Transformers for Computer Vision

Alexey Dosovitskiy

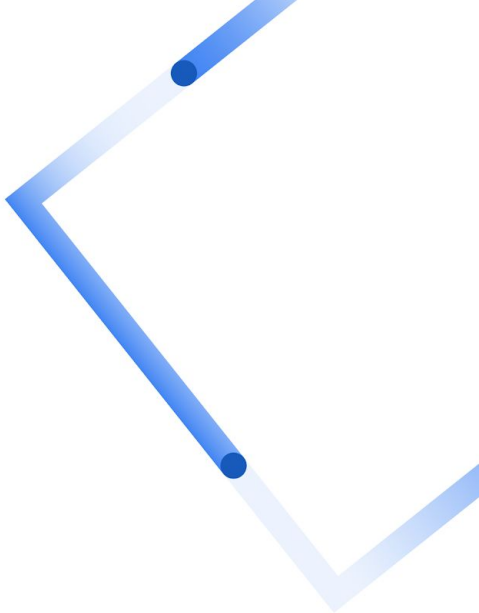
EEML summer school
July 7th 2021, Budapest (virtually)

Google Research

Outline

- Background
  - Models
  - Large-scale transfer
- Transformers for image classification
- Beyond classification
- Beyond images

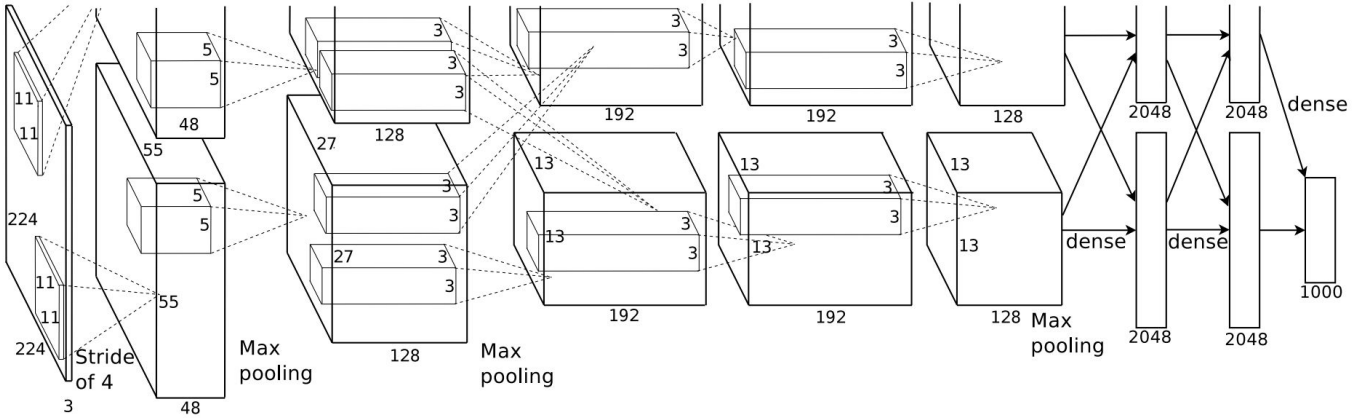
Background: Models



Google Research

AlexNet

- AlexNet (2012) - first big success of deep learning in vision*

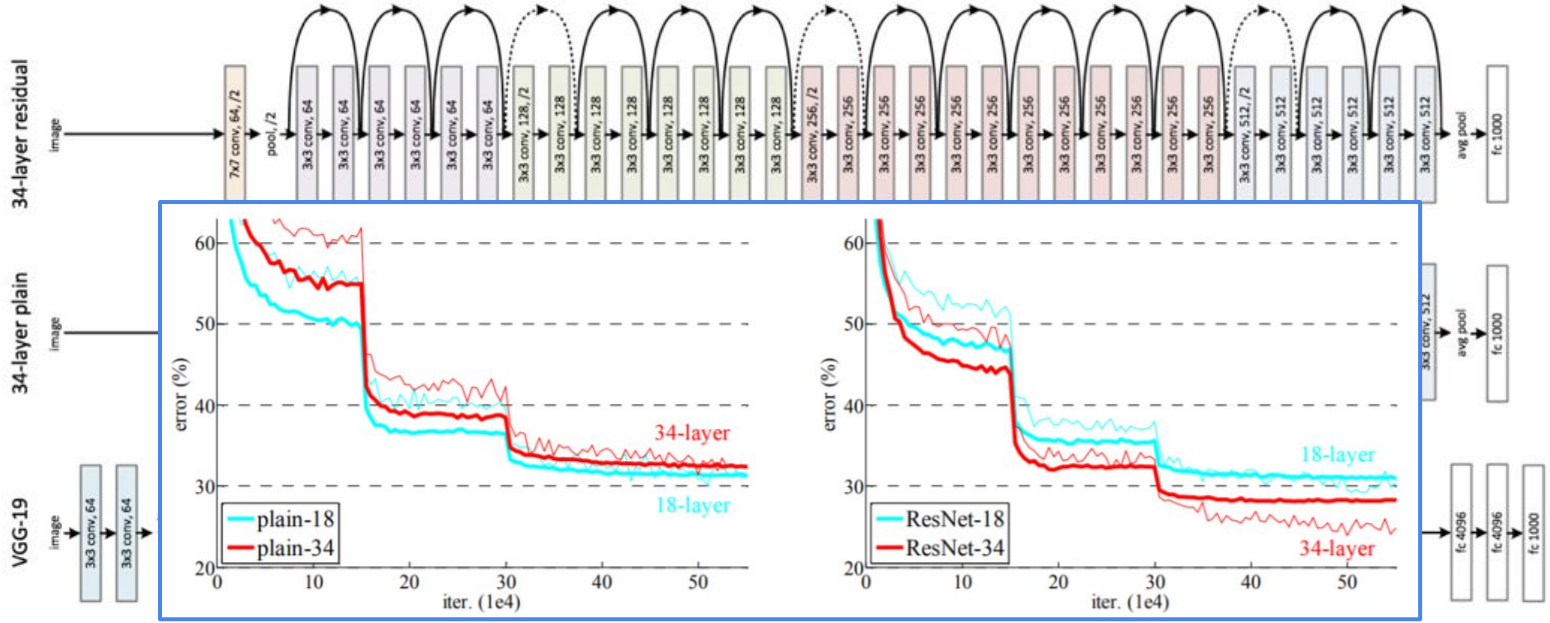


* ConvNets had previously shown good results on specialized dataset like handwritten digits (LeCun et al.) or traffic signs (Ciersan et al.), but not on large and diverse “natural” datasets

Krizhevsky et al., *ImageNet Classification with Deep Convolutional Neural Networks*, NIPS 2012

ResNet

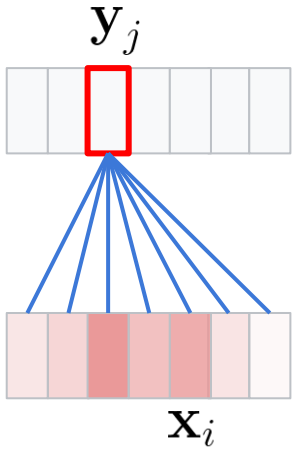
- ResNet (2015) - make deep models train well by adding residual connections



Kaiming He et al., *Deep Residual Learning for Image Recognition*, CVPR 2016

Self-attention

- Each of the tokens (=vectors) attends to all tokens
  - Extra tricks: learned key, query, and value projections, inverse-sqrt scaling in the softmax, and multi-headed attention (omit for simplicity)
- It's a set operation (permutation-invariant)
  - ...and hence need "position embeddings" to "remember" the spatial structure
- It's a global operation
  - Aggregates information from all tokens



$$\boldsymbol{\alpha}_j = softmax(\frac{K\mathbf{x}_1 \cdot Q\mathbf{x}_j}{\sqrt{d_K}}, \ldots, \frac{K\mathbf{x}_n \cdot Q\mathbf{x}_j}{\sqrt{d_K}})$$

$$\mathbf{y}_j = \sum_{i=1}^{n} \alpha_{ji} V\mathbf{x}_i$$

Simplified! Multi-headed attention not shown

Vaswani et al., *Attention Is All You Need*, NeurIPS 2017

Transformer

- Transformer “encoder”
  - A stack of alternating self-attention and MLP blocks
  - Residuals and LayerNorm

- Transformer “decoder” (not shown)
  - A slightly more involved architecture useful when the output space is different from the input space (e.g. translation)

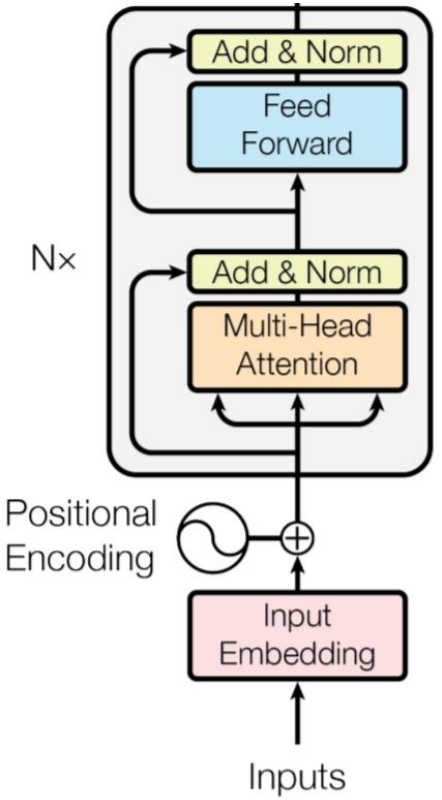
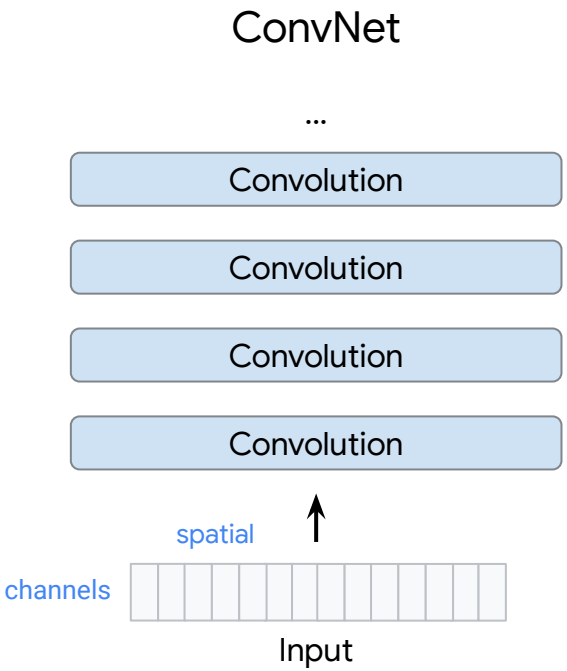


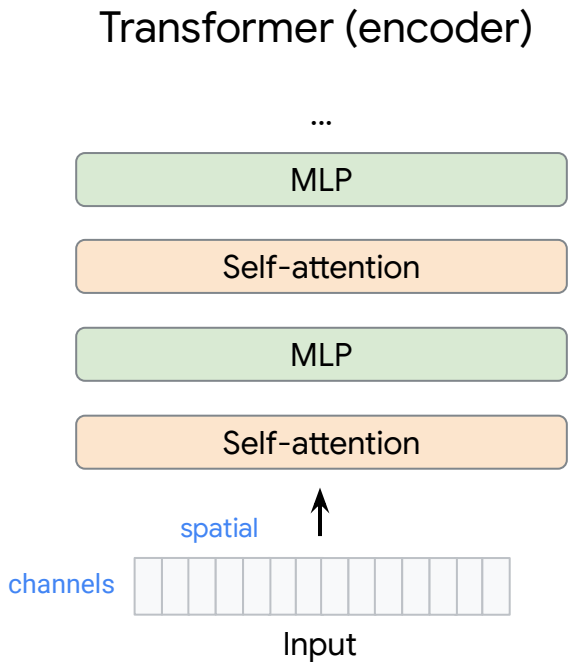
Figure from Vaswani et al.

