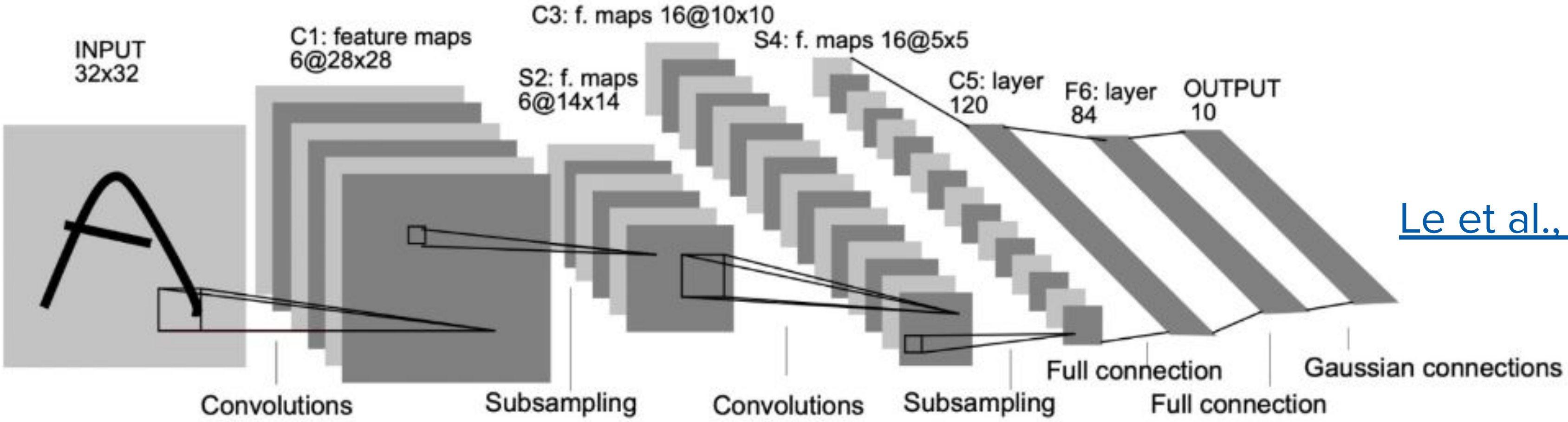
CNN Evolution

Deep Learning

Aziz Temirkhanov
Lambda, HSE

LeNet

- Introducing Convolutions into Deep Learning tasks
- Subsampling (pooling)
- FC Network as head



Le et al., 1998

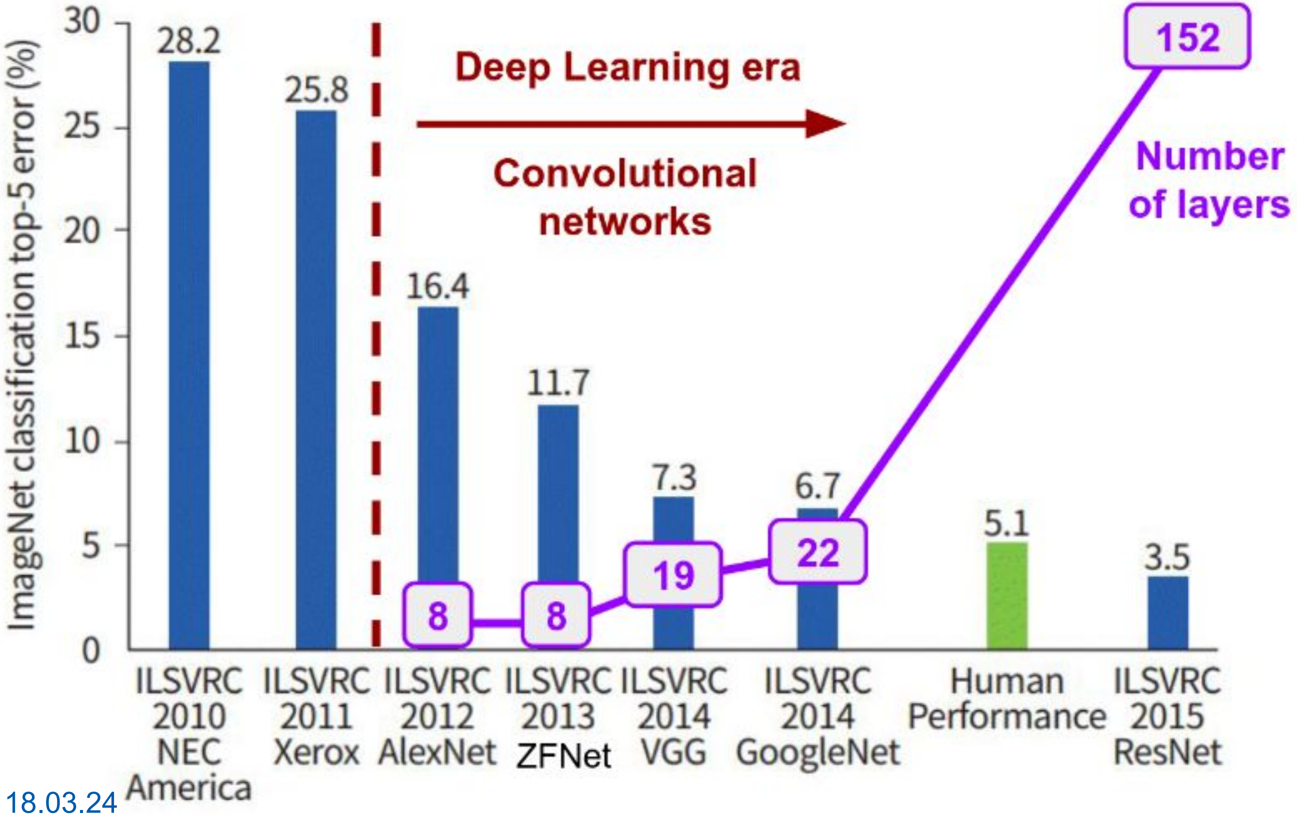
ImageNet Large Scale Visual Recognition Challenge

- 1000 classes
- Over 1M images (currently 14M+)
- Web Scraped data, annotated with Amazon MTurk
- Image Classification task that rapidly speed up the development of CV

O.Russakovsky et al., 2015

IMAGENET

ILSVRC


ImageNet classification top-5 error (%)
30
25
20
15
10
5
0
28.2
25.8
16.4
11.7
7.3
6.7
5.1
3.5
Deep Learning era
Convolutional networks
152
Number of layers
8
8
19
22
ILSVRC 2010 NEC America
ILSVRC 2011 Xerox
ILSVRC 2012 AlexNet
ILSVRC 2013 ZFNet
ILSVRC 2014 VGG
ILSVRC 2014 GoogleNet
Human Performance
ILSVRC 2015 ResNet

