How Transformers Are Changing the Nature of Deep Learning Models

Tom Michiels

Principal System Architect

Synopsys

Outline

- The surprising rise of transformers in vision
- The structure of attention and transformer
- Transformers applied to vision
- Why transformers are here to stay for vision

SYNOPSYS®

CNNs Dominating 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 Dominating Vision Tasks Since 2012

RCNN FRCNN SSD YOLOV2 Mask RCNN YoloV3 YoloV4/V5 EfficientDet

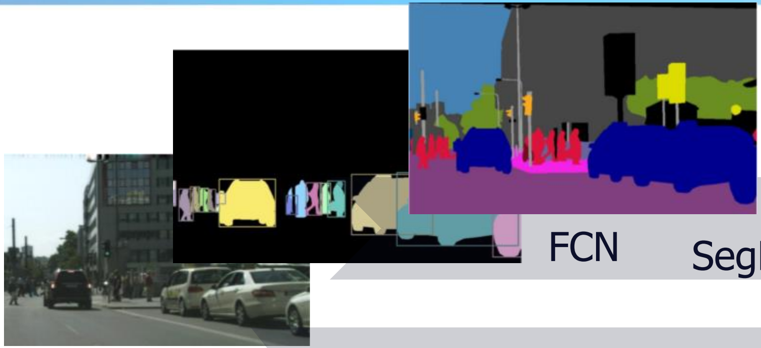
AlexNet VGG ResNet MobileNet V1 MobileNet V2 EfficientNet MobileNet v3

2012 … 2014 2015 2016 2017 2018 2019 2020 2021 2022

Object Detection

CNNs Dominating Vision Tasks Since 2012

embedded VISION SUMMIT



FCN SegNet DeepLab DSNet

RCNN FRCNN SSD YOLOV2 Mask RCNN YoloV3 YoloV4/V5 EfficientDet

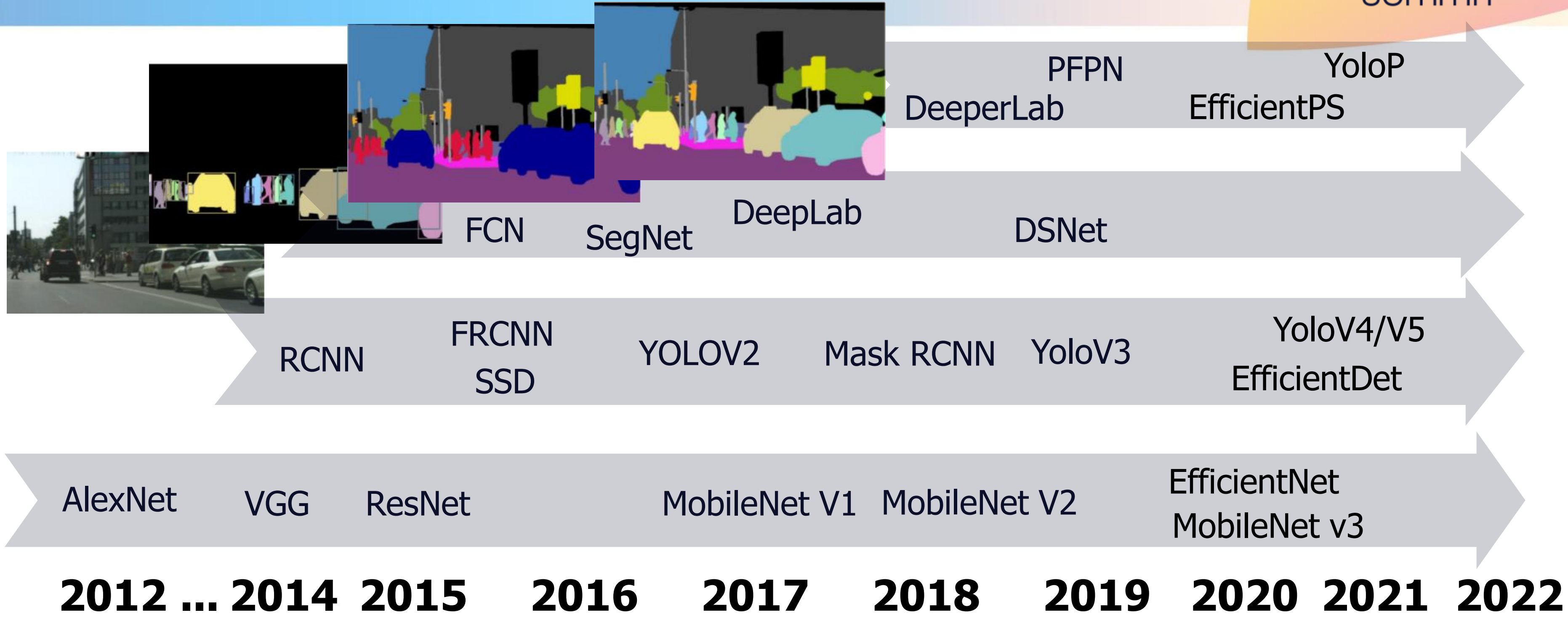
AlexNet VGG ResNet MobileNet V1 MobileNet V2 EfficientNet MobileNet v3

