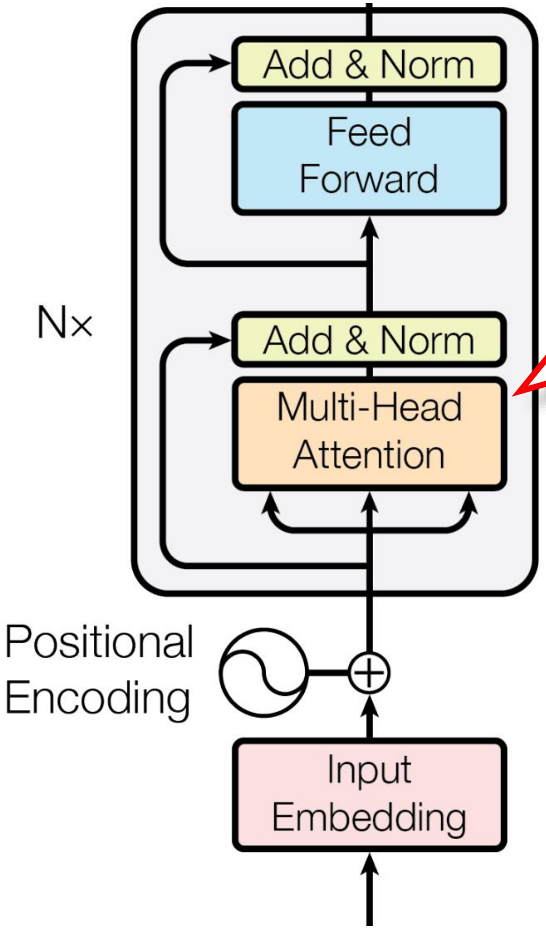


Attention: Mix in Features of Other Tokens

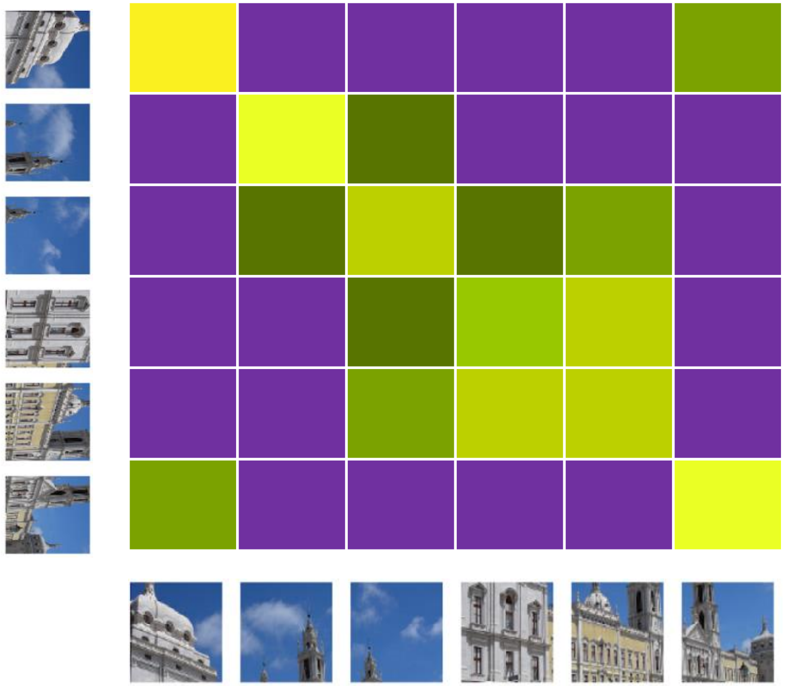


The Structure of a Transformer: Attention