AlexNet

- Max Pooling, ReLU
- Dropout and Image Augmentation

AlexNet		
Top-1 acc	**Top-5 acc**	**#params**
56.5	79.0	61.1M

******Tables taken from torchvision models*

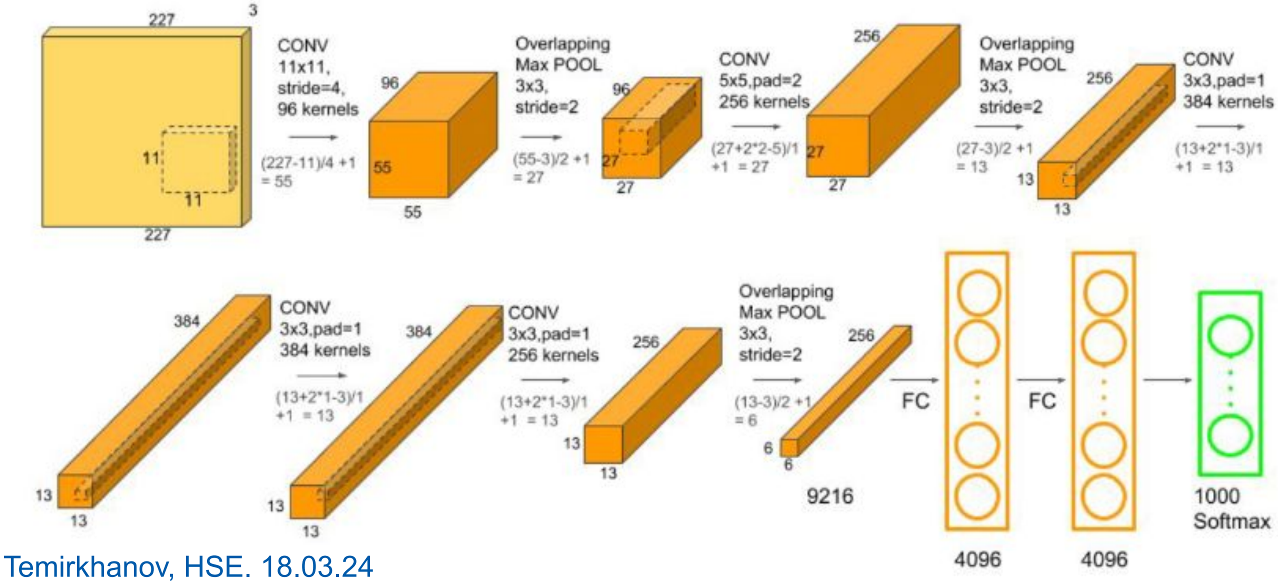
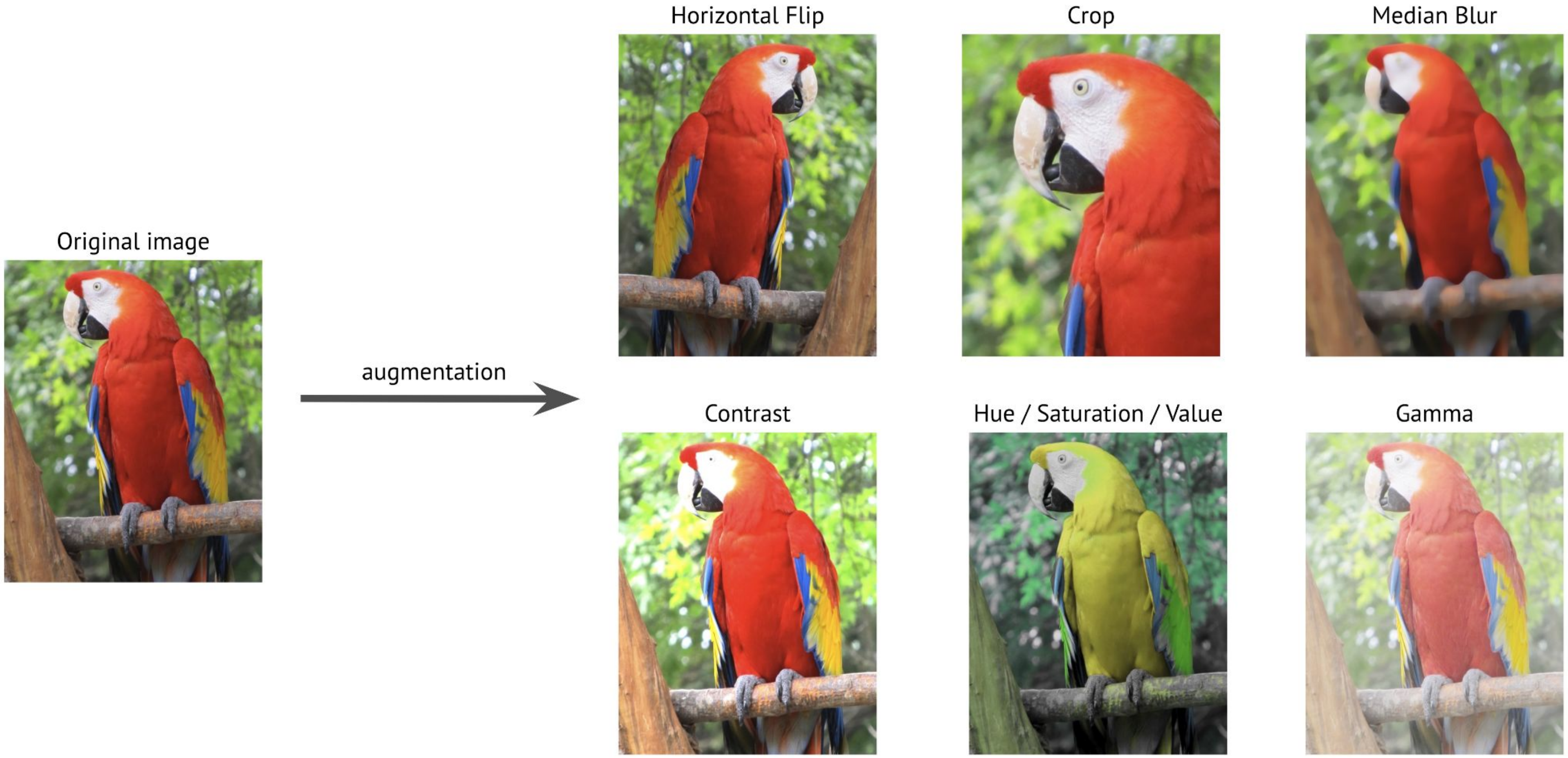