2012 … 2014 2015 2016 2017 2018 2019 2020 2021 2022

SYNOPSYS®

Semantic Segmentation

CNNs Dominating Vision Tasks Since 2012

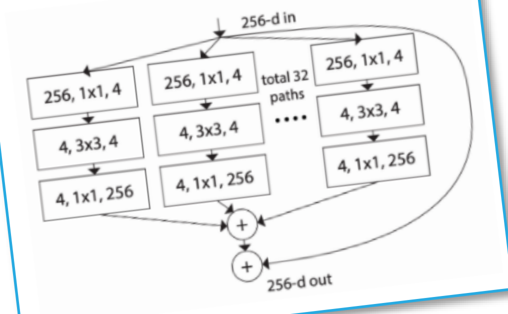


Panoptic Vision

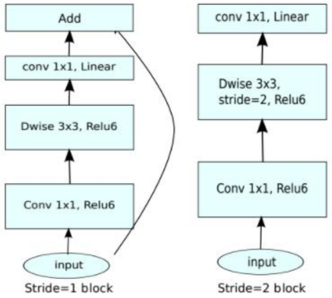
A Decade of CNN Development…

embedded VISION SUMMIT

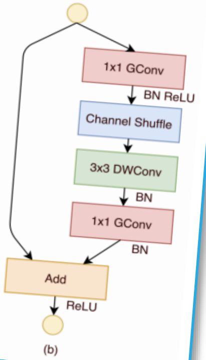
Inception



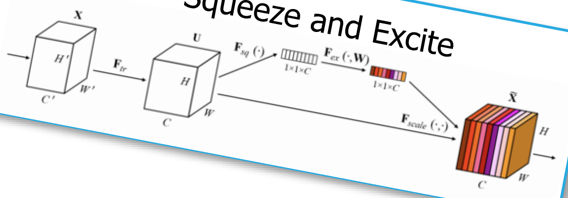
Inverted Residual Blocks



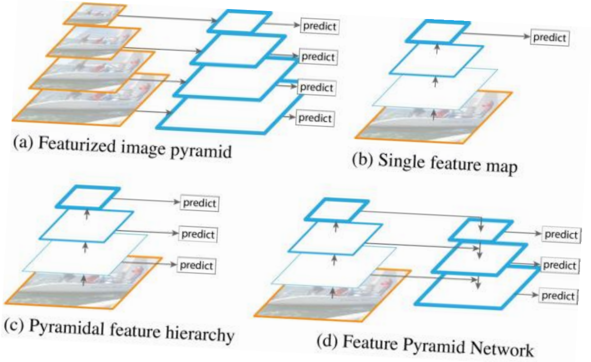
Shufflenet



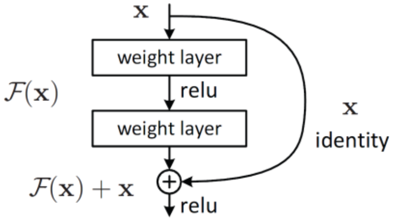
Squeeze and Excite



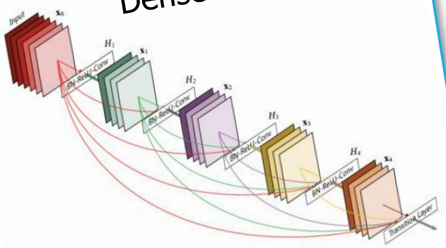
Feature Pyramid



Residual Connection



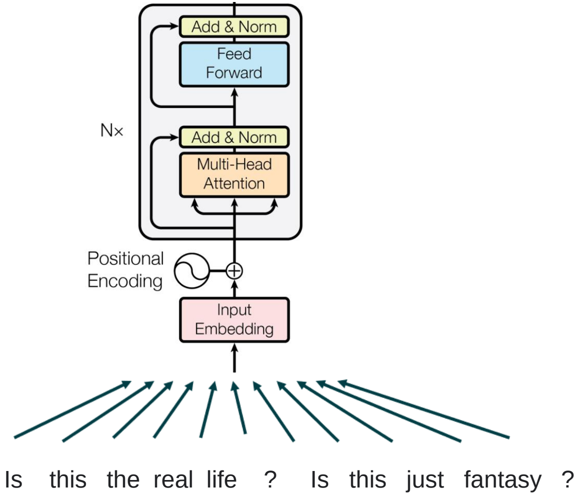
DenseNet



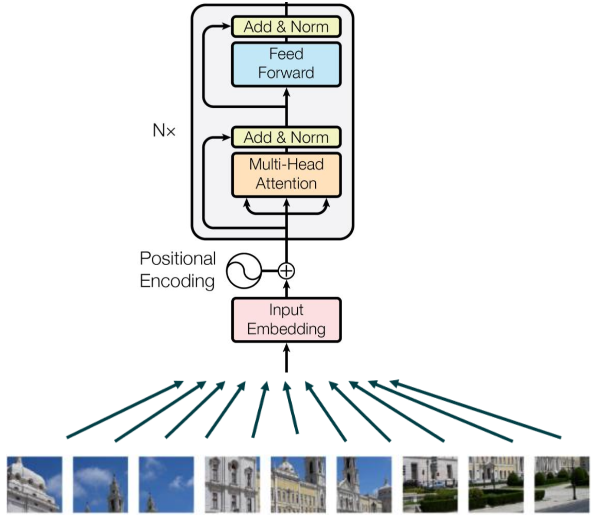
SYNOPSYS®

Beaten in Accuracy by Transformers

Transformer, a model designed for natural language processing

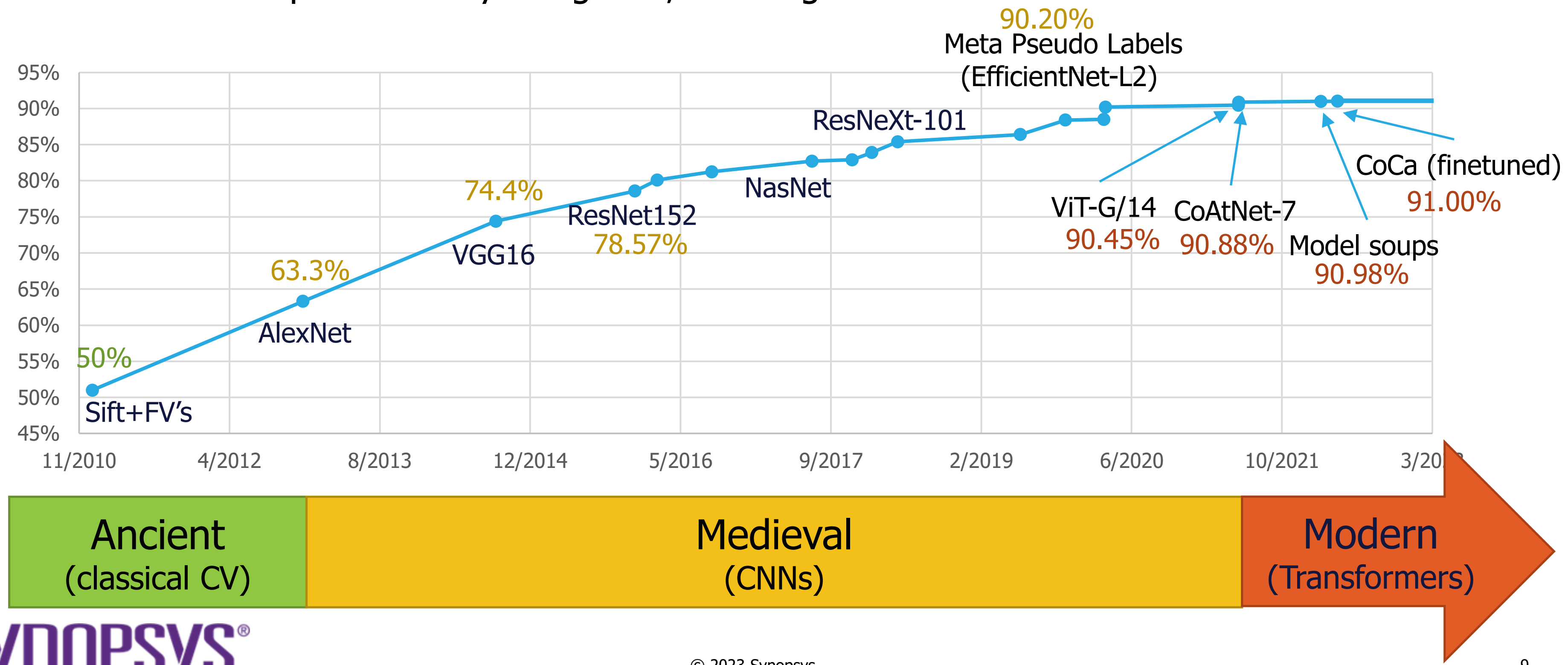


… without any modifications applied to image patches, beats the highly specialized CNNs in accuracy



Accuracy Records on ImageNet

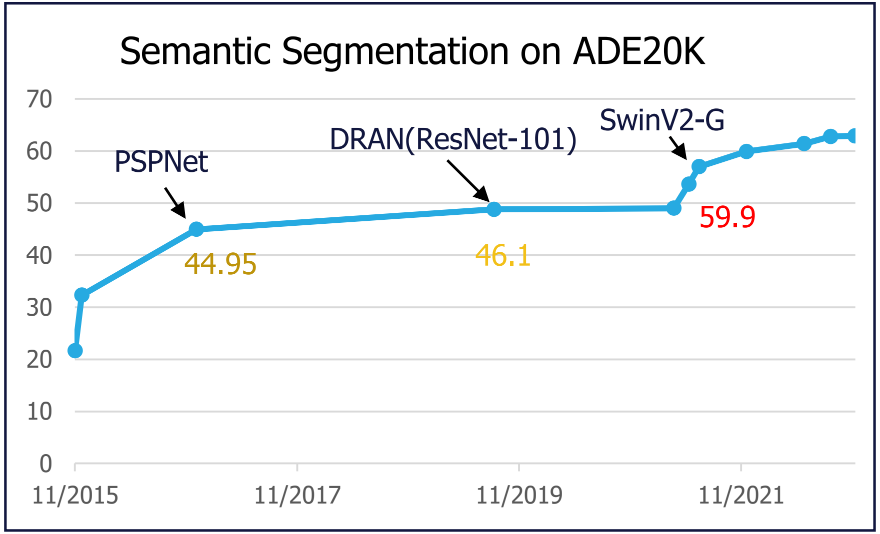
State-of-the-art Top-1 Accuracy ImageNet, entering a new era?