Multi-Head Attention

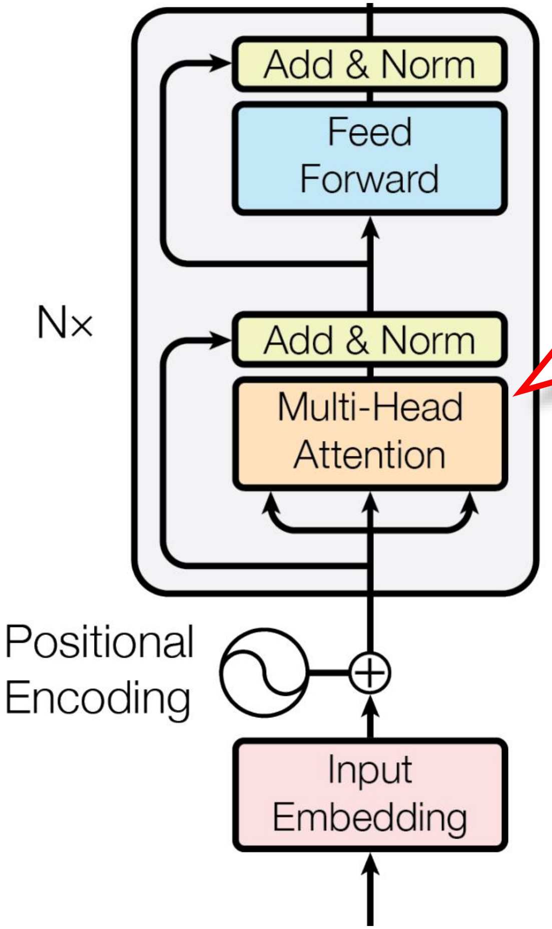


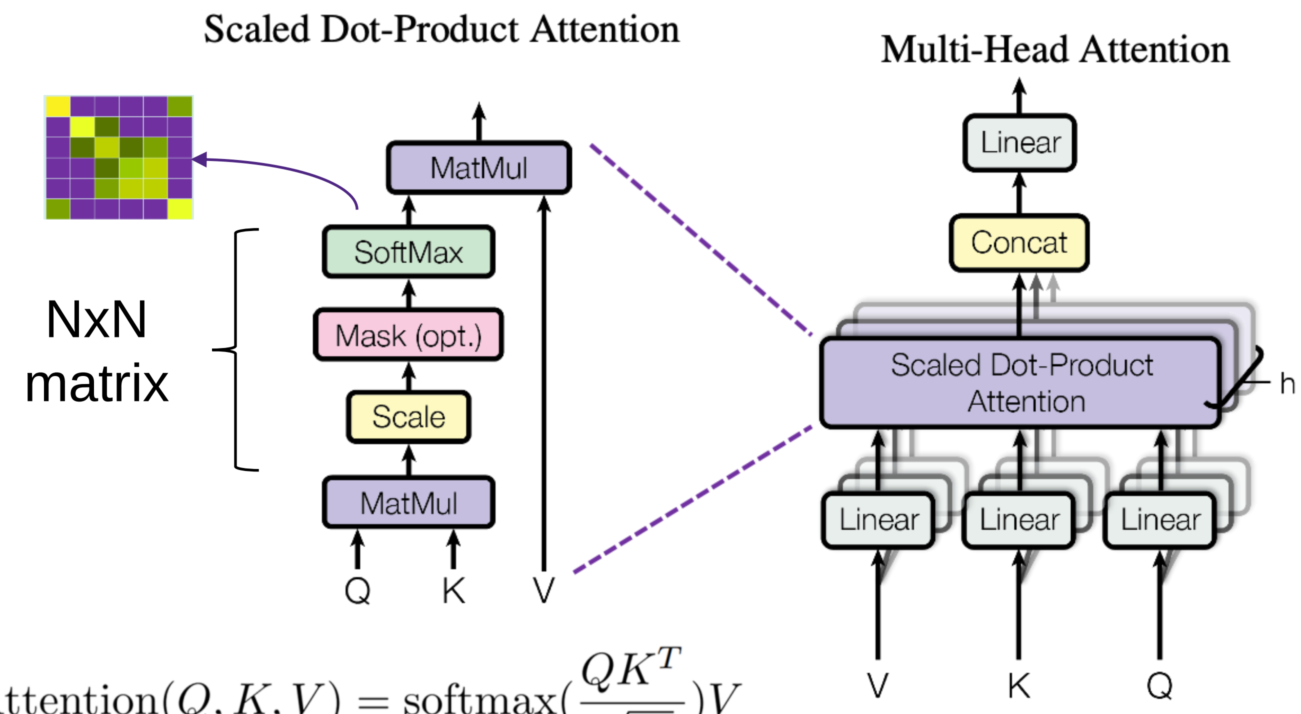
Attention: Mix in Features of Other Tokens



