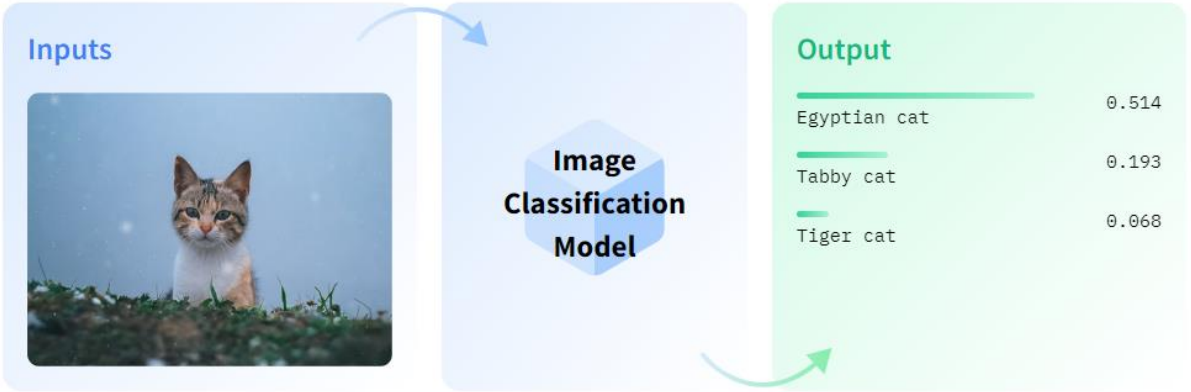
Image Classification models Survey

From then to now

Presenter : Yasin Fakhar

What is image classification

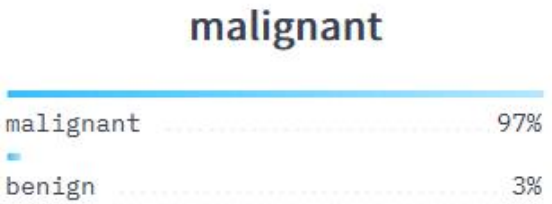
Image classification is the task of assigning a label or class to an entire image. Images are expected to have only one class for each image. Image classification models take an image as input and return a prediction about which class the image belongs to.

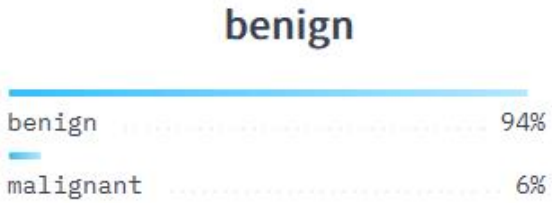


https://huggingface.co/tasks/image-classification

Use case

- Detection of cancer cells in pathology slides
- Keyword Classification
- Image Search
- Face recognition in security

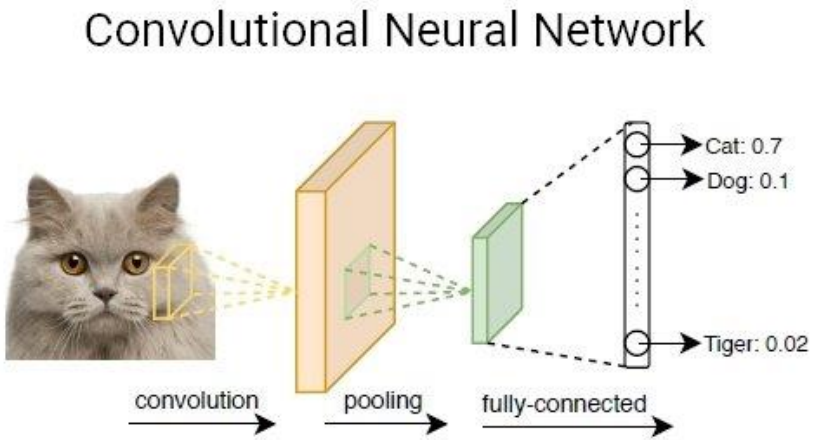




Convolution Neural Network

among deep neural networks (DNN), the convolutional neural network (CNN) has demonstrated excellent results in computer vision tasks, especially in image classification.

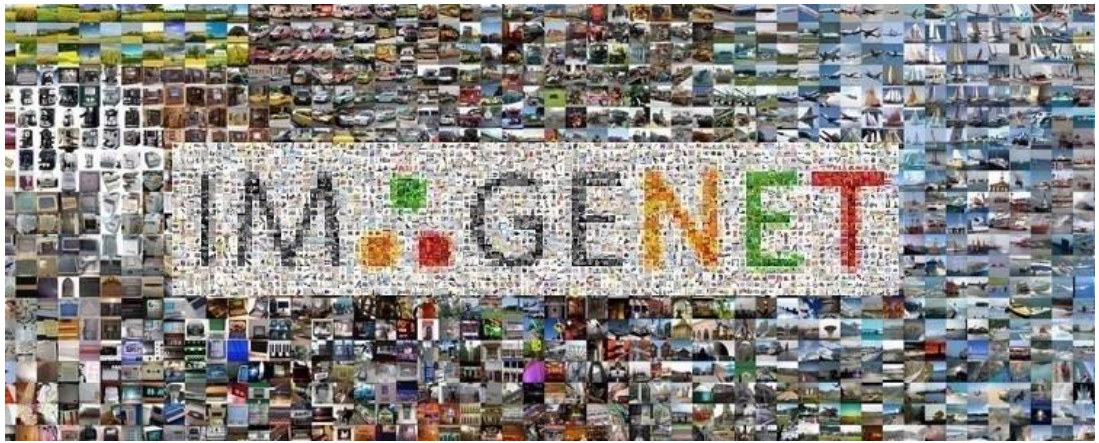
Convolutional Neural Network (CNN, or ConvNet) is a special type of multi-layer neural network inspired by the mechanism of the optical and neural systems of humans.



