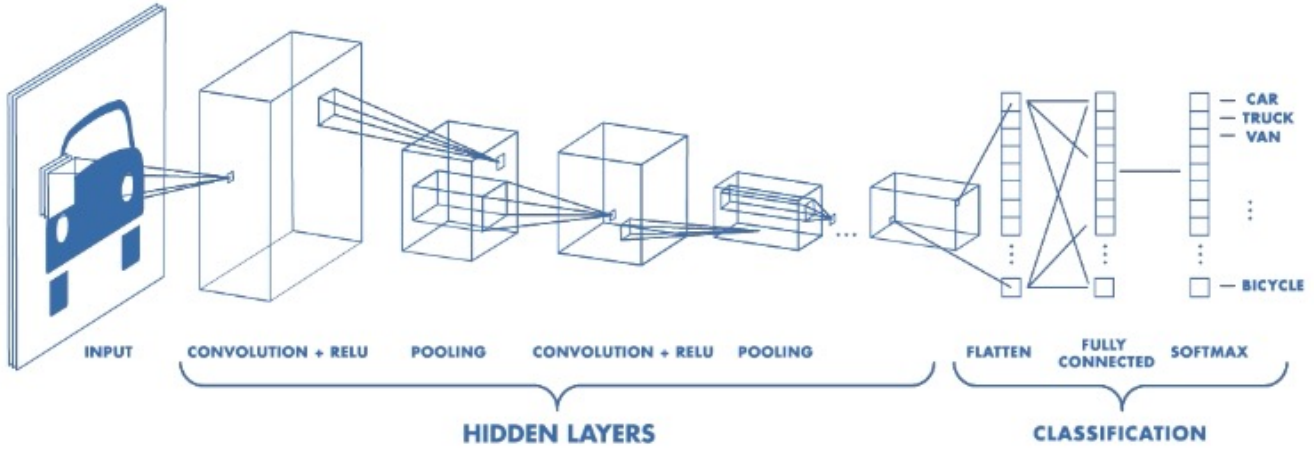
COMP9517

Computer Vision

2024 Term 3 Week 7

Dr Sonit Singh

UNSW SYDNEY



Deep Learning

Image Classification using CNNs

Introduction

Image Classification

- Image classification with linear classifier

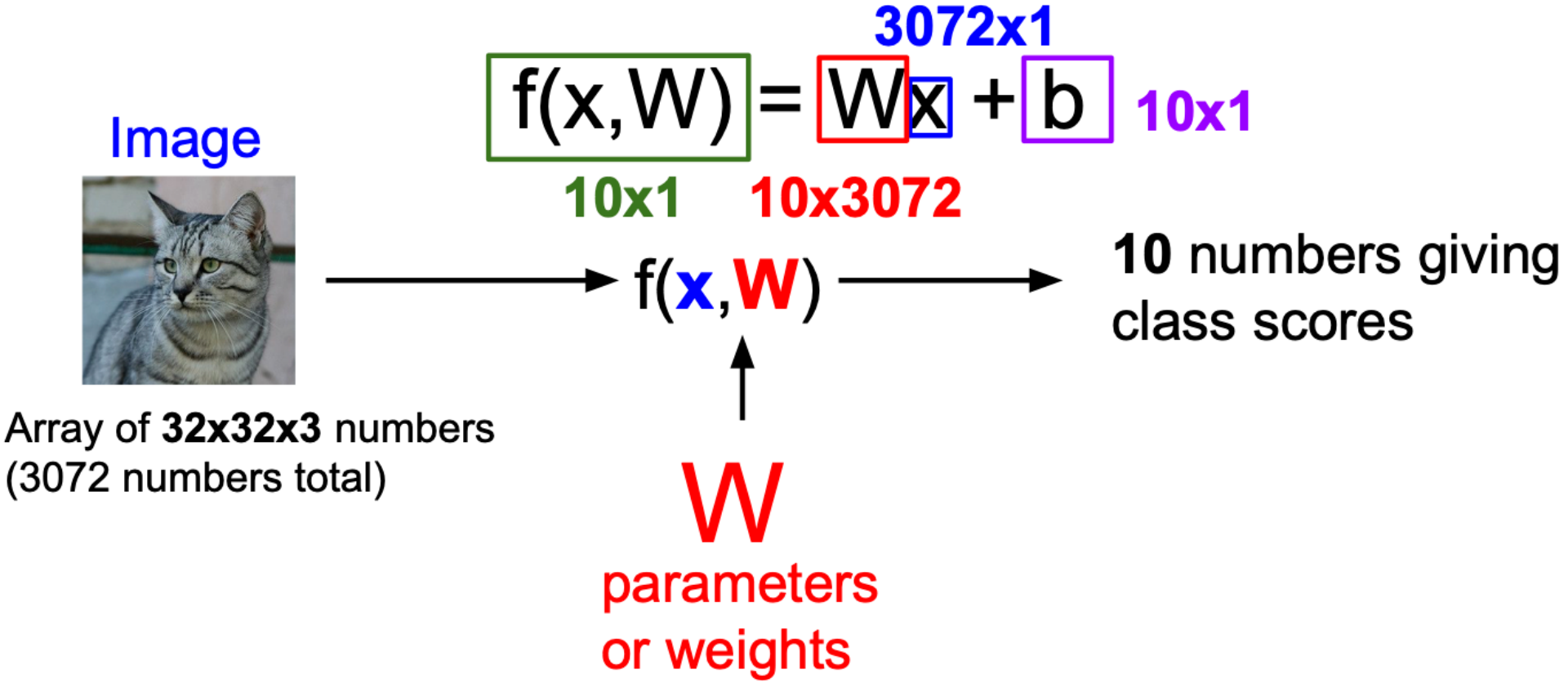
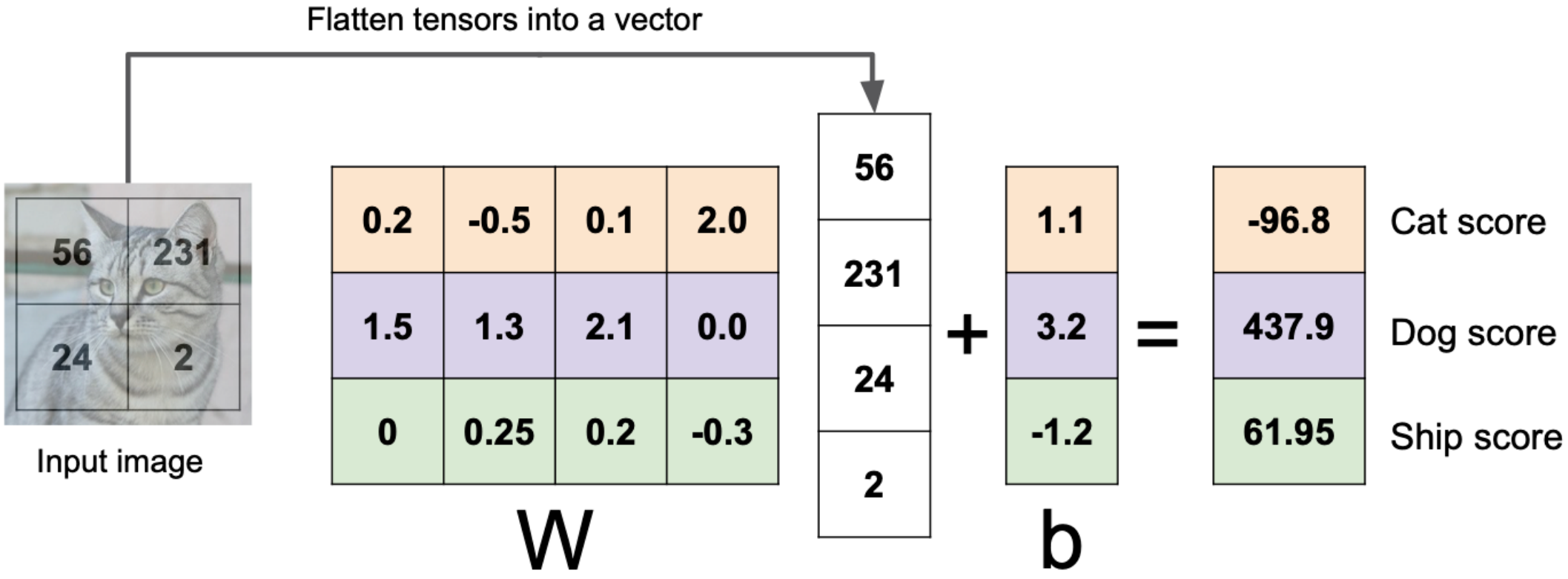


Image Classification

- An image example with 4 pixels and 3 classes.



