Deep Learning

Deeper, Better, _____ , Stronger than Machine Learning

Valentin Barriere

Universidad de Chile – DCC

CC6204, Primavera 2024

Convolutional Architecture

Outline : ImageNet Moment

ImageNet Moment

CNN classical Architectures

ImageNet Moment: The boom of Neural Nets in 2012

We have seen that

- NNs learn themselves what to extract in terms of interesting features
- NNs compose lower-level features to create higher-level features
- The bigger the training set is, the more complex the task can be and the bigger the model can be!

Let's see now how it started to take-off

ImageNet Moment: The boom of Neural Nets in 2012

We have seen that

- NNs learn themselves what to extract in terms of interesting features
- NNs compose lower-level features to create higher-level features
- The **bigger the training set is**, the more complex the task can be and the bigger the model can be!

Let's see now how it started to take-off: "the ImageNet moment"

Figure 1: ImageNet is large-scale dataset for image classification

What is ImageNet?

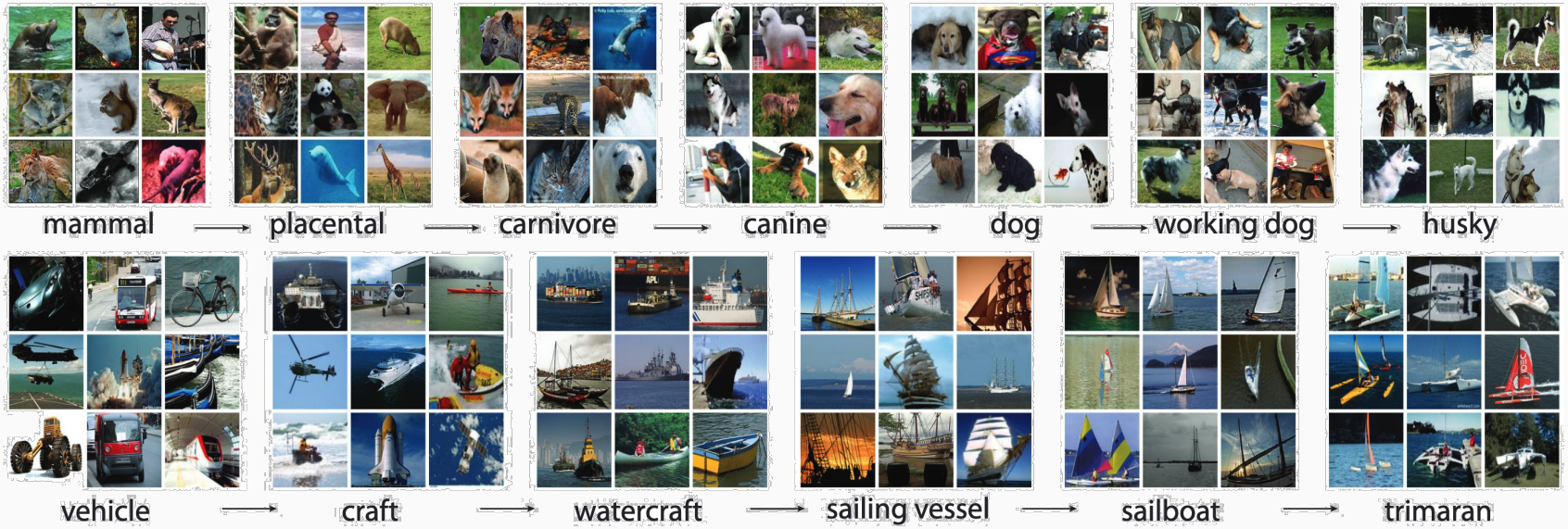


Figure 2: ImageNet dataset: 14M images for 21k hierarchical classes [1]

- Many old-school models used a subpart of it called Imagenet Large Scale Visual Recognition Challenge dataset (ILSVR) to pre-train:
  - ILSVR has 1.2M (200K test) images, for 1k classes
  - Use top-1 and top-5 accuracy as metrics for the challenge

What is ImageNet?

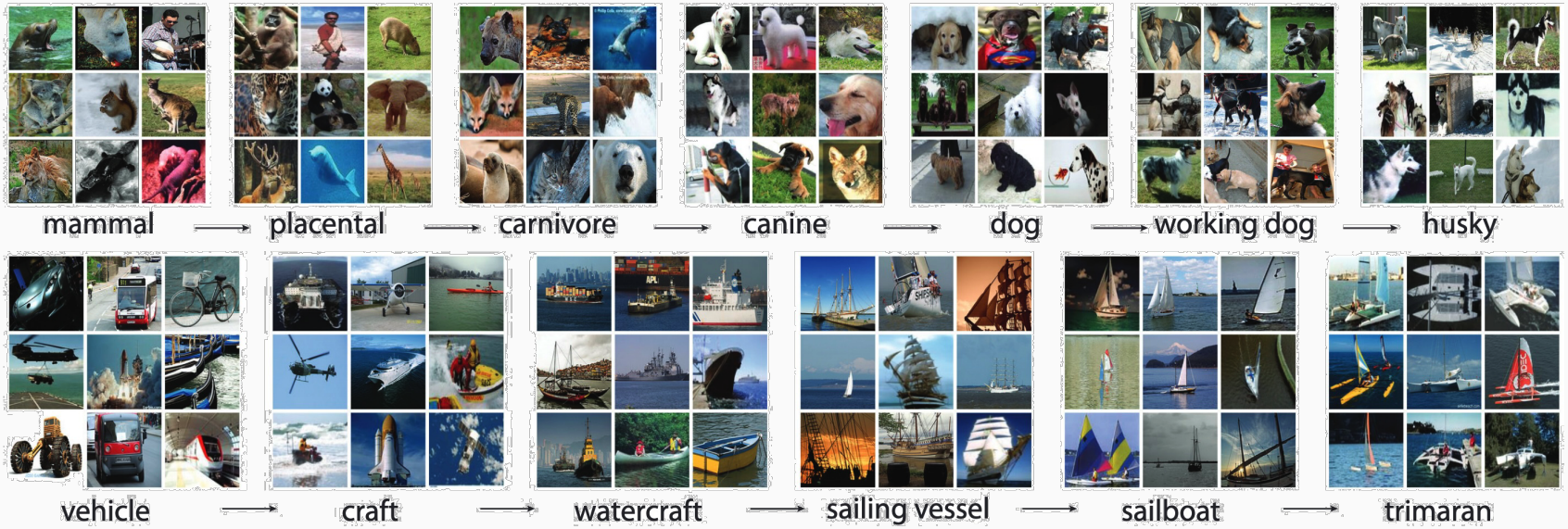
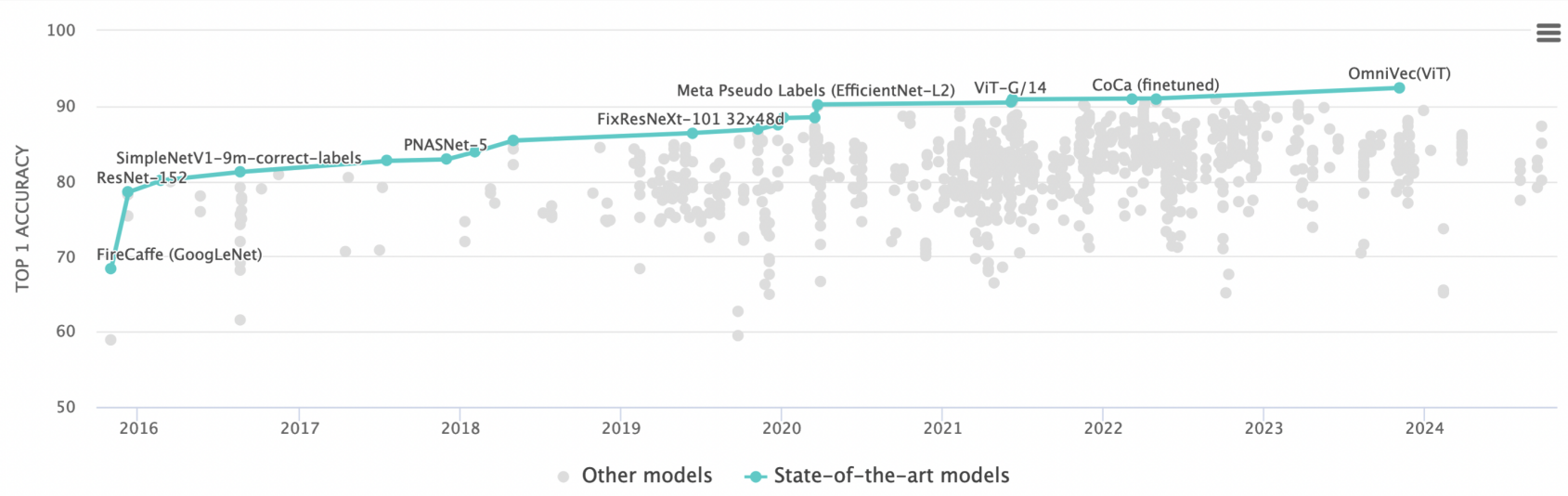


Figure 2: ImageNet dataset: 14M images for 21k hierarchical classes [1]

- Many old-school models used a subpart of it called Imagenet Large Scale Visual Recognition Challenge dataset (ILSVR) to pre-train:
  - ILSVR has 1.2M (200K test) images, for 1k classes
  - Use top-1 and top-5 accuracy as metrics for the challenge
- The full dataset has been pre-processed later to allow for good pre-training [7]

ImageNet

Still used nowadays as a benchmark to evaluate models performances!