https://viso.ai/computer-vision/image-classification/

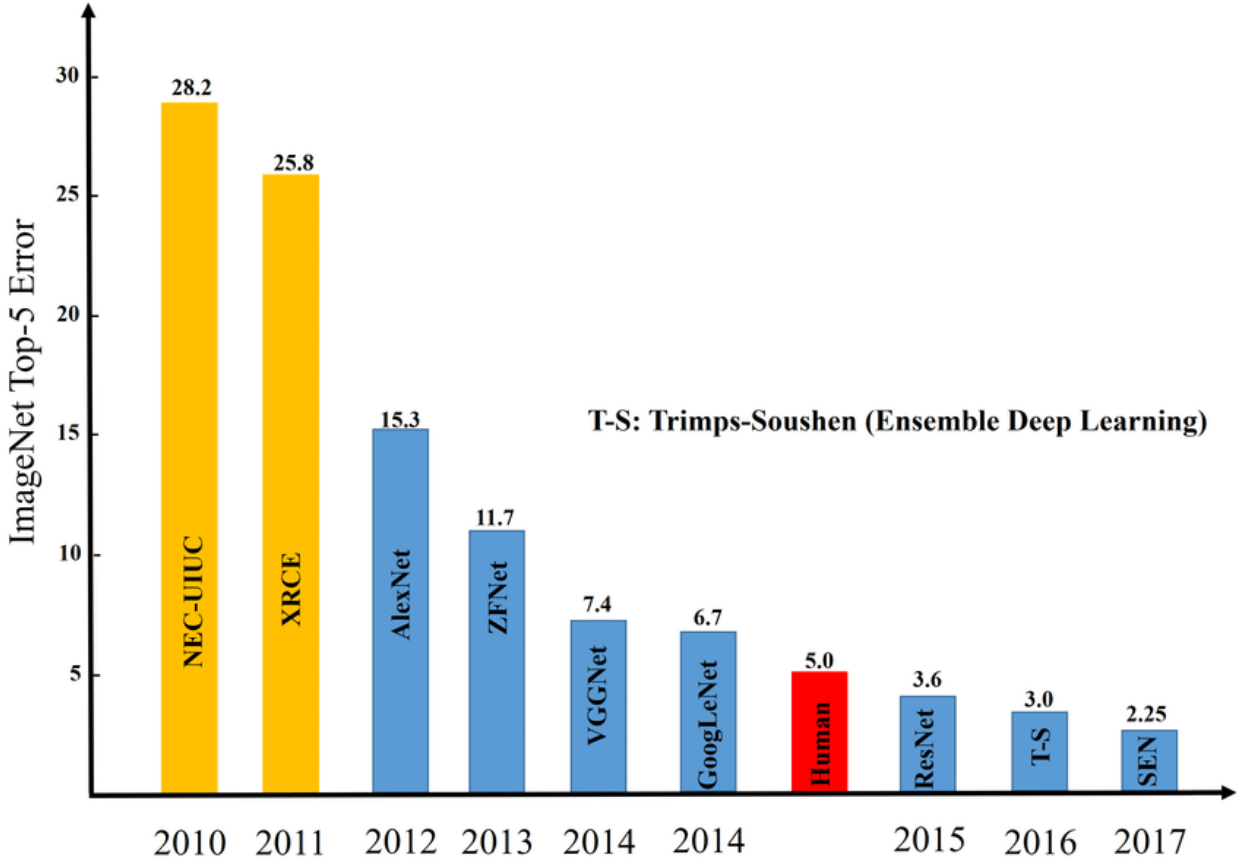
ImageNet challenge

ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images.



ImageNet challenge

This dataset spans 1000 object classes and contains 1,281,167 training images, 50,000 validation images and 100,000 test images.