Image Augmentation

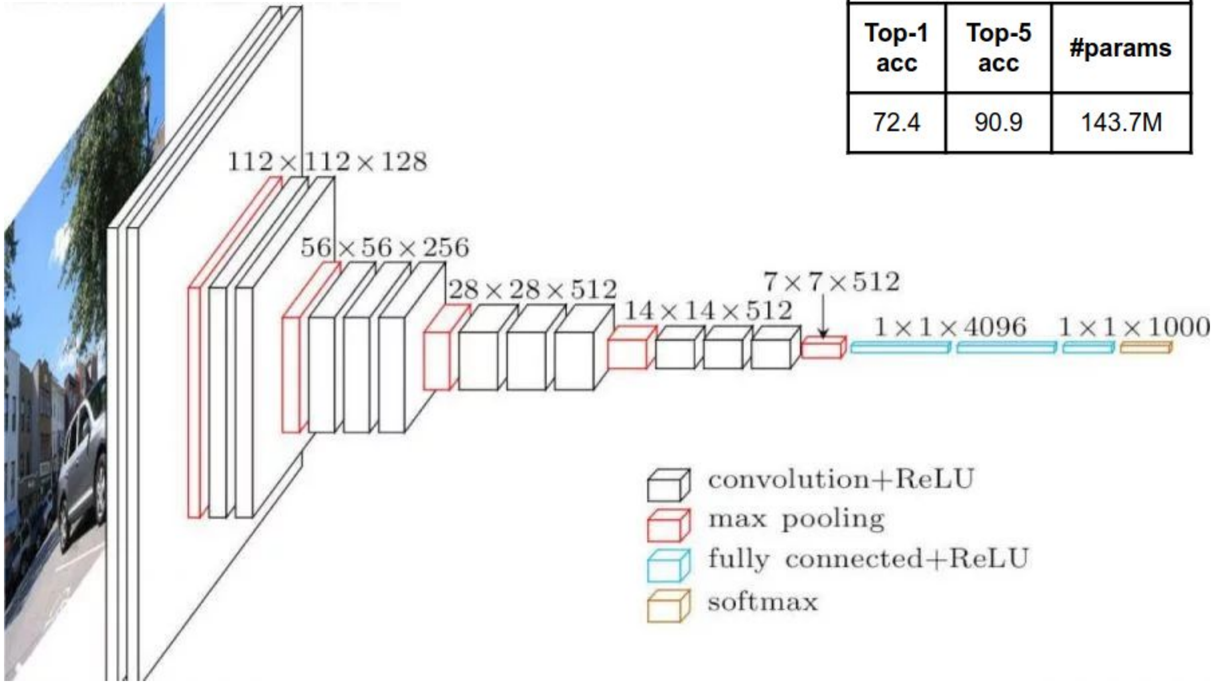


VGG

- Visual Geometry Group
- VGG16 and VGG19 — 16 or 19 layers
- Hard to train — vanishing gradient
- Trained in several stages

Simonyan and Zisserman, 2014

VGG-19		
Top-1 acc	**Top-5 acc**	**#params**
72.4	90.9	143.7M

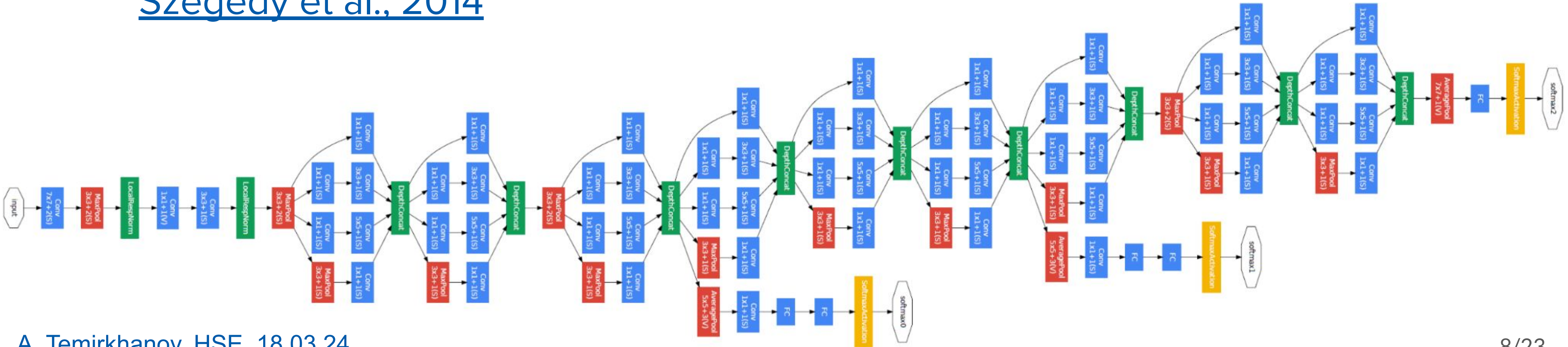


Inception

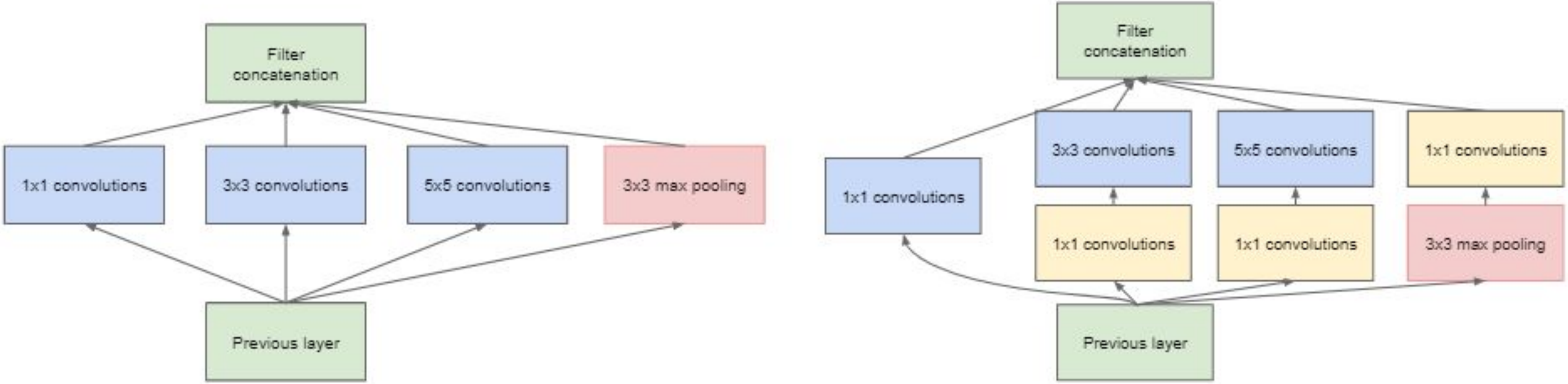
- Google LeNet or Inception
- Introduces a Inception block that computes several convolution simultaneously (i.e. in parallel)
- Does not train end-to-end, use Auxiliary Classifier