The Structure of a Transformer: Attention

Multi-Head Attention





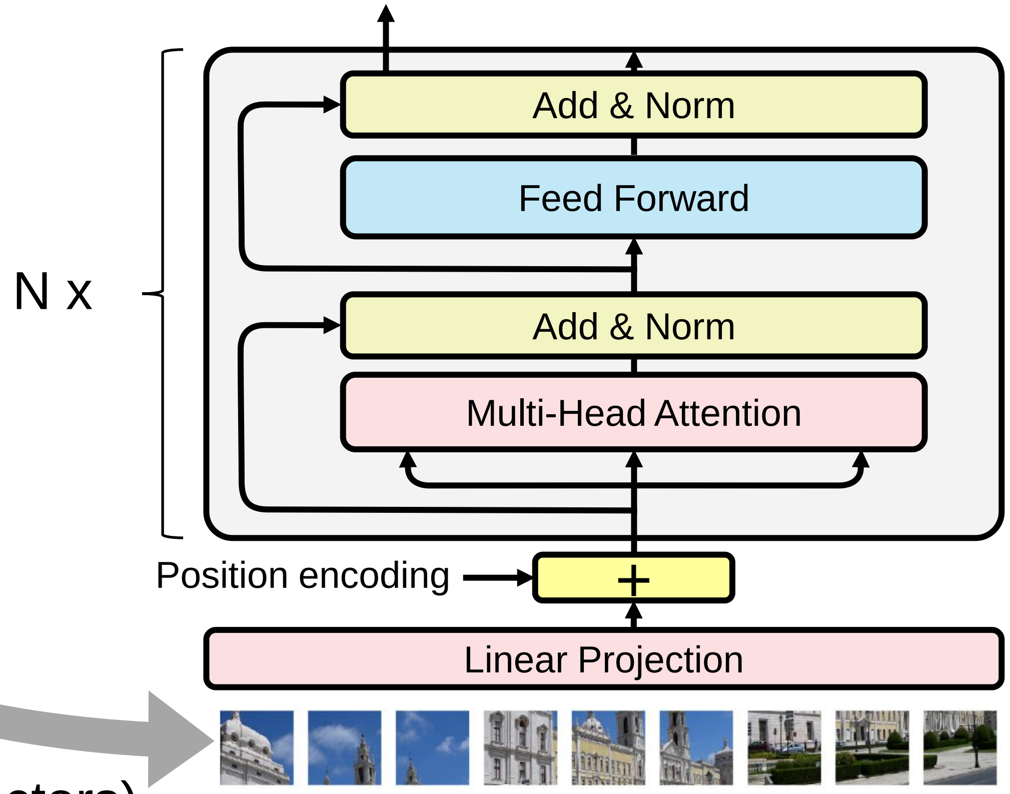
$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

N: number of tokens

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



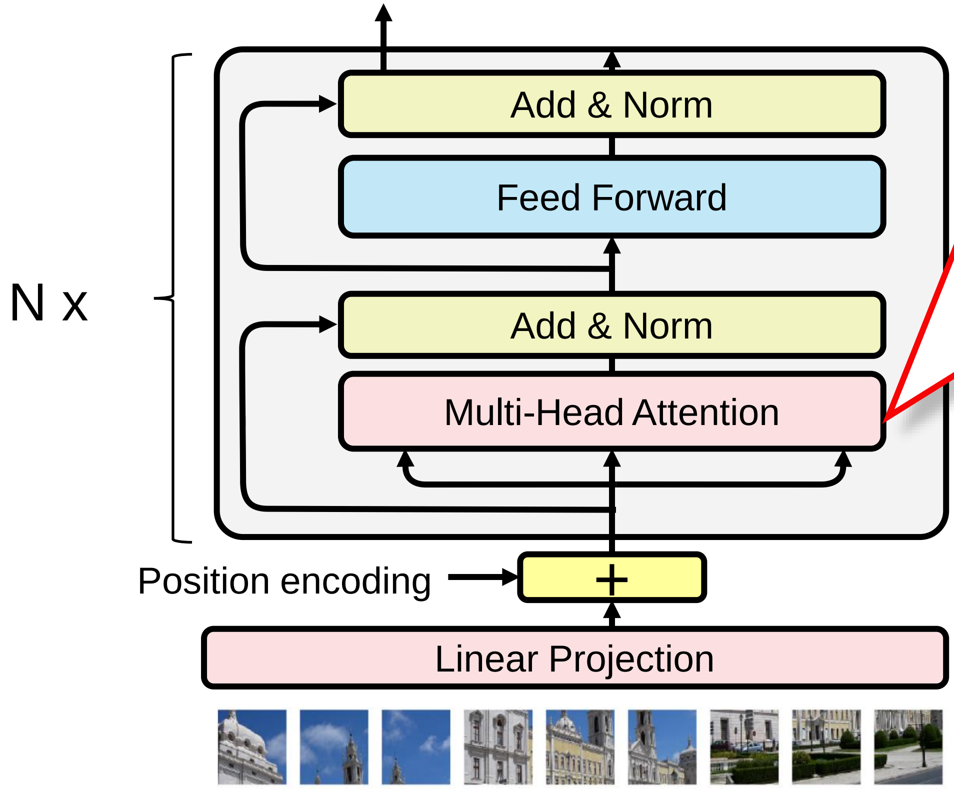
Pixels in a tile are flattened into tokens (vectors) that feed in the transformer