Vaswani et al., *Attention Is All You Need*, NeurIPS 2017

ConvNet vs Transformer



Convolutions (with kernels > 1x1) mix both the channels and the spatial locations



MLPs (=1x1 convs) only mix the channels, per location

Self-attention mixes the spatial locations (and channels a bit)

BERT

- Transformers pre-trained self-supervised perform great on many NLP tasks
  - Masked language modeling (MLM)
  - Next sentence prediction (NSP)

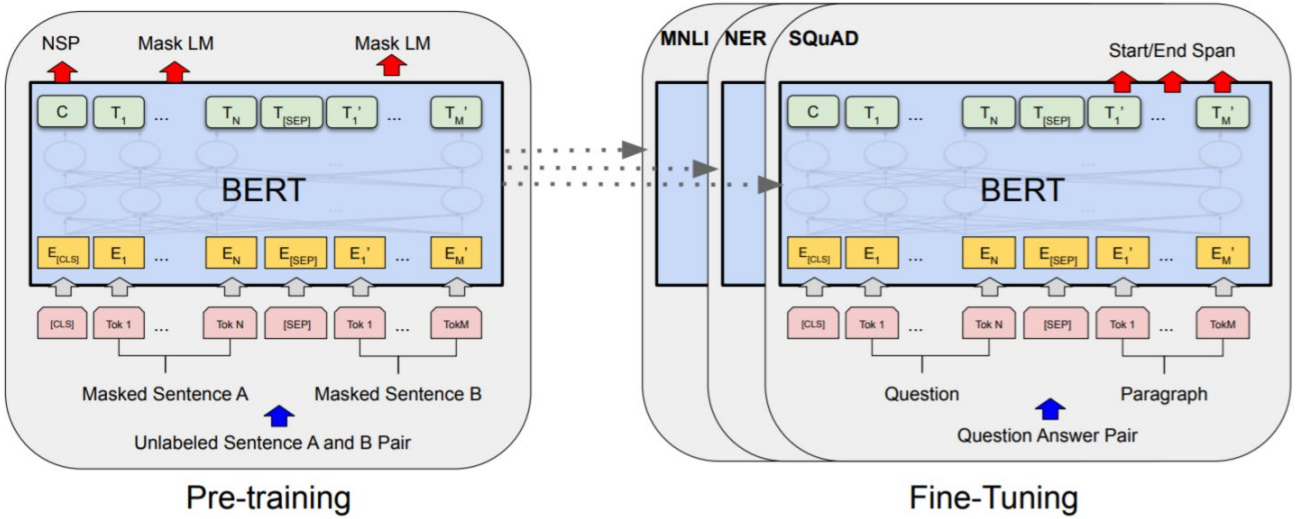


Figure from Devlin et al.

Devlin et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, arXiv 2018

T5, GPT-3

- T5 (Text-to-Text Transfer Transformer)
  - Formulate many NLP tasks as text-to-text
  - Pre-train a large transformer BERT-style and show that it transfers really well
- GPT-3 (Generative Pre-Training)
  - Same basic approach, but generative pre-training and even larger model
  - Zero-shot transfer to many tasks: no need for fine-tuning!

Large-scale self-supervised pre-training “solved”* NLP

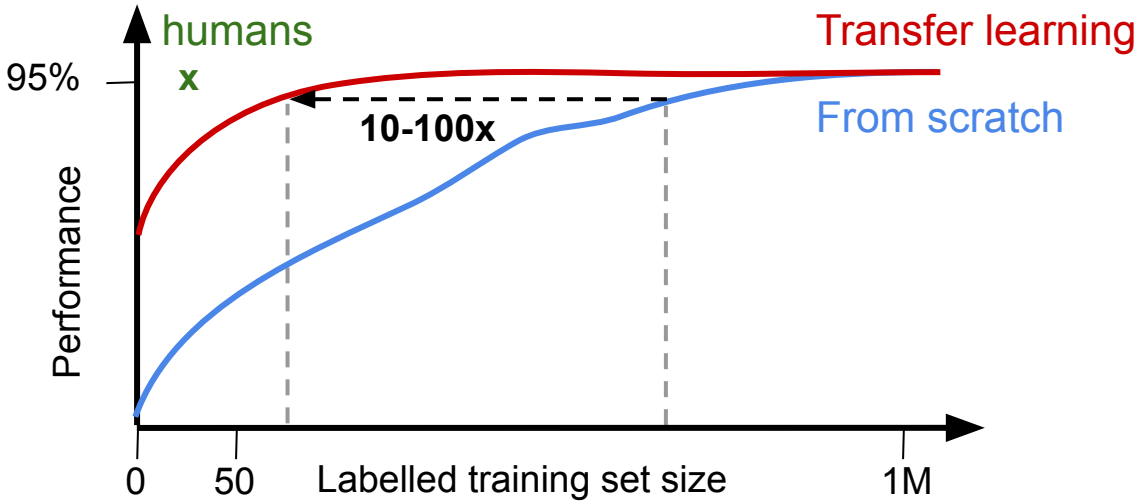
*at least made some really impressive progress

Raffel et al., *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*, JMLR 2020
Brown et al., *Language Models are Few-Shot Learners*, NeurIPS 2020

Background: Large-scale transfer for image classification

Google Research

Why transfer?



This and some of the following slides by Neil Houlsby

Large scale representation learning

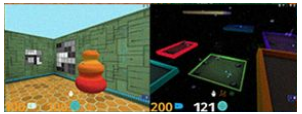
Use **any** available data

Expensive, but **only once**

Transfer to new tasks

Cheap, needs **little data**

Works on many tasks/metrics (hopefully!)



“Pre-training” or “Upstream” task

“Downstream” task

Google Research

Pre-training Datasets

"ImageNet"
ILSVRC2012 ImageNet challenge

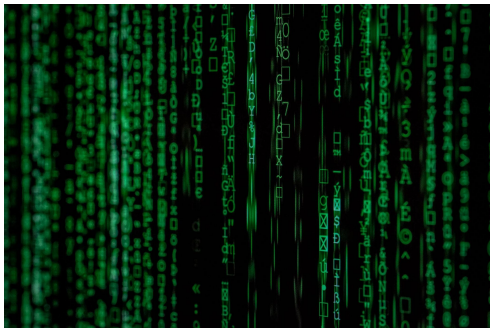
1.3M images
1k classes

"ImageNet-21k"
The full Imagenet dataset

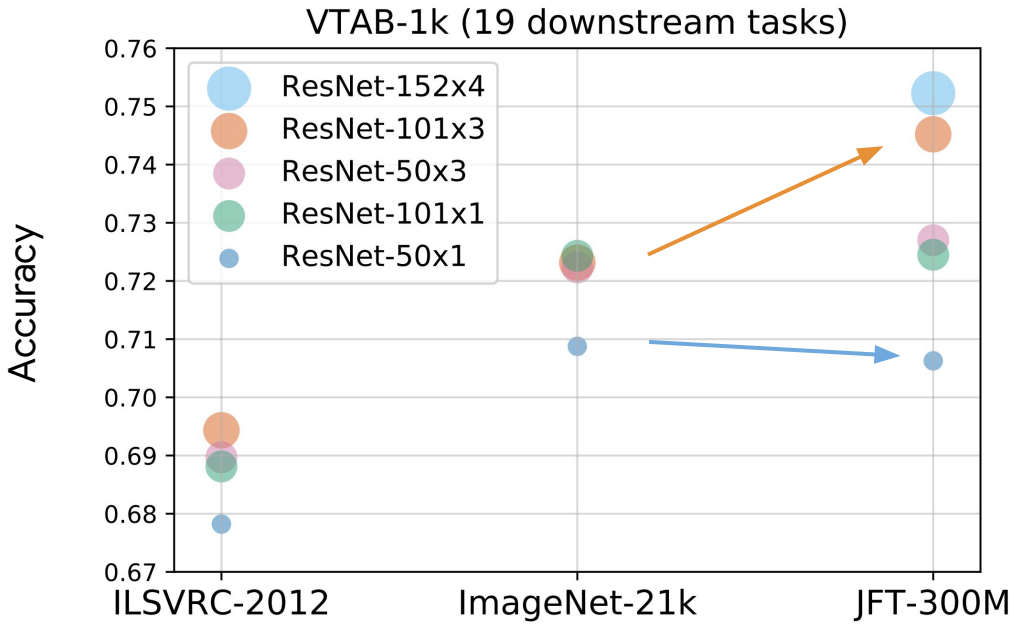
14M images
21k classes

"JFT-300M"
Google internal dataset

300M images
18k classes

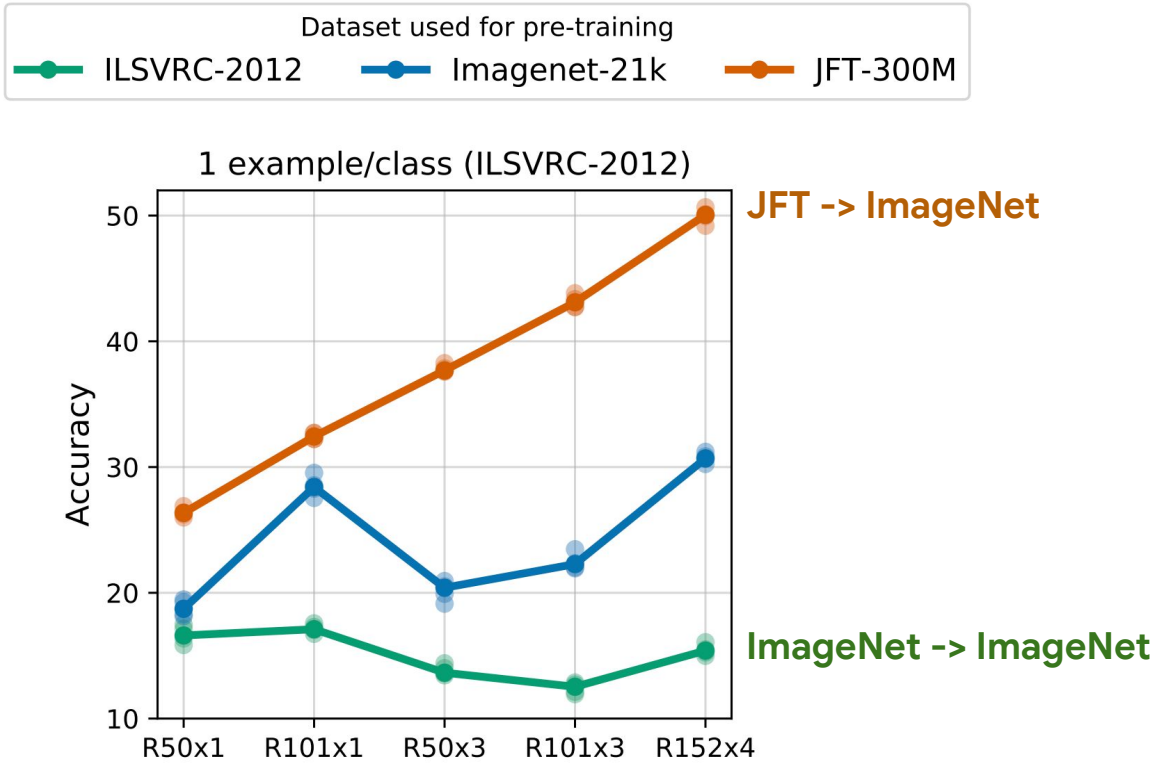


Scaling classic ResNet: Big Transfer “BiT”



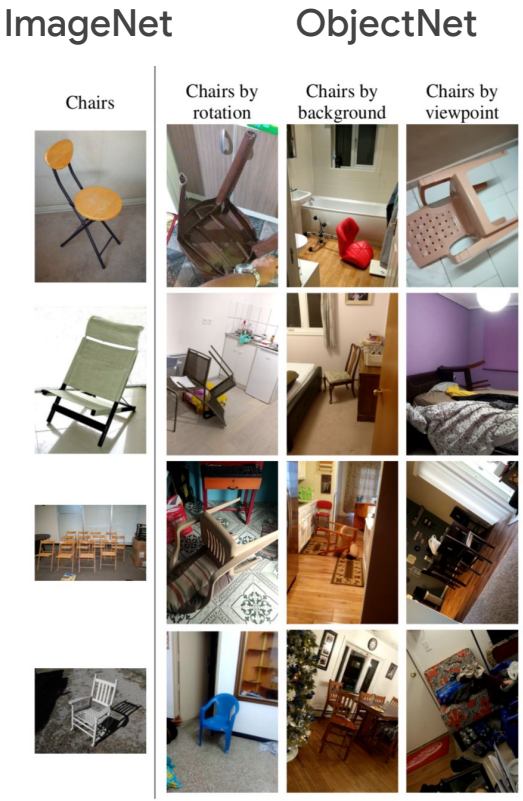
Kolesnikov et al., *Big Transfer (BiT): General Visual Representation Learning*. ECCV 2020