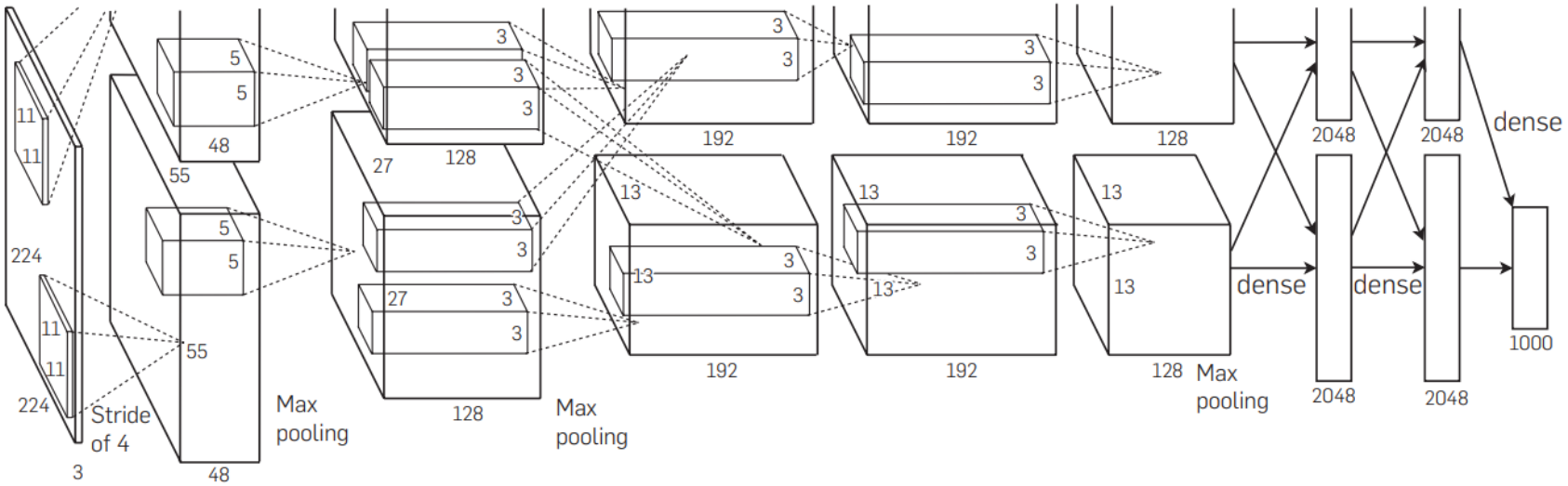
https://www.researchgate.net/figure/Main-breakthroughs-in-ImageNet-image-classification-challenge_fig1_335937276

Alex Net

AlexNet competed in the ImageNet Large Scale Visual Recognition Challenge on September 30, 2012. The network achieved a top-5 error of 15.3%

https://en.wikipedia.org/wiki/AlexNet

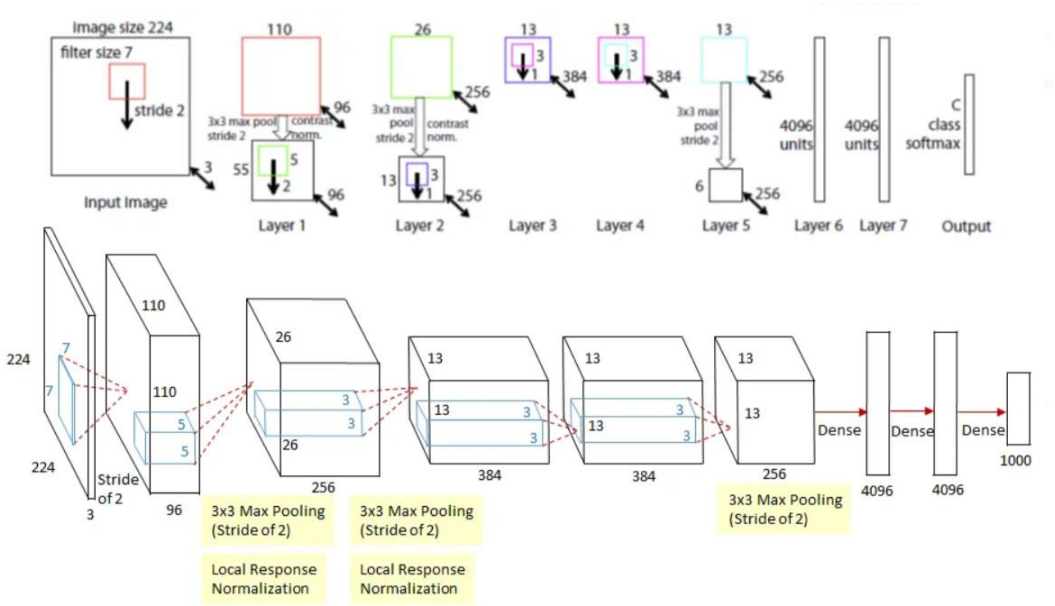
Alex Net



Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Communications of the ACM* 60.6 (2017): 84-90.

Alex Net

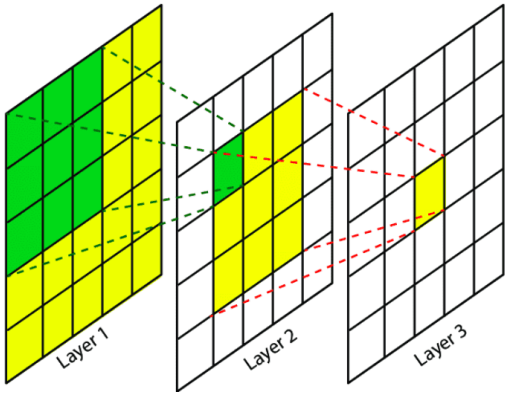
ZFNet is a kind of winner of the ILSVRC (ImageNet Large Scale Visual Recognition Competition) 2013, which is an image classification competition, which has significantly improvement over AlexNet,
the winner of ILSVRC 2012.