More details [here]

Deeper and deeper

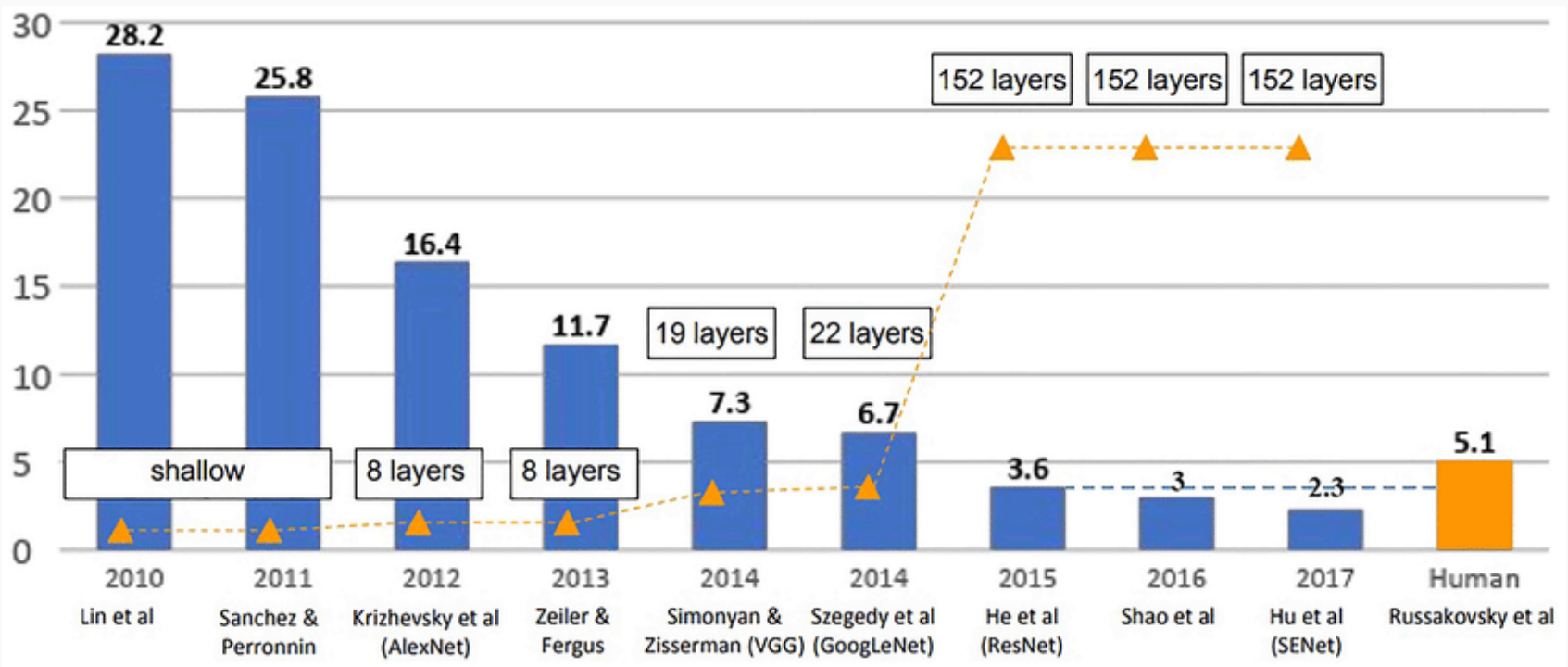


Figure 3: The winners of ILSRVC through the years, with the associated depth

- In 2012: AlexNet is the first NN winning the competition, reducing drastically the relative error by 36%

Deeper and deeper

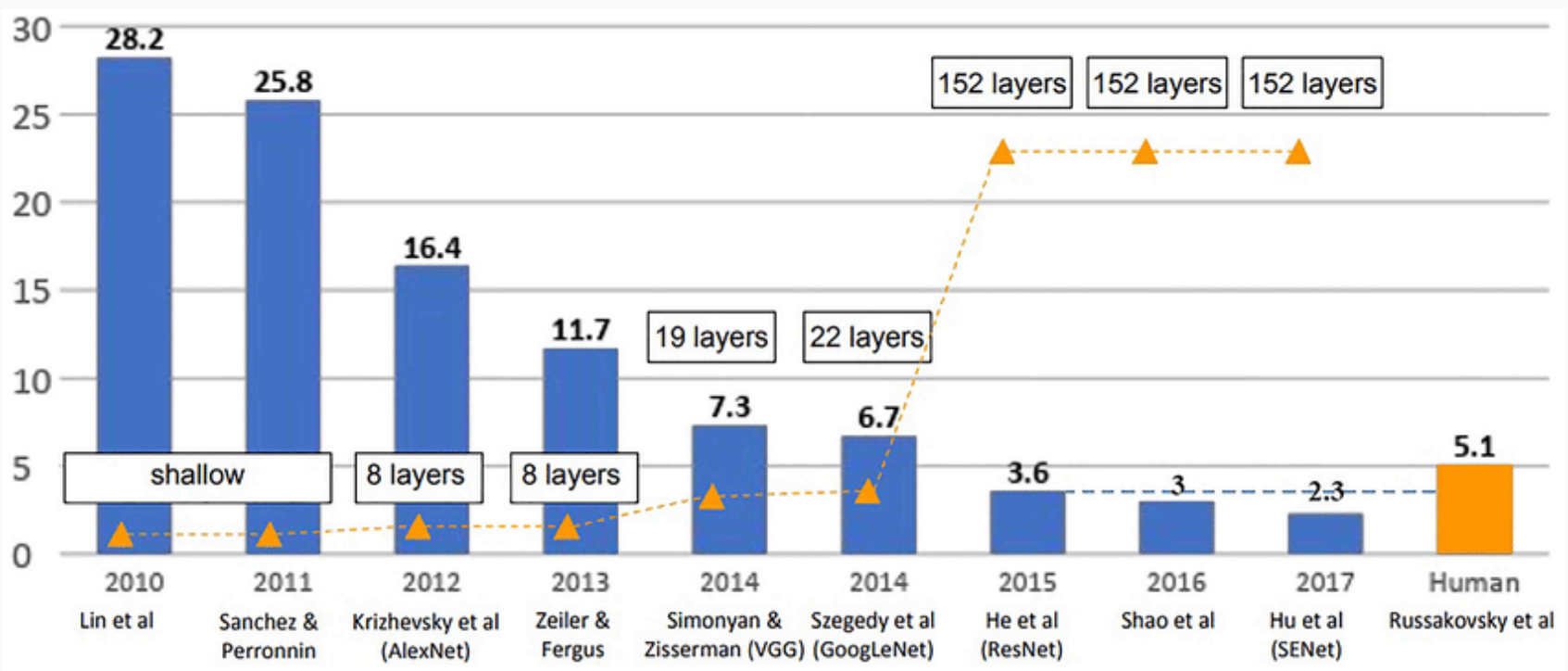


Figure 3: The winners of ILSRVC through the years, with the associated depth

- In 2012: AlexNet is the first NN winning the competition, reducing drastically the relative error by 36%
- In 2015: Resnet
  - Reduces drastically the relative error by 46%
  - It increased drastically the depth of the network to 152 layers, this moment was called the "Revolution of Depth"

Outline : CNN classical Architectures

ImageNet Moment

CNN classical Architectures

LeNet (from Yann Lecun) [5]

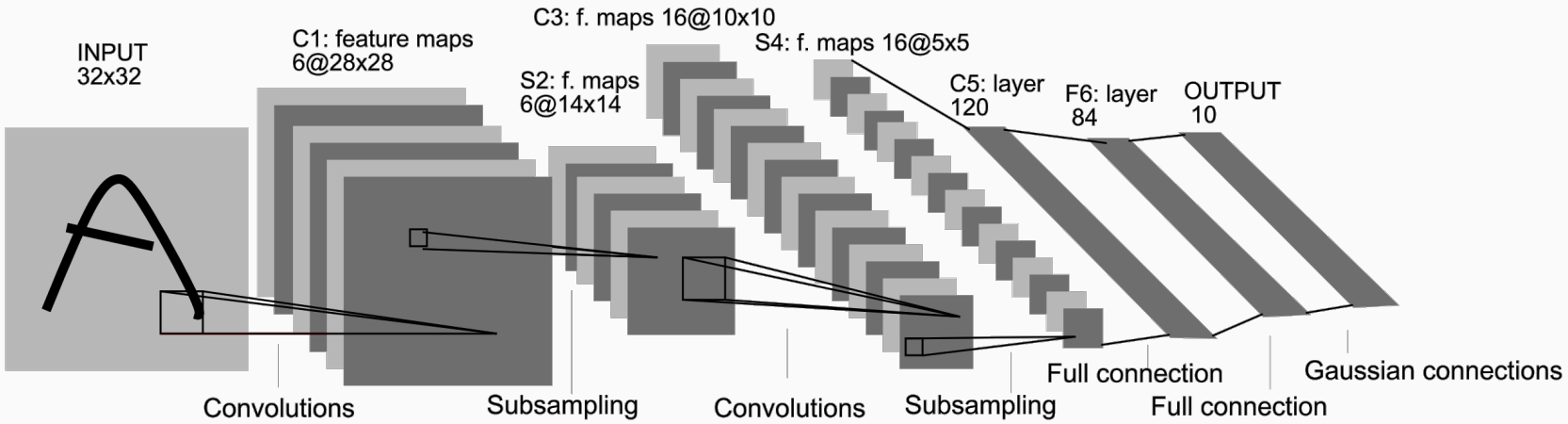


Figure 4: LeNet-5 used for handwritting digits recognition (MNIST dataset)