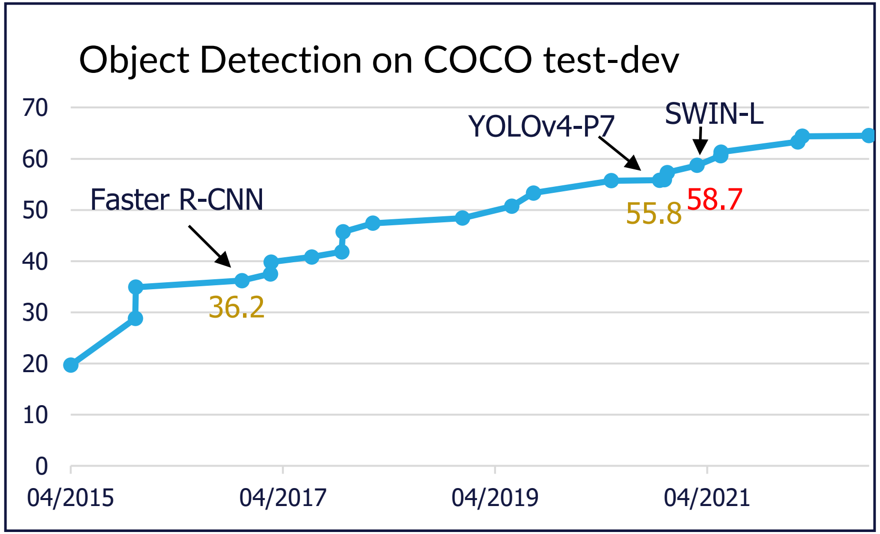


SYNOPSYS®

Transformers in Other Vision Tasks

embedded VISION SUMMIT

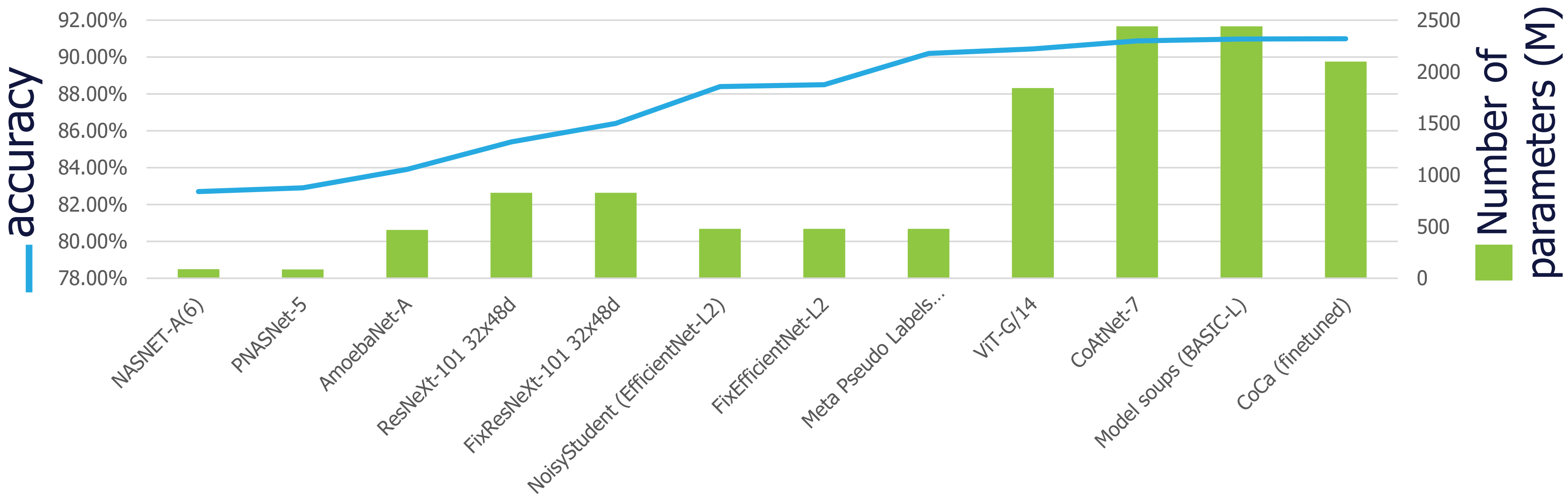




State-of-the-Art of other Vision tasks are dominated by transformers

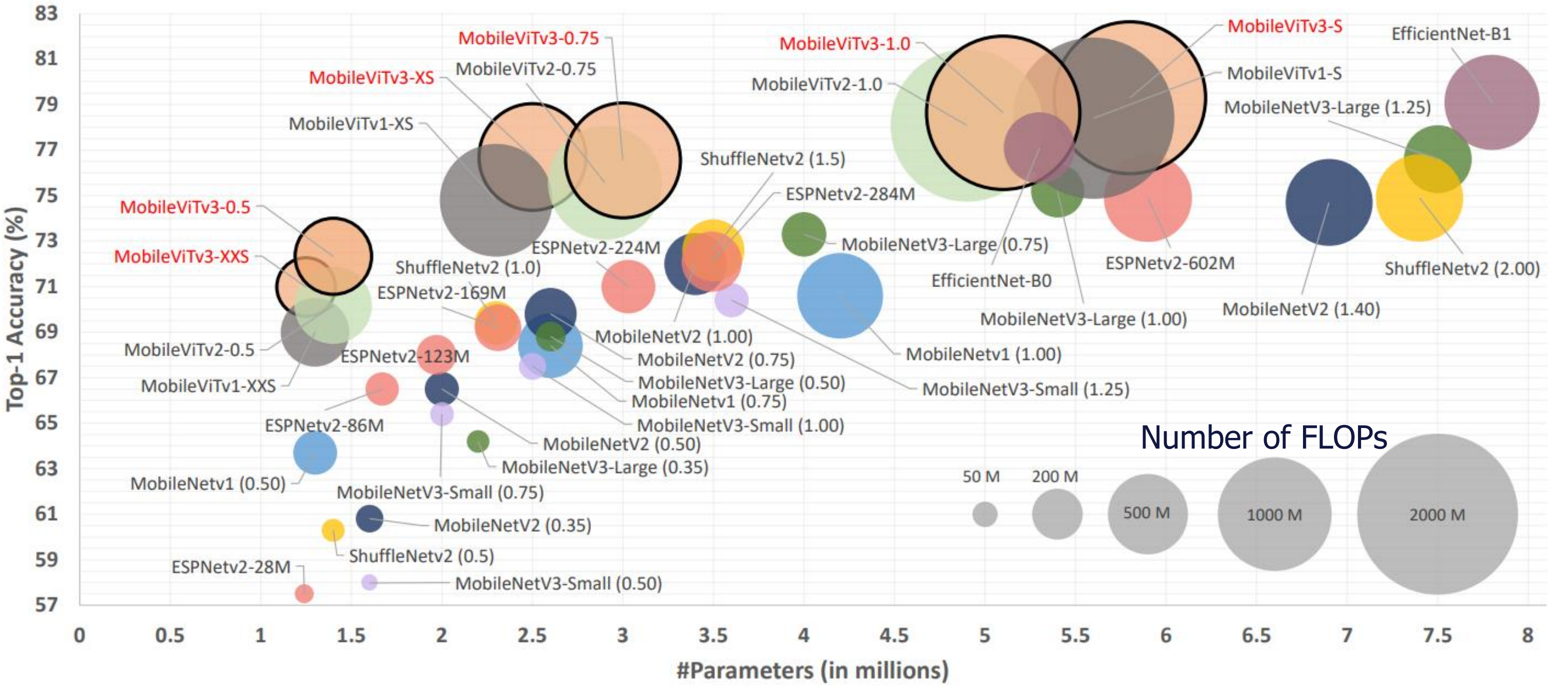
SYNOPSYS®

But These State-of-the-Art Models Are Huge!



Is the state-of-the-art really relevant for embedded applications?