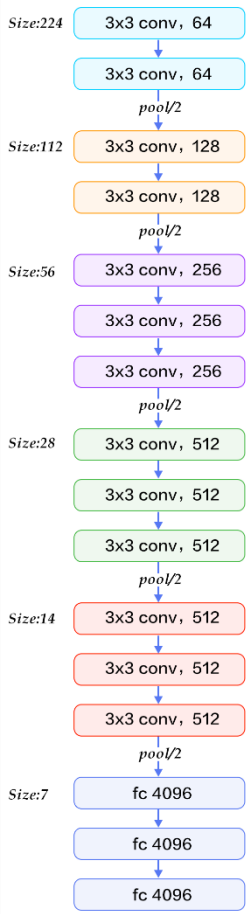


VGG Net

Second place of ILSVRC-14 in 2014

- Smaller filter with more layers
- Use just 3*3 filters with stride 1

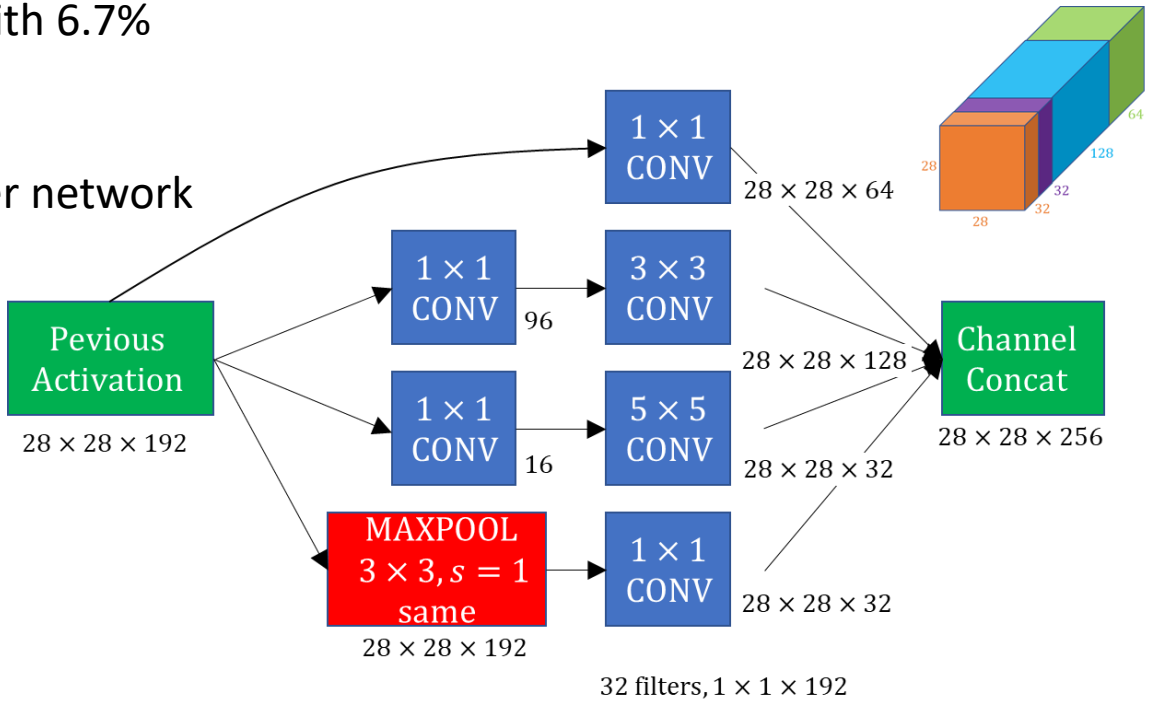




Google Net

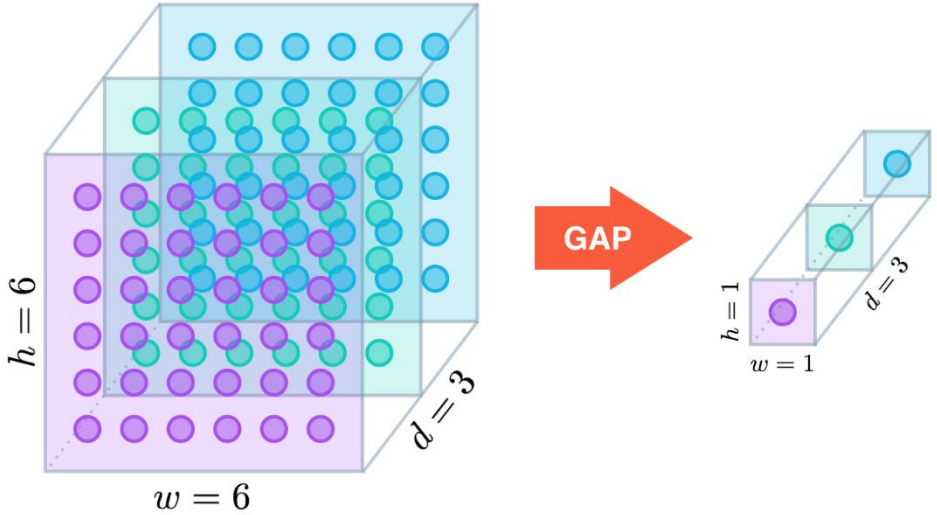
Winner of ILSVRC-14 in 2014 with 6.7%
Top five error