Szegedy et al., 2014

GoogLeNet		
Top-1 acc	**Top-5 acc**	**#params**
69.8	89.5	6.6M



Inception Block

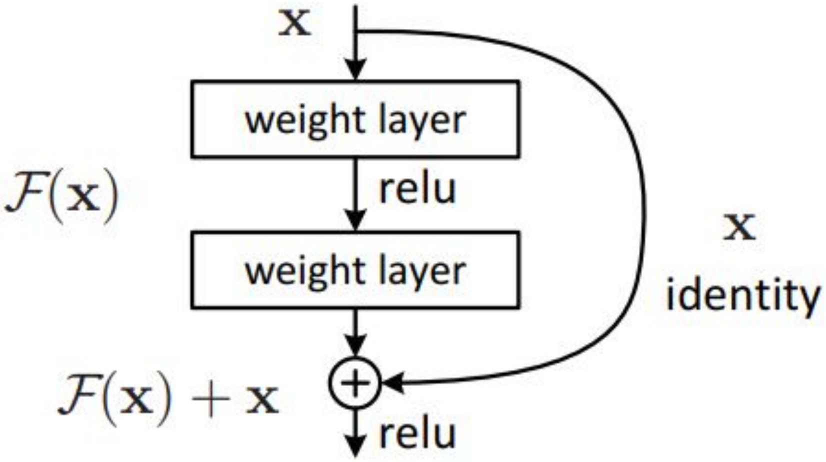


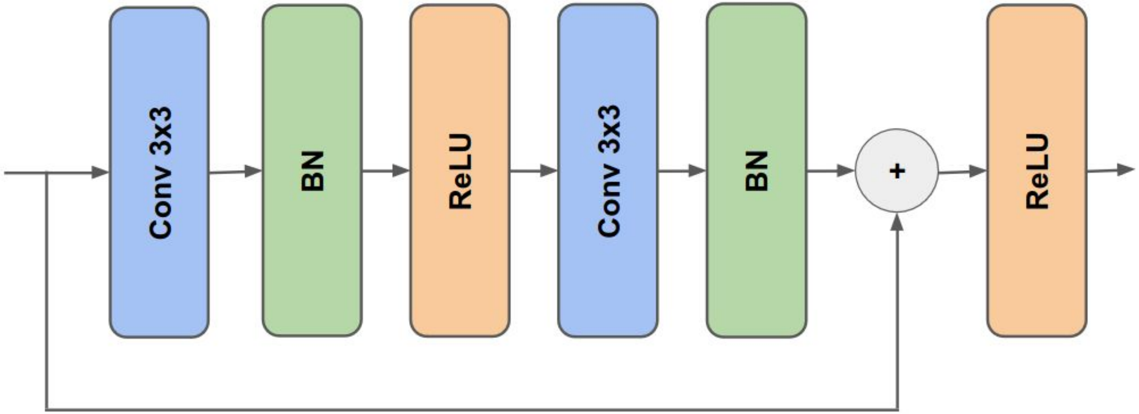
(a) Inception module, naïve version

(b) Inception module with dimension reductions

Skip Connection

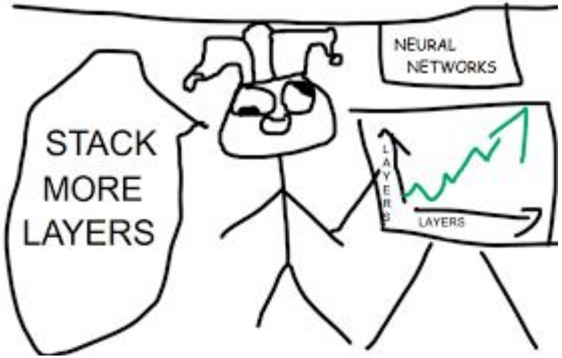
- Residual block or skip connection
- Mitigates Vanishing Gradient problem
- Thus, can stack much more layers!



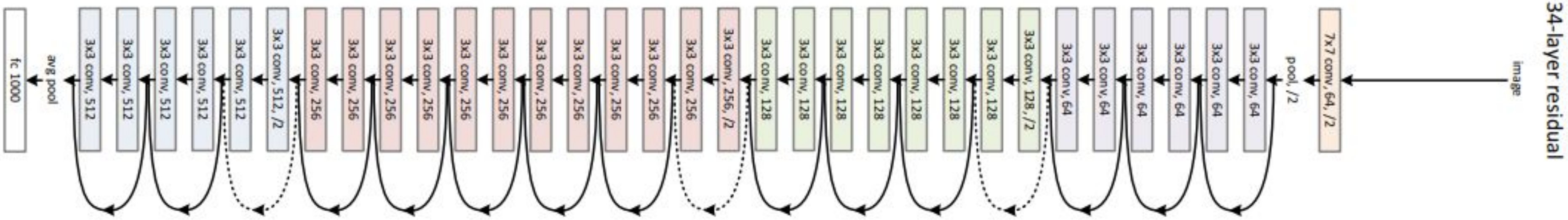


ResNet Family

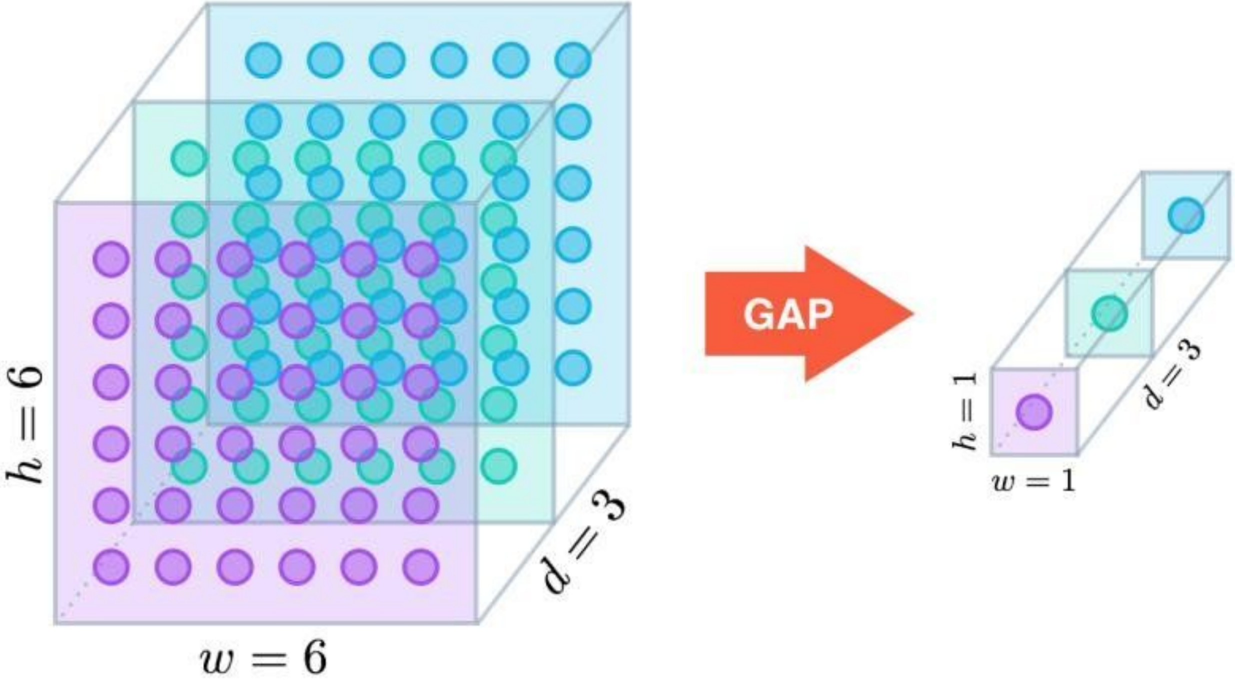
- Introduce Skip Connection (Residual) Block
- Stack more layers!
- BN to stabilize training
- No max pooling
- Global Average Pooling
- ResNet18, ResNet34, ResNet50, ResNet101, Resnet152