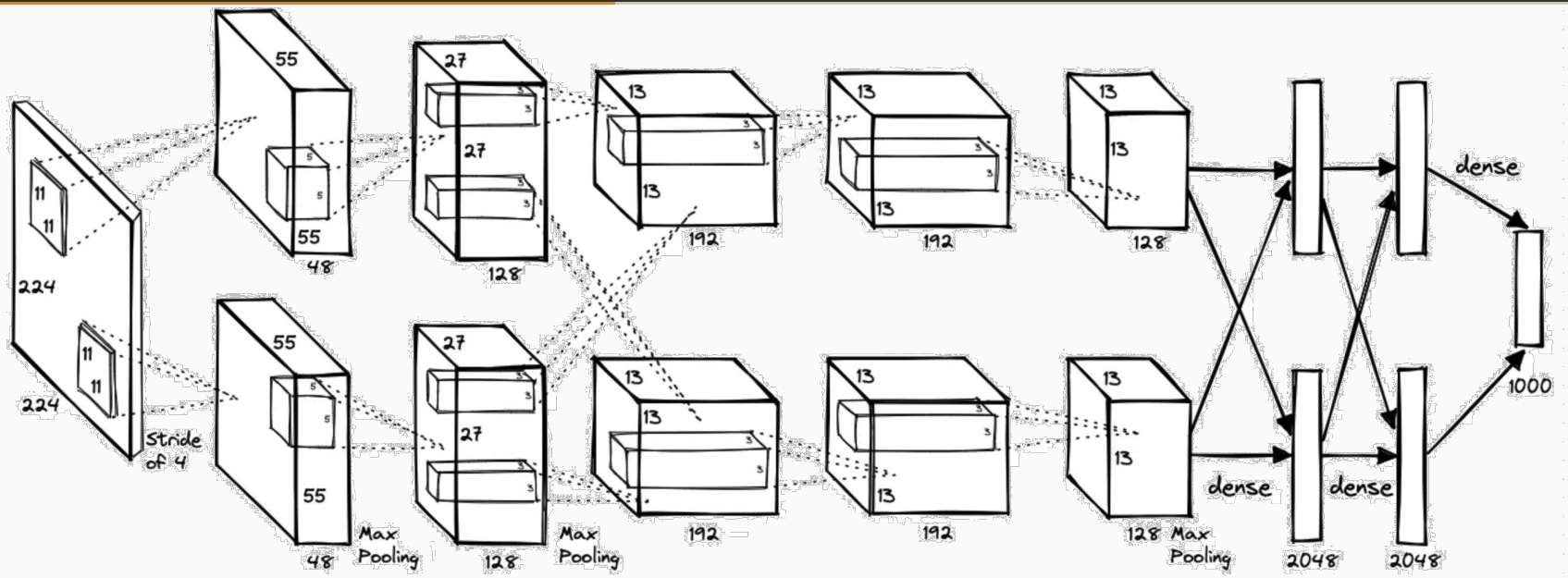
- The first famous CNN
- 2 times ConvLayer with 5x5 filters and stride 1
- Mean pooling 2x2 with stride 2
- Fully-connected layers (MLP)

Overall: AlexNet

Model Name	# of Parameters	Top-1 Accuracy	Year
AlexNet	**60 M**	**63.3 %**	**2012**
Inception V1	5 M	69.8 %	2014
VGG 16	138 M	74.4 %	2014
VGG 19	144 M	74.5 %	2014
Inception V2	11.2 M	74.8 %	2015
ResNet-50	26 M	77.15 %	2015
ResNet-152	60 M	78.57 %	2015
Inception V3	27 M	78.8 %	2015
DenseNet-121	8 M	74.98 %	2016
DenseNet-264	22 M	77.85 %	2016
EfficientNet-B0	5.3 M	76.3 %	2019
EfficientNet-B4	19 M	82.6 %	2019
EfficientNet-B7	66 M	84.4 %	2019
BiT-L (ResNet)	928 M	87.54 %	2019
Meta Pseudo Labels	480 M	90.2 %	2021

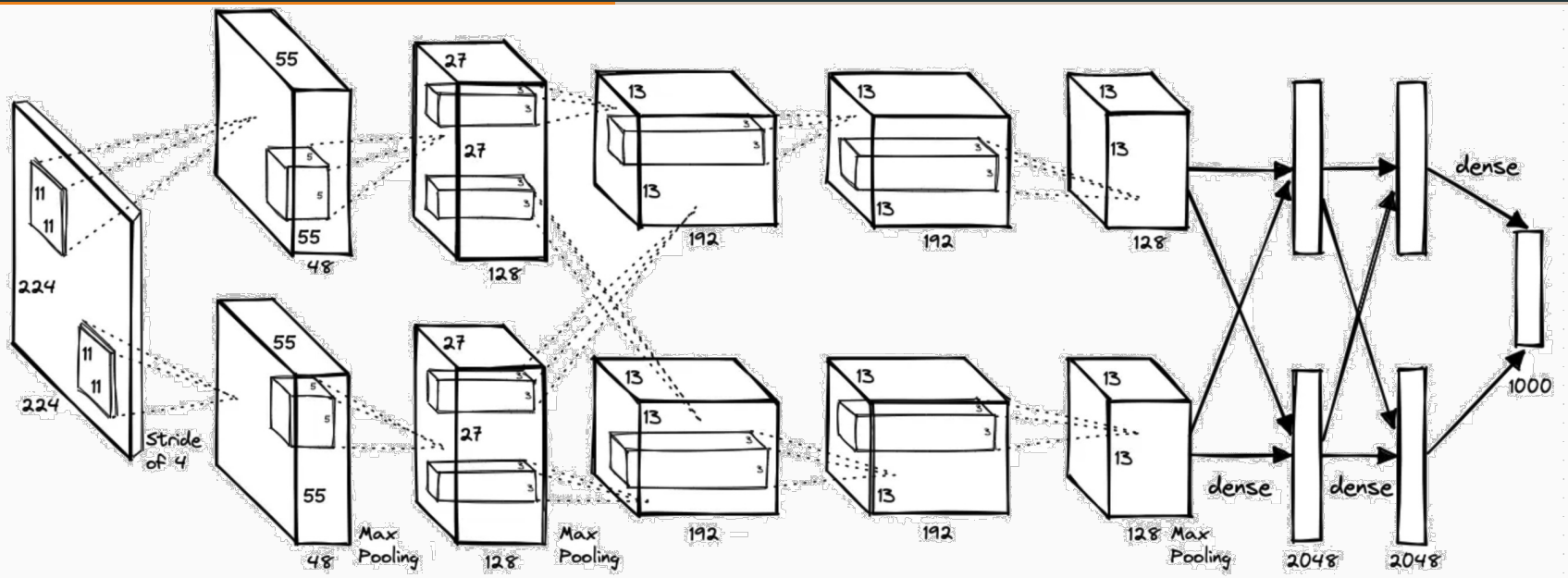
Table 1: Comparison of Different Models on ImageNet

AlexNet (from Alex Krizhevsky) [4]



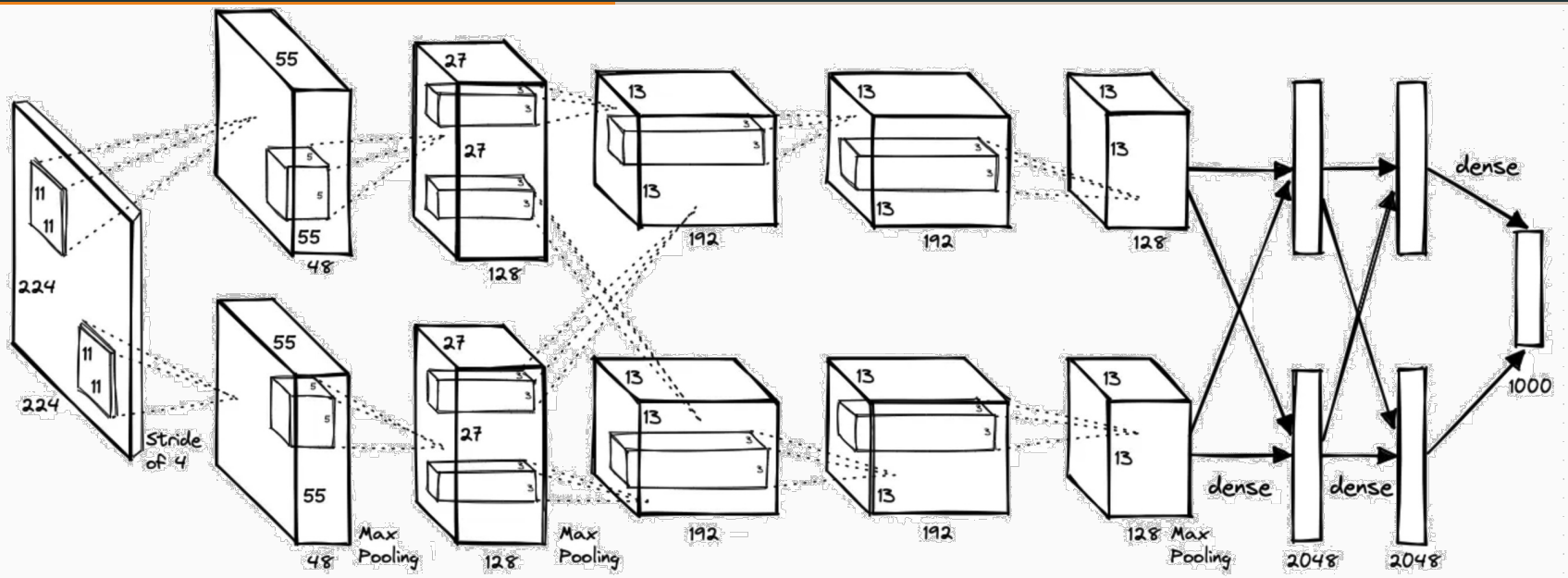
- Much deeper than previous CNN architectures

AlexNet (from Alex Krizhevsky) [4]