- More parameter with deeper network
- Use inception module

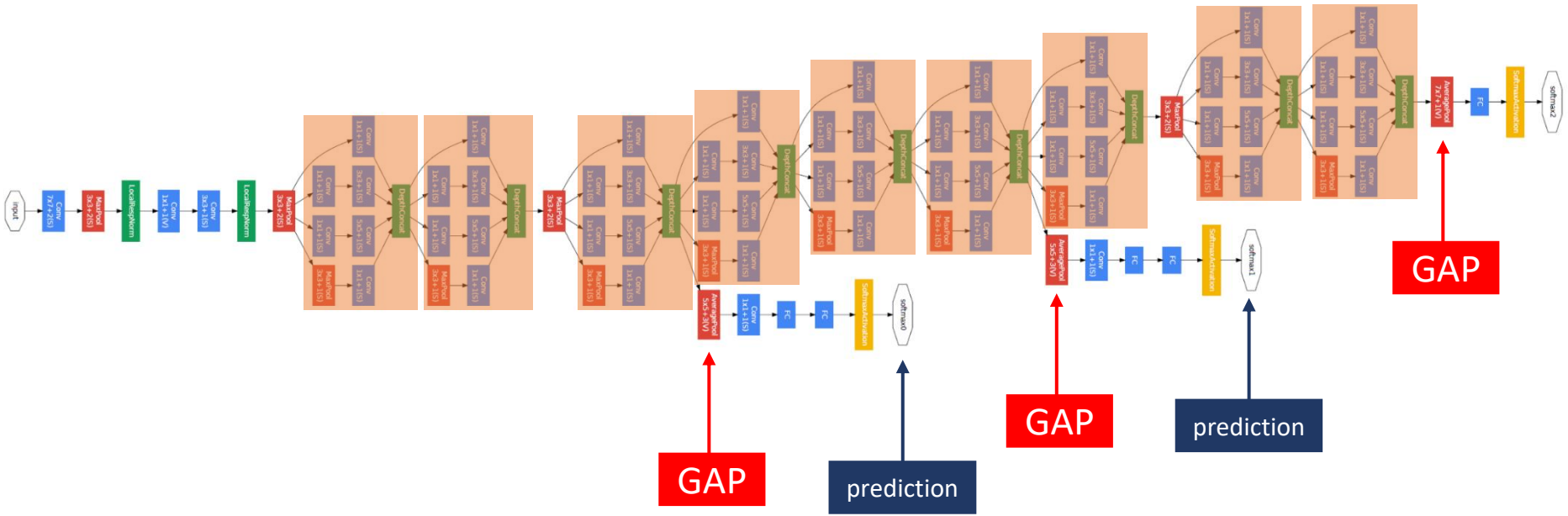


Global Average Pooling

Prevent over fitting



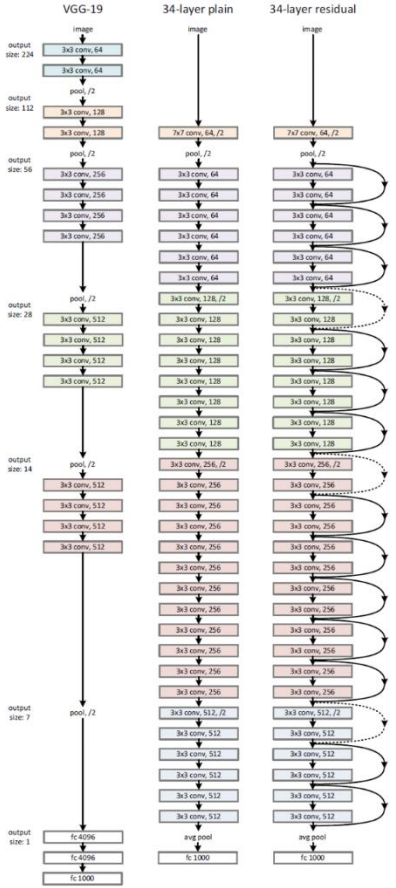
Google Net



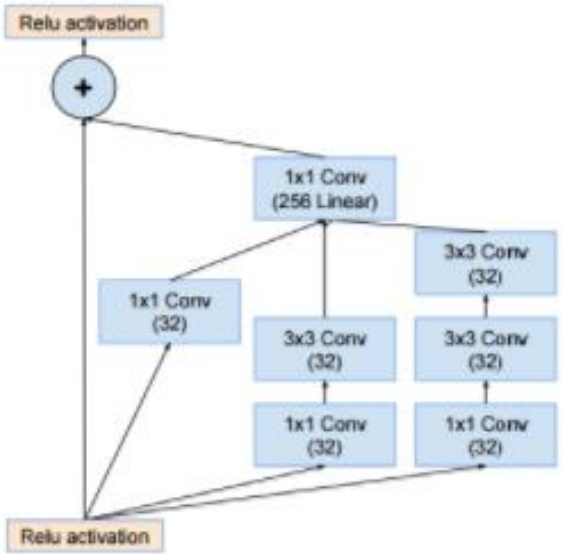
ResNet

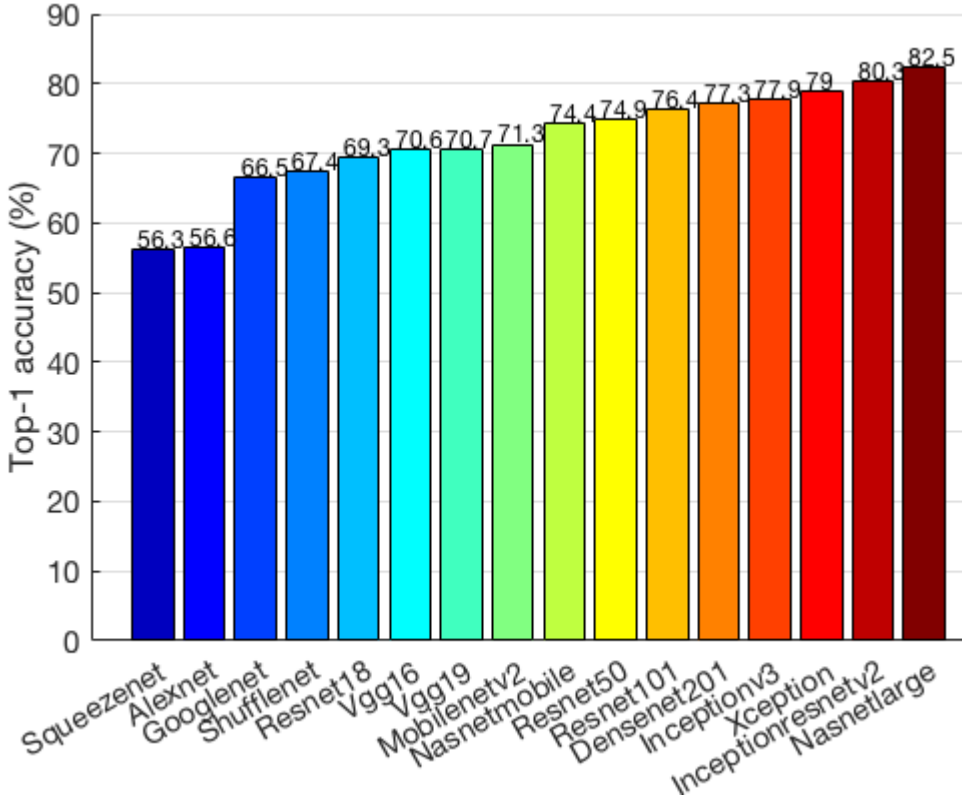
Winner of ILSVRC-15 in 2015 with 3.57%
Top five error

- 34,50,101,152 layers
- Using residual block
- Using global average pooling



ResNet + Google Net + VGG = Google Net inception v4




Top-1 accuracy (%)
0
10
20
30
40
50
60
70
80
90
56.3
56.6
66.5
67.4
69.3
70.6
70.7