Compact Transformers versus CNNs



source MobileViTv3 https://arxiv.org/abs/2209.15159

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

embedded VISION SUMMIT

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

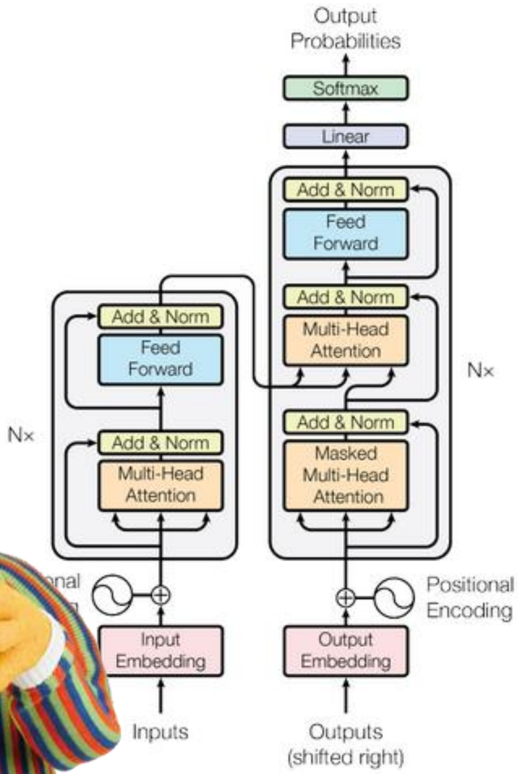
Model	# Params ⇓	FLOPs ⇓	Top-1 ⇑	Inference Time (ms) iPhone12 - CPU	Inference Time (ms) iPhone12 - Neural Engine	
MobileNetv2	3.5 M	**0.3 G**	73.3	**7.50 ms**	**0.92 ms**	CPU/NNE = 8.1X
DeIT	5.7 M	1.3 G	72.2	28.15 ms	10.99 ms	
PiT	4.9 M	0.7 G	73.0	24.03 ms	10.56 ms	
MobileViT (Ours)	**2.3 M**	0.7 G	**74.8**	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 CNNs
  - The authors expect that future AI accelerators will better support transformers

SYNOPSYS®

The Structure of Attention and Transformer

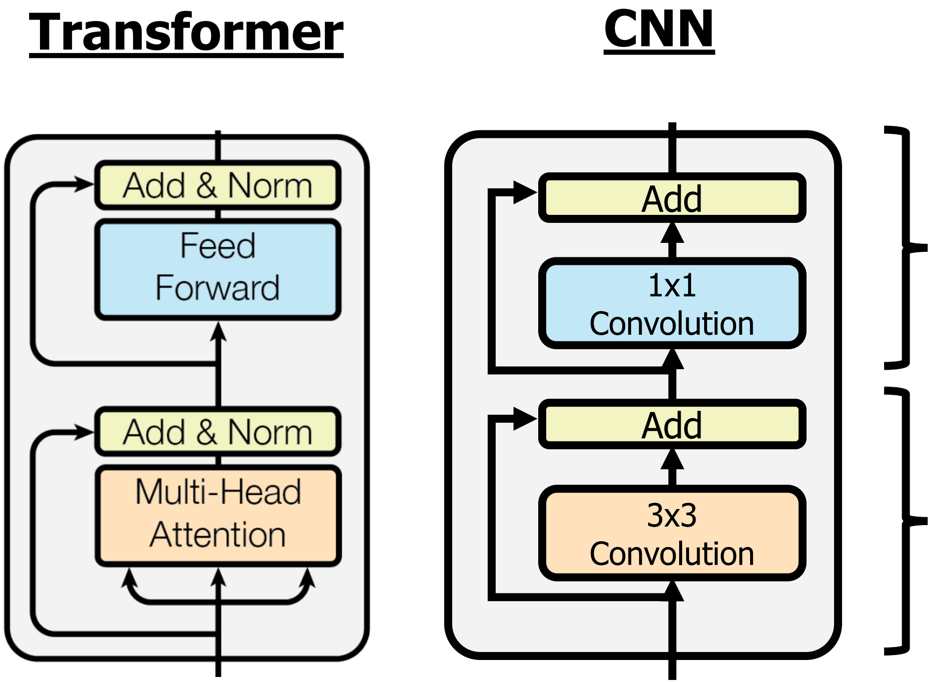
Bert and Transformers



- 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

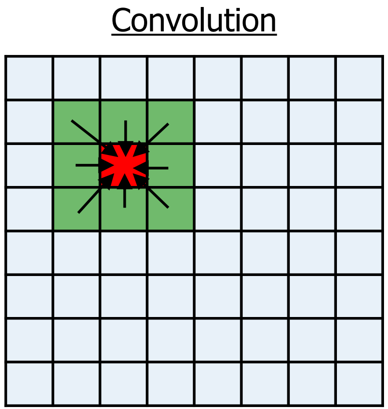
Convolutions, Feed Forward, and Multi-Head Attention

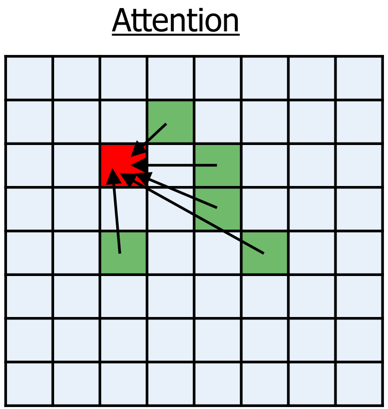


- The feed forward layer of the transformer is identical to a 1x1 convolution
- In this part of the model, no information is flowing between tokens/pixels
- Multi-head attention and 3x3 convolution layers are the layers responsible for mixing information between tokens/pixels

Convolutions as Hard-Coded Attention

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



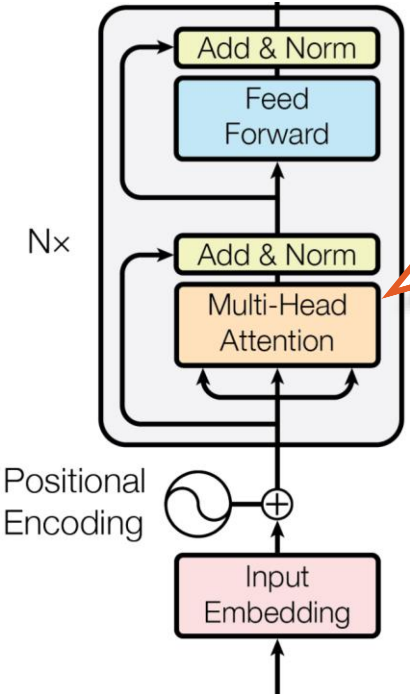


Convolutions mix in features from tokens based on fixed spatial location

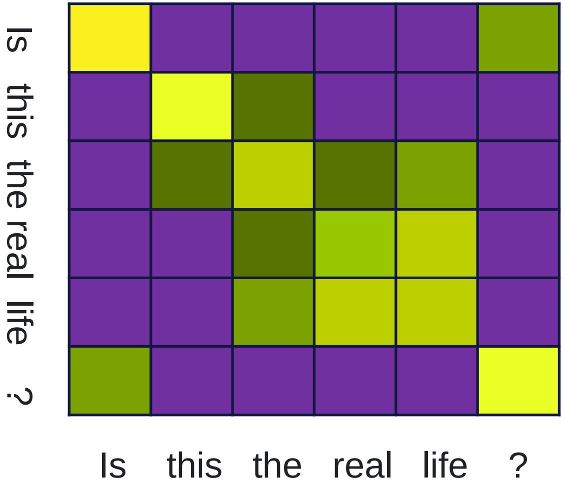
Attention mix in features from tokens based on learned attention