Large pre-training for few-shot tasks



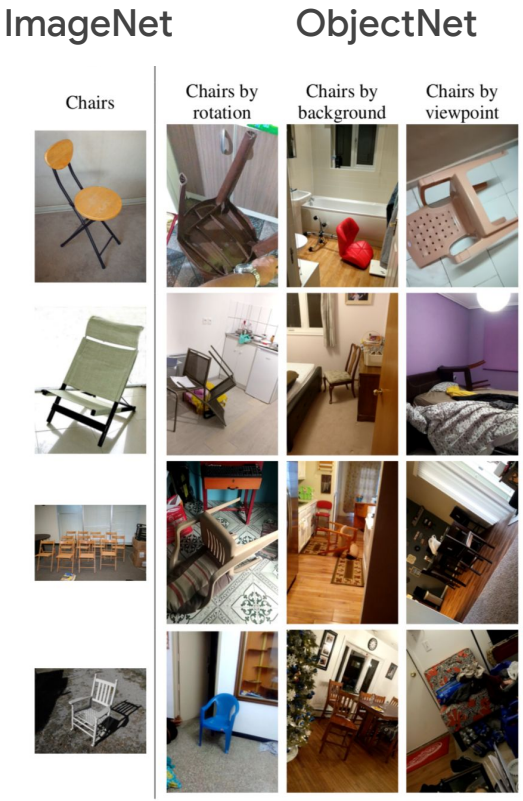
Kolesnikov et al., *Big Transfer (BiT): General Visual Representation Learning*. ECCV 2020

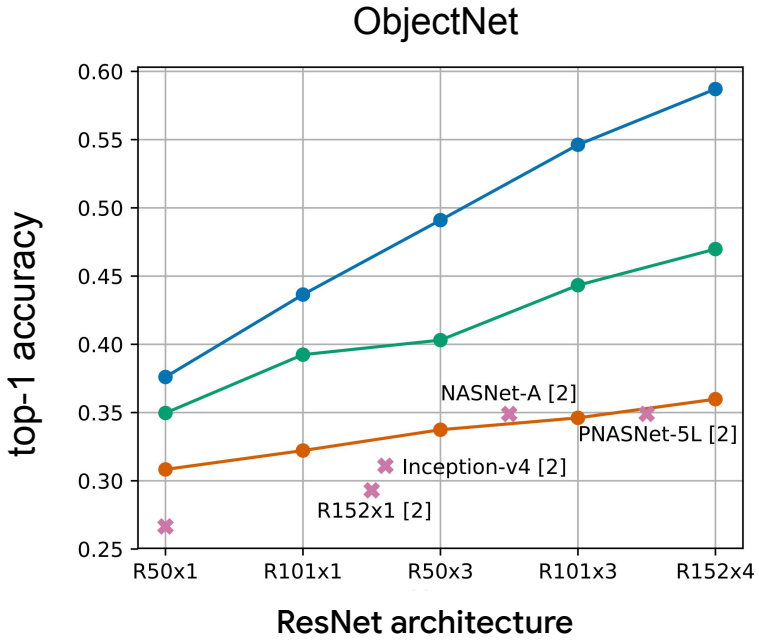
Out-of-distribution evaluation



Djolonga et al., *On Robustness and Transferability of Convolutional Neural Networks*, CVPR 2021

Out-of-distribution evaluation





Num training images

300M images (JFT)

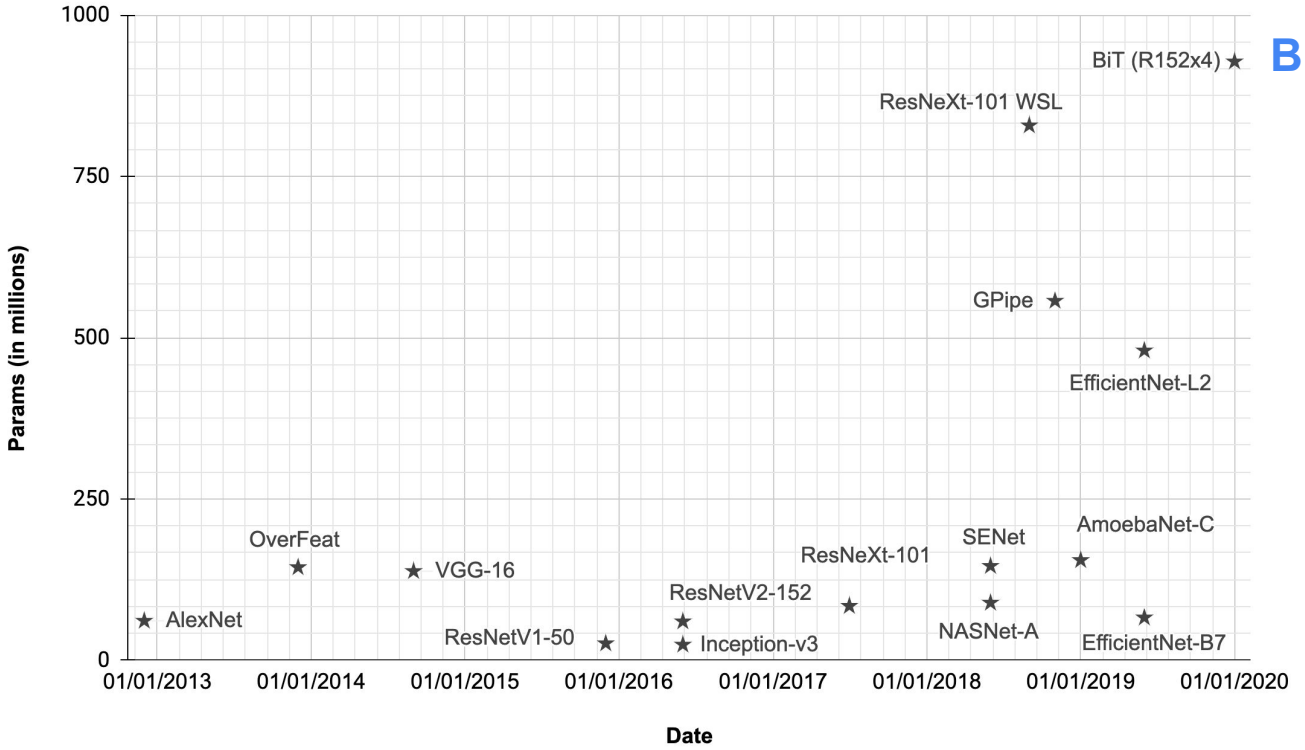
14M images (ImageNet-21K)

1.3M images (ImageNet-1K)

x Baselines

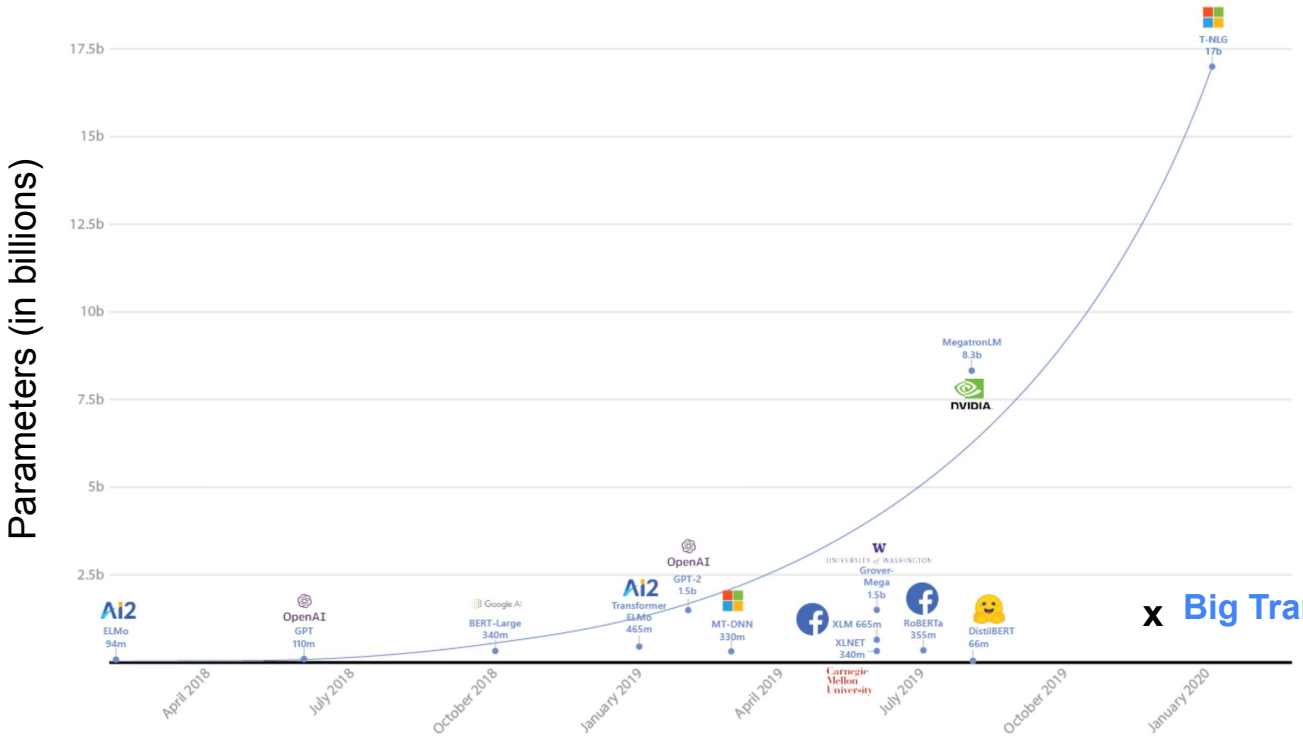
Djolonga et al., *On Robustness and Transferability of Convolutional Neural Networks*, CVPR 2021

History of Large Vision Models



Big Transfer

Large Language Models



GPT3
GShard
LaMDA
...

x **Big Transfer**

Figure: [MSR Blog]

Transformers for image classification

Google Research

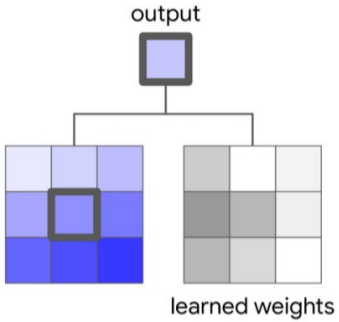
Transformers for vision?

- “LSTM ⟶ Transformer” ~ “ConvNet ⟶ ??? ”
- Issue with self-attention for vision: computation is quadratic in the input sequence length, quickly gets very expensive (with > few thousand tokens)
  - For ImageNet: 224x224 pixels ⟶ ~50,000 sequence length
  - Even worse for higher resolution and video

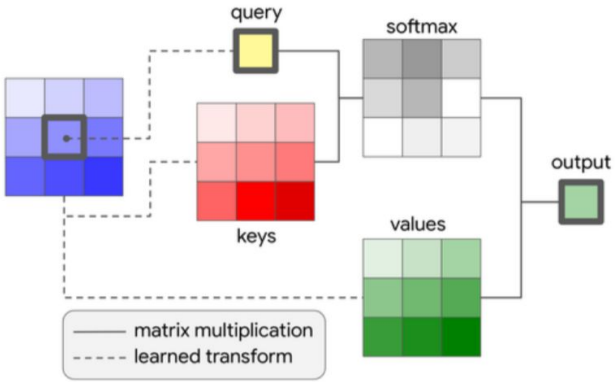
How can we deal with this quadratic complexity?