Role of CNNs in Image Classification

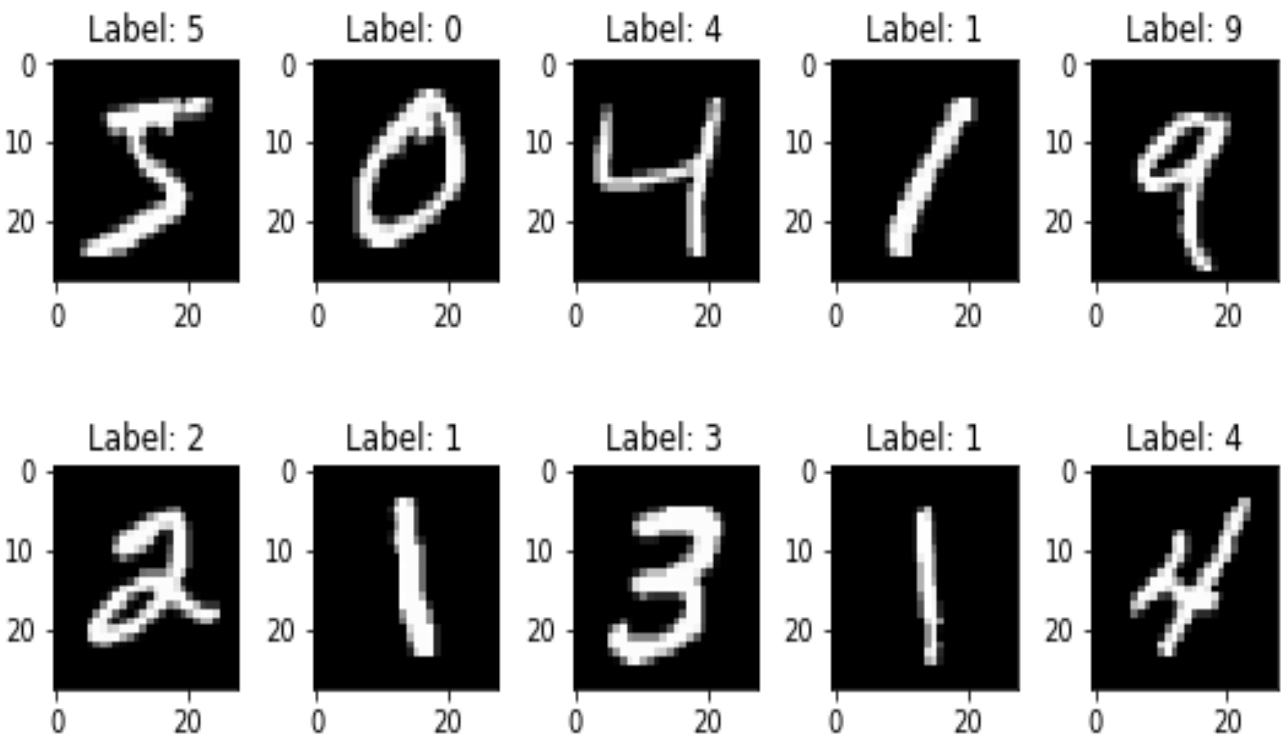
- CNNs are designed specifically for processing grid-like data, making them well-suited for images.
- They can automatically learn relevant features from raw pixel data, eliminating the need for handcrafted feature engineering.
- CNNs use convolutional layers to detect local patterns, followed by pooling layers to reduce spatial dimensions and prevent overfitting.
- The fully connected layers at the end of a CNN make the final classification decision based on the learned features.

Advantages of CNNs in Image Classification

- Automatic Feature Extraction
- Hierarchical Feature Learning
- Weight Sharing and Parameter Efficiency
- Transfer Learning and Pretrained Model
- Translation Invariance
- Superior Performance
- Robustness to Variations
- Scalability

Datasets

MNIST


Label: 5
Label: 0
Label: 4
Label: 1
Label: 9
Label: 2
Label: 1
Label: 3
Label: 1
Label: 4

MNIST

Overview:

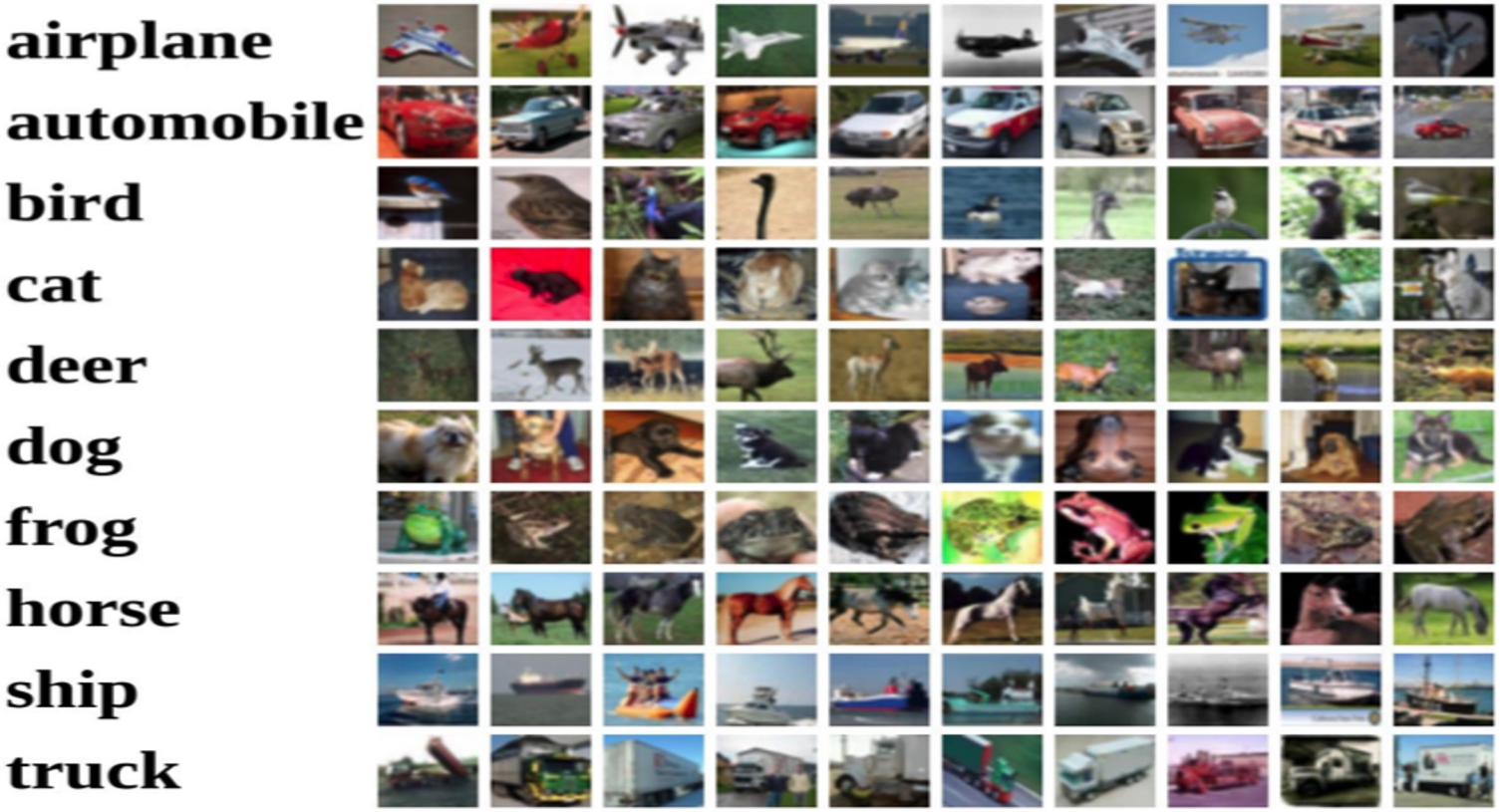
- MNIST, short for "Modified National Institute of Standards and Technology," is a widely used dataset in the field of machine learning and computer vision.
- It consists of a collection of handwritten digits that are extensively used for tasks such as digit recognition and image classification.

Key Characteristics:

- 70,000 grayscale images.
- 28x28 pixels in size.
- Each image contains a single digit (0 through 9).
- Labeled dataset: Each image is associated with a corresponding digit label.

CIFAR-10

CIFAR10 [Alex Krizhevsky, “Learning Multiple Layers of Features from Tiny Images”, Technical Report, 2009.]



10 classes
50,000 training images
10,000 testing images

CIFAR-10

Overview:

- CIFAR-10 stands for the “Canadian Institute For Advanced Research - 10,” and it is a widely used dataset for image classification tasks in the field of machine learning and computer vision.