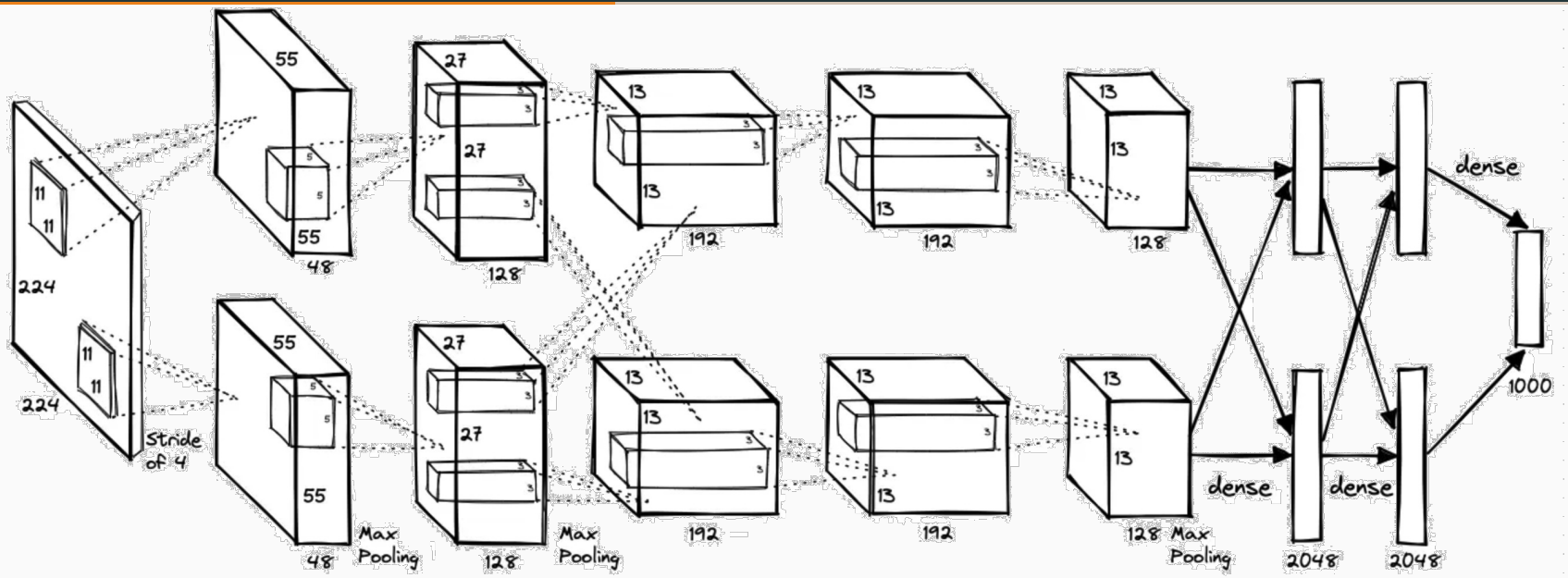
- Much deeper than previous CNN architectures
- Introduced ReLU activation to alleviate vanishing gradient problem

AlexNet (from Alex Krizhevsky) [4]



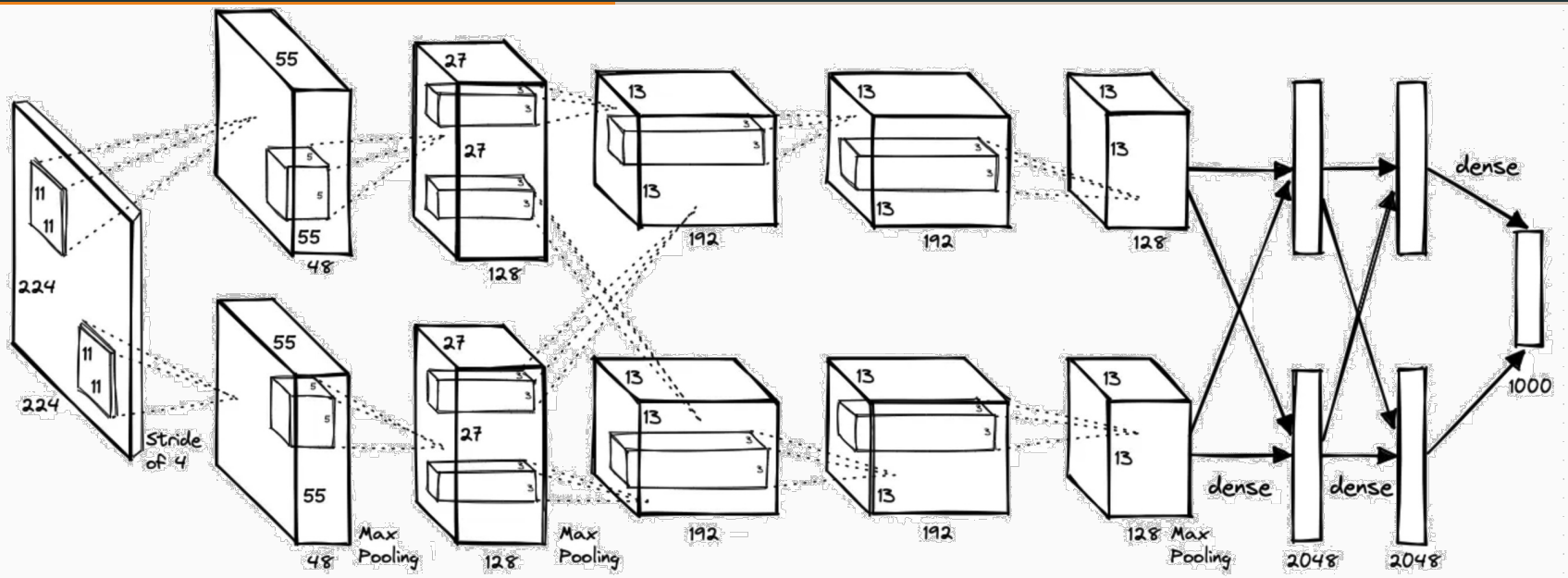
- Much deeper than previous CNN architectures
- Introduced ReLU activation to alleviate vanishing gradient problem
- Parallelized two GPUs to train the model, with two streams combining at specific layers

AlexNet (from Alex Krizhevsky) [4]



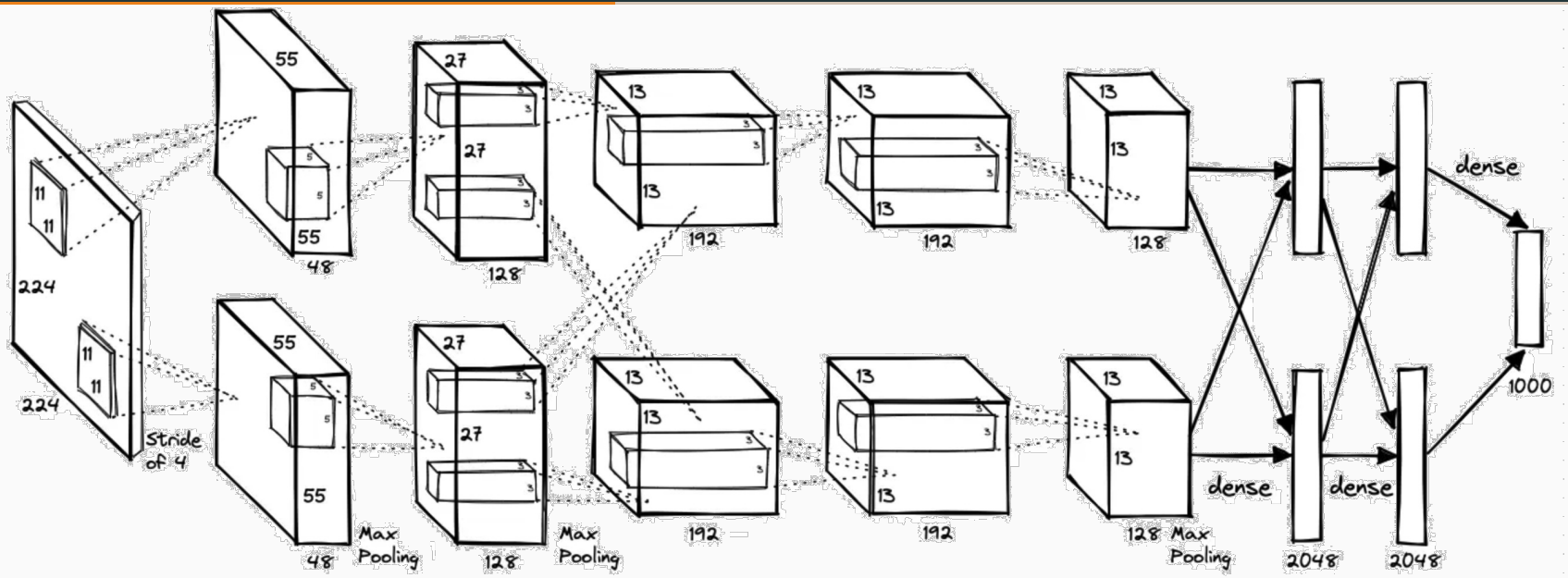
- Much deeper than previous CNN architectures
- Introduced ReLU activation to alleviate vanishing gradient problem
- Parallelized two GPUs to train the model, with two streams combining at specific layers
- Heavy Dropout and Data Augmentation improve robustness

AlexNet (from Alex Krizhevsky) [4]



- Much deeper than previous CNN architectures
- Introduced ReLU activation to alleviate vanishing gradient problem
- Parallelized two GPUs to train the model, with two streams combining at specific layers
- Heavy Dropout and Data Augmentation improve robustness
- Overlapping max pooling reduces side effect and improves robustness

AlexNet (from Alex Krizhevsky) [4]