He et al., 2015

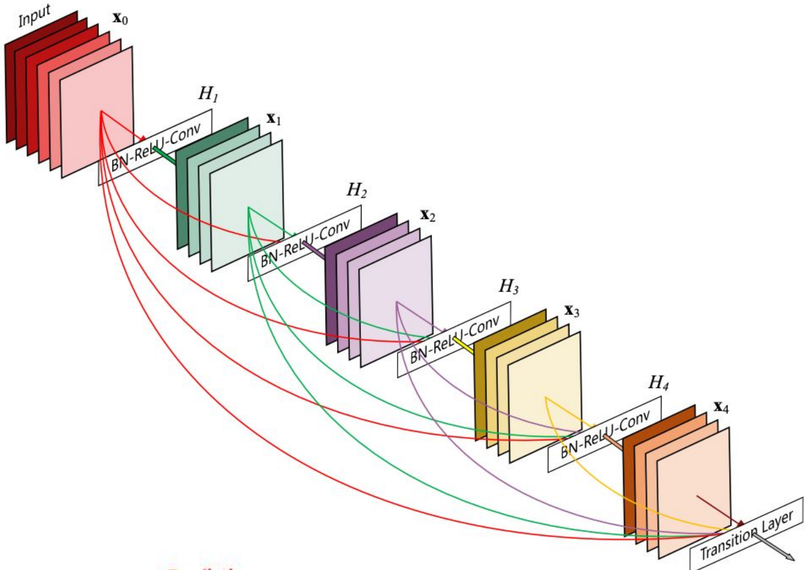


Global Average Pooling

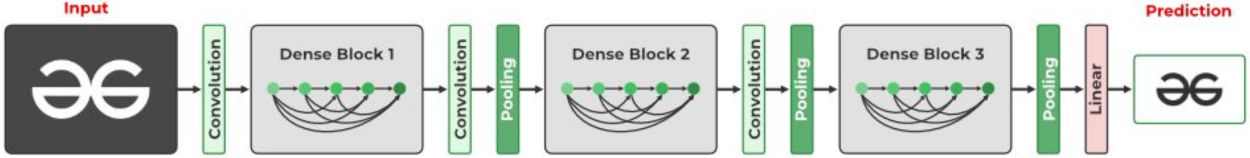


DenseNet

- Introduces a DenseBlock
- Any 2 layers are connected
- Channel-wise feature map concatenation



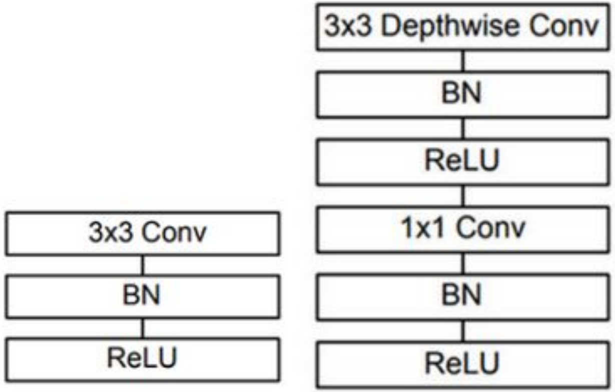
[Huang et al., 2016]