Key Characteristics:

- CIFAR-10 consists of 60,000 color images.
- These images are divided into 10 different classes, each representing a distinct object or category.
- The dataset is split into 50,000 training images and 10,000 testing images.
- Each image is 32x32 pixels in size.
- The 10 classes include objects like airplanes, automobiles, birds, cats, and more.
- CIFAR-100 version

ImageNet

Consists of 14 million images, more than 21,000 classes, and about 1 million images have bounding box annotations

- Annotated by humans using crowdsourcing platform “Amazon Mechanical Turk”

- ImageNet Large-Scale Visual Recognition Challenge (ILSVRC)
- annual competition to foster the development and benchmarking of state-of-the-art algorithms in Computer Vision
- Led to improvement in architectures and techniques at the intersection of CV and DL

Image Credit: Synced. https://syncedreview.com/2020/06/23/google-deepmind-researchers-revamp-imagenet/

COCO

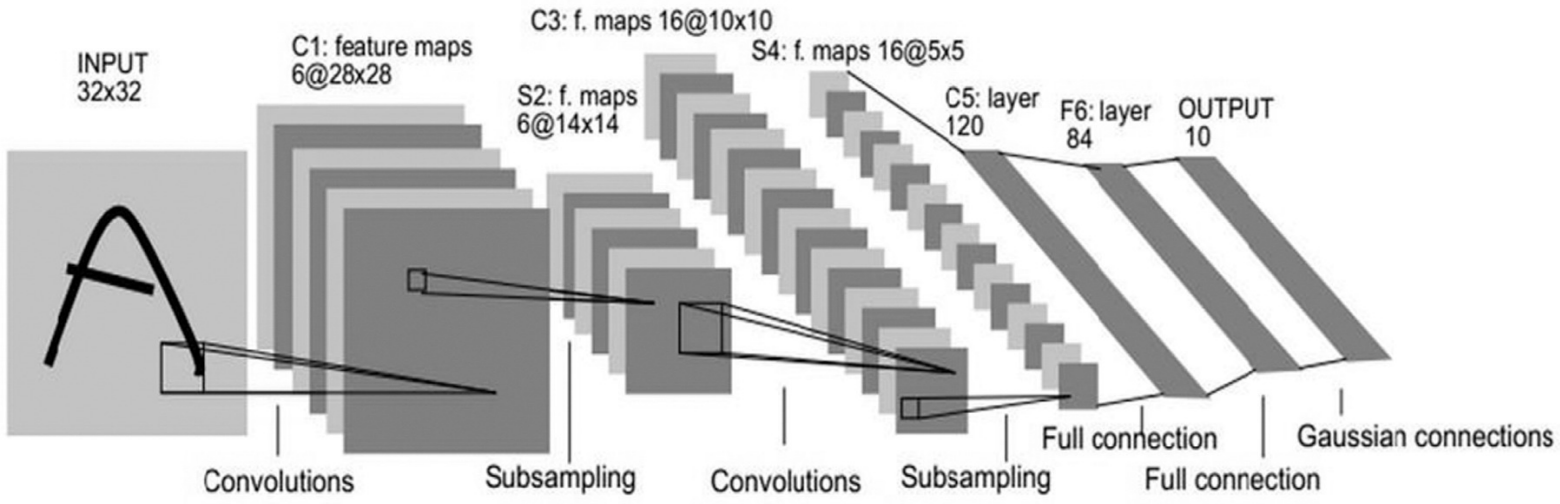
Common Objects in Context (COCO) dataset is a large-scale object detection, segmentation, and captioning dataset.

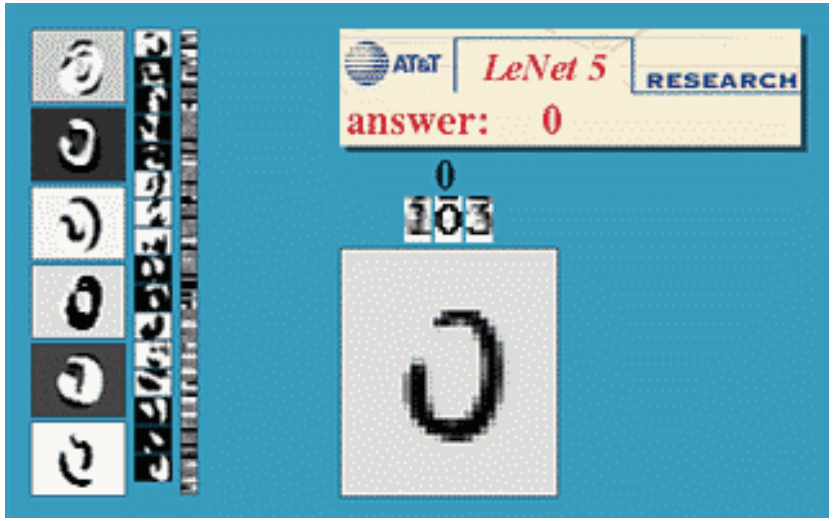
- https://cocodataset.org/#home

CNN Architectures

LeNet-5

- First developed by Yann Lecun in 1989 for digit recognition
  - First time backprop is used to automatically learn visual features
  - Two convolutional layers, three fully connected layers (32 x 32 input, 6 and 12 FMs, 5 x 5 filters)
  - Stride = 2 is used to reduce image dimensions
  - Scaled tanh activation function
  - Uniform random weight initialization

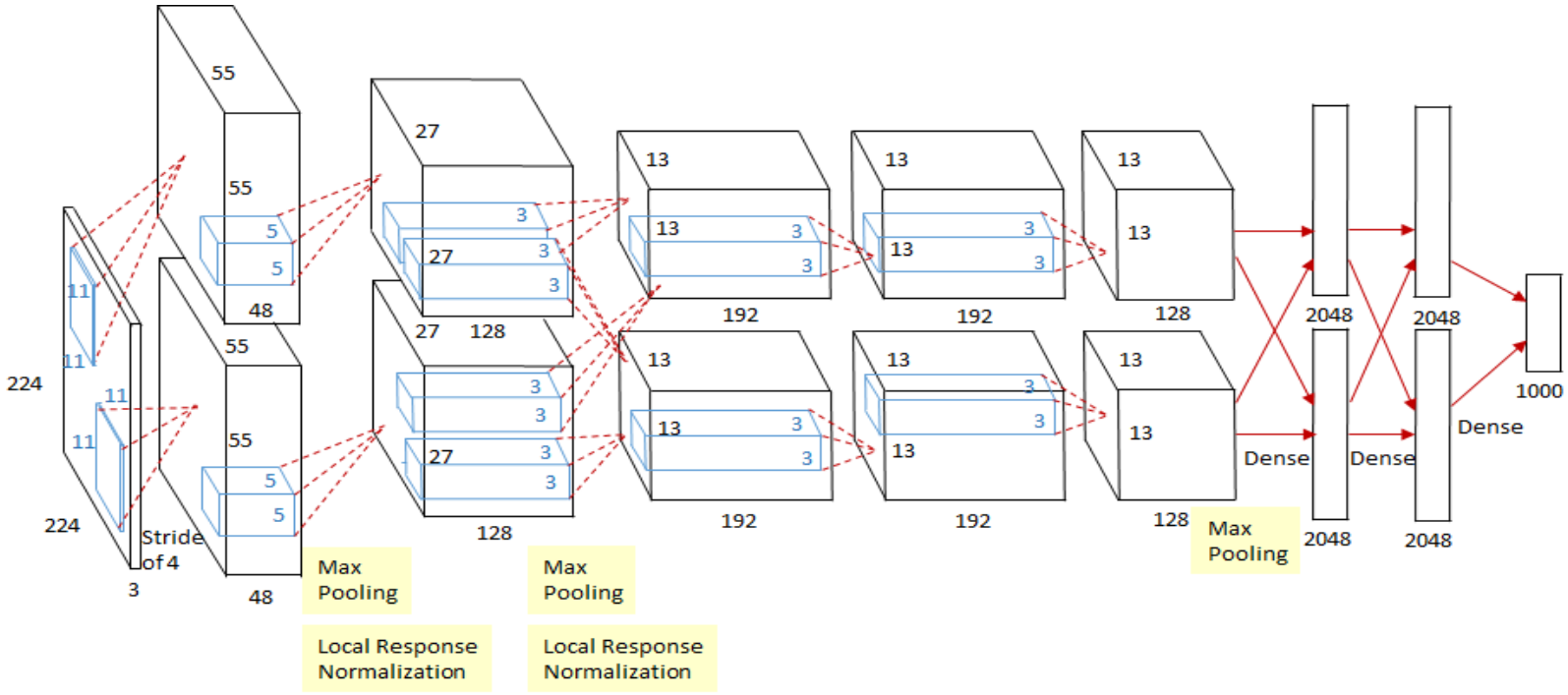




Source: Lecun et al. (1989). Gradient-based learning applied to document recognition.