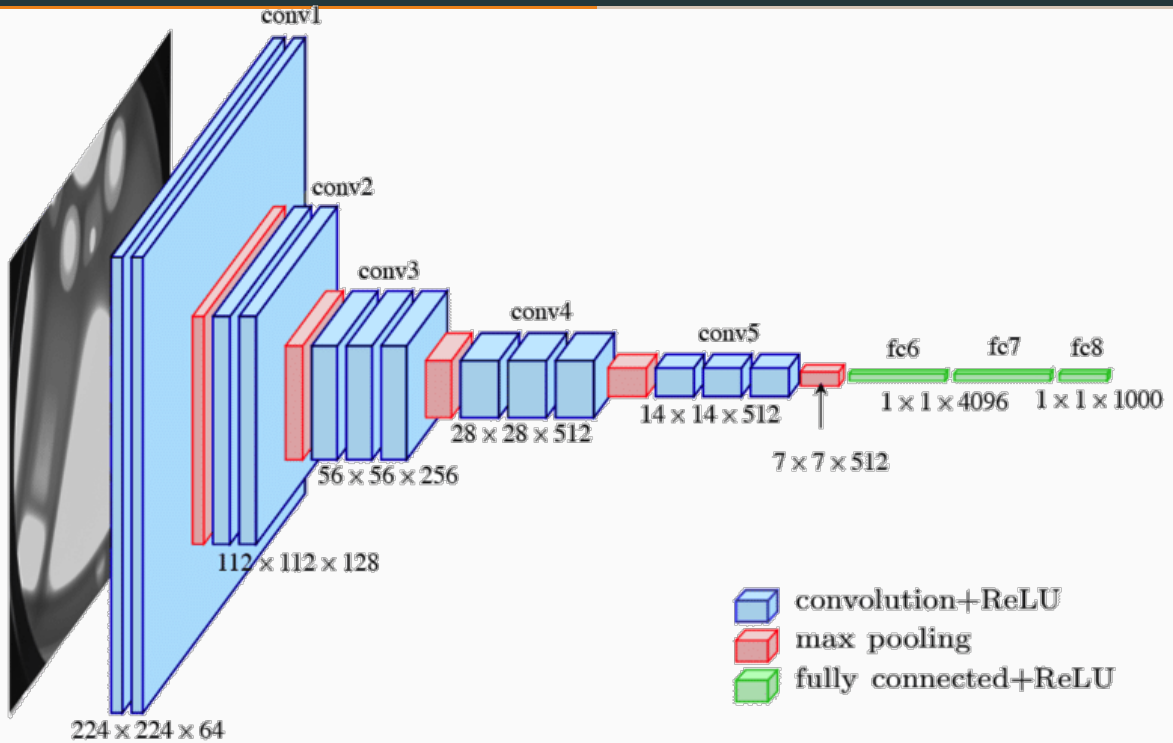
- Much deeper than previous CNN architectures
- Introduced ReLU activation to alleviate vanishing gradient problem
- Parallelized two GPUs to train the model, with two streams combining at specific layers
- Heavy Dropout and Data Augmentation improve robustness
- Overlapping max pooling reduces side effect and improves robustness
- Layer Response Normalization (LRN) mimicking biological neurons to improve robustness

Overall: VGG

Model Name	# of Parameters	Top-1 Accuracy	Year
AlexNet	60 M	63.3 %	2012
Inception V1	5 M	69.8 %	2014
VGG 16	**138 M**	**74.4 %**	**2014**
VGG 19	**144 M**	**74.5 %**	**2014**
Inception V2	11.2 M	74.8 %	2015
ResNet-50	26 M	77.15 %	2015
ResNet-152	60 M	78.57 %	2015
Inception V3	27 M	78.8 %	2015
DenseNet-121	8 M	74.98 %	2016
DenseNet-264	22 M	77.85 %	2016
EfficientNet-B0	5.3 M	76.3 %	2019
EfficientNet-B4	19 M	82.6 %	2019
EfficientNet-B7	66 M	84.4 %	2019
BiT-L (ResNet)	928 M	87.54 %	2019
Meta Pseudo Labels	480 M	90.2 %	2021

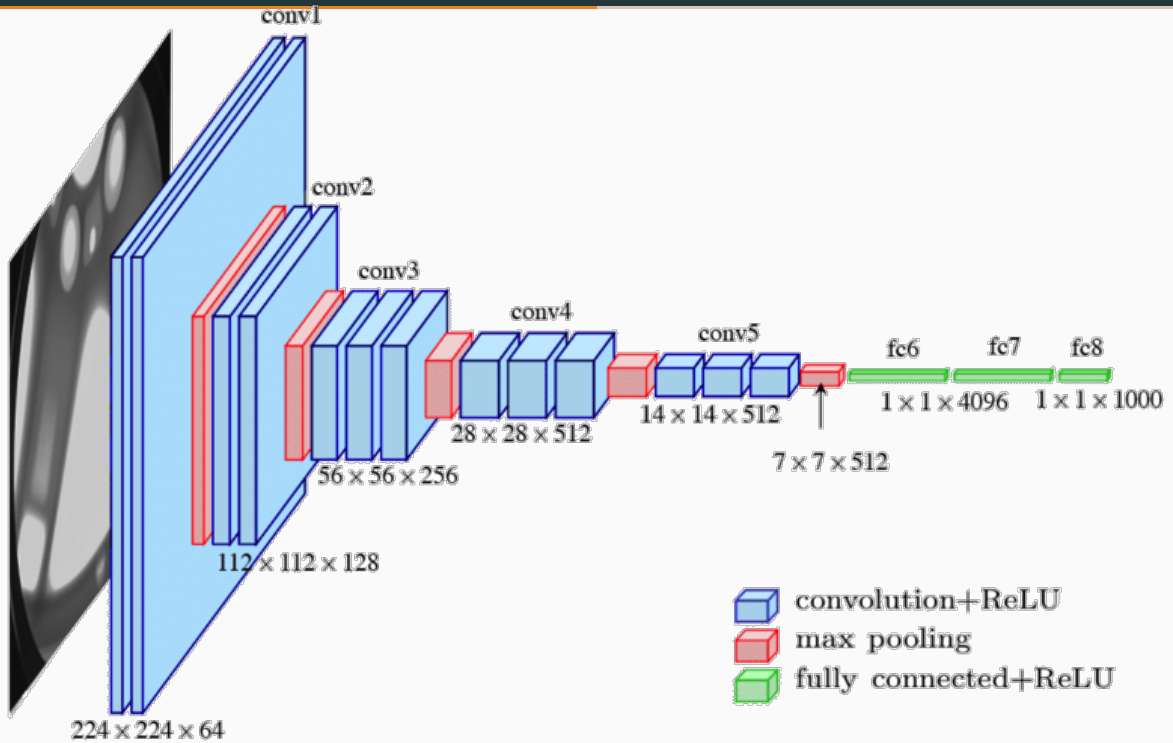
Table 2: Comparison of Different Models on ImageNet

VGG (from Visual Geometry Group) [8]



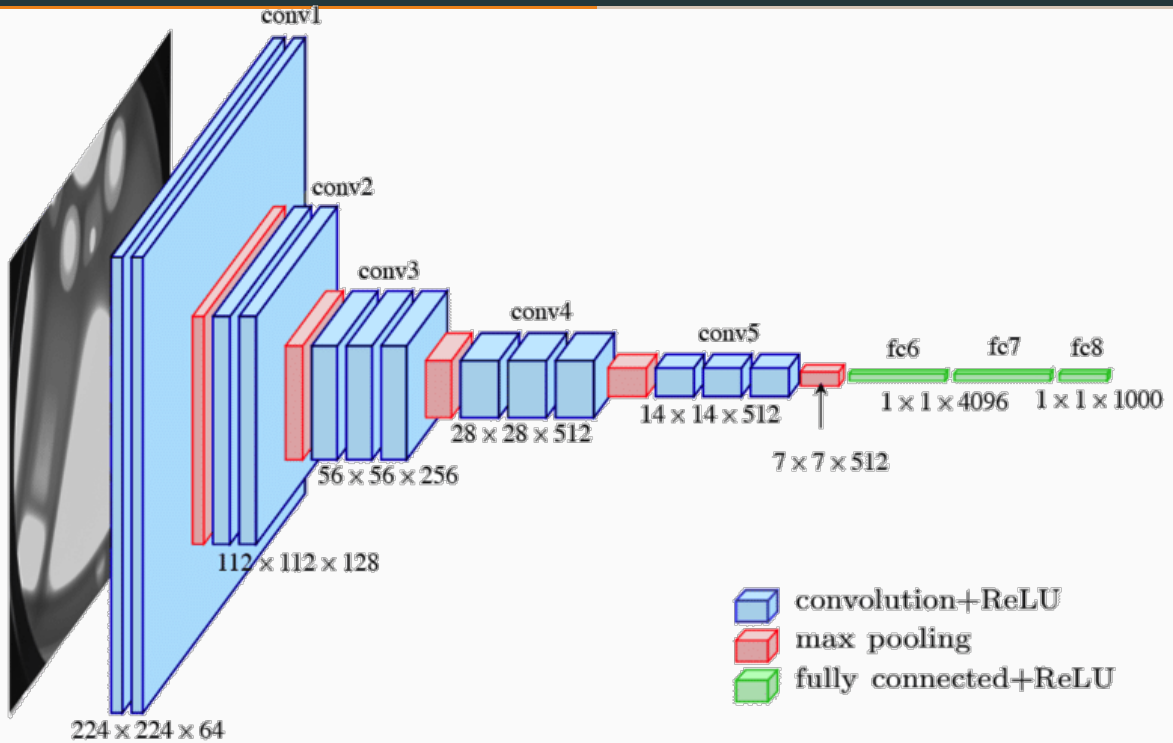
- Use **only** small (3x3) filters stacked together:

VGG (from Visual Geometry Group) [8]