Local Self-Attention

Convolution



Local self-attention

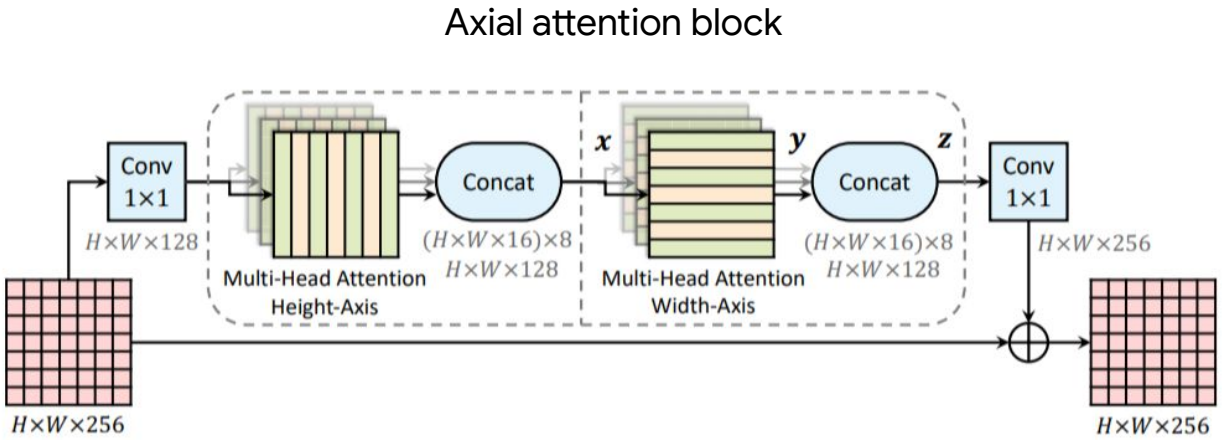


Idea: Make self-attention local, use it instead of convs in a ResNet

Figures from Ramachandran et al.

Hu et al., Local Relation Networks for Image Recognition, ICCV 2019
Ramachandran et al., Stand-Alone Self-Attention in Vision Models, NeurIPS 2019
Zhao et al., Exploring Self-attention for Image Recognition, CVPR 2020

Axial Self-Attention



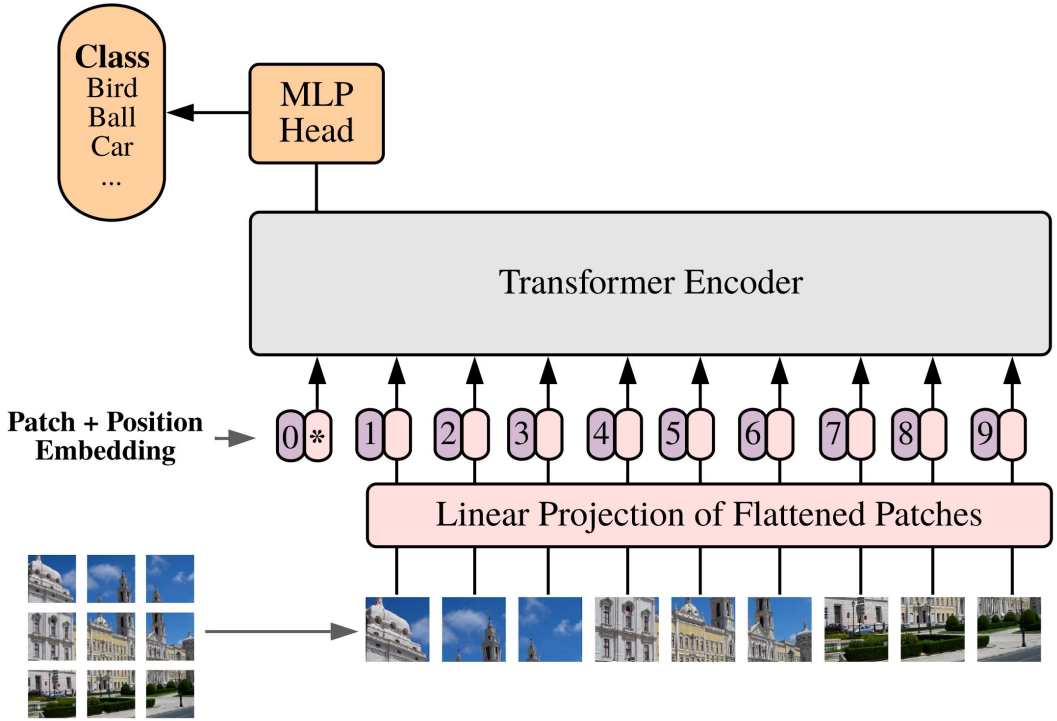
Idea: Make self-attention 1D (a.k.a. axial), use it instead of convs

Figure from Wang et al.

Wang et al., *Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation*, ECCV 2020

Vision Transformer (ViT)

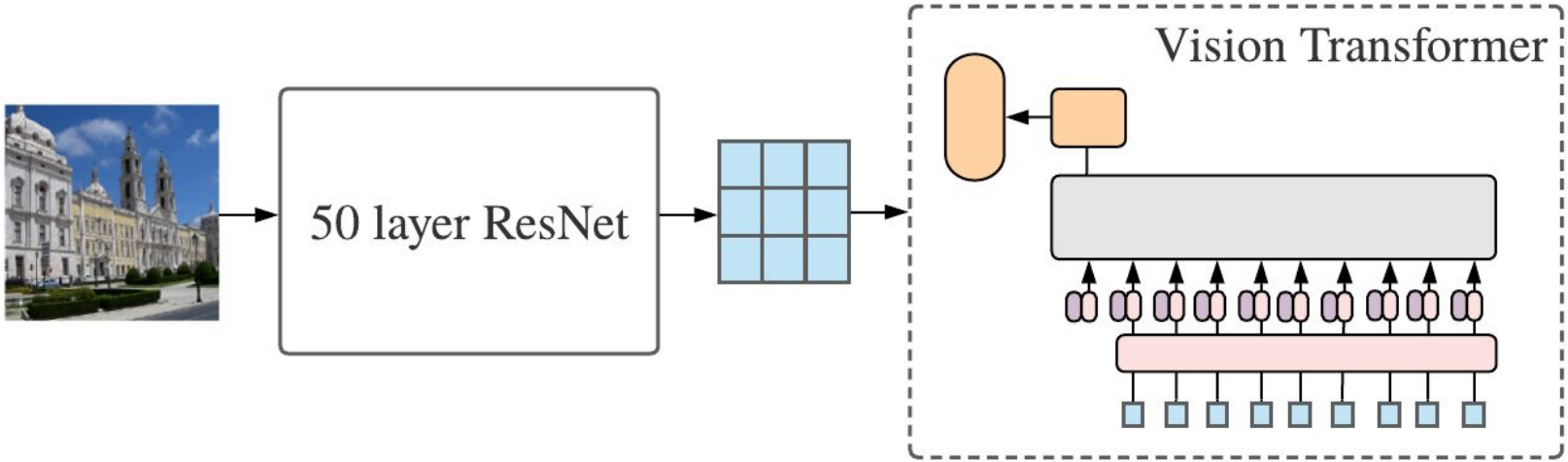
Idea: Take a transformer and apply it directly to image patches



Cordonnier et al., *On the Relationship between Self-Attention and Convolutional Layers*, ICLR 2020

Dosovitskiy et al., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, ICLR 2021

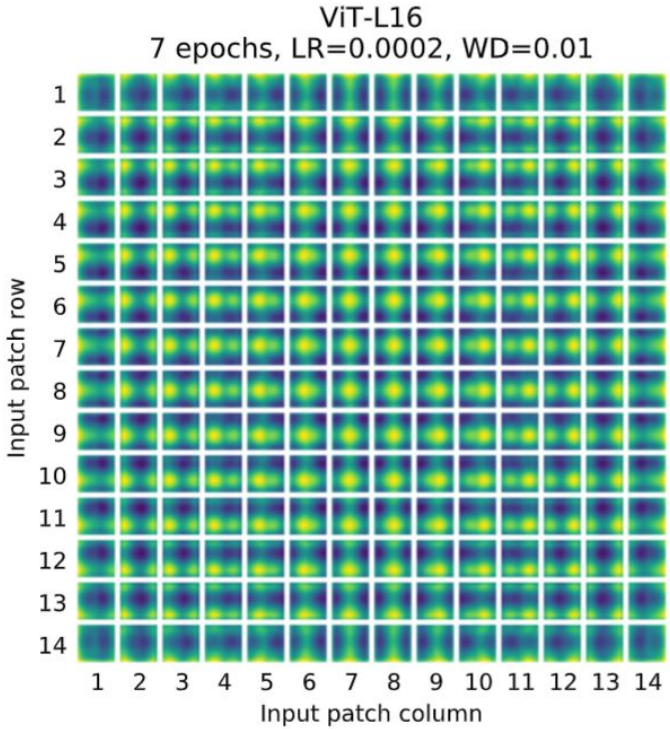
ResNet-ViT Hybrid



Bichen Wu et al. *Visual Transformers: Token-based Image Representation and Processing for Computer Vision*, arXiv 2020

Dosovitskiy et al., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, ICLR 2021

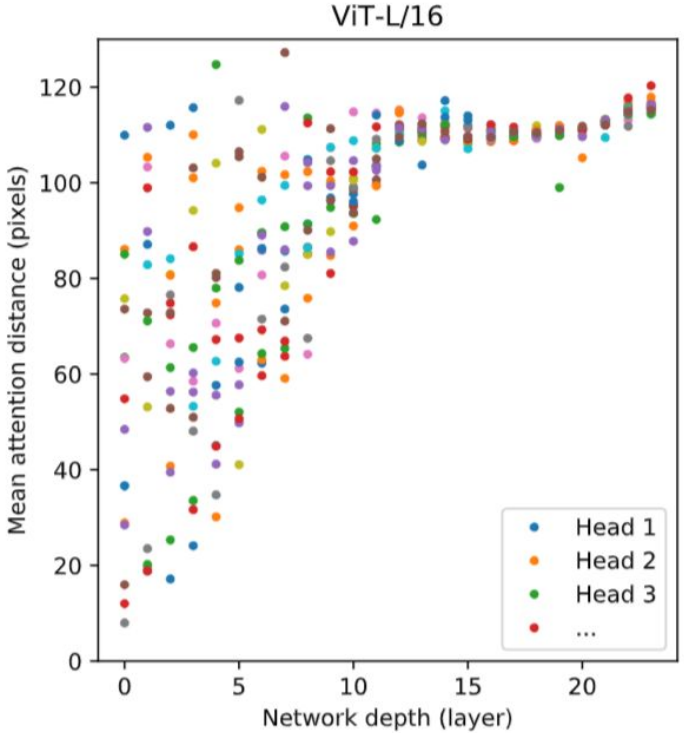
Analysis: Learned Position Embeddings



Conclusion: Learns intuitive local structures, but also deviates from locality in interesting ways

Dosovitskiy et al., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, ICLR 2021

Analysis: “Receptive Field Size”



Conclusion: Initial layers are partially local, deeper layers are global

Dosovitskiy et al., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, ICLR 2021

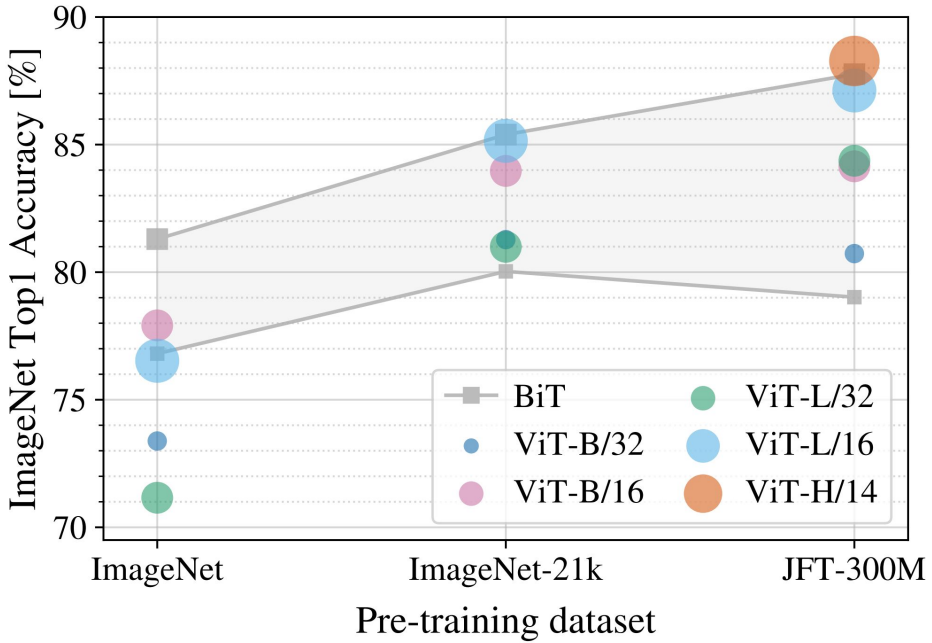
Scaling with Data

Key
ViT = Vision Transformer
BiT = Big Transfer (~ResNet)

ViT overfits on ImageNet, but shines on larger datasets

* with heavy regularization ViT has been shown to also work on ImageNet (Touvron et al.)