AlexNet

- Added super important heuristics
  - ReLU non-linearity
  - Local Response Normalization
  - Data Augmentation
  - Dropout
- **Winner of 2012 ILSVRC**



IMAGENET

ImageNet Large Scale Visual Recognition Challenge (ILSVRC)
https://www.image-net.org/challenges/LSVRC/

Image Credit: Kdnuggets – Deep Learning's Most Important Ideas. https://www.kdnuggets.com/2020/09/deep-learnings-most-important-ideas.html

VGG

- Developed at Visual Geometry Group (Oxford) by Simonyan and Zisserman
- 1st runner up (Classification) and Winner (localization) of ILSVRC 2014 competition
- VGG-19 comprises of 144 million parameters

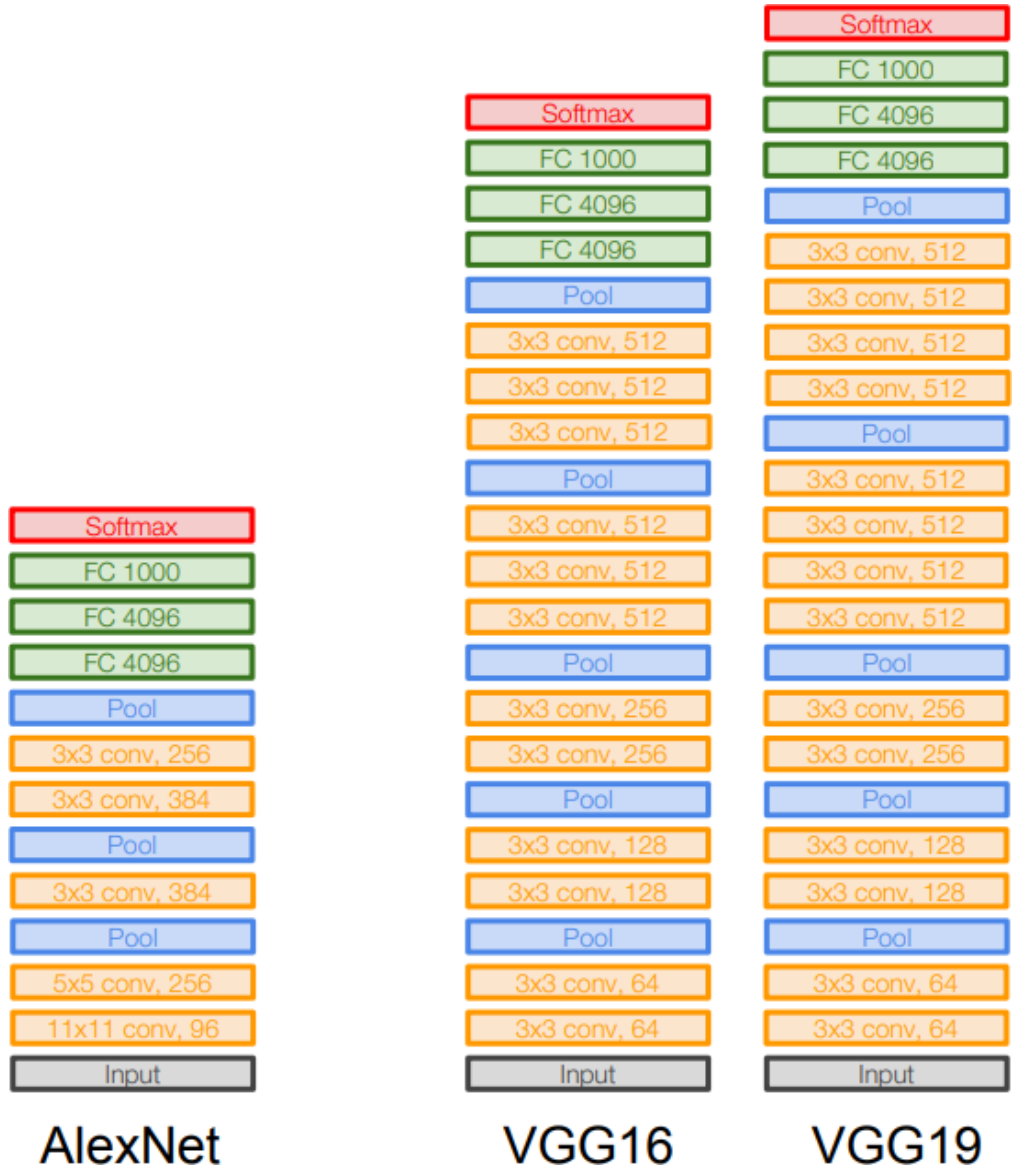
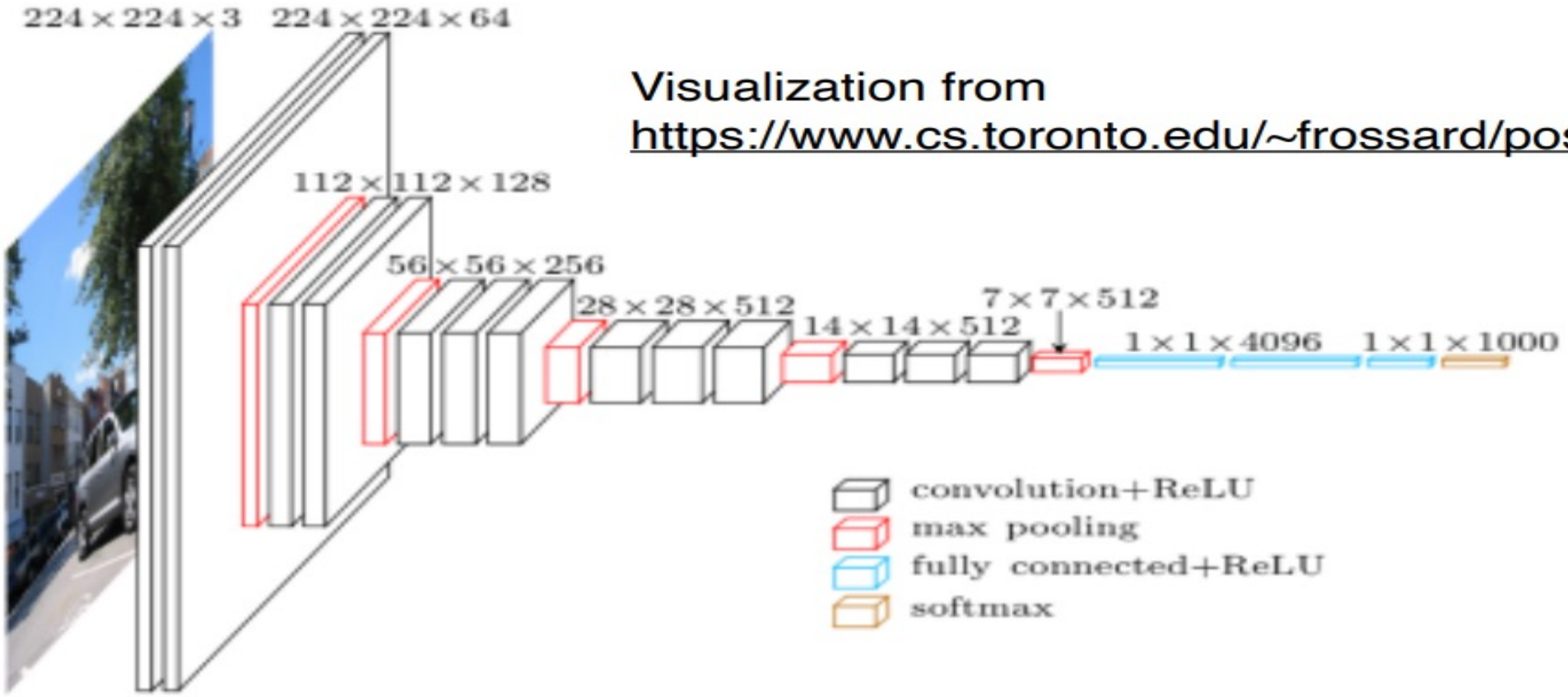


Image Credit: Medium. https://medium.com/coinmonks/paper-review-of-vggnet-1st-runner-up-of-ilsvlc-2014-image-classification-d02355543a11