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

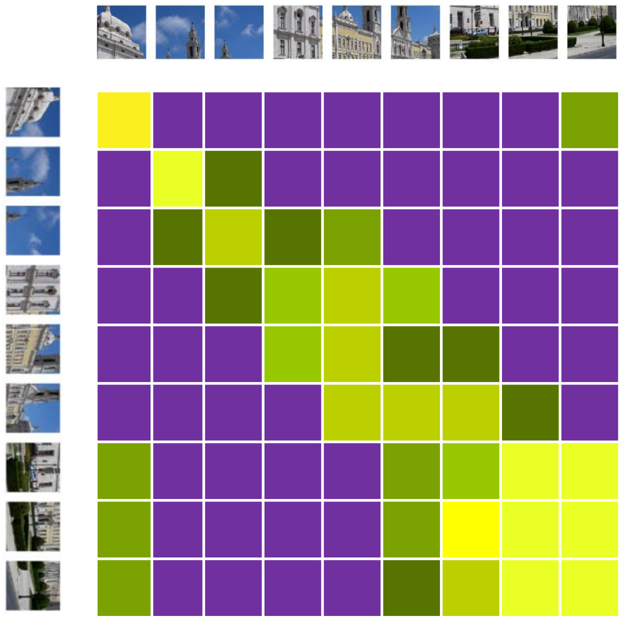
They are beating convolutional neural networks in **accuracy** and **training time,** but **not in inference time.**

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

Vision Transformer → Increasing Resolution



Attention matrix scales quadratically with the number of patches



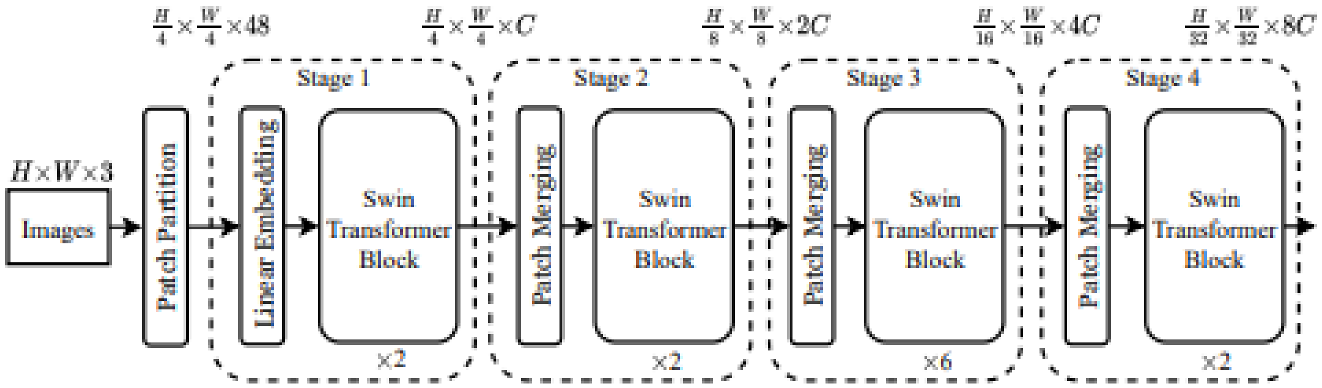
N x N matrix

Where N = the number of tokens/patches

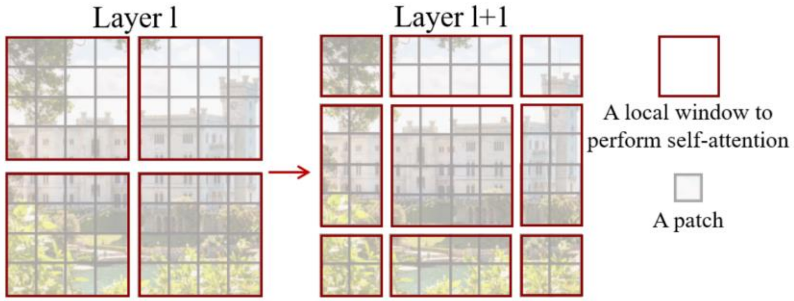
Swin Transformers

Hierarchical Vision Transformer using Shifted Windows (*)