The Structure of a Transformer: Attention

Multi-Head Attention

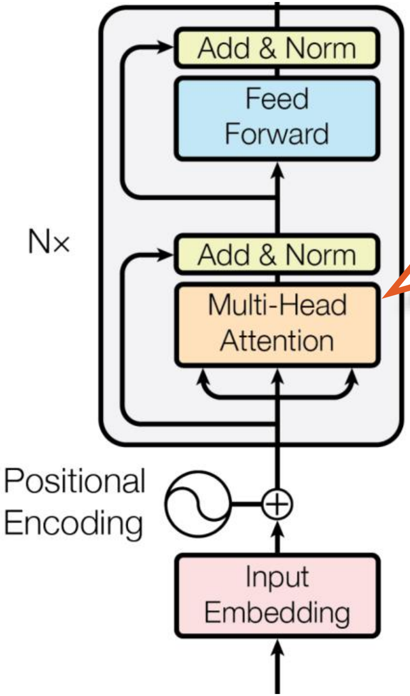


Attention: Mix in features of other tokens

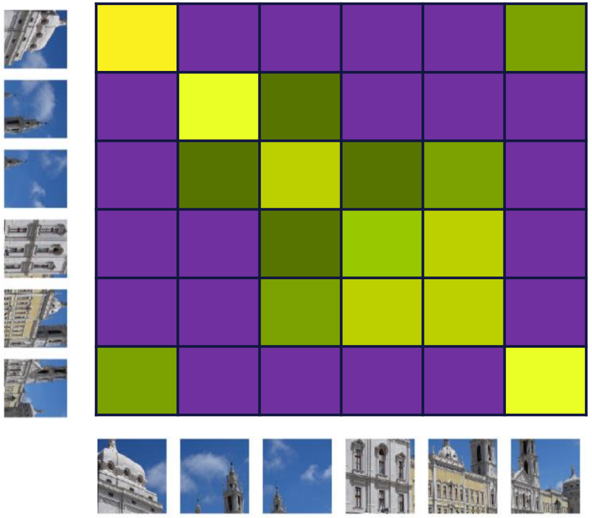


The Structure of a Transformer: Attention

Multi-Head Attention

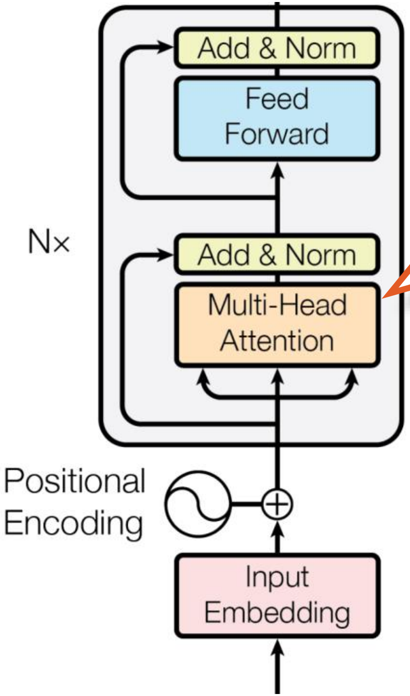


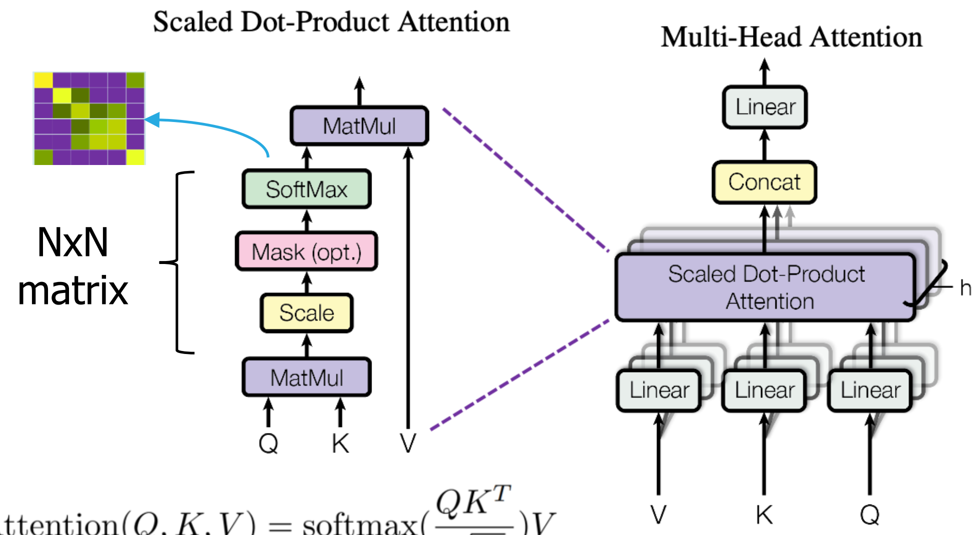
Attention: Mix in features of other tokens



The Structure of a Transformer: Attention

Multi-Head Attention



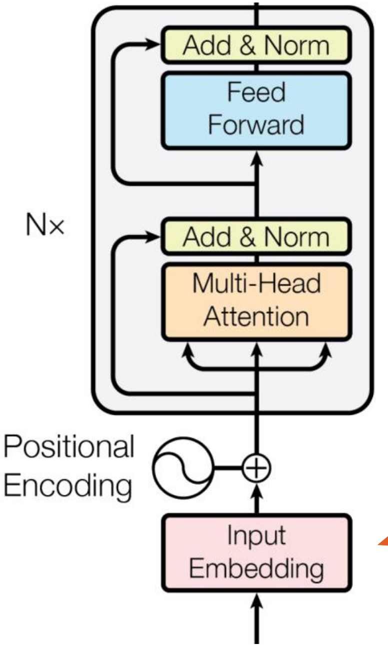


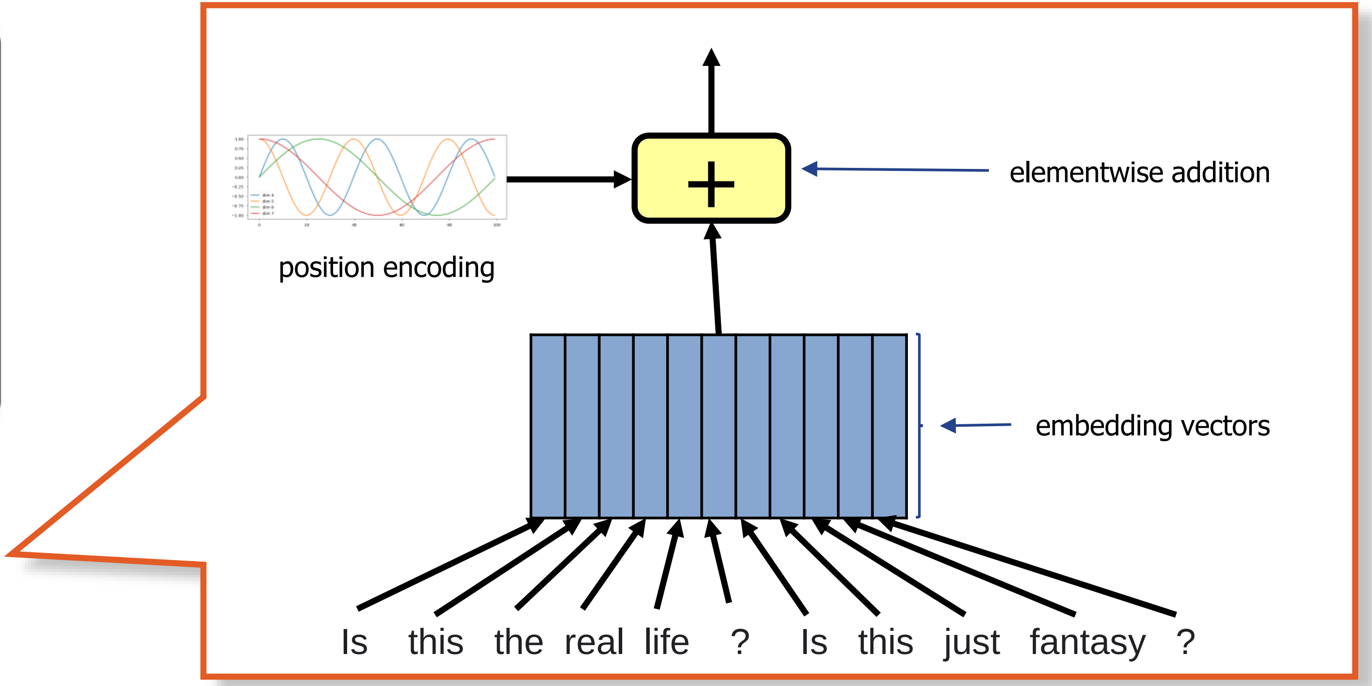
$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