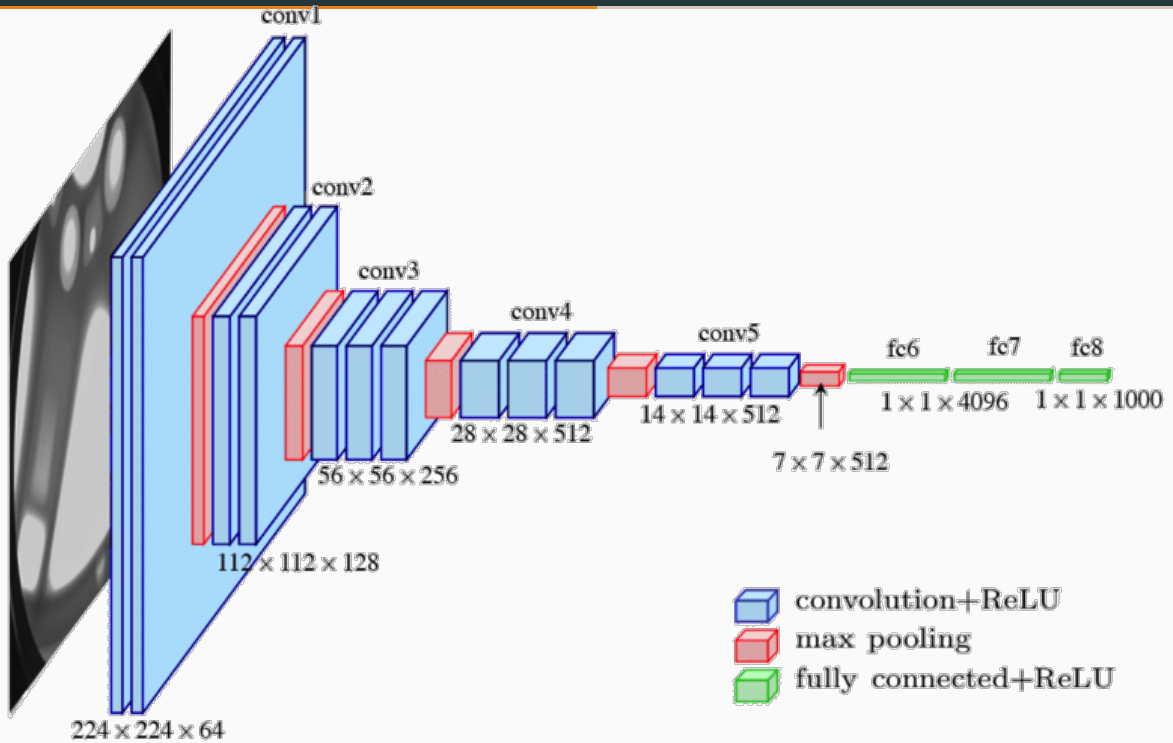
- Use **only** small (3x3) filters stacked together:
  - Same receptive field as bigger filters (like 11x11 from AlexNet) but fewer params

VGG (from Visual Geometry Group) [8]



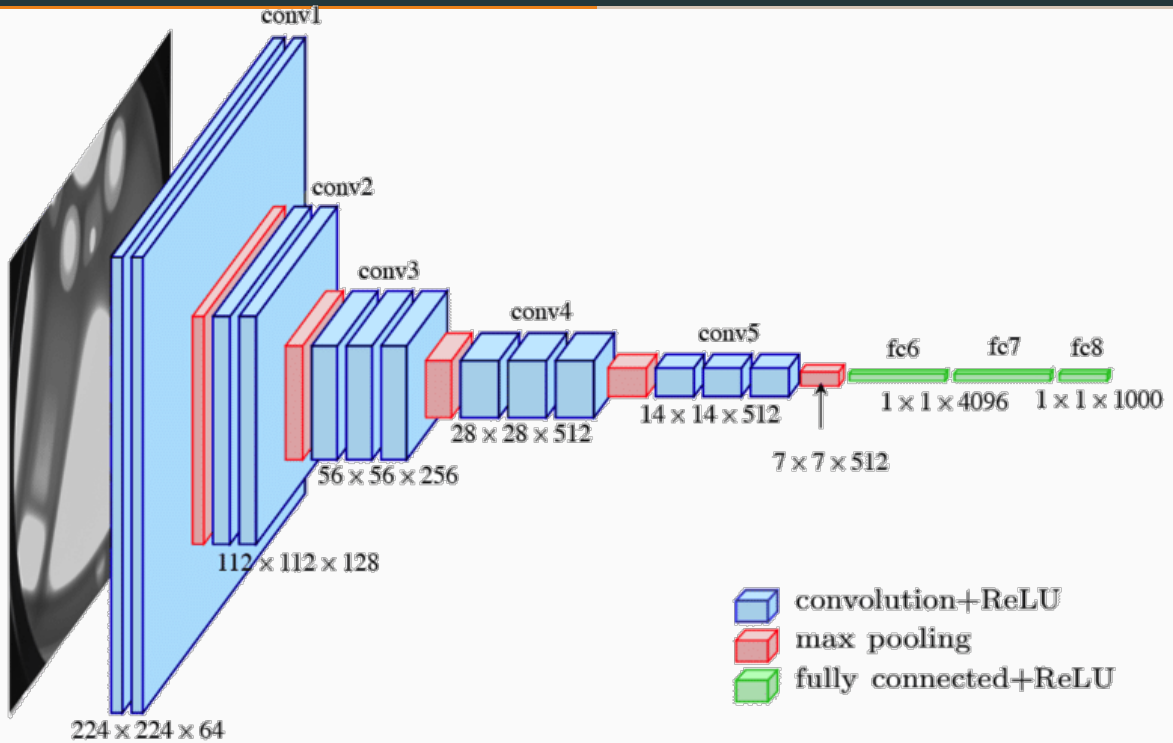
- Use **only** small (3x3) filters stacked together:
  - Same receptive field as bigger filters (like 11x11 from AlexNet) but fewer params
  - Deeper (16/19 vs 8 for AlexNet) net implies more **hierarchical** feature learning ⇒ more complex patterns

VGG (from Visual Geometry Group) [8]



- Use **only** small (3x3) filters stacked together:
  - Same receptive field as bigger filters (like 11x11 from AlexNet) but fewer params
  - Deeper (16/19 vs 8 for AlexNet) net implies more **hierarchical** feature learning ⇒ more complex patterns
  - Maxpool after few convolutional layers (let convLayers process)

VGG (from Visual Geometry Group) [8]



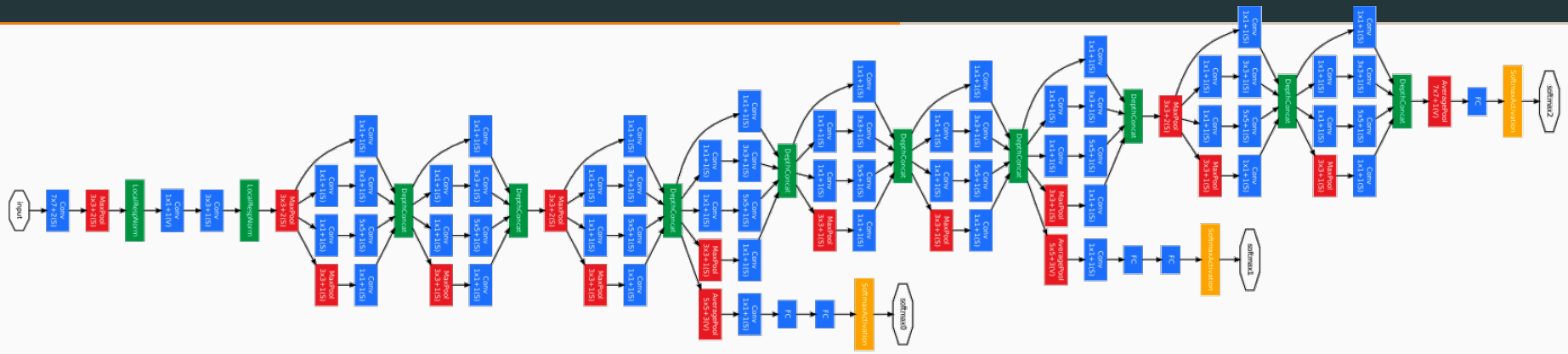
- Use **only** small (3x3) filters stacked together:
  - Same receptive field as bigger filters (like 11x11 from AlexNet) but fewer params
  - Deeper (16/19 vs 8 for AlexNet) net implies more **hierarchical** feature learning ⇒ more complex patterns
  - Maxpool after few convolutional layers (let convLayers process)
- More Feature Maps: starting with 64 channels, while deeper layers increased to 128/256/512 channels (richer feature learning)

Overall: GooLeNet

Model Name	# of Parameters	Top-1 Accuracy	Year
AlexNet	60 M	63.3 %	2012
Inception V1	**5 M**	**69.8 %**	**2014**
VGG 16	138 M	74.4 %	2014
VGG 19	144 M	74.5 %	2014
Inception V2	**11.2 M**	**74.8 %**	**2015**
ResNet-50	26 M	77.15 %	2015
ResNet-152	60 M	78.57 %	2015
Inception V3	27 M	78.8 %	2015
DenseNet-121	8 M	74.98 %	2016
DenseNet-264	22 M	77.85 %	2016
EfficientNet-B0	5.3 M	76.3 %	2019
EfficientNet-B4	19 M	82.6 %	2019
EfficientNet-B7	66 M	84.4 %	2019
BiT-L (ResNet)	928 M	87.54 %	2019
Meta Pseudo Labels	480 M	90.2 %	2021

Table 3: Comparison of Different Models on ImageNet

GooLeNet (from Google/LeNet) [9]