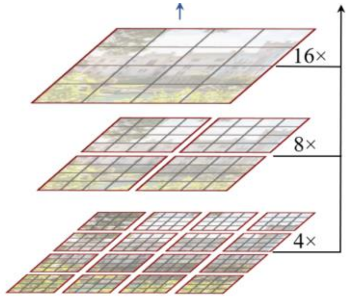
Adaptation makes Transformers scale for larger images:
1. Shifted Window Attention
2. Patch-Merging



State of the Art for
- Object Detection (COCO)
- Semantic Segmentation (ADE20K)



Shifted Window Attention



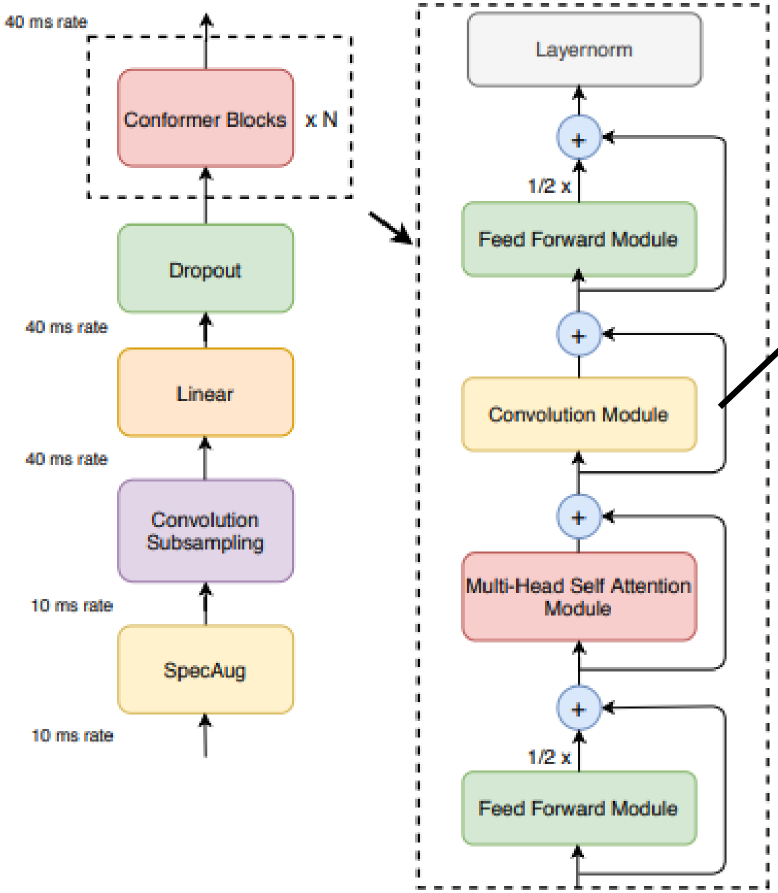
Patch-Merging

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

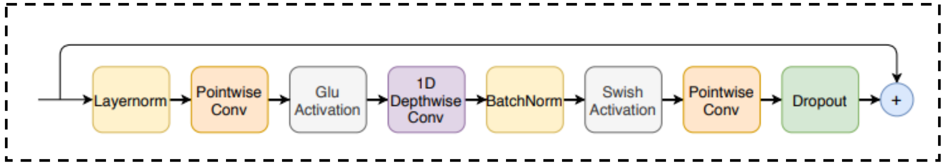
Other Application Domains:
Speech Recognition, Action Recognition

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:



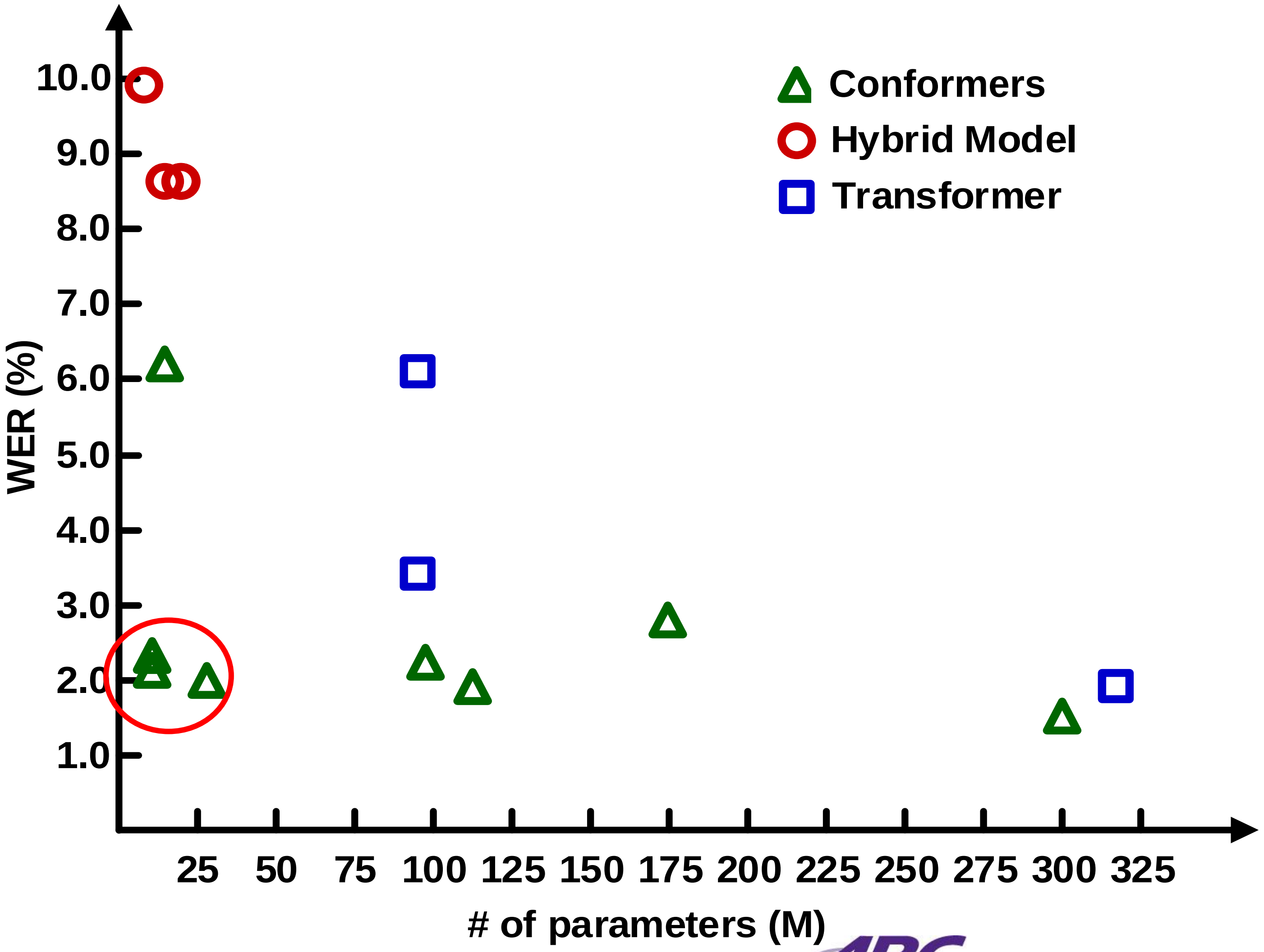
- Compared to RNN, LSTM, DW-Conv and Transformers, Conformers give excellent accuracy / size ratio

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

https://arxiv.org/abs/2005.08100

Speech Recognition – contd.