N: number of tokens

The Structure of a Transformer: Embedding

Embedding of input tokens and the positional encoding



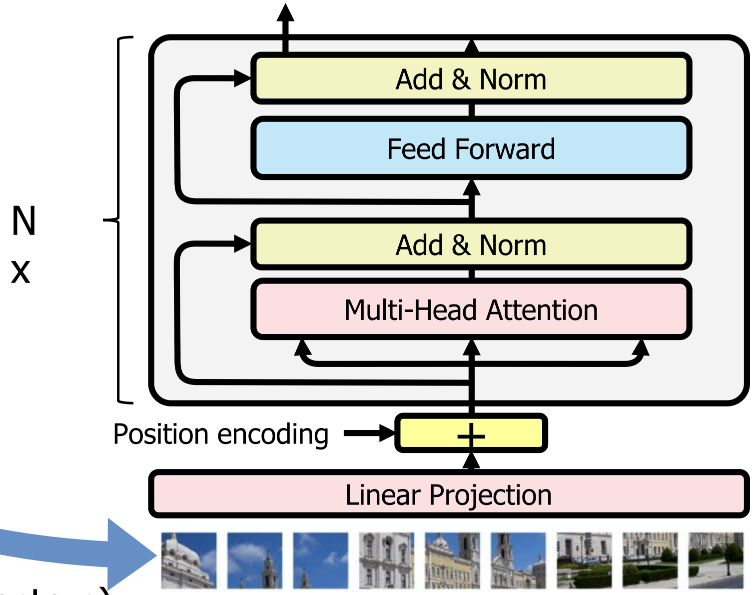


Applying Transformers to Vision Tasks

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



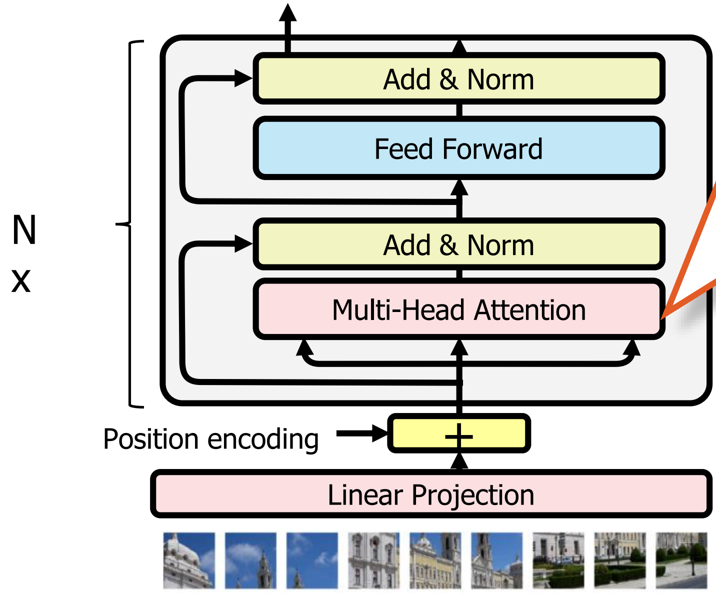
Pixels in a tile are flattened into tokens (vectors) that feed in the transformer

Vision transformers are **best-known method for image classification**

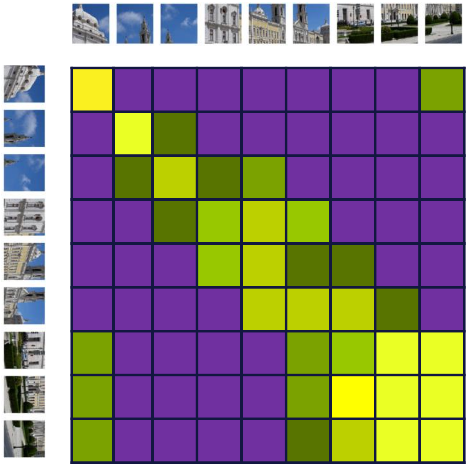
They are beating convolutional neural networks in **accuracy** and **training time,** but **not in inference time**

(*) 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

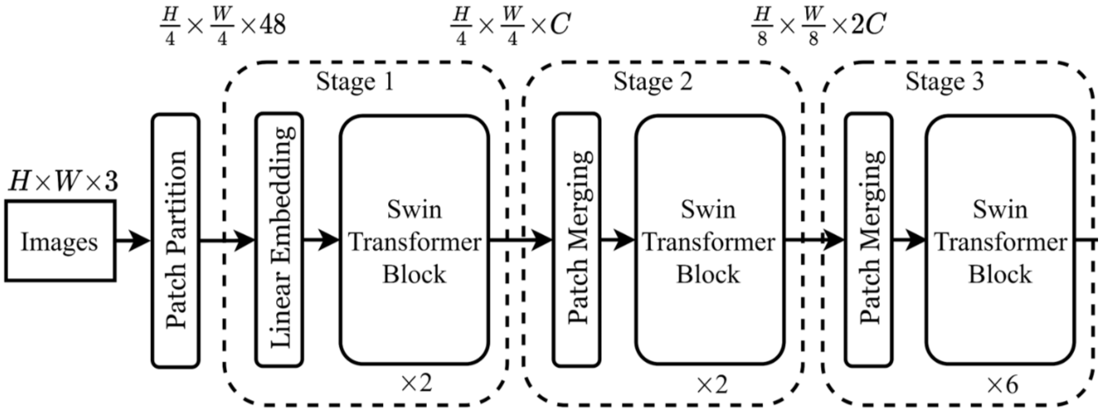
Swin Transformers

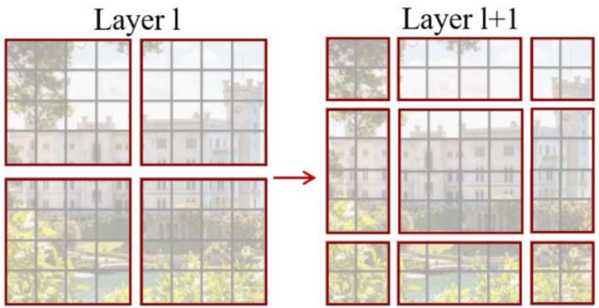
Hierarchical Vision Transformer Using Shifted Windows (*)

Adaptation makes transformers scale for larger images:
1. Shifted window attention
2. Patch-merging

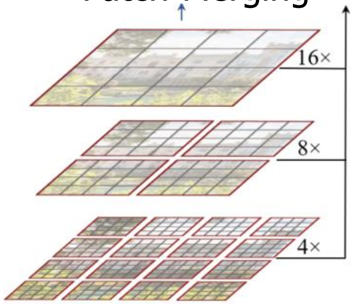
State of the art for
- Object detection (COCO)
- Semantic segmentation (ADE20K)