71.3
74.4
74.9
76.4
77.3
77.9
79
80.3
82.5
Squeezenet
Alexnet
Googlenet
Shufflenet
Resnet18
Vgg16
Vgg19
Mobilenetv2
Nasnetmobile
Resnet50
Resnet101
Densenet201
Inceptionv3
Xception
Inceptionresnetv2
Nasnetlarge

Vision Transformer (ViT)

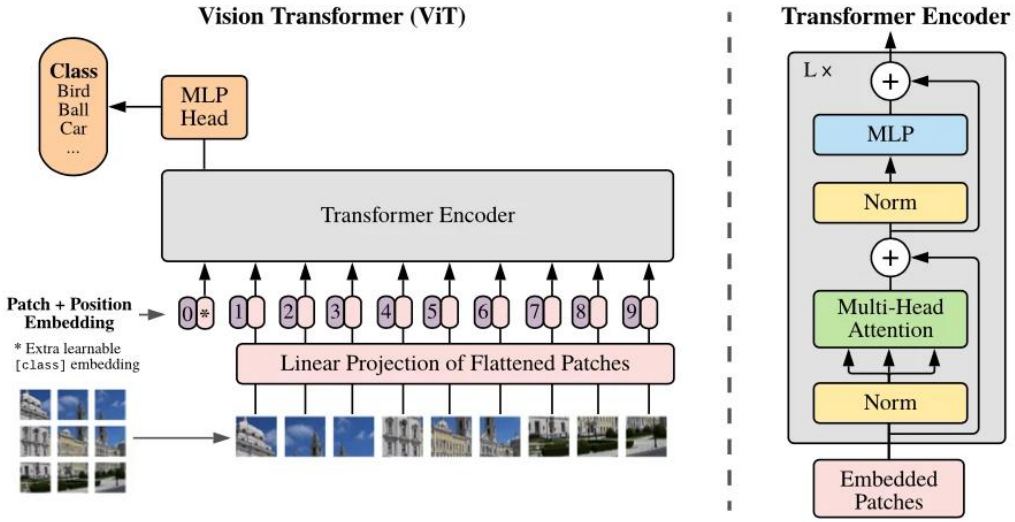
ICLR 2021 titled “An Image is Worth 16*16 Words: Transformers for Image Recognition at Scale” by google brain team

A transformer in machine learning is a deep learning model that uses the mechanisms of attention, differentially weighing the significance of each part of the input data.