** training ViT on ImageNet with the sharpness-aware minimizer (SAM) also works very well (Chen et al.)

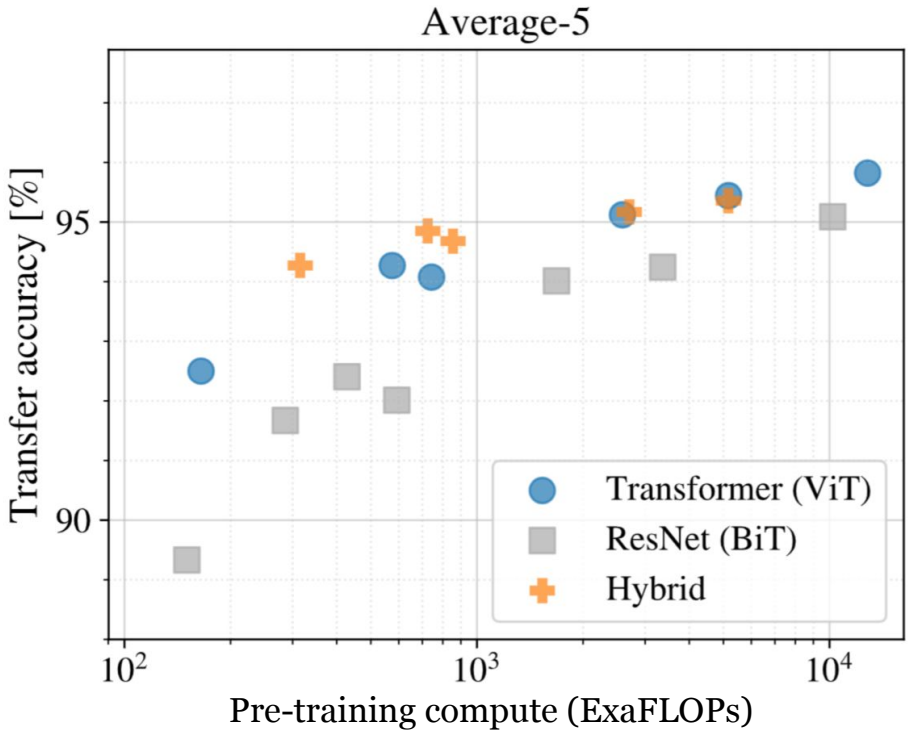


Dosovitskiy et al., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, ICLR 2021
Xiangning Chen et al., *When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations*, arXiv 2021
Touvron et al., *Training data-efficient image transformers & distillation through attention*, arXiv 2020

Scaling with Compute

Given sufficient data, ViT gives good performance/FLOP

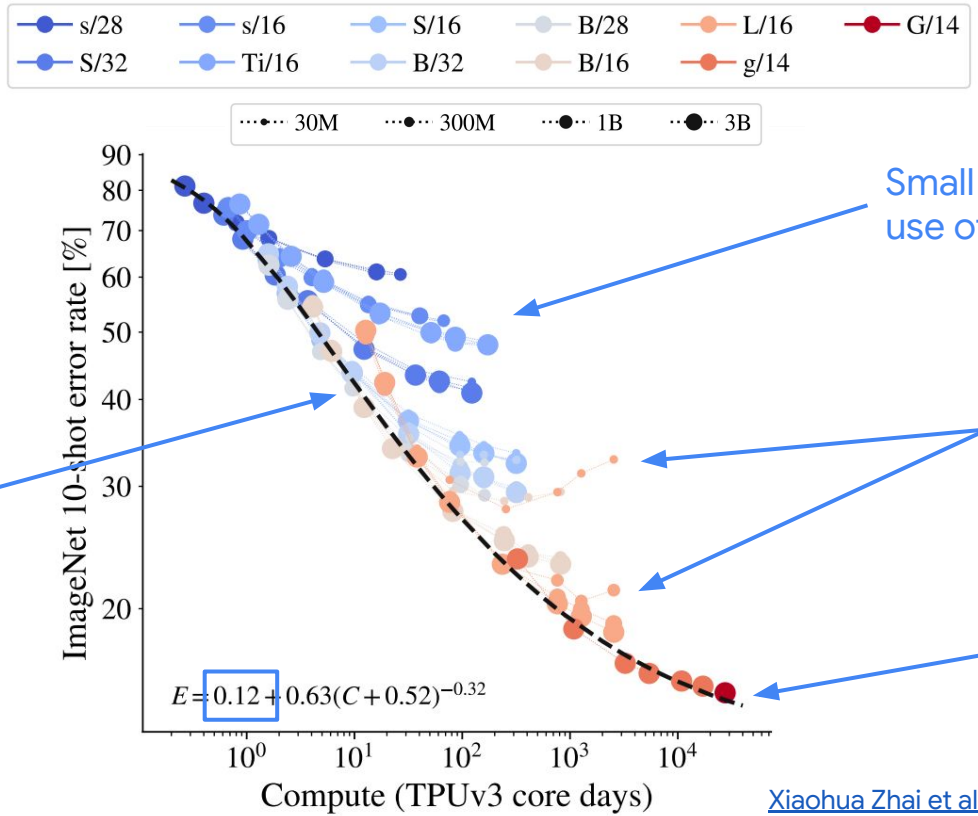
Hybrids yield benefits only for smaller models



Dosovitskiy et al., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, ICLR 2021

Scaling Laws

How many images do you need for a big model & vice-versa?



Small models can't make use of >30M examples

Even 300M examples are insufficient for large models

Power-law behaviour $E=aC^{-b}$

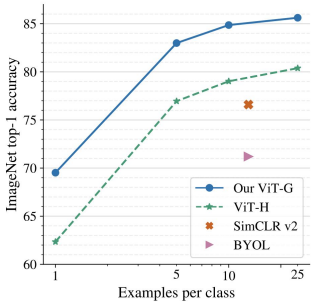
Saturates before y=0

Xiaohua Zhai et al., *Scaling Vision Transformers*, arXiv 2021

Scaling Laws

How many images do you need for a big model & vice-versa?

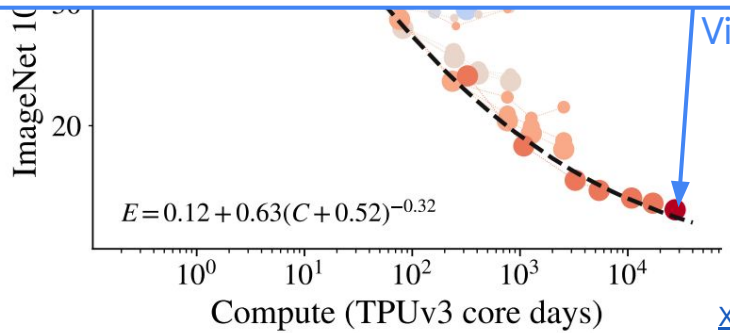
84.86% top-1 accuracy on 10-shot ImageNet (**<1% of train set**)



ImageNet **SOTA** | ImageNet (**OOD**/re-label) variants | 19 **diverse** tasks

Benchmark	ImageNet	INet V2	INet ReaL	ObjectNet	VTAB (light)
NS (Eff.-L2) [39]	88.3	80.2	-	68.5	-
MPL (Eff.-L2) [24]	90.2	-	**91.02**	-	-
CLIP (ViT-L/14) [26]	85.4	75.9	-	**72.3**	-
ALIGN (Eff.-L2) [16]	88.6	70.1	-	-	-
BiT-L (ResNet) [18]	87.54	-	90.54	58.7	76.29
ViT-H/14 [11]	88.55	-	90.72	-	77.63
Our ViT-G/14	**90.45±0.03**	**83.33±0.03**	90.81±0.01	70.53±0.52	**78.29±0.53**

ViT-G/14: 2B params, 3B images

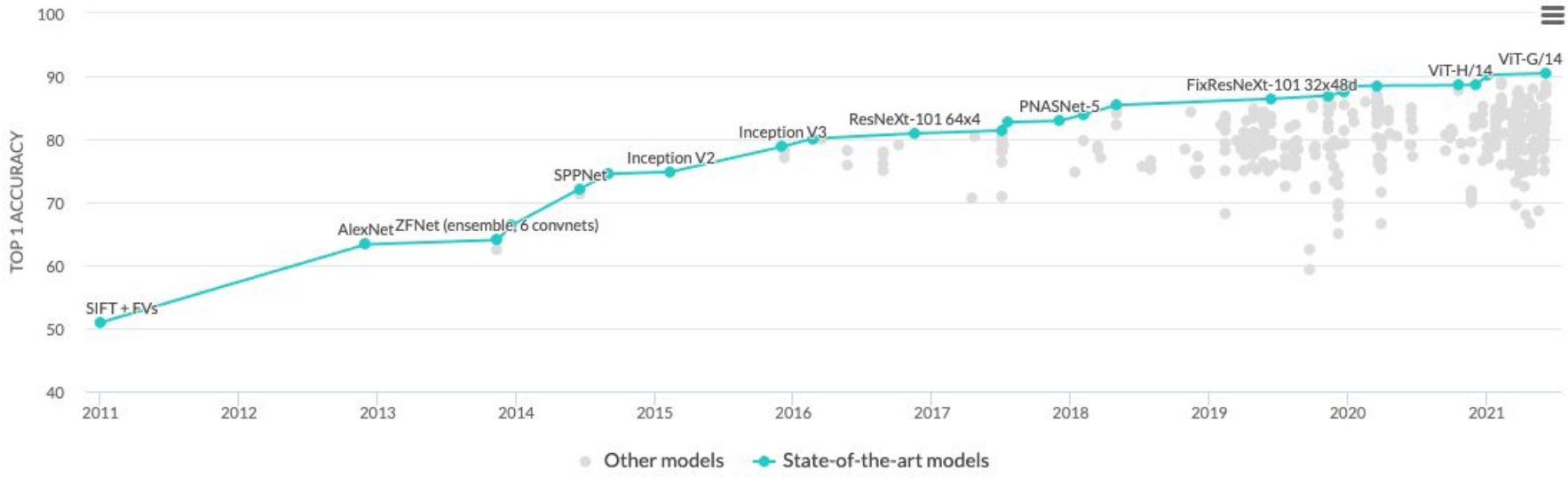


Xiaohua Zhai et al., *Scaling Vision Transformers*, arXiv 2021

Image Classification on ImageNet

Leaderboard Dataset

View Top 1 Accuracy by Date for All models



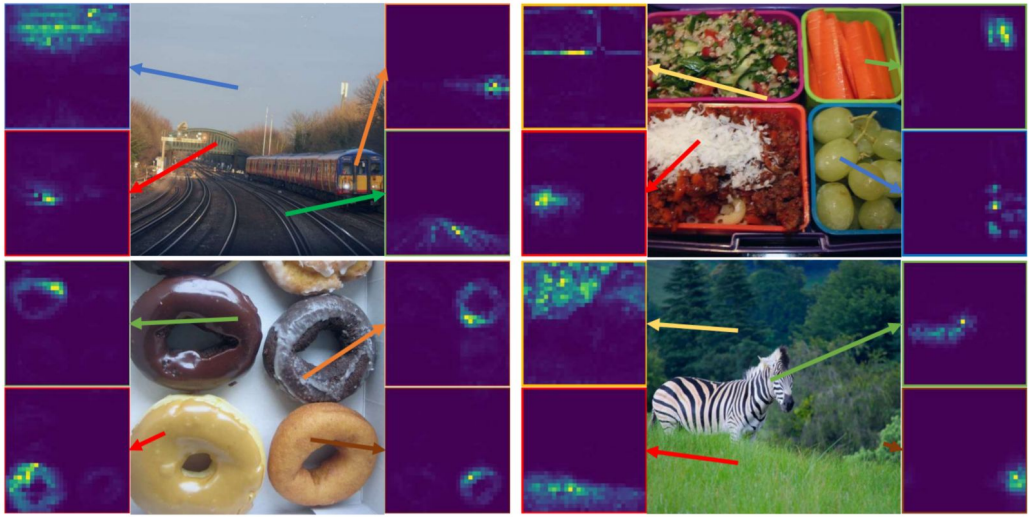
Self-Supervised Pre-training and Localization