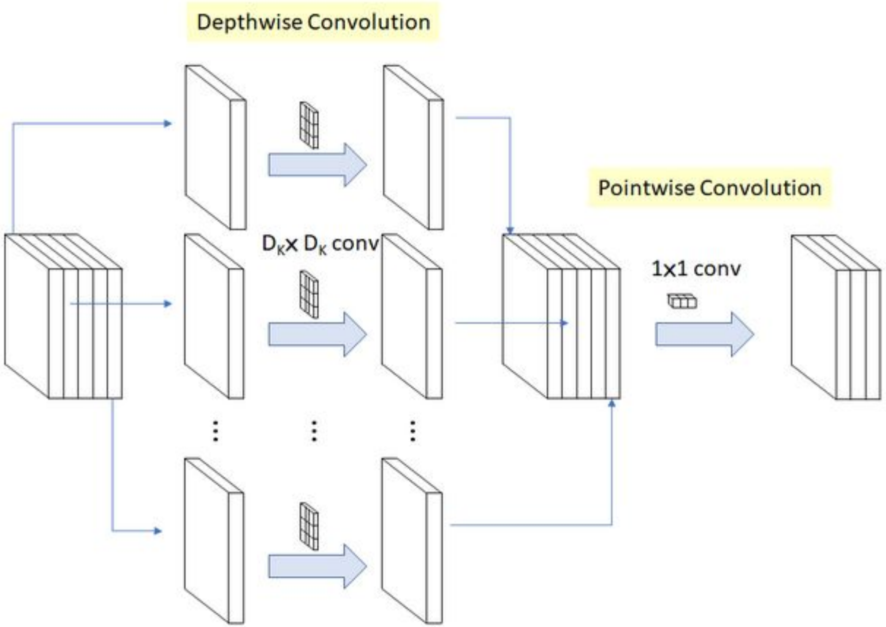


MobileNet

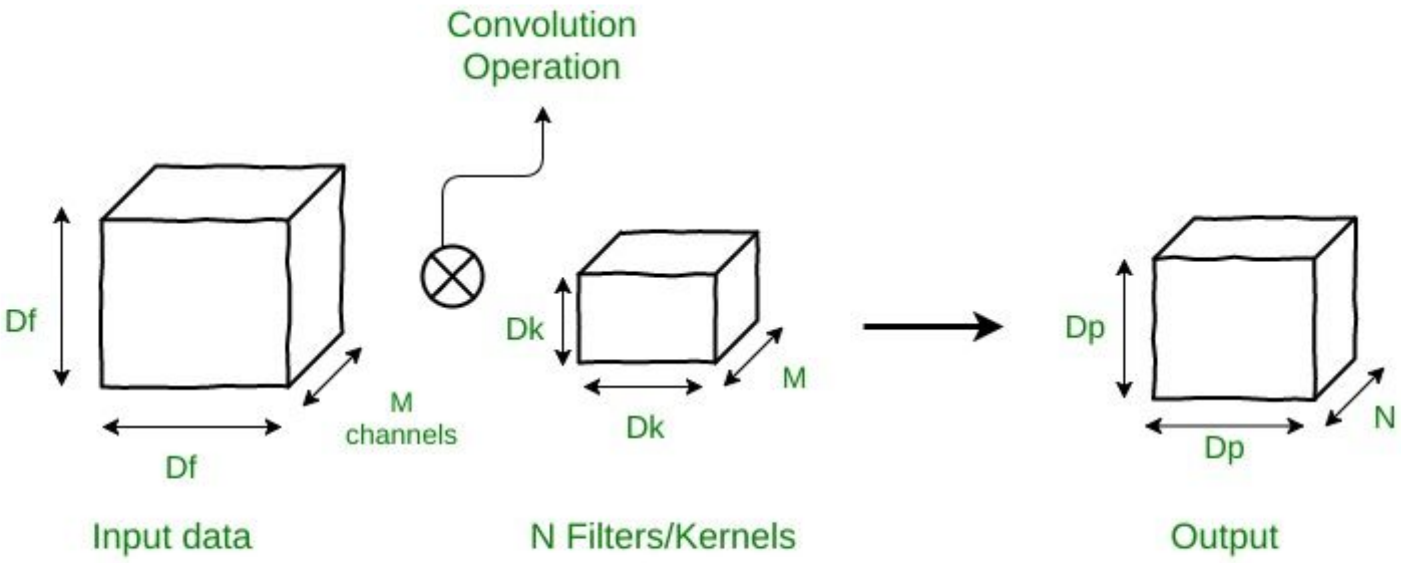
- Optimized for mobile devices
- Depthwise and pointwise convolutions