VGG-16

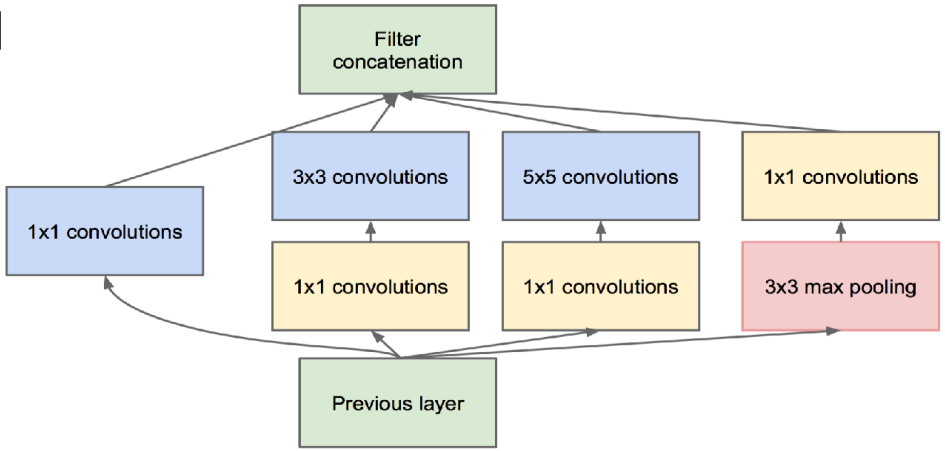


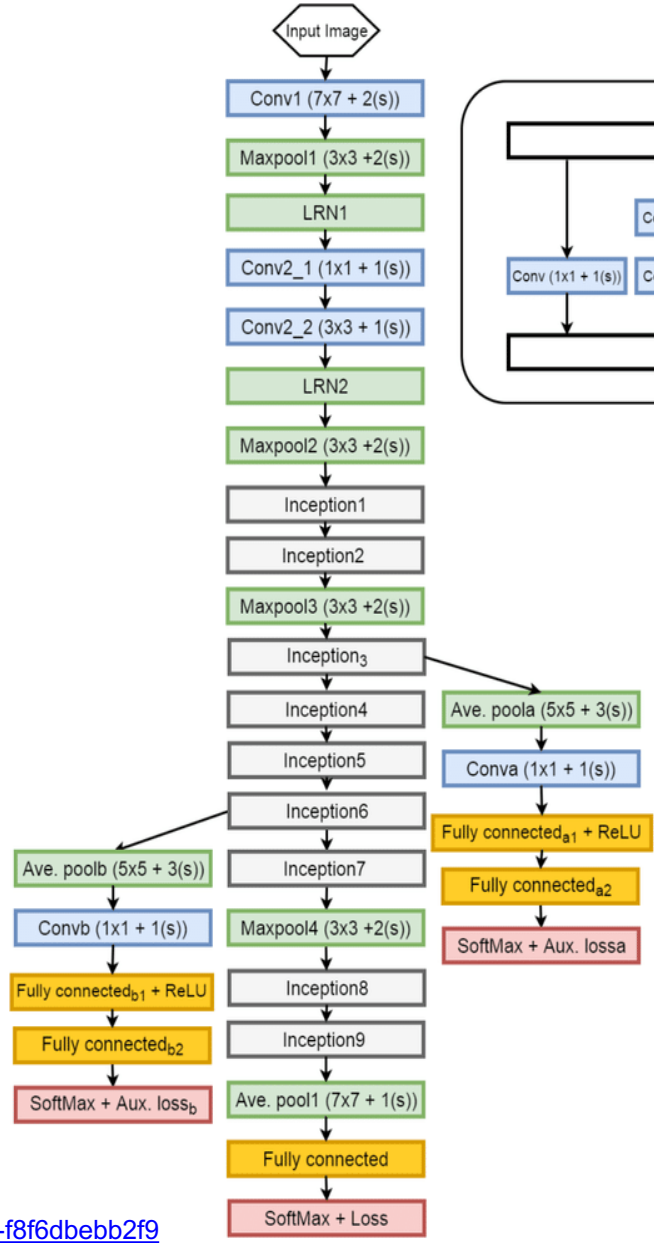
Visualization from
https://www.cs.toronto.edu/~frossard/post/vgg16/

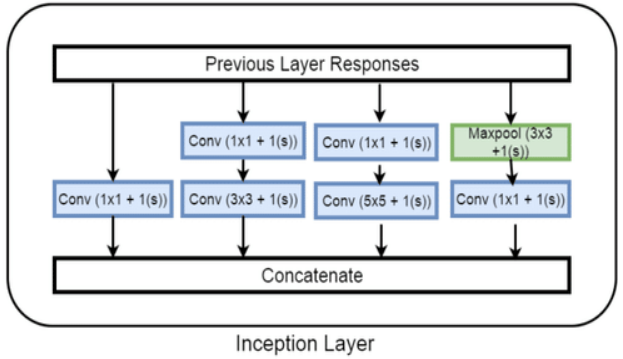
Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

GoogLeNet

- A 22-layer CNN developed by researchers at Google
- Deeper networks prone to overfitting and suffer from exploding or vanishing gradient problem
- Core idea "**Inception module**"
  - **Multi-branch, Multi-size kernel**
- Adding **Auxiliary loss as an extra supervision**
- Winner of 201





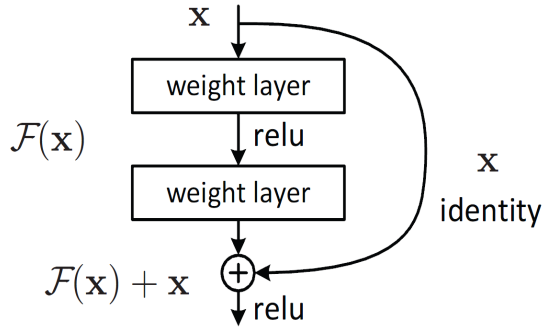


Inception Layer

Source: Convolutional Neural Networks. https://medium.com/@rajat.k.91/convolutional-neural-networks-why-what-and-how-f8f6dbebb2f9

ResNet

- Developed by researchers at Microsoft (Kaiming et al.)
- Core idea **"residual connections" or "skip connections"** to preserve the gradient
- The identity matrix transmits forward the input data that avoids the loose of information (the data vanishing problem)



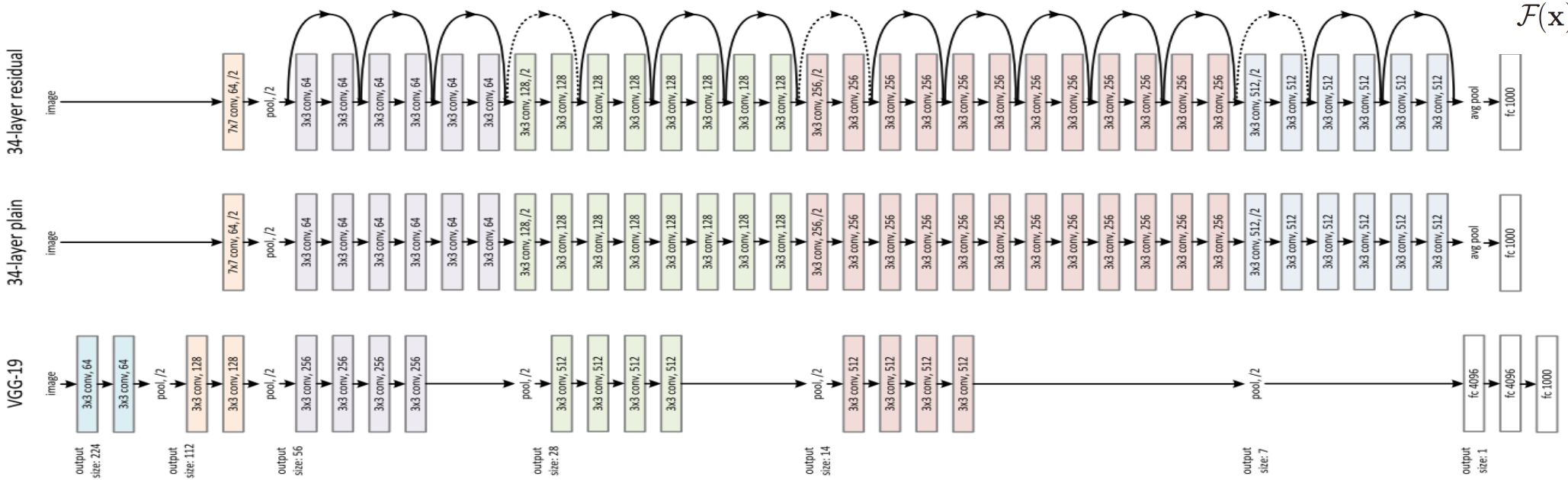
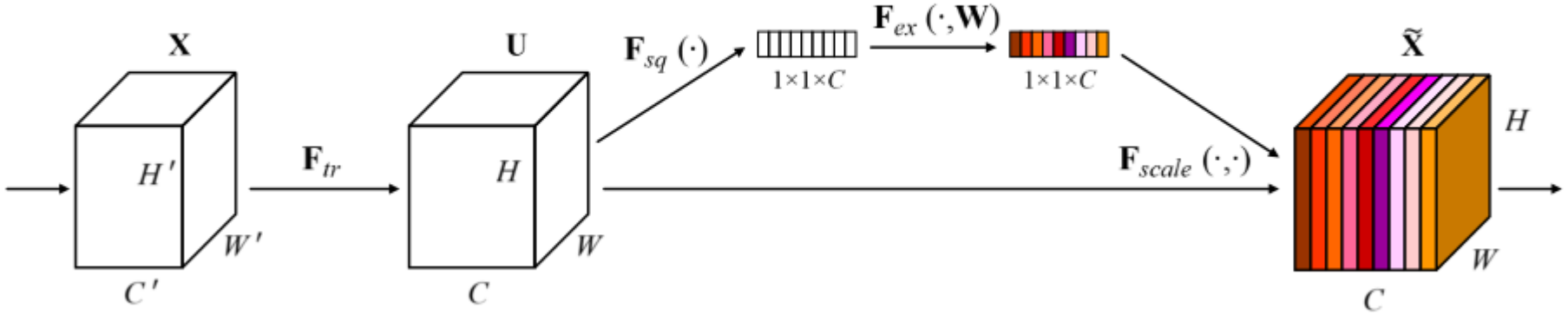


Image Credit: Medium. https://medium.com/@pierre_guillou/understand-how-works-resnet-without-talking-about-residual-64698f157e0c

SENet (Squeeze-and-Excitation Network)