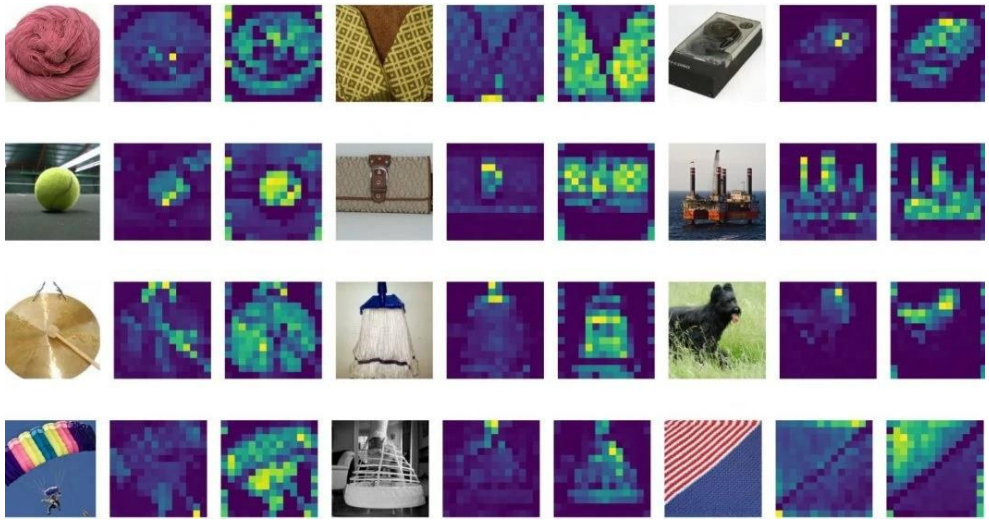
https://viso.ai/deep-learning/vision-transformer-vit/

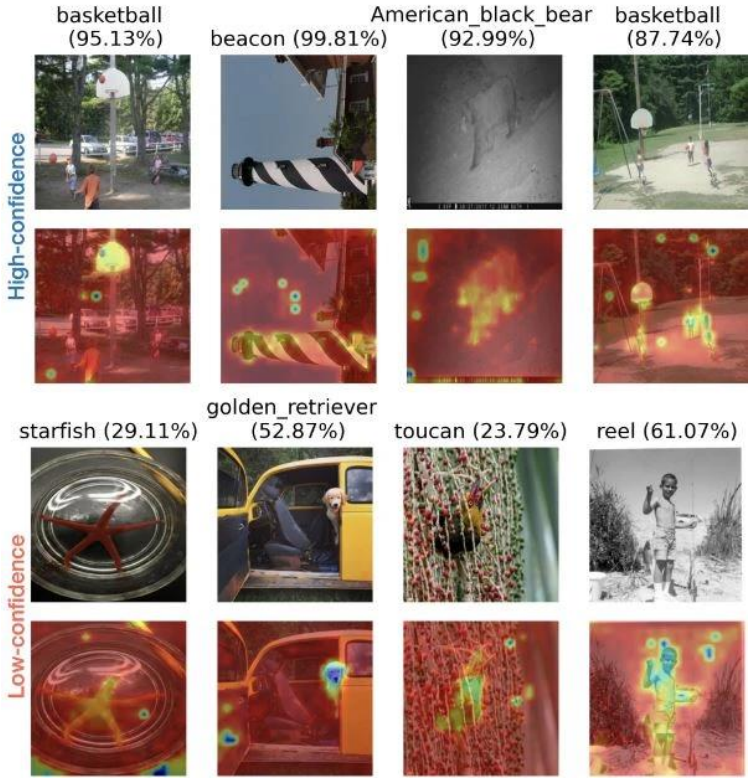
Vision Transformer (ViT)

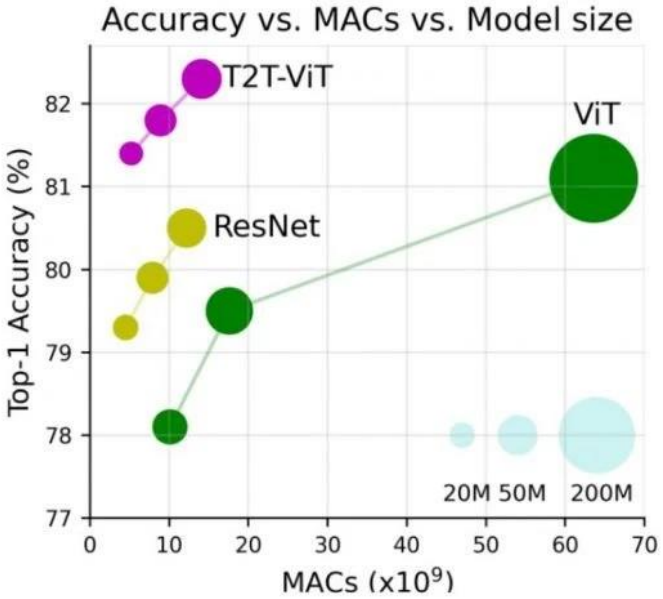
The ViT model represents an input image as a series of image patches, like the series of word embeddings used when using transformers to text, and directly predicts class labels for the image.

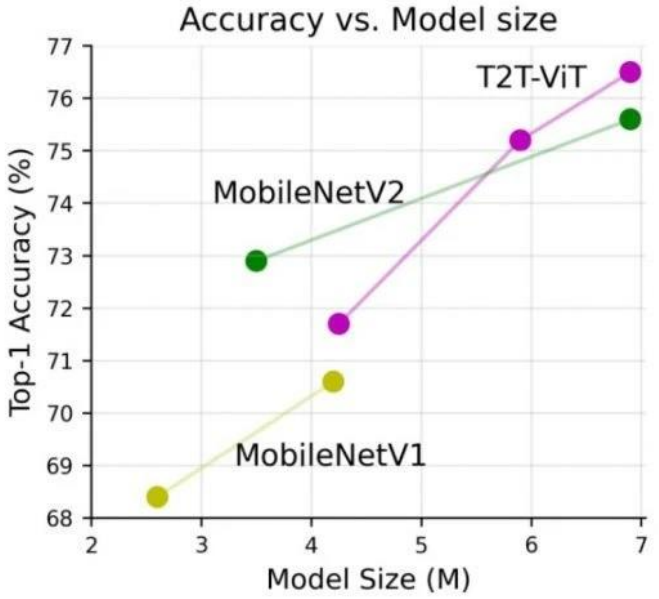


Vision Transformer (ViT)