BEiT: pre-train ViT on “inpainting” a.k.a. masked modeling, BERT-style

Trick 1: use VQ-VAE codes from DALL-E to represent patches

Trick 2: mask blocks of patches

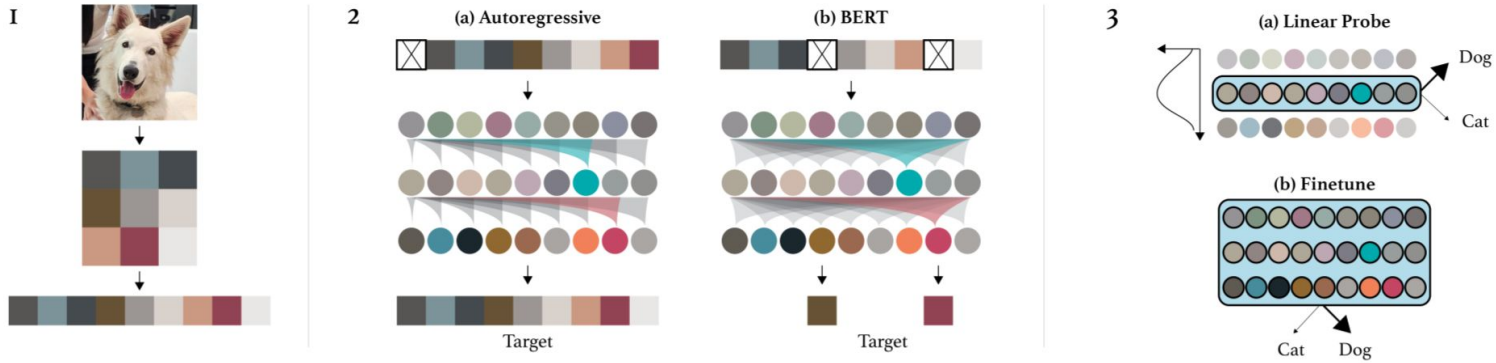
Works well on classification and seems to get some notion of object localization

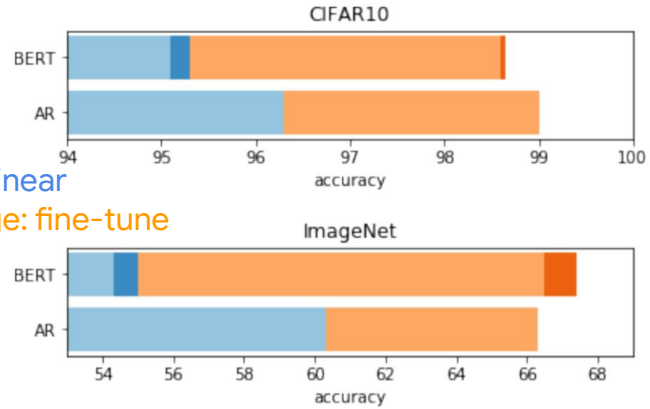


Figures from Hangbo Bao et al.

Hangbo Bao et al., *BEiT: BERT Pre-Training of Image Transformers*, arXiv 2021

Generative Pre-training





Blue: linear
Orange: fine-tune

- Language-style generative/BERT pre-training
- Operate at small res and potentially on quantized patches to make it faster
- Works surprisingly well, but is slow

Figures from Mark Chen et al.

Mark Chen et al., *Generative Pretraining From Pixels*, ICML 2020
Mark Chen et al., *Image GPT*, OpenAI blog 2020
Ramesh et al., *DALL·E: Creating Images from Text*, OpenAI blog 2021

Questions?

Beyond classification

Google Research

Background: Localization Tasks

Figure from Kirillov et al., Panoptic Segmentation, CVPR 2019

Background: Localization Tasks

(a) image (b) semantic segmentation

Figure from Kirillov et al., Panoptic Segmentation, CVPR 2019

Background: Localization Tasks

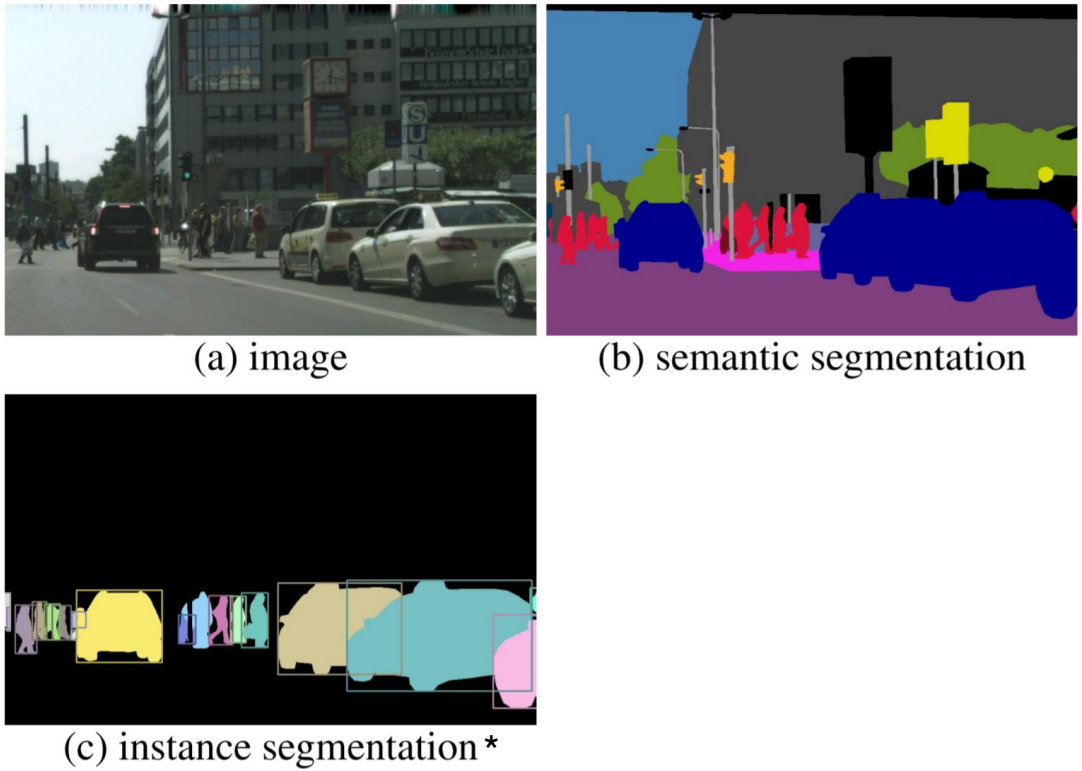


Figure from Kirillov et al., Panoptic Segmentation, CVPR 2019

* object detection = only predicting the bounding boxes

Background: Localization Tasks

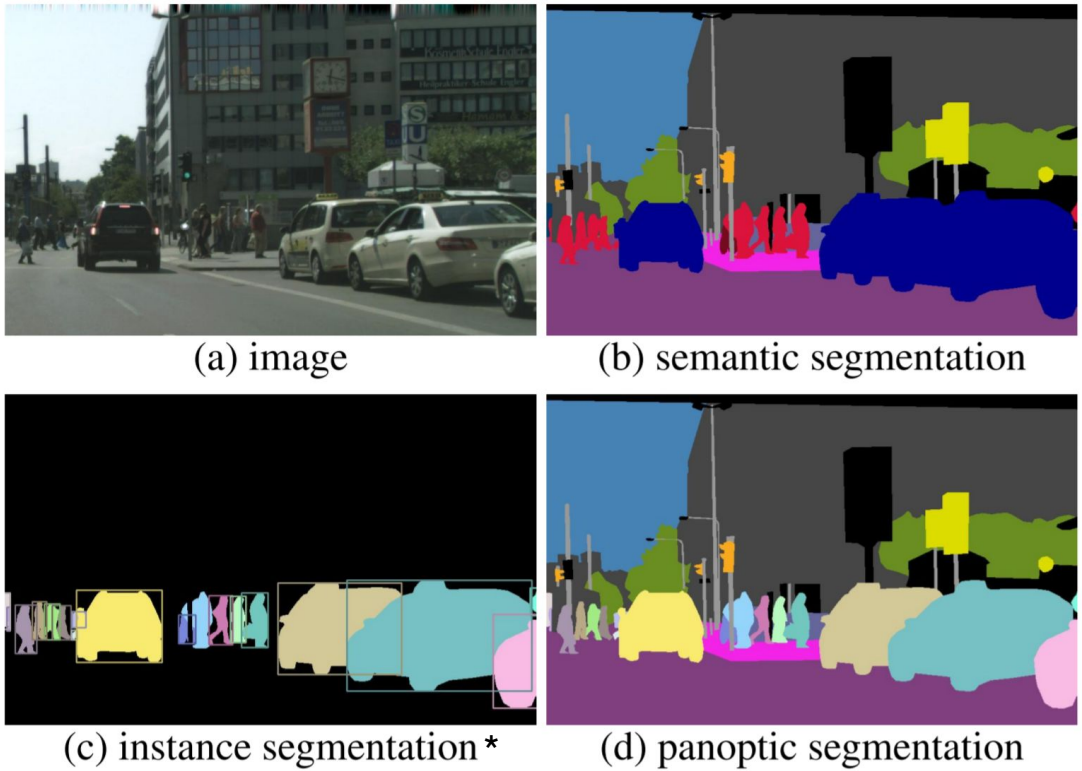


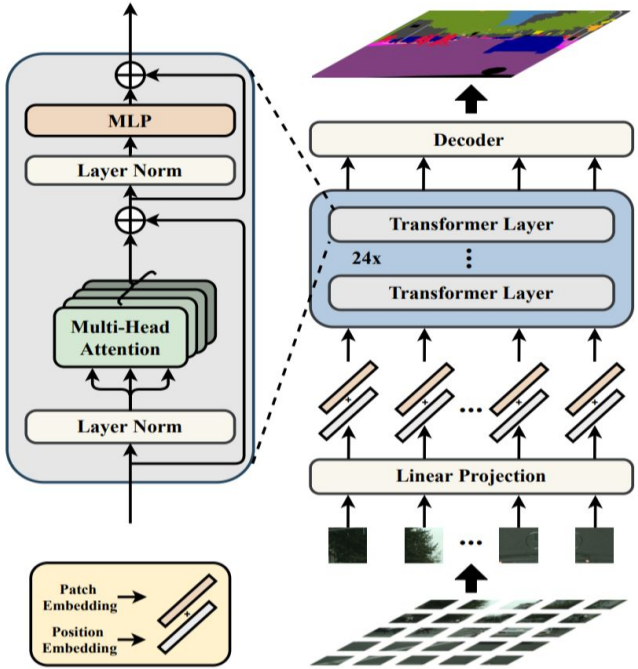
Figure from Kirillov et al., Panoptic Segmentation, CVPR 2019

* object detection = only predicting the bounding boxes

Segmentation Transformer (SETR)

Basically ViT for segmentation, with a decoder on top

Competitive with SOTA, despite using lower resolution

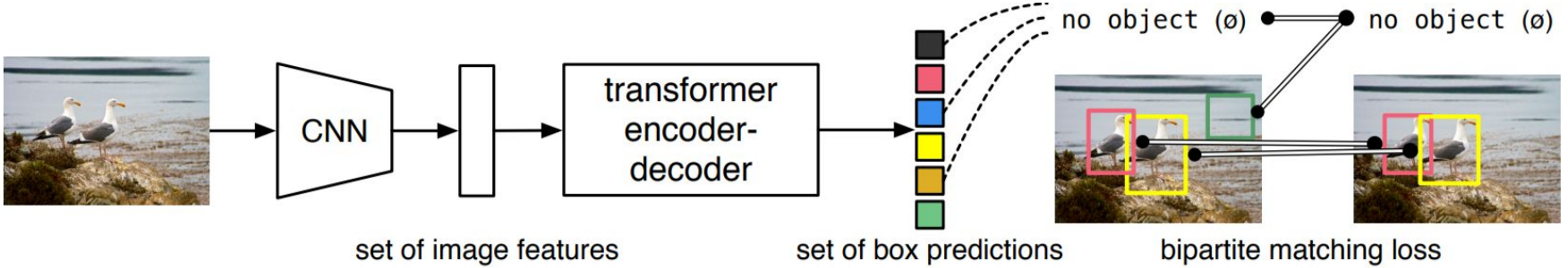


Method	Backbone	mIoU
PSPNet [60]	ResNet-101	78.40
DenseASPP [50]	DenseNet-161	80.60
BiSeNet [52]	ResNet-101	78.90
PSANet [61]	ResNet-101	80.10
DANet [18]	ResNet-101	81.50
OCNet [54]	ResNet-101	80.10
CCNet [26]	ResNet-101	81.90
Axial-DeepLab-L [47]	Axial-ResNet-L	79.50
Axial-DeepLab-XL [47]	Axial-ResNet-XL	79.90
SETR-*PUP* (100k)	T-Large	81.08
SETR-*PUP*‡	T-Large	81.64

Figures from Sixiao Zheng et al.