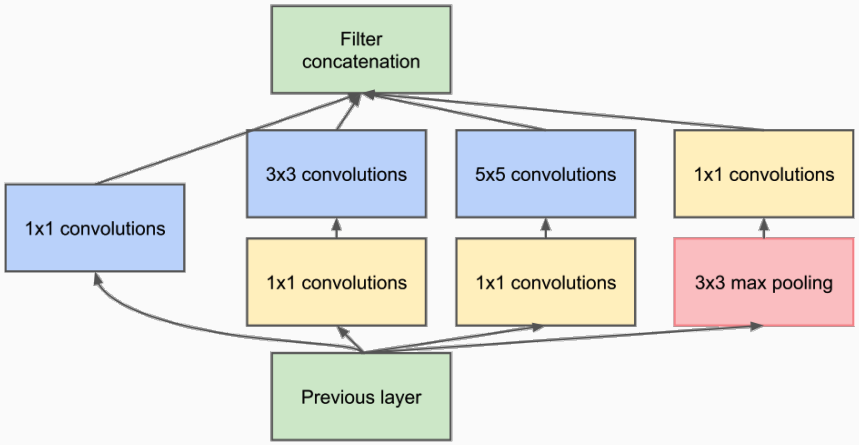


Figure 5: Inception module

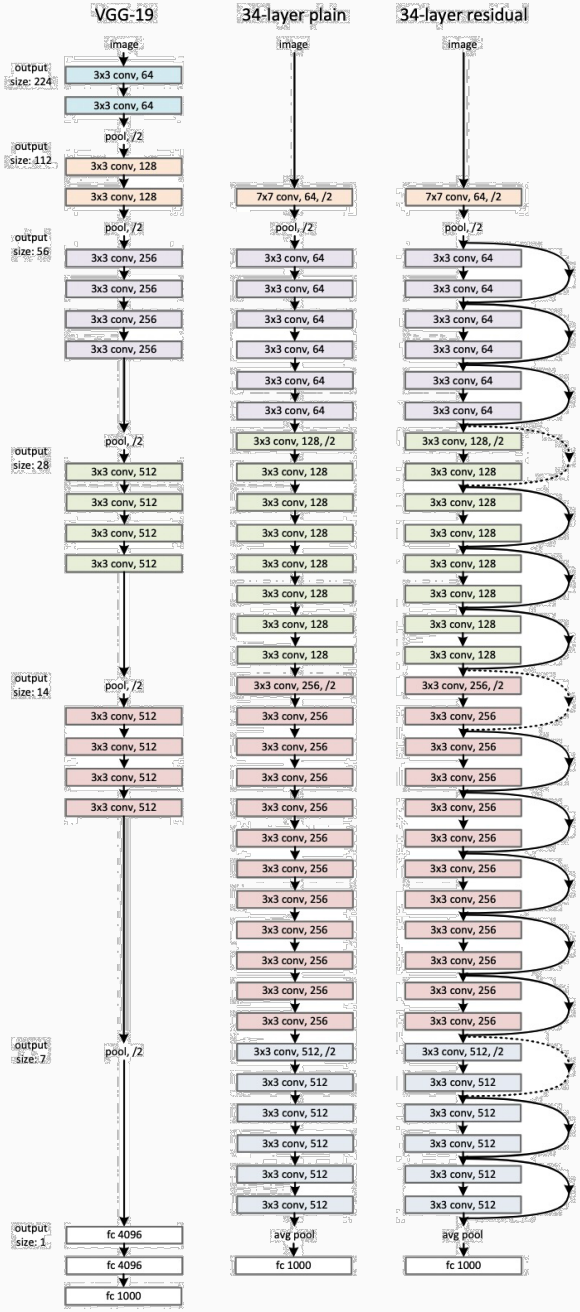
- 22 layers
- Inception Modules: multi-scale feature extraction through parallel convolutions
- 1x1 convolutions to reduce the number of channels (linear combinations of features from previous layers)
- Auxiliary classifiers to learn more robust features at different depths

Overall: ResNet

Model Name	# of Parameters	Top-1 Accuracy	Year
AlexNet	60 M	63.3 %	2012
Inception V1	5 M	69.8 %	2014
VGG 16	138 M	74.4 %	2014
VGG 19	144 M	74.5 %	2014
Inception V2	11.2 M	74.8 %	2015
ResNet-50	**26 M**	**77.15 %**	**2015**
ResNet-152	**60 M**	**78.57 %**	**2015**
Inception V3	27 M	78.8 %	2015
DenseNet-121	8 M	74.98 %	2016
DenseNet-264	22 M	77.85 %	2016
EfficientNet-B0	5.3 M	76.3 %	2019
EfficientNet-B4	19 M	82.6 %	2019
EfficientNet-B7	66 M	84.4 %	2019
BiT-L (ResNet)	928 M	87.54 %	2019
Meta Pseudo Labels	480 M	90.2 %	2021

Table 4: Comparison of Different Models on ImageNet

ResNet (from Residuals) [2]



- Residual Connections (or skip connections)
- Revolution Depth: way deeper architecture
- Modular architectures: each residual block consists
  - A few convolutional layers (usually two or three)
  - Batch normalization
  - ReLU activation functions
  - Skip connection
- Batch normalization helps stabilize training by normalizing layer inputs, which improves convergence speed and generalization
- Note that no dropout was used...

ResNet (from Residuals) [2]

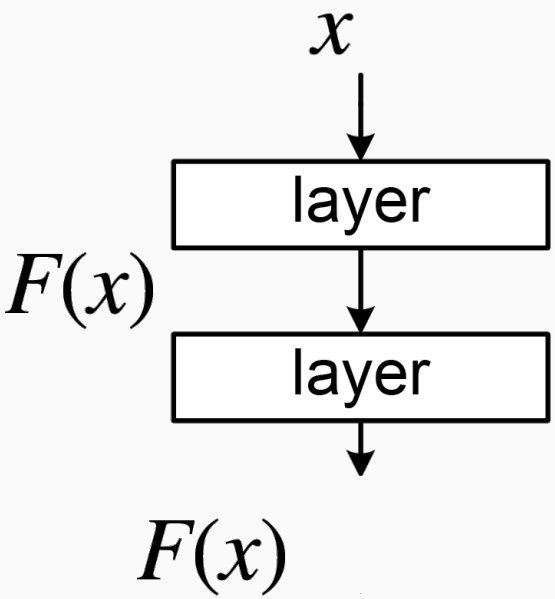


Figure 6: Residual connections are crucial in the ResNet architecture

ResNet (from Residuals) [2]

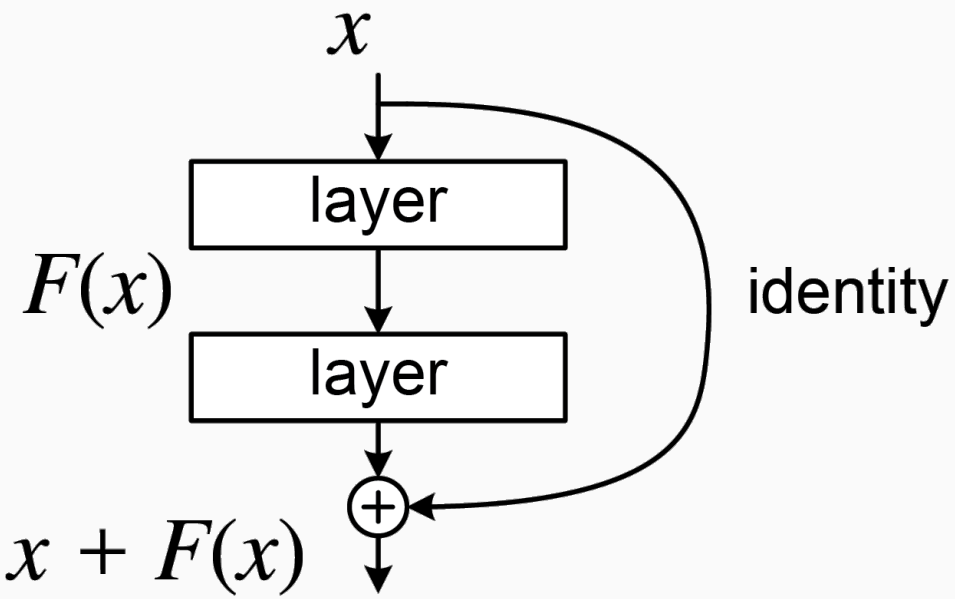


Figure 6: Residual connections are crucial in the ResNet architecture

ResNet (from Residuals) [2]

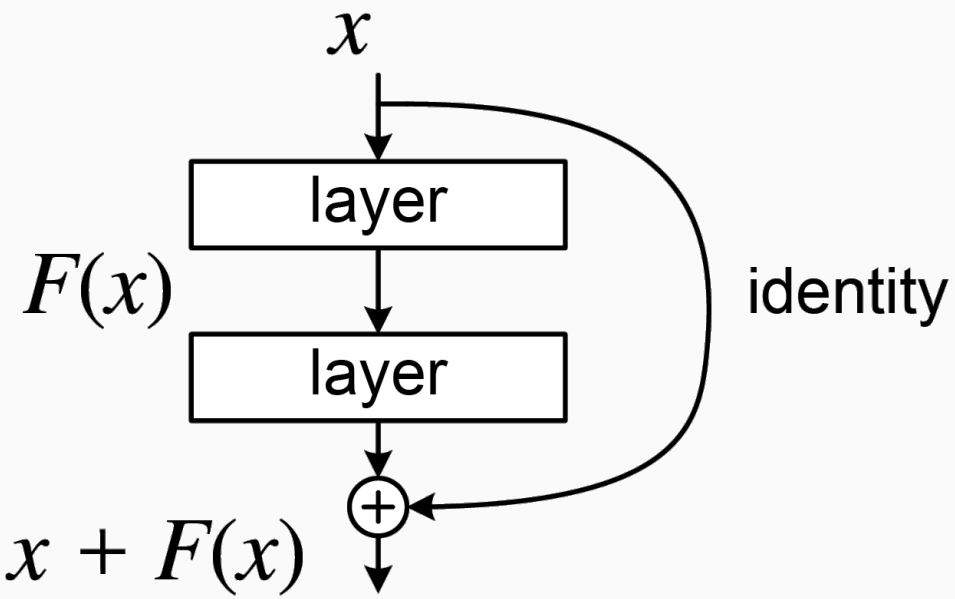


Figure 6: Residual connections are crucial in the ResNet architecture

- The network learns the **residual mapping** between the input and output where where $F(x)$ is the residual function to be learned
- Network focus on learning the **differences** or changes rather than the full mapping
- Help mitigate the **vanishing gradient problem** of very deep networks where gradients become too small to update weights effectively

ResNet (from Residuals) [2]

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Figure 7: Several size of architectures where trained on ImageNet. Building blocks are shown in brackets, with the numbers of blocks stacked.