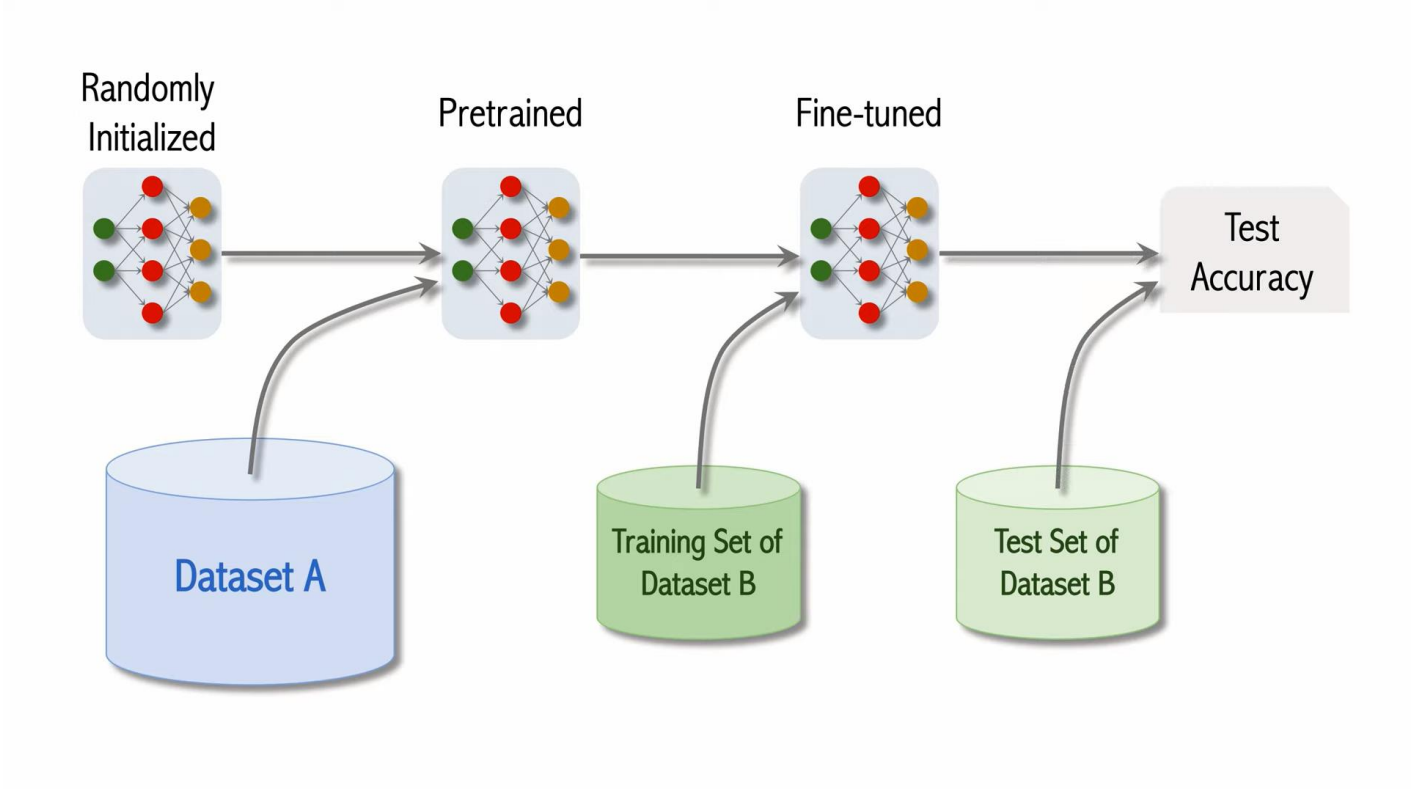



Accuracy vs. MACs vs. Model size
T2T-ViT
ViT
ResNet
20M 50M 200M
Top-1 Accuracy (%)
MACs (x10^9)
77
78
79
80
81
82
0
10
20
30
40
50
60
70


Accuracy vs. Model size
T2T-ViT
MobileNetV2
MobileNetV1
Top-1 Accuracy (%)
Model Size (M)
68
69
70
71
72
73
74
75
76
77
2
3
4
5
6
7

	# of Images	# of Classes
ImageNet (Small)	1.3 Million	1 Thousand
ImageNet-21K (Medium)	14 Million	21 Thousand
JFT (Big)	300 Million	18 Thousand

https://www.youtube.com/watch?v=HZ4j_U3FC94


Randomly Initialized
Pretrained
Fine-tuned
Test Accuracy
Dataset A
Training Set of Dataset B
Test Set of Dataset B

- Pretrain the model on Dataset A, fine-tune the model on Dataset B, and evaluate the model on Dataset B.

- Pretrained on ImageNet (small), ViT is slightly worse than ResNet.
- Pretrained on ImageNet-21K (medium), ViT is comparable to ResNet.
- Pretrained on JFT (large), ViT is slightly better than ResNet.

END

Thanks for your attention
Because attention is all you need ☺