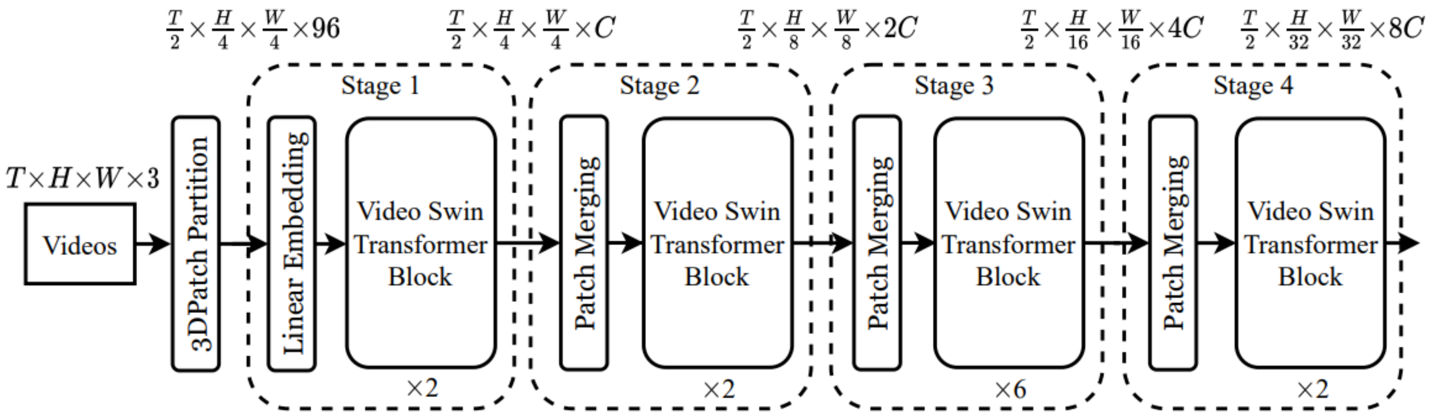
Compared to RNN, LSTM, DW-Conv and Transformers, Conformers give excellent accuracy / size ratio