Overall: DenseNet

Model Name	# of Parameters	Top-1 Accuracy	Year
AlexNet	60 M	63.3 %	2012
Inception V1	5 M	69.8 %	2014
VGG 16	138 M	74.4 %	2014
VGG 19	144 M	74.5 %	2014
Inception V2	11.2 M	74.8 %	2015
ResNet-50	26 M	77.15 %	2015
ResNet-152	60 M	78.57 %	2015
Inception V3	27 M	78.8 %	2015
DenseNet-121	**8 M**	**74.98 %**	**2016**
DenseNet-264	**22 M**	**77.85 %**	**2016**
EfficientNet-B0	5.3 M	76.3 %	2019
EfficientNet-B4	19 M	82.6 %	2019
EfficientNet-B7	66 M	84.4 %	2019
BiT-L (ResNet)	928 M	87.54 %	2019
Meta Pseudo Labels	480 M	90.2 %	2021

Table 5: Comparison of Different Models on ImageNet

DenseNet (from Dense connections) [3]

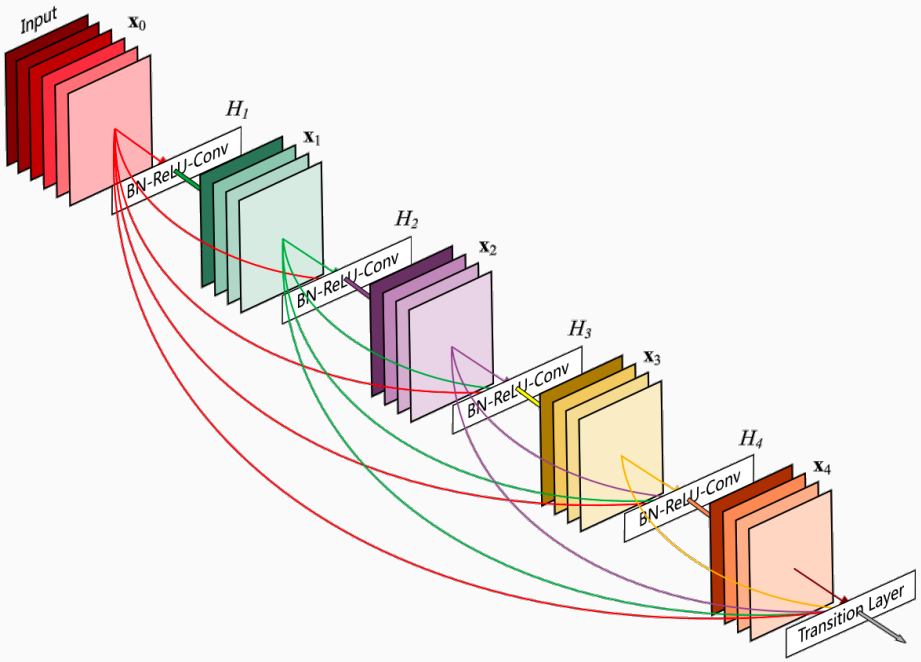


Figure 8: Each layer takes all preceding feature-maps as input

- **Dense Connectivity**: Each layer is connected to every other layer, mitigates the vanishing gradient problem (like skip-connections)
- **Parameter Efficiency**: reuse of features across layers $\Rightarrow$ no redundant feature maps $\Rightarrow$ more computationally efficient.
- **Transition Layers**: control the growth of feature maps and downsample spatial dimensions

Overall: EfficientNet

Model Name	# of Parameters	Top-1 Accuracy	Year
AlexNet	60 M	63.3 %	2012
Inception V1	5 M	69.8 %	2014
VGG 16	138 M	74.4 %	2014
VGG 19	144 M	74.5 %	2014
Inception V2	11.2 M	74.8 %	2015
ResNet-50	26 M	77.15 %	2015
ResNet-152	60 M	78.57 %	2015
Inception V3	27 M	78.8 %	2015
DenseNet-121	8 M	74.98 %	2016
DenseNet-264	22 M	77.85 %	2016
EfficientNet-B0	**5.3 M**	**76.3 %**	**2019**
EfficientNet-B4	**19 M**	**82.6 %**	**2019**
EfficientNet-B7	**66 M**	**84.4 %**	**2019**
BiT-L (ResNet)	928 M	87.54 %	2019
Meta Pseudo Labels	480 M	90.2 %	2021

Table 6: Comparison of Different Models on ImageNet

EfficientNet (from Efficient Architecture) [6]

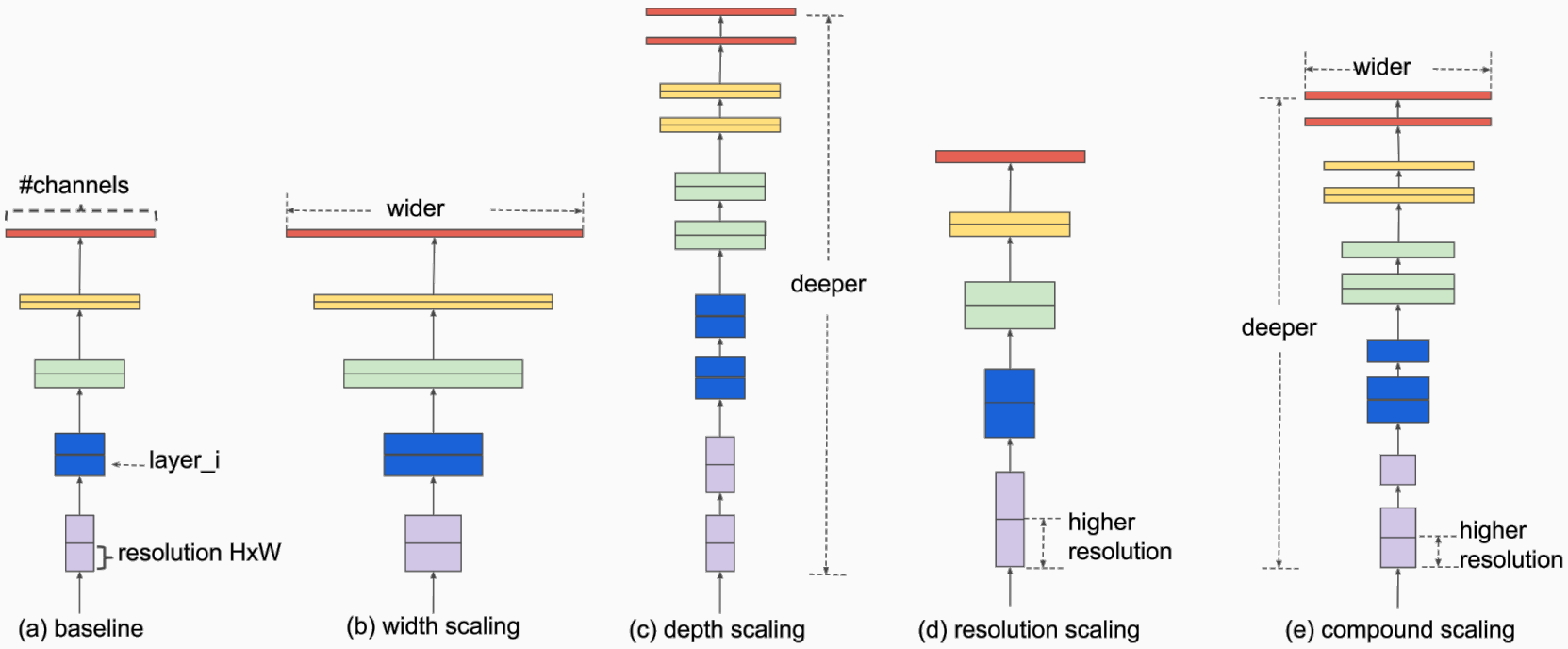


Figure 9: Compound scaling method uniformly scales all three dimensions with a fixed ratio compared to conventional scaling that increase only one dimension

- Systematic way to scale up network **depth, width, and resolution simultaneously** in a balanced manner
- Use of AutoML (Neural Architecture Search) to find the optimal architecture, more effective design compared to manual tuning

Overall

Model Name	# of Parameters	Top-1 Accuracy	Year
AlexNet	60 M	63.3 %	2012
Inception V1	5 M	69.8 %	2014
VGG 16	138 M	74.4 %	2014
VGG 19	144 M	74.5 %	2014
Inception V2	11.2 M	74.8 %	2015
ResNet-50	26 M	77.15 %	2015
ResNet-152	60 M	78.57 %	2015
Inception V3	27 M	78.8 %	2015
DenseNet-121	8 M	74.98 %	2016
DenseNet-264	22 M	77.85 %	2016
EfficientNet-B0	5.3 M	76.3 %	2019
EfficientNet-B4	19 M	82.6 %	2019
EfficientNet-B7	66 M	84.4 %	2019
BiT-L (ResNet)	928 M	87.54 %	2019
Meta Pseudo Labels	480 M	90.2 %	2021

Table 7: Comparison of Different Models on ImageNet

Questions?

References i

J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei.
ImageNet: A Large-Scale Hierarchical Image Database.
In 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pages 248–255. IEEE, 2009.

K. He, X. Zhang, S. Ren, and J. Sun.
Deep residual learning for image recognition.
In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 2016-Decem, pages 770–778, 2016.

G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger.
Densely Connected Convolutional Networks.
In CVPR, 2017.

References ii

A. Krizhevsky, I. Sutskever, and G. E. Hinton.
ImageNet Classification with Deep Convolutional Neural Networks.
In Nips, 2012.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner.
Gradient-based learning applied to document recognition.
Proceedings of the IEEE, 86(11):2278–2324, 1998.

Q. V. L. Mingxing Tan.
EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks Mingxing.
In Proceedings of the 36th International Conference on Machine Learning (ICML), 2019.

References iii

T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor.
ImageNet-21K Pretraining for the Masses.
In NeurIPS, pages 1–12, 2021.

K. Simonyan and A. Zisserman.
Very deep convolutional networks for large-scale image recognition.
In 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, pages 1–14, 2015.

C. Szegedy, S. Reed, P. Sermanet, V. Vanhoucke, and A. Rabinovich.
Going deeper with convolutions.
In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015.