Sixiao Zheng et al., *Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers*, arXiv 2020

Detection Transformer (DETR)



ResNet + Transformer, trained to directly predict a set of bounding boxes

Beats a well-tuned Faster RCNN on detection, also works well on panoptic segmentation

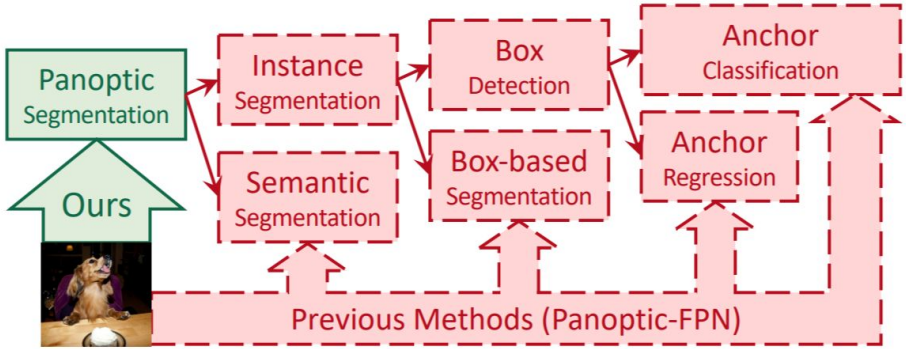
Deformable DETR is yet much better

Model	GFLOPS/FPS	#params	AP
Faster RCNN-DC5	320/16	166M	39.0
Faster RCNN-FPN	180/26	42M	40.2
Faster RCNN-R101-FPN	246/20	60M	42.0
Faster RCNN-DC5+	320/16	166M	41.1
Faster RCNN-FPN+	180/26	42M	42.0
Faster RCNN-R101-FPN+	246/20	60M	44.0
DETR	86/28	41M	42.0
DETR-DC5	187/12	41M	43.3
DETR-R101	152/20	60M	43.5
DETR-DC5-R101	253/10	60M	**44.9**

Figures from Carion et al.

Carion et al., *End-to-End Object Detection with Transformers*, ECCV 2020

MaX-DeepLab (Panoptic Segmentation)



Method	Backbone	TTA	PQ	PQ^{Th}	PQ^{St}
Box-based panoptic segmentation methods					
Panoptic-FPN [47]	RN-101		40.9	48.3	29.7
DETR [10]	RN-101		46.0	-	-
UPSNet [98]	DCN-101 [25]	✓	46.6	53.2	36.7
DetectoRS [76]	RX-101 [97]	✓	49.6	57.8	37.1
Box-free panoptic segmentation methods					
Panoptic-DeepLab [21]	X-71 [24, 75]	✓	41.4	45.1	35.9
Axial-DeepLab-L [89]	-		43.6	48.9	35.6
Axial-DeepLab-L [89]	-	✓	44.2	49.2	36.8
MaX-DeepLab-S	-		49.0	54.0	41.6
MaX-DeepLab-L	-		51.3	57.2	42.4

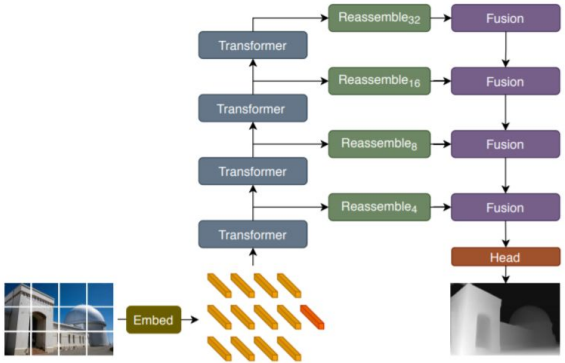
Architecture similar to DETR, but ResNet & Transformer in parallel and is geared towards panoptic segmentation (trained with PQ-inspired loss directly).

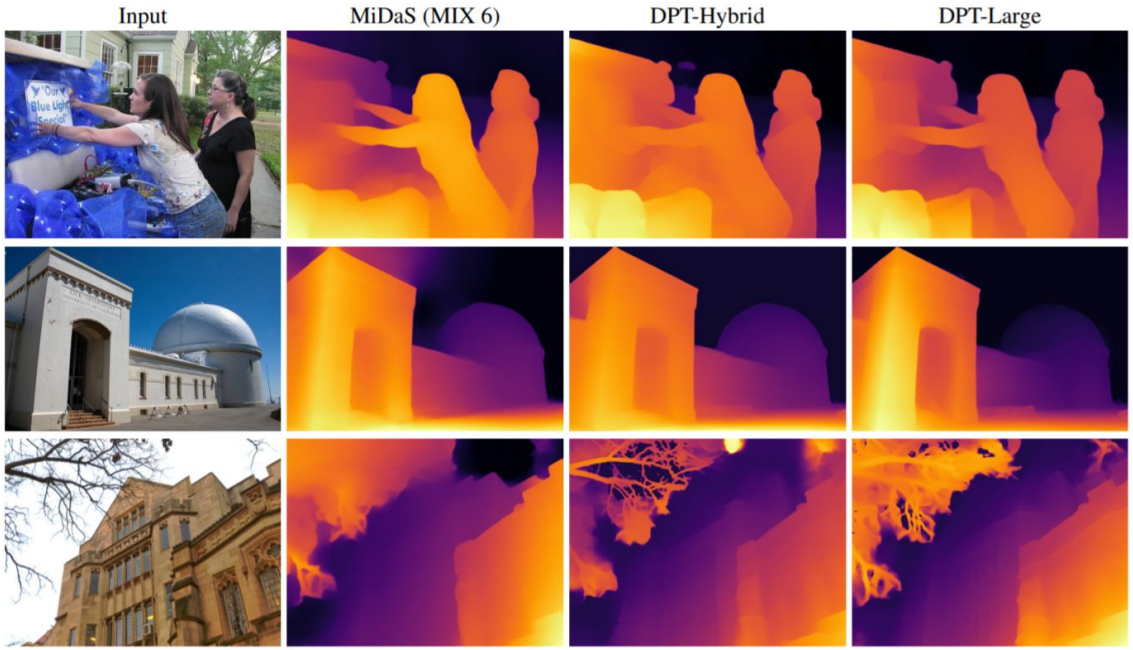
Figures from Huiyu Wang et al.

Huiyu Wang et al., *Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation*, ECCV 2020
Huiyu Wang et al., *MaX-DeepLab: End-to-End Panoptic Segmentation with Mask Transformers*, arXiv 2020

Monocular Depth Estimation

- A ViT-based model with an added top-down “decoder”
- Up to 28% improvement on monocular depth estimation compared to ConvNets





Figures from Ranftl et al.

Ranftl et al., *Vision Transformers for Dense Prediction*, arXiv 2021

Beyond images

Google Research

Point Transformer

Transformer operates on sets → a good fit for point clouds!

Hourglass-shaped model with local vector self-attention

SOTA or close on several benchmarks

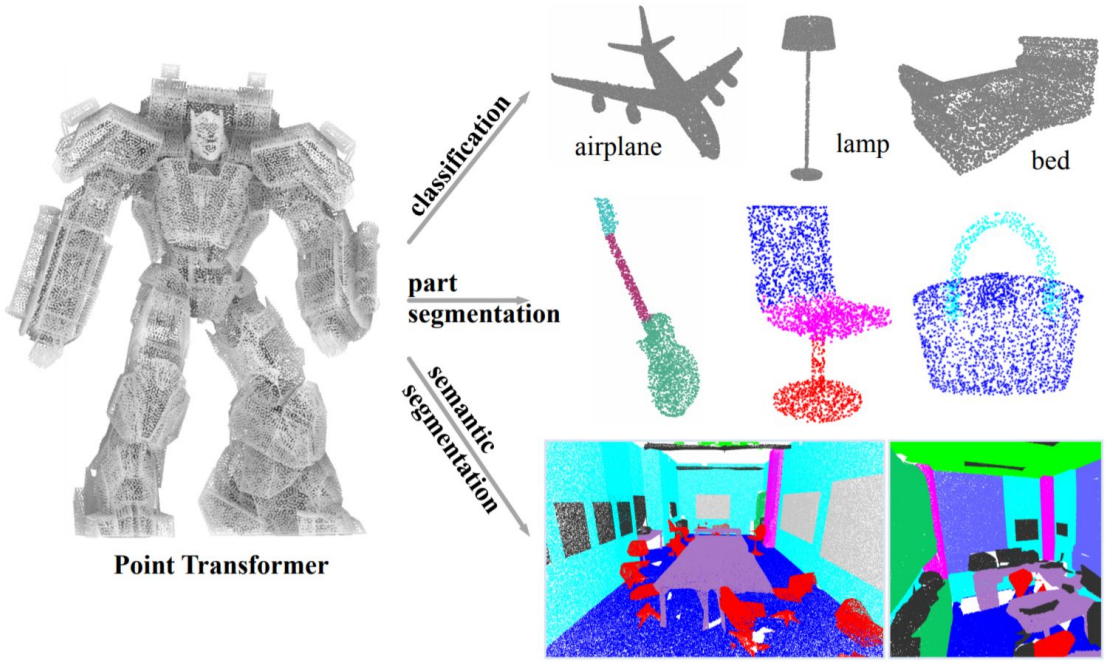
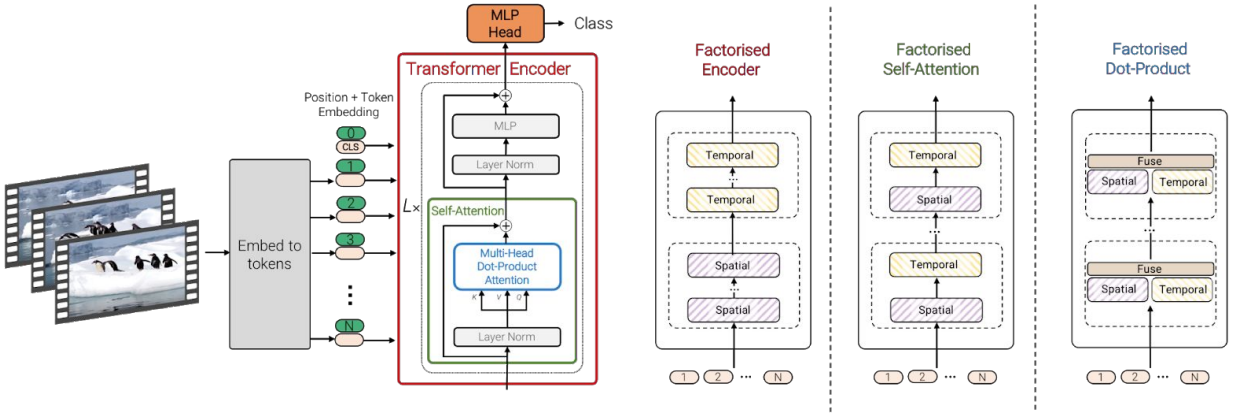


Figure from Hengshuang Zhao et al.