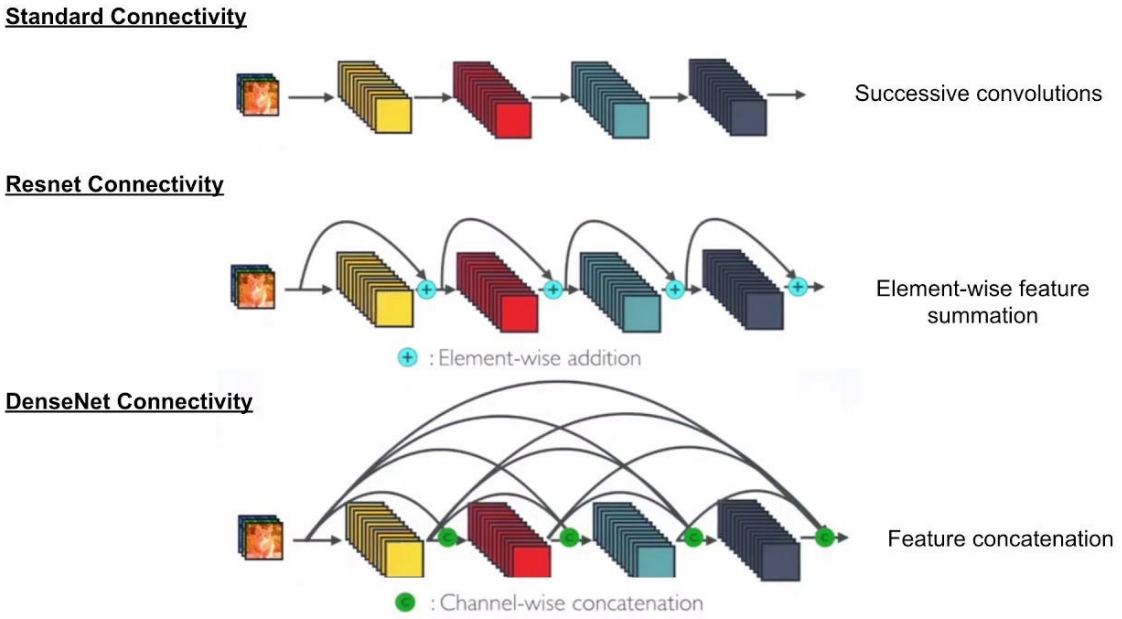
- CNNs fuse the spatial and channel information to extract features to solve the task
- Before this, networks weights each of its channels equally when creating the output feature maps
- SENets added a content aware mechanism to weight each channel adaptively
- SE block helps to improve representation power of the network, able to better map the channel dependency along with access to global information



Source: Convolutional Neural Networks. https://medium.com/@rajat.k.91/convolutional-neural-networks-why-what-and-how-f8f6dbebb2f9

DenseNet

- "Dense block" vs "Res block"
- More flexible connections
- Transition layer is used between dense blocks
  - Reduce dimensionality and computation



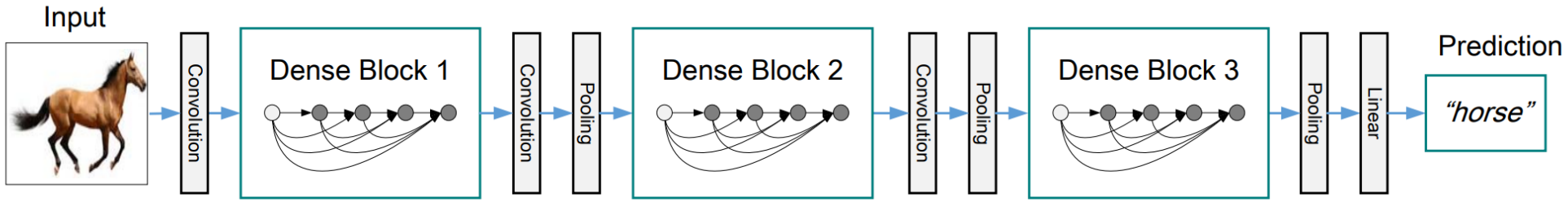


Figure 2: A deep DenseNet with three dense blocks. The layers between two adjacent blocks are referred to as transition layers and change feature-map sizes via convolution and pooling.

Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 2017.

EfficientNet

- Efficient utilizes a **compound scaling method to uniformly scale depth, width, and resolution, providing high accuracy with computational efficiency**.

- Instead of randomly scaling up width, depth or resolution, compound scaling uniformly scales each dimension with a certain fixed set of scaling coefficients.

- If the input image is bigger, then the network needs more layers to increase the receptive field and more channels to capture more fine-grained patterns on the bigger image.

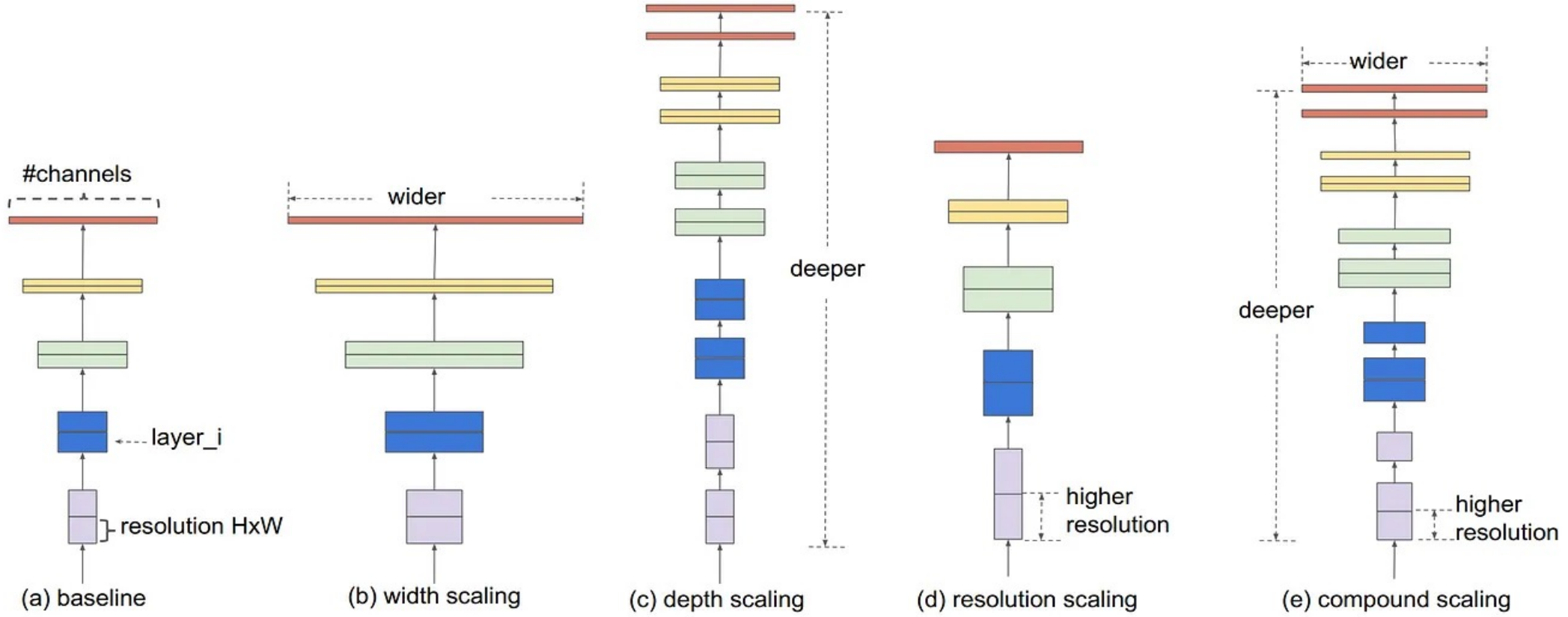
$$\text{Depth } d = \alpha^{\emptyset}, \text{Width } w = \beta^{\emptyset}, \text{Resolution } r = \gamma^{\emptyset}, (1)$$

$$\text{such that } \alpha . \beta^2 . \gamma^2 \approx 2$$

$$\alpha \geq 1, \beta \geq 1, \gamma \geq 1$$

Tan, et al. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks " *ICML 2019*.