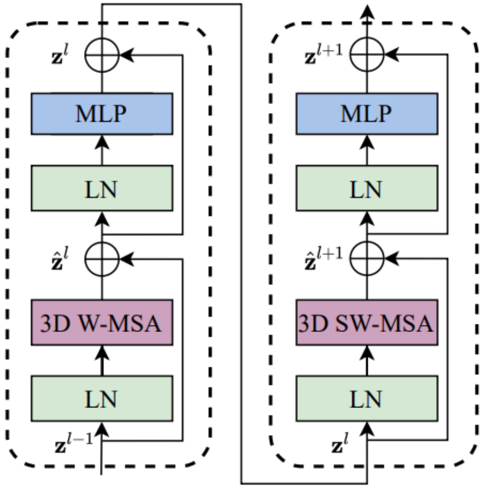


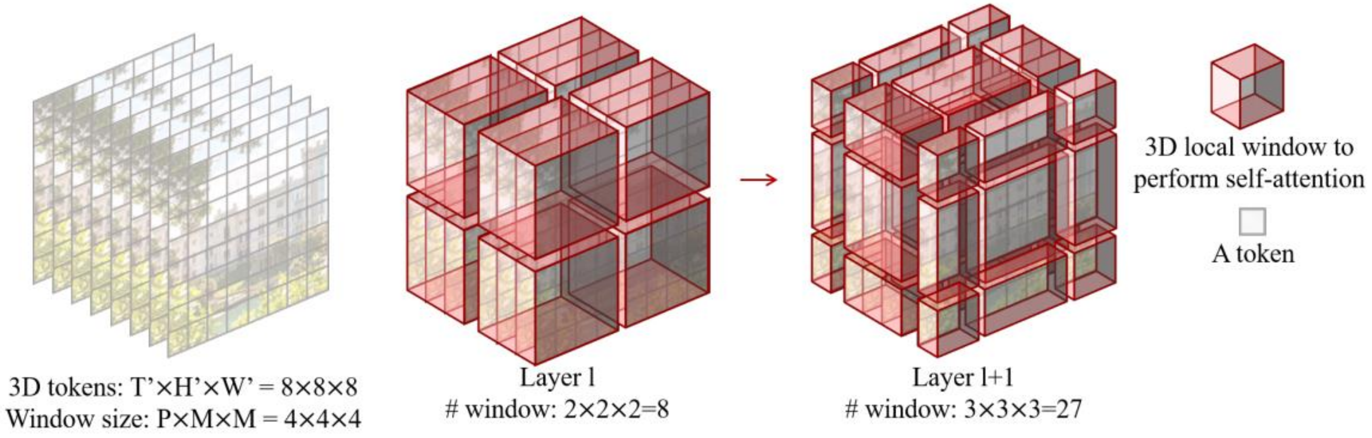
https://arxiv.org/abs/2005.08100

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

https://arxiv.org/abs/2106.13230

Why Attention and Transformers are Here to Stay for Vision

Visual Perception beyond Segmentation & Object Detection

Today

Panoptic Segmentation

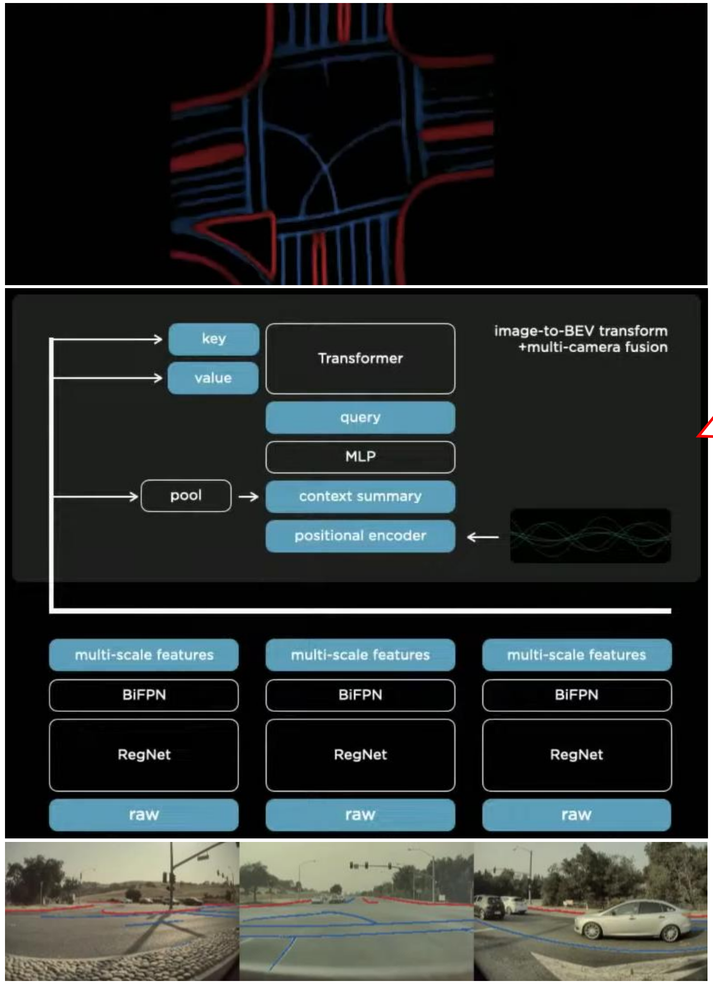
2022-…

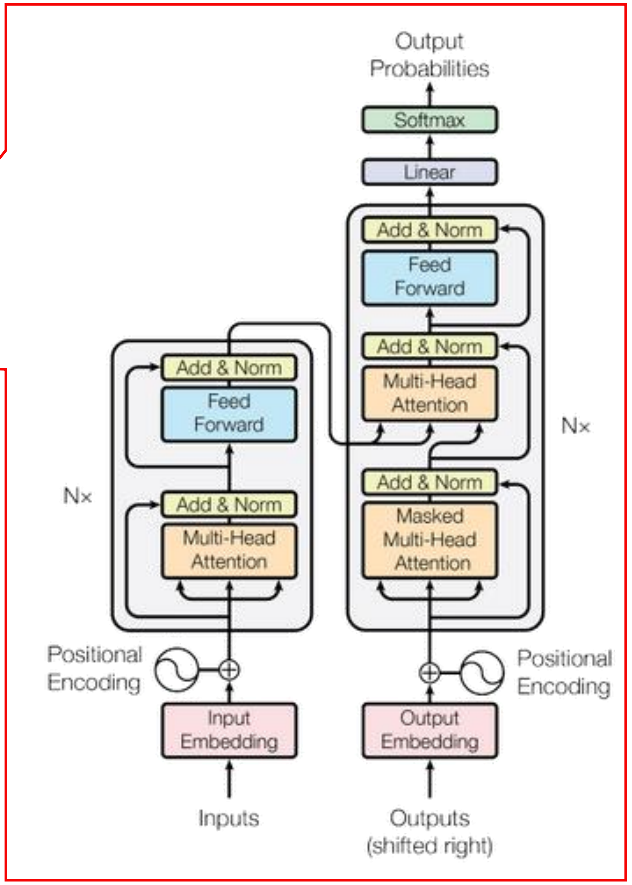
What is happening in this scene?

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

Using Transformers Make Predictions in Vector Space

Tesla AI Day, Aug-2021





- 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

Source: Tesla AI Day, 21-Aug-2021:
https://www.youtube.com/watch?v=fdtC1AxFNkk

Why Transformers are Here to Stay in Vision

- Attention based networks outperform CNN-only networks on accuracy
  - Highest accuracy required for high-end applications
  - Initially at a high compute cost
- 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

Thank You