Howard et al., 2017

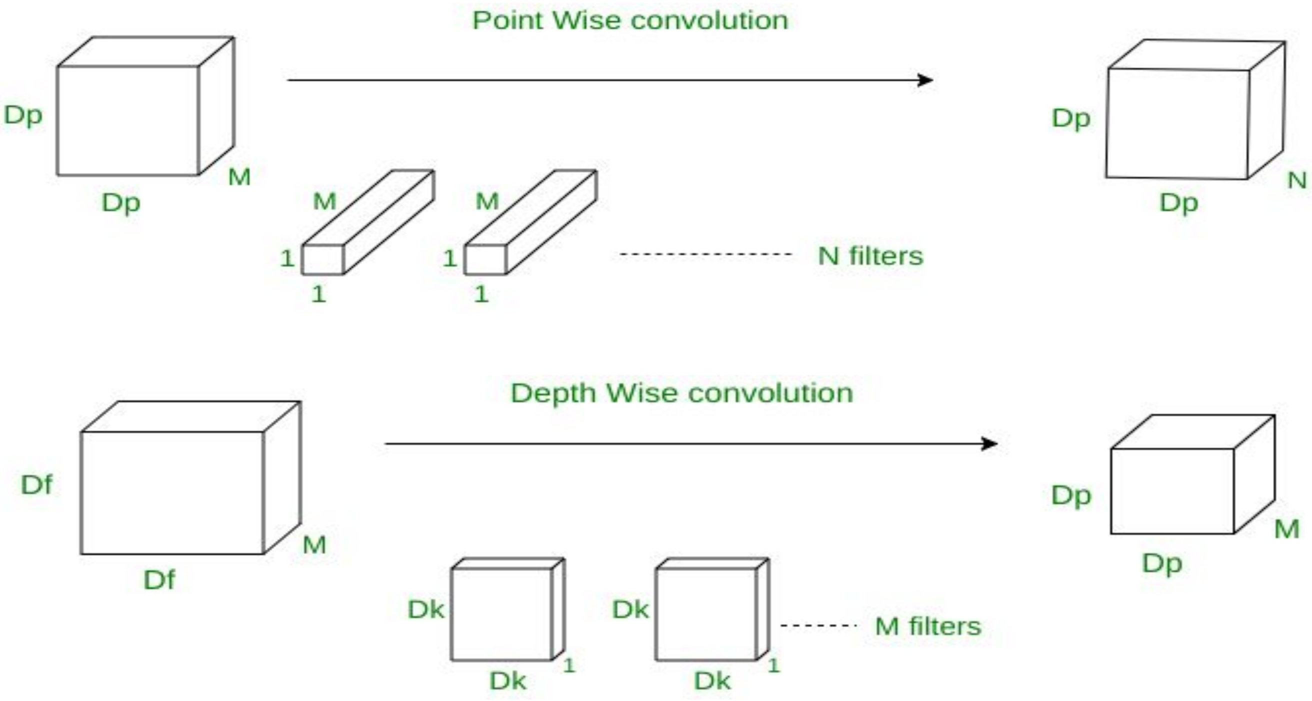




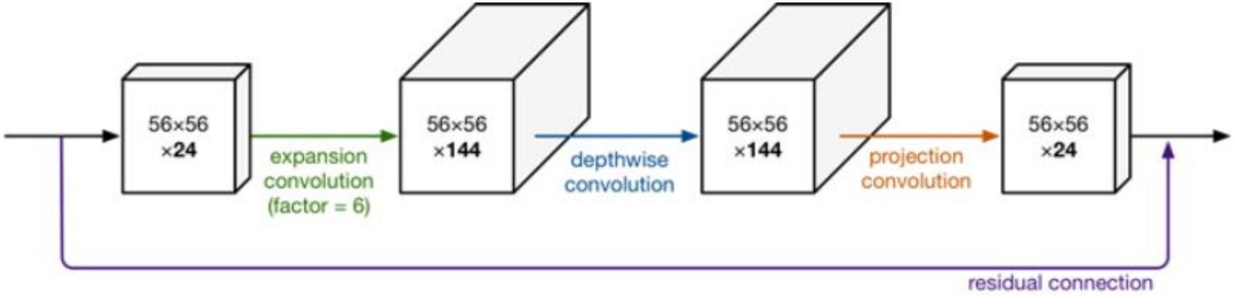
Convolution Recap

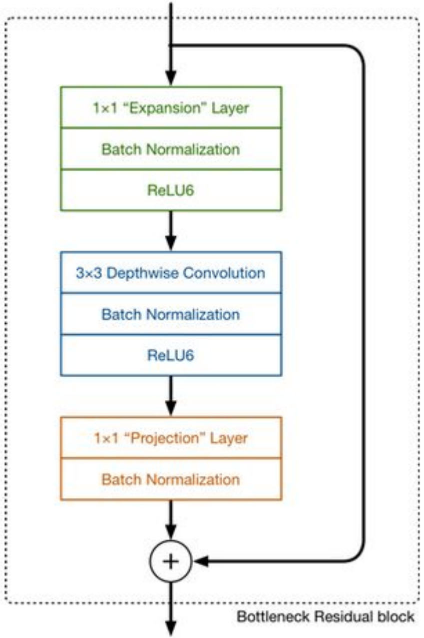


Depth Wise and Point Wise Convolutions



MobileNet v2 and v3





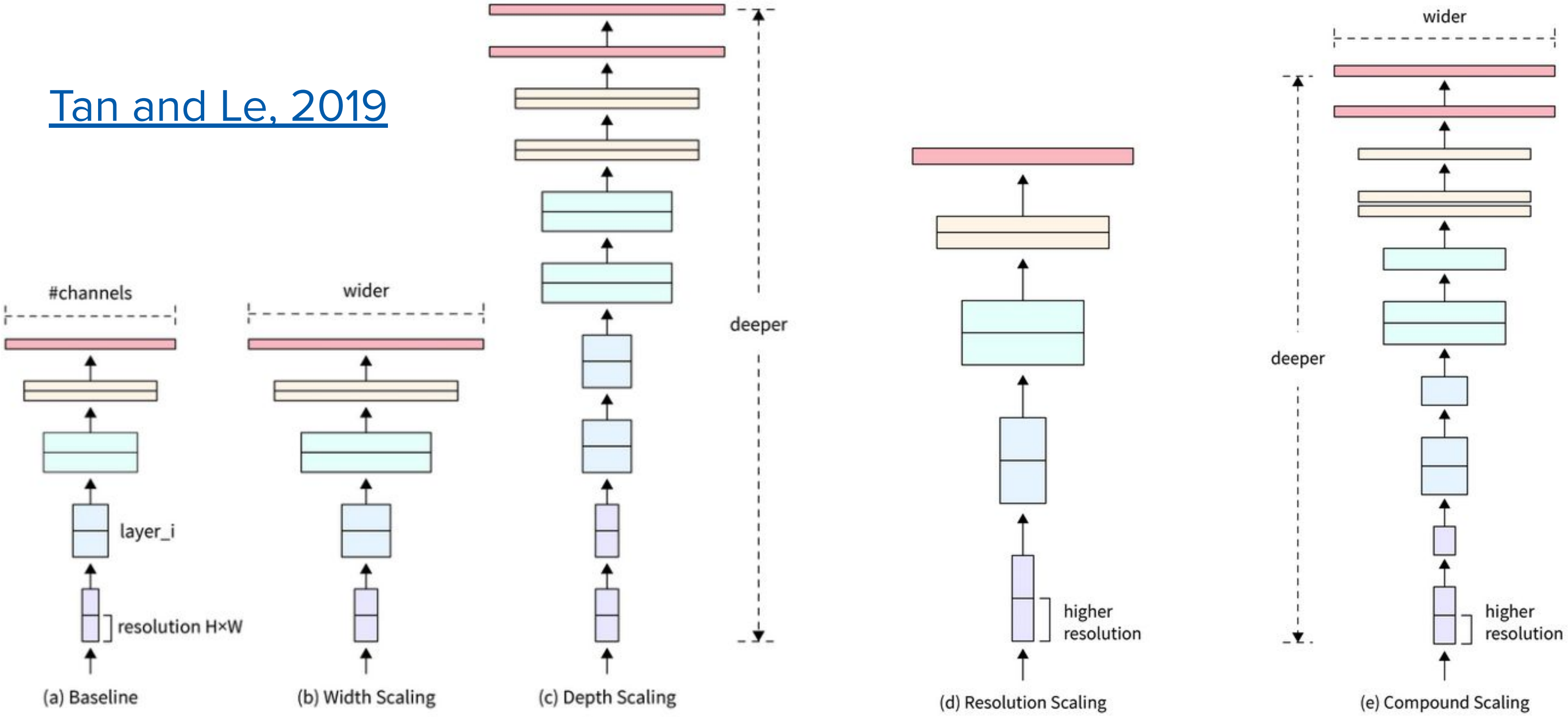
- Reduce the number of channels
- Residual Connections
- NAS for v3 version

Sandler et al., 2018

Howard et al., 2019

EfficientNet

Tan and Le, 2019



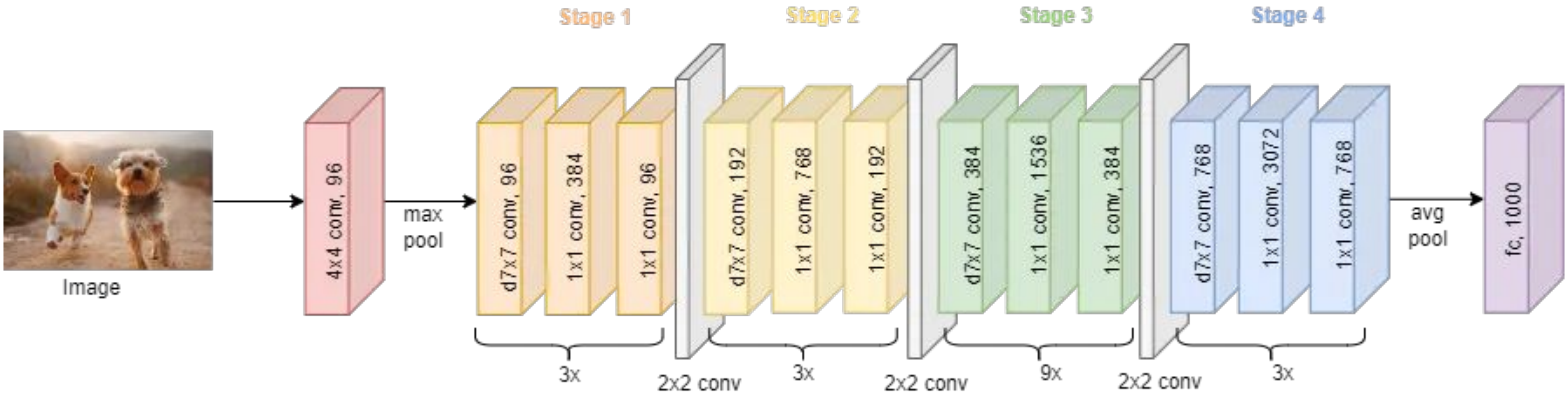
Transformers Era?