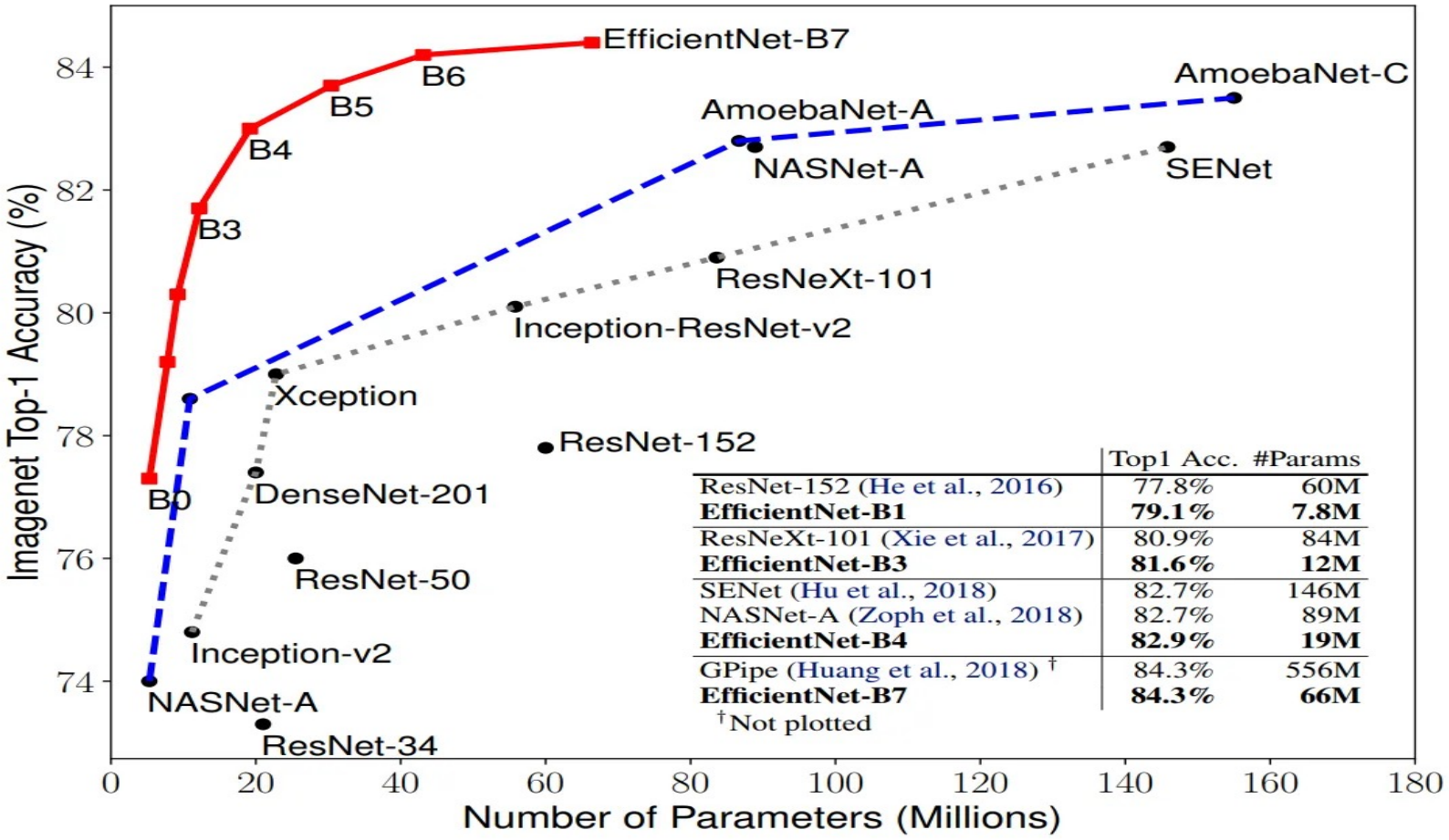
EfficientNet



Tan, et al. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks " *ICML 2019*.

EfficientNet

- EfficientNet size and performance on the ImageNet dataset



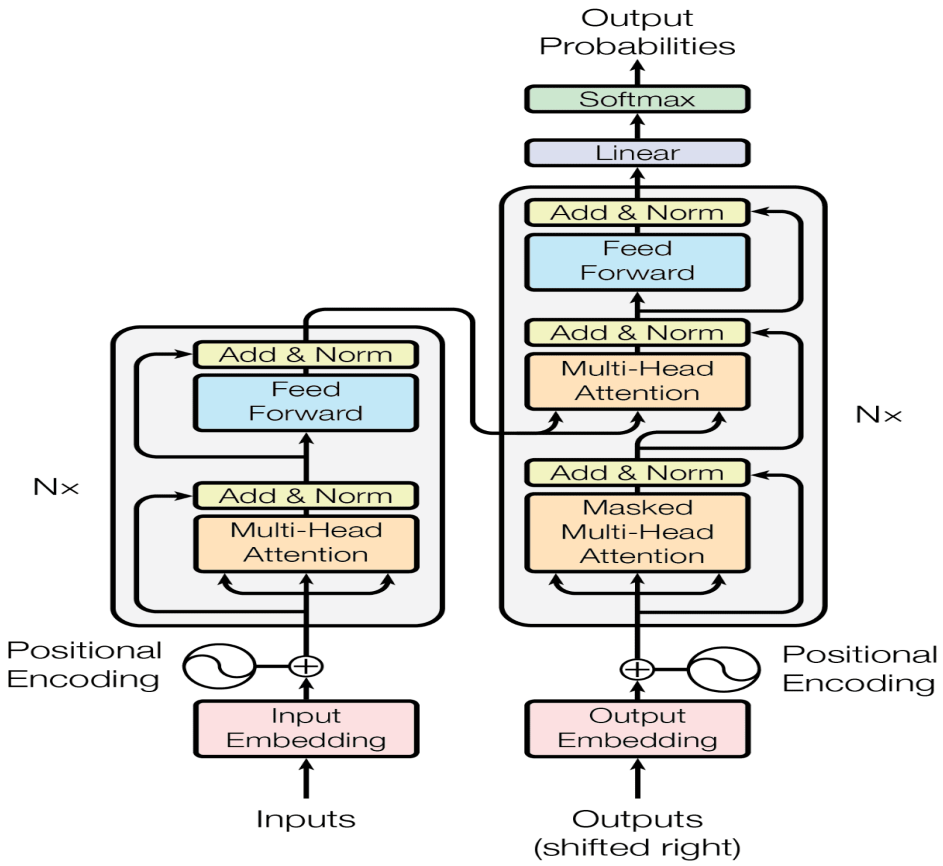
	Top1 Acc.	#Params
ResNet-152 (He et al., 2016)	77.8%	60M
EfficientNet-B1	**79.1%**	**7.8M**
ResNeXt-101 (Xie et al., 2017)	80.9%	84M
EfficientNet-B3	**81.6%**	**12M**
SENet (Hu et al., 2018)	82.7%	146M
NASNet-A (Zoph et al., 2018)	82.7%	89M
EfficientNet-B4	**82.9%**	**19M**
GPipe (Huang et al., 2018) †	84.3%	556M
EfficientNet-B7	**84.3%**	**66M**

†Not plotted

Tan, et al. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks " *ICML 2019*.

Vision Transformer (ViT)