Hengshuang Zhao et al., *Point Transformer*, arXiv 2020

Video Vision Transformer (ViViT)

- Factorized attention to handle **larger (video) inputs**
- Bring large-scale image-based pre-training to video
- SOTA across 6 common classification benchmarks



Kinetics 400

Method	Top 1	Top 5
blVNet [16]	73.5	91.2
STM [30]	73.7	91.6
TEA [39]	76.1	92.5
TSM-ResNeXt-101 [40]	76.3	–
I3D NL [72]	77.7	93.3
CorrNet-101 [67]	79.2	–
ip-CSN-152 [63]	79.2	93.8
LGD-3D R101 [48]	79.4	94.4
SlowFast R101-NL [18]	79.8	93.9
X3D-XXL [17]	80.4	94.6
TimeSformer-L [2]	80.7	94.7
ViViT-L/16x2	80.6	94.7
ViViT-L/16x2 320	**81.3**	**94.7**
Methods with large-scale pretraining		
ip-CSN-152 [63] (IG [41])	82.5	95.3
ViViT-L/16x2 (JFT)	82.8	95.5
ViViT-L/16x2 320 (JFT)	83.5	95.5
ViViT-H/16x2 (JFT)	**84.8**	**95.8**

Arnab et al., ViViT: A Video Vision Transformer, arXiv 2021

Multimodal Learning: Vision-and-Language

Embed image (bounding boxes) and text, jointly process with a transformer, train on 4 self-supervised tasks

SOTA on many vision-and-language tasks

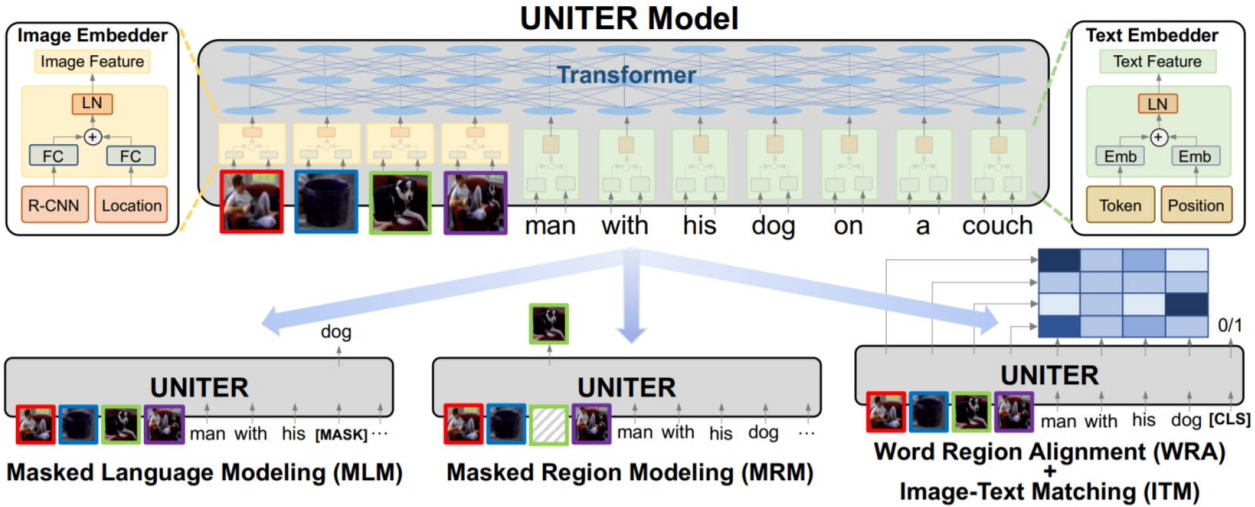


Figure from Yen-Chun Chen et al.

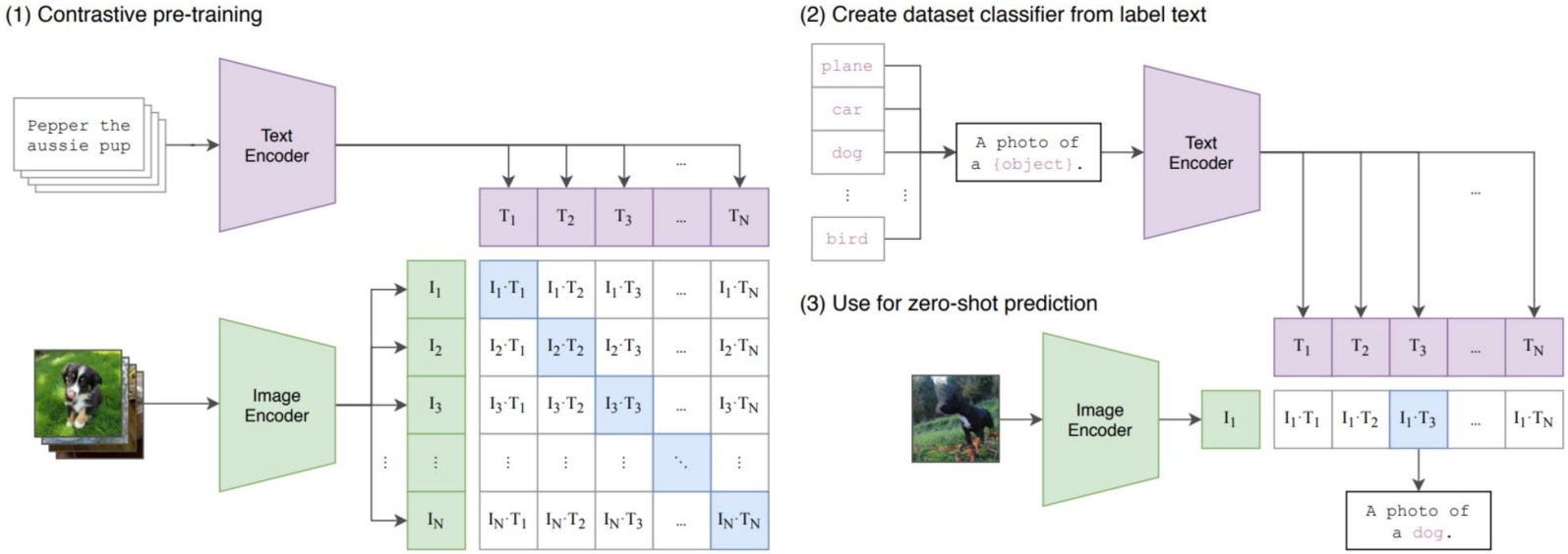
Gen Li et al., *Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training*, AAAI 2020

Luowei Zhou et al., *Unified Vision-Language Pre-Training for Image Captioning and VQA*, AAAI 2020

Weijie Su et al., *VL-BERT: Pre-training of Generic Visual-Linguistic Representations*, ICLR 2020

Yen-Chun Chen et al., *UNITER: UNiversal Image-TExt Representation Learning*, ECCV 2020

Learning Visual Representations with Text



Desai & Johnson, *VirTex: Learning Visual Representations from Textual Annotations*, arXiv 2020

Sariyildiz et al., *Learning Visual Representations with Caption Annotations*, ECCV 2020

Yuhao Zhang et al., *Contrastive Learning of Medical Visual Representations from Paired Images and Text*, arXiv 2020

Radford et al., Learning Transferable Visual Models From Natural Language Supervision, 2021

Figure from Radford et al.

Learning Visual Representations with Text

76.2% (!) zero-shot top-1 accuracy on ImageNet

Much more robust that supervised ImageNet models

Good on many other datasets too

	aYahoo	ImageNet	SUN
Visual N-Grams	72.4	11.5	23.0
CLIP	**98.4**	**76.2**	**58.5**

Figure from Radford et al.

Radford et al., *Learning Transferable Visual Models From Natural Language Supervision*, arXiv 2021

Video-Audio-Text Transformer (VATT)

- Multimodal videos (video + audio + text transcript) are a great source of training data
- All modalities processed with a single transformer
- Trained contrastively, similar to CLIP
- (Nearly) SOTA on many tasks on video and audio, decent performance on images

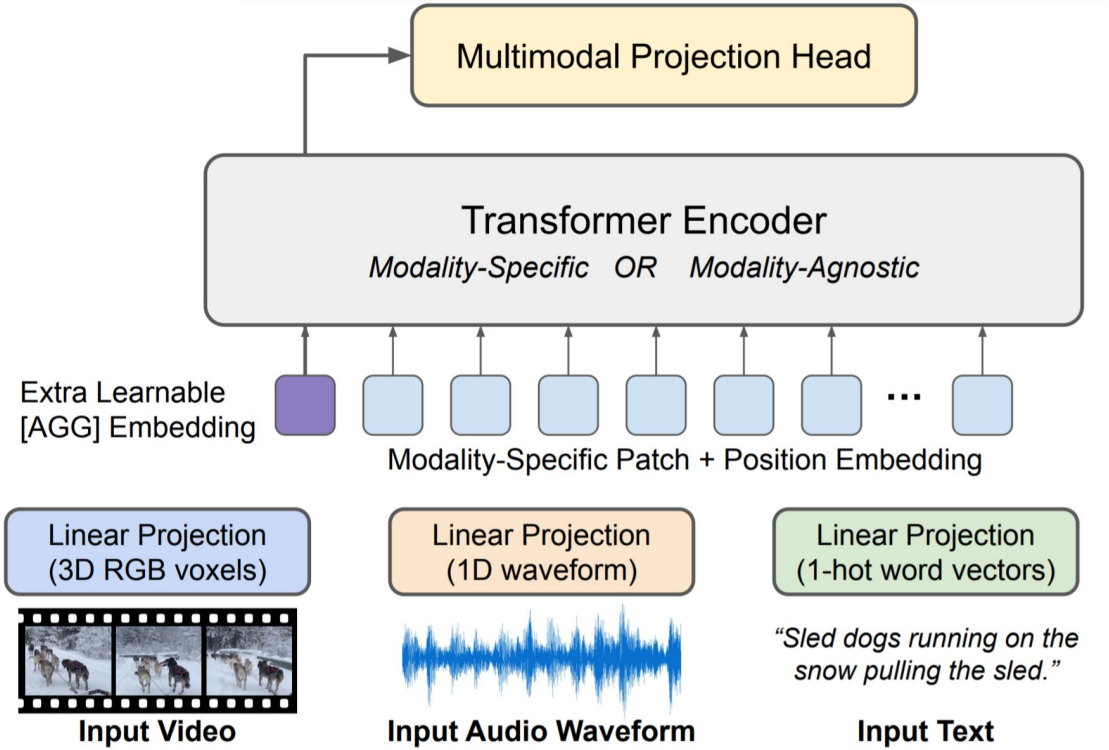


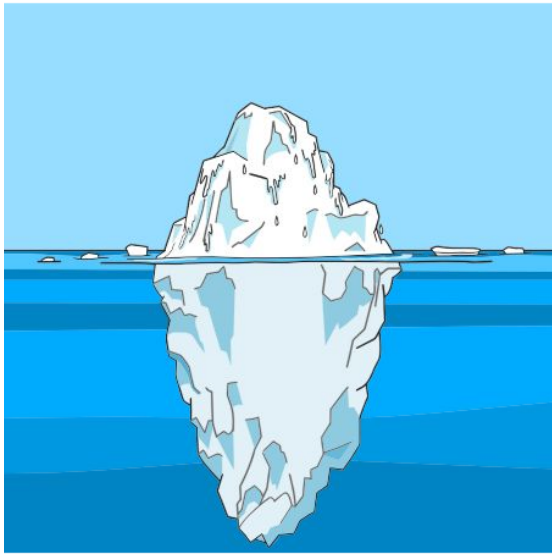
Figure from Akbari et al.

Akbari et al., *VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text*, arXiv 2021

Summary

This was just the tip of the iceberg...

- “Pyramid-shaped”: Pyramid ViT, LeViT, Swin Transformer, Pooling ViT, Hierarchical ViT, ...
- Other variants: Transformer in Transformer,
- Using ideas from ConvNets: ConViT, LocalViT, CeiT, CvT
- Deeper: DeepViT, CaiT, ...
- For dense tasks: Segmenter, SegFormer, ...
- For video: TimeSFormer, STAM, ...
- For high-res: Vision Longformer, ...
- Self-supervised: MoCo ViT, SiT, DINO, BEiT, ...
- ... and many more (a nice overview blog post from April)



https://openclipart.org/detail/299871/tip-of-the-iceberg

Summary

- Transformers are coming to vision
  - Large-scale training and transfer for image classification
  - But also detection, segmentation, depth estimation, etc
  - ... moreover, point clouds, videos, and multimodal learning

A lot of exciting work, but vision still not solved...

A lot remains to be done!