- Introduced in 2017, Transformer architecture (attention mechanism) had rapidly took over NLP domen
- Main rule of ML — apply any good idea from different domain to your task
- Several years later ViT arrived at the scene
- But still, CNN is widely used
- Easier and faster to train, better as baseline model and out-of-box model

Not quite

- ConvNext: A ConvNet for the 2020s
- NFNet: ConvNet without normalizations

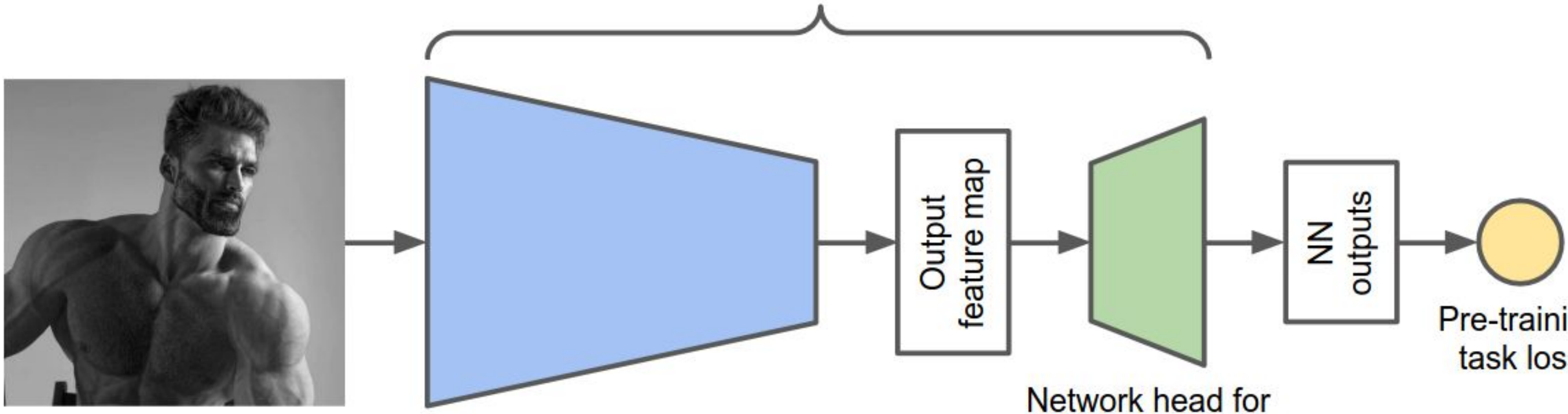
Liu et al, 2022
Brok et al, 2021
Smith et al., 2023



Downstream tasks

Transfer Learning: Pre-Training

Train both backbone and head



Pre-training (large) dataset image

Network backbone (body), convolutional encoder

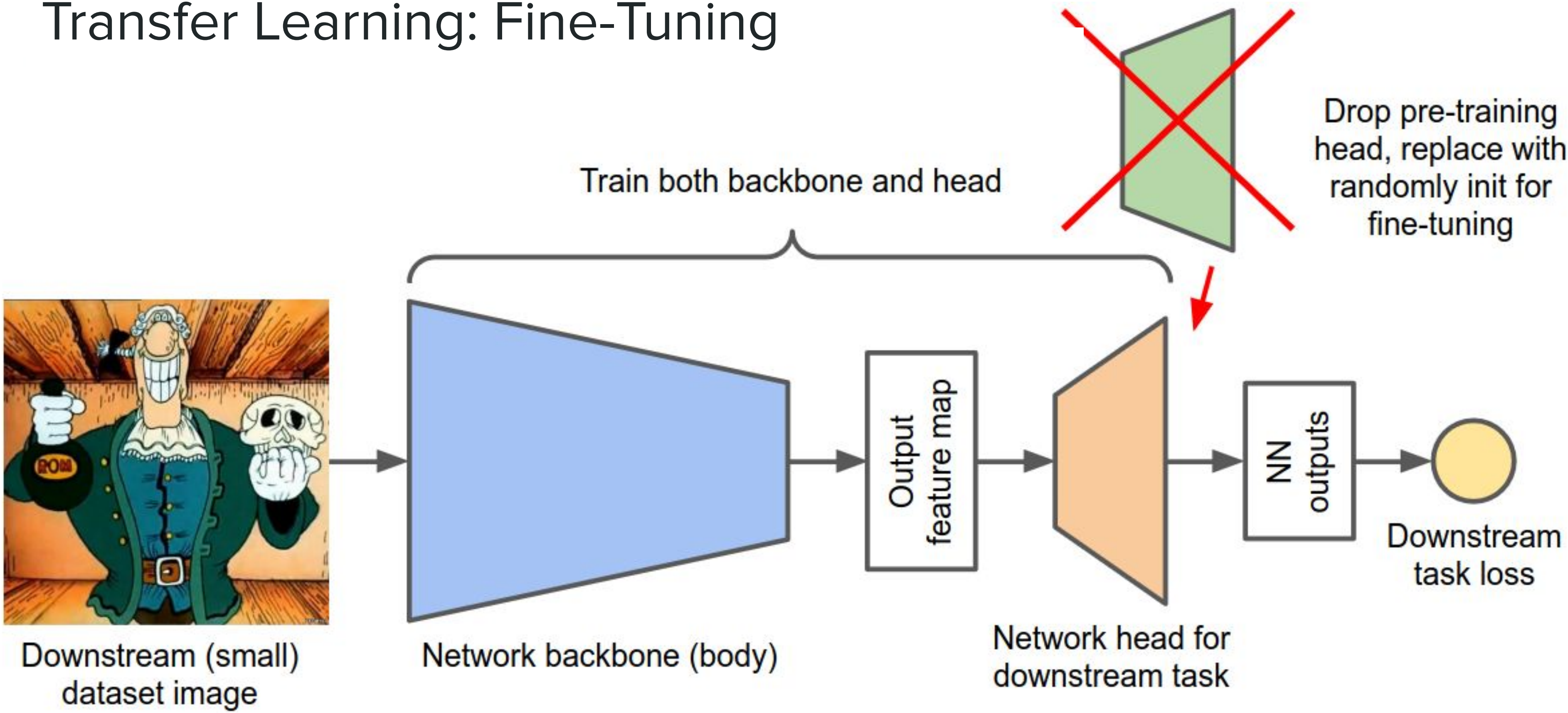
Output feature map

Network head for pre-training task, usually linear layer

NN outputs

Pre-training task loss

Transfer Learning: Fine-Tuning



Train both backbone and head

Drop pre-training head, replace with randomly init for fine-tuning

Output feature map

NN outputs

Downstream (small) dataset image

Network backbone (body)

Network head for downstream task

Downstream task loss