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



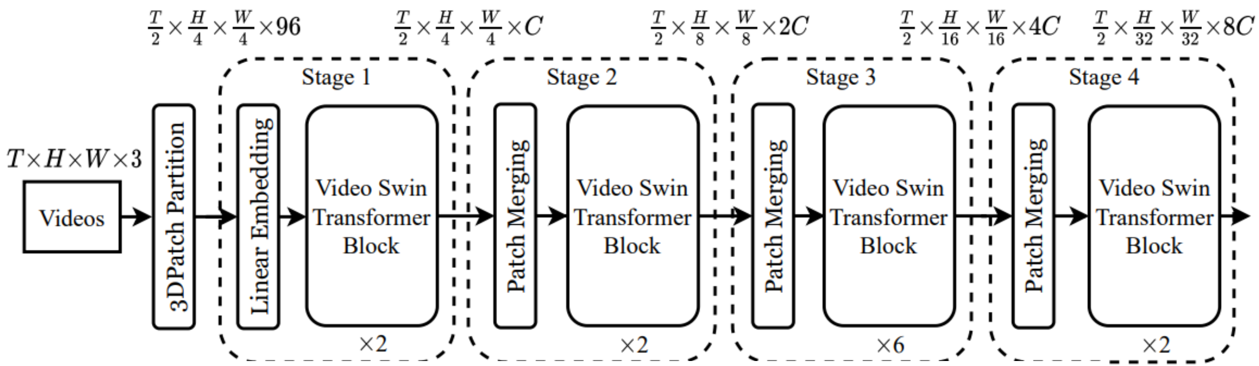


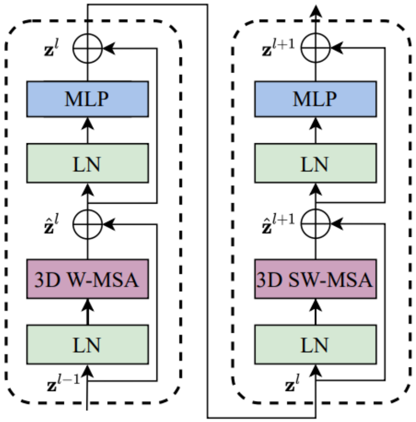
Patch-Merging

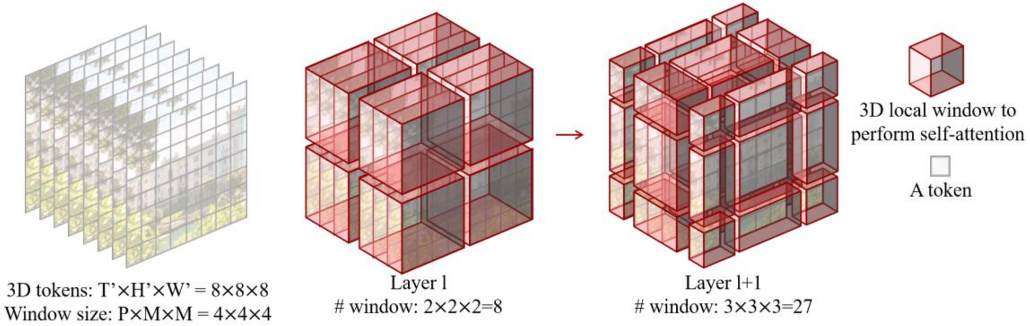


Action Classification with Transformers

Video Swin Transformer







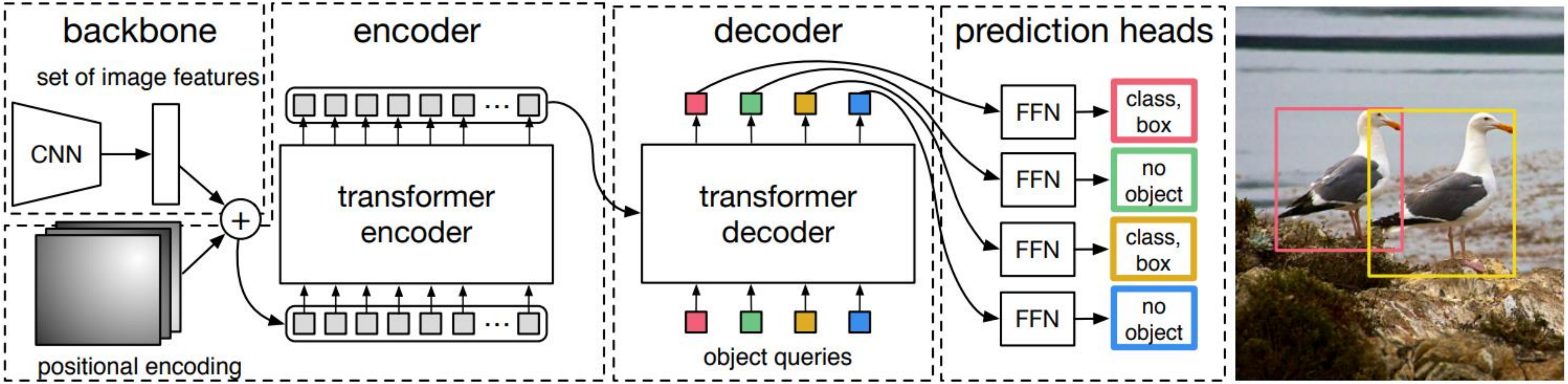
Video Swin Transformers extend the (shifted) window to three dimensions (2D spatial + time)

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

https://arxiv.org/abs/2106.13230

Object Detection with Transformers

End-to-End Object Detection with Transformers (Facebook 2020)

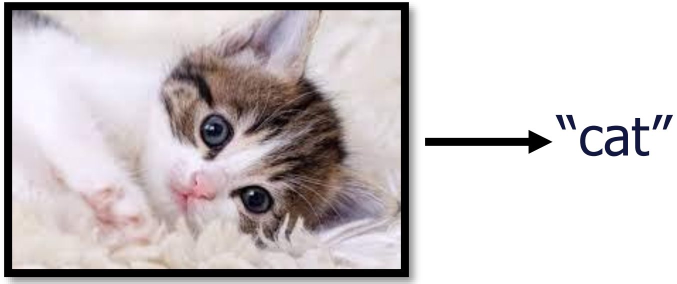


DETR uses a CNN (ResNet-50) as a backbone
Off-the-shelf transformer encoder and decoder
Trained Object Queries retrieve possible candidates for objects

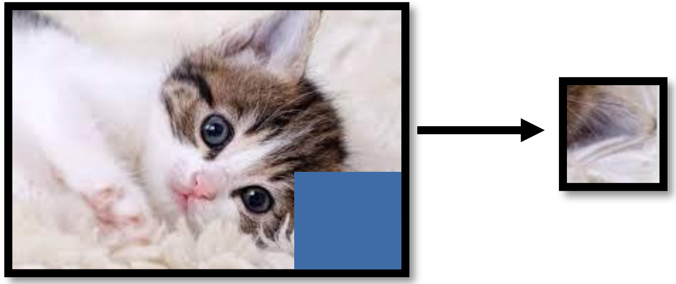
https://arxiv.org/abs/2005.12872

Training Vision Transformers

- **More data** required to train a transformer to overcome the lack of inductive bias of convolution
- Vision Transformers take **significantly less training time** than comparable CNN's
- **Self-supervised Pre-Training** for Vision Transformers



Supervised learning

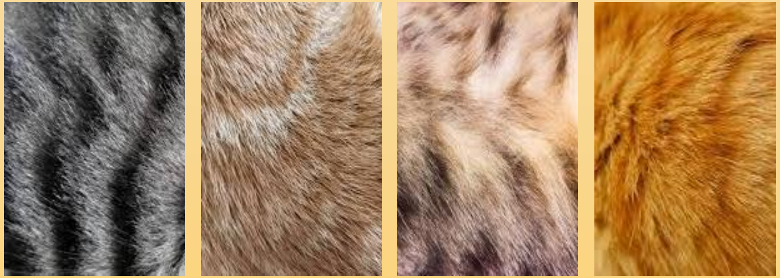


Self-Supervised learning

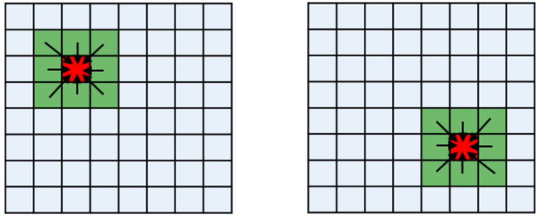
Why Attention and Transformers are Here to Stay for Vision

Inductive Bias of CNNs

Recognizing Cat Fur



Recognizing a whole Cat

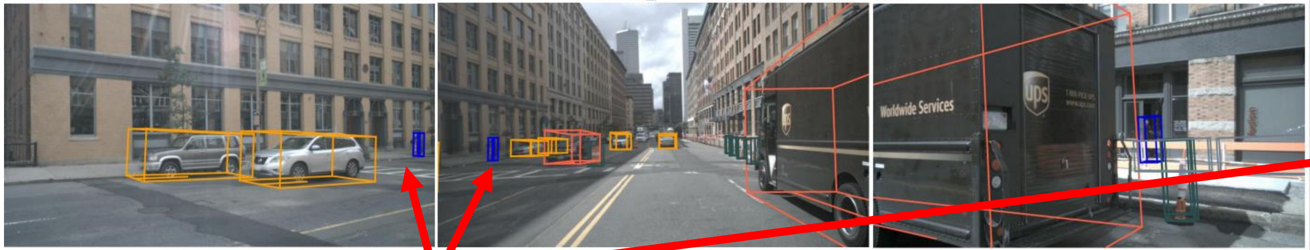


Convolution use the same weights
for processing every pixel

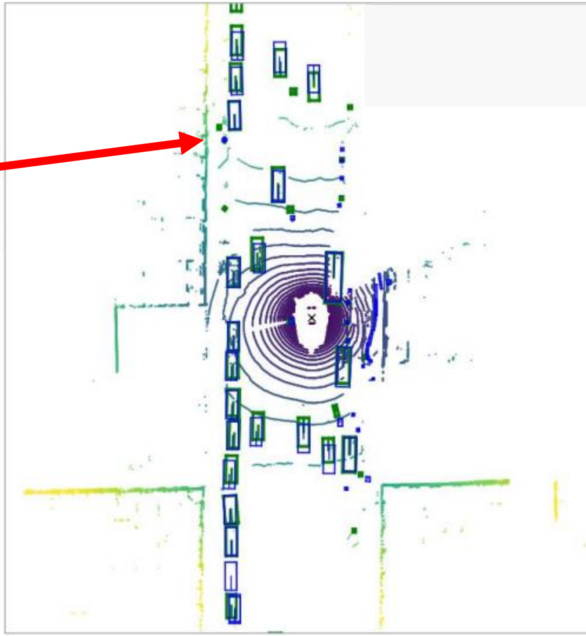
The inductive bias of a convolution, is more helpful for recognizing low level features like cat fur than for recognizing more complex objects like cats

Inductive Bias of CNNs and Sensor Fusion

Camera front Left   Camera Front   Camera Front Right



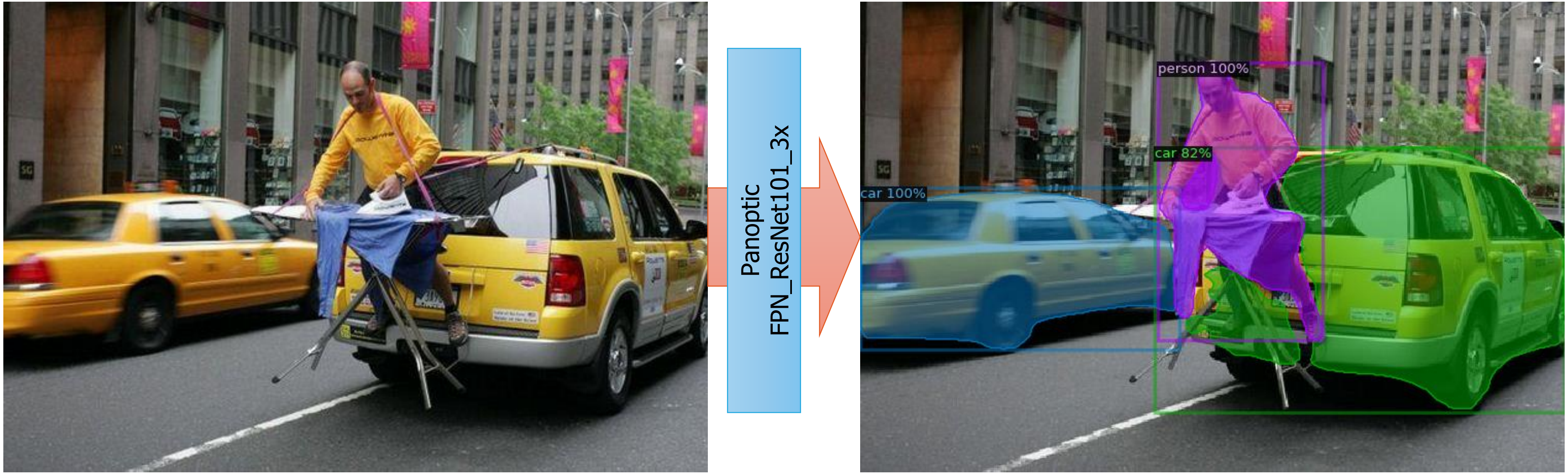
Lidar TOP



Fused in a single object

- Many real-life applications use multiple sensors
- Transformers excel in sensor fusion
- The inductive bias of CNNs designed for grid-like structures limits their effectiveness in sensor fusion

Panoptic Segmentation Using CNNs



Panoptic Segmentation combines predictions from both instance and semantic segmentation in a unified output

Image source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg
Model Used: Detectron2 - COCO-PanopticSegmentation/panoptic_fpn_R_101_3x

Multimodal Transformers

LLaVA Language and

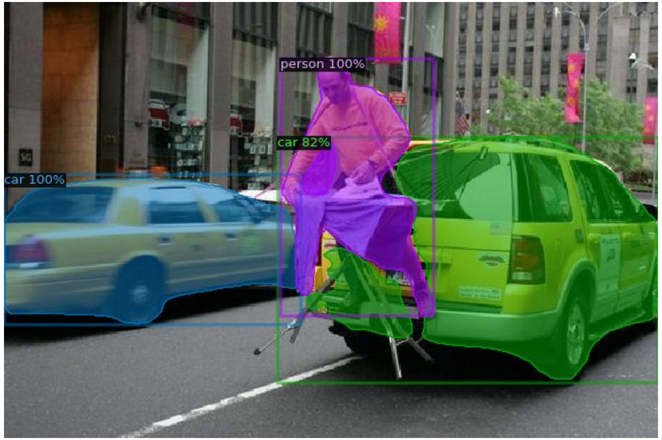
PN_ResNet101_3x

User: What is unusual about this image?

LLaVA: The unusual aspect of the image is that a man is ironing clothes on the back of a yellow minivan while it is on the road. This is an unconventional and unsafe place to perform such an activity, as ironing clothes typically requires a stable surface and appropriate equipment. Ironing clothes in a moving vehicle could lead to potential hazards for both the person doing the ironing and other road users.

https://arxiv.org/pdf/2304.08485.pdf

How Much Intelligence Do We Need?



Panoptic Segmentation

User: What is unusual about this image?

LLaVA: The unusual aspect of the image is that a man is ironing clothes on the back of a yellow minivan while it is on the road. This is an unconventional and unsafe place to perform such an activity, as ironing clothes typically requires a stable surface and appropriate equipment. ….

Multi-modal language/vision

How much (artificial) intelligence is required for real-life applications?

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(**)
- Real-life vision: demands beyond CNN inductive bias
  - Scene understanding needs common-sense knowledge that may not be learned by vision alone
  - Sensor fusion: complex geometrical mappings are ill-suited for CNN bias

(*) 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

ARC NPX6 NPU IP

www.synopsys.com/npx

Visit Synopsys Booth 309

- **Partner demo:** Visionary.ai True Night Vision SW ISP
- Meet with Synopsys executives and experts
  - Discuss emerging neural network architectures like transformers and vision/object detection for safety-critical automotive SoCs
  - Learn about the latest in practical technology to bring visual intelligence into embedded systems, mobile apps, cars, and PCs