- First application of Transformer model to Images



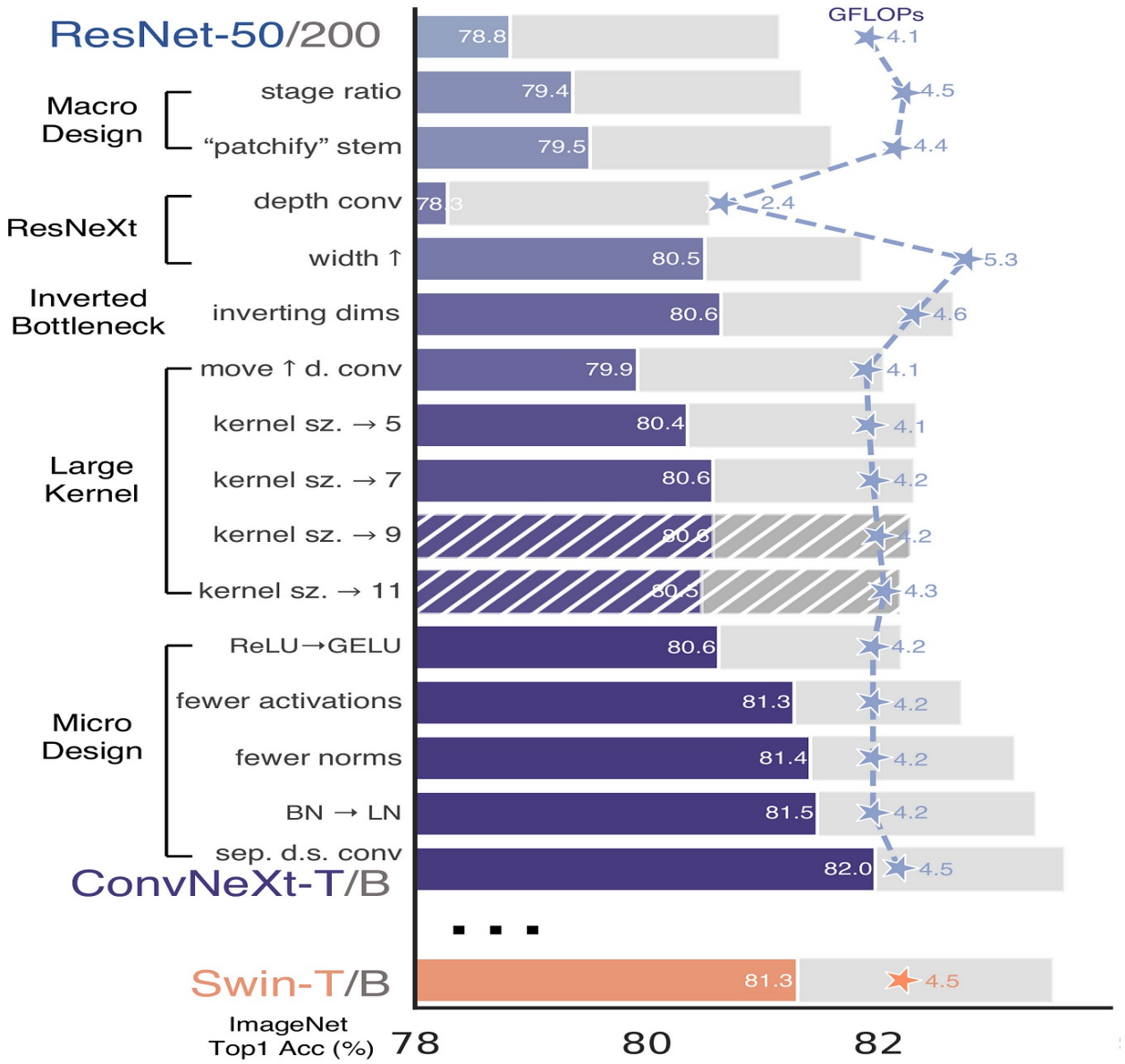
Vision Transformers

Transformers | Davide Coccomini | 2021

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

ConvNeXt

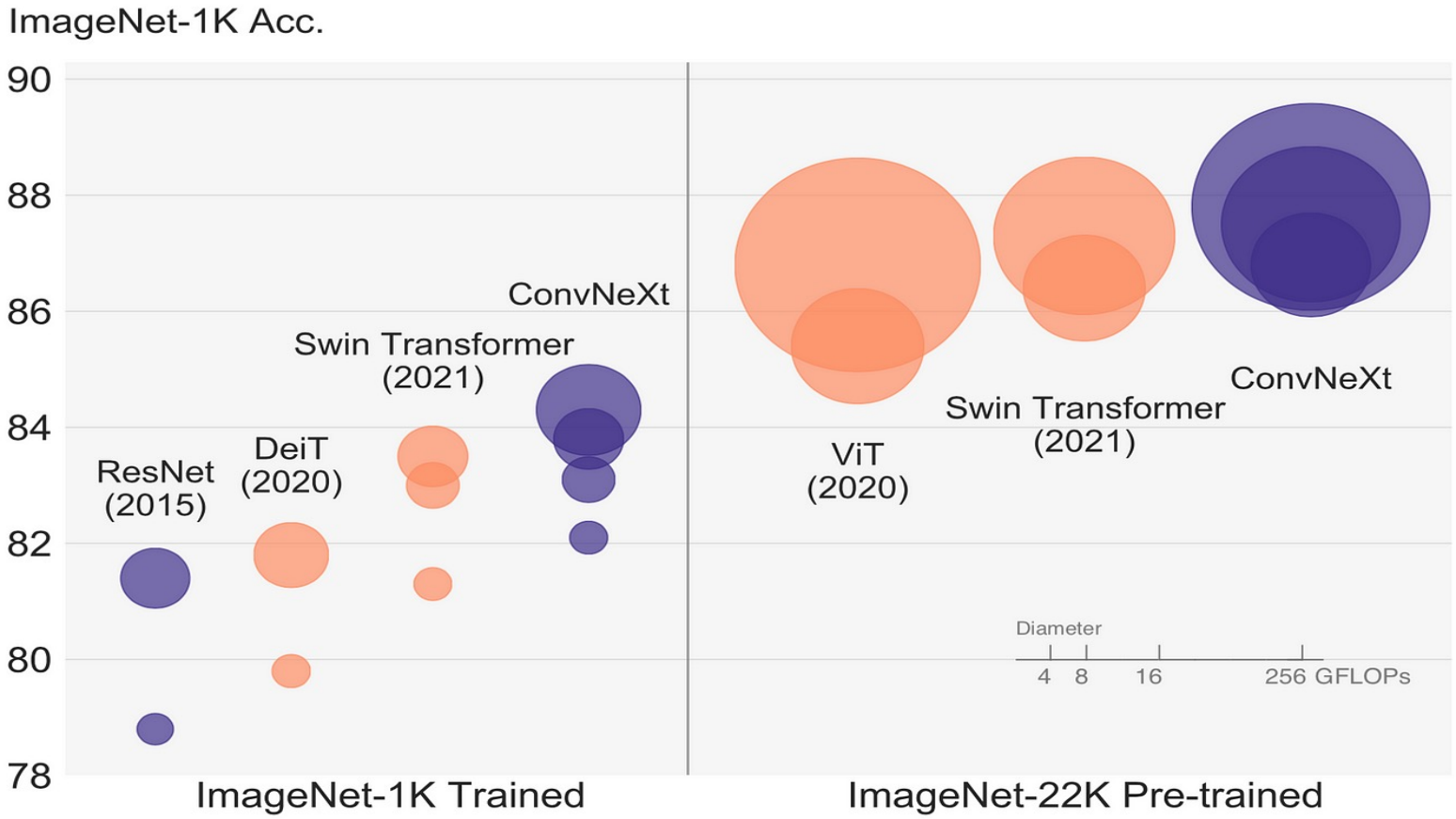
- A ConvNet for the 2020s
- With the rise of Transformer model, shift towards ViT and SwinTransformer
- **Introduced image-specific inductive biases**



Liu et al. "A ConvNet for the 2020s" *CVPR 2022*. https://openaccess.thecvf.com/content/CVPR2022/papers/Liu_A_ConvNet_for_the_2020s_CVPR_2022_paper.pdf .

ConvNeXt

- ConvNeXt performance



Liu et al. "A ConvNet for the 2020s" *CVPR 2022.* https://openaccess.thecvf.com/content/CVPR2022/papers/Liu_A_ConvNet_for_the_2020s_CVPR_2022_paper.pdf .

Recent trends in Deep Learning

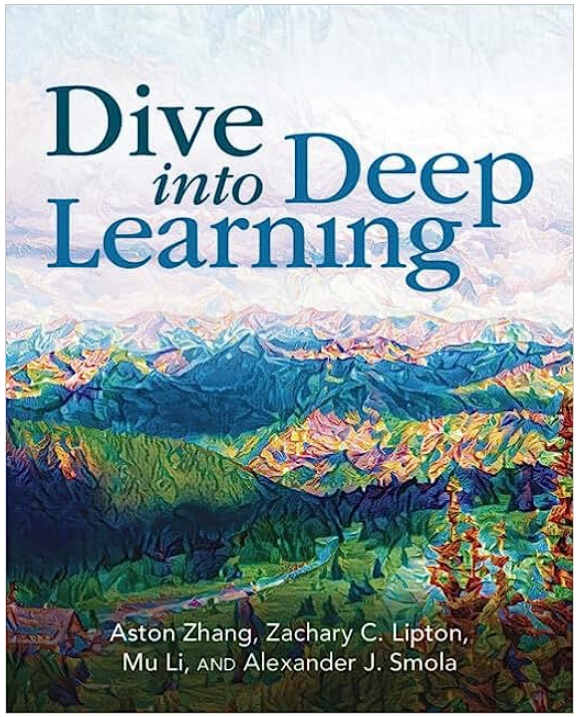
- Unsupervised Learning
- Self-supervised Learning
- Transformers and LLMs: The Game Changers
- Explainable AI (XAI): Demystifying the Black Box
- Federated Learning: Privacy-Preserving AI
- Edge AI: Bringing Intelligence to the Source
- Neuromorphic Computing: Mimicking the human brain
- AI Ethics and Governance: Building Responsible AI

https://www.linkedin.com/pulse/future-deep-learning-trends-emerging-technologies-devfiinc-rxzec/

Demo Examples

Demo Example 1

- AlexNet – PyTorch
  - https://d2l.ai/chapter_convolutional-modern/alexnet.html
- VGG16
  - https://d2l.ai/chapter_convolutional-modern/vgg.html
- ResNet
  - https://d2l.ai/chapter_convolutional-modern/resnet.html
  - https://colab.research.google.com/github/d2l-ai/d2l-pytorch-colab/blob/master/chapter_convolutional-modern/resnet.ipynb



https://d2l.ai/index.html

Key takeaways

- Training methodology
  - Split data into training (such as 70%), validation (10%), and testing (20%)
  - Take care of data leakage
  - Check distribution of classes, work on balanced datasets (ideally)
  - Find and develop baseline models at first
  - Tune hyperparameters on validation set. Save best model and do inference on test set (once)
  - Don’t use off-the-shelf model blindly. Do ablation studies to know its working
- Data augmentation techniques are not standardized
  - Get input from experts to know what data augmentations make sense in the domain
- Results
  - Use multiple metrics rather a single metric to report results (often they are complementary)
  - Show both qualitative and quantitative results (e.g., image segmentation)

Acknowledgements

- Slides from
  - https://syncedreview.com/2020/06/23/google-deepmind-researchers-revamp-imagenet/
  - Lecun et al. (1989). Gradient-based learning applied to document recognition.
  - Deep Learning's Most Important Ideas. https://www.kdnuggets.com/2020/09/deep-learnings-most-important-ideas.html
  - https://medium.com/coinmonks/paper-review-of-vggnet-1st-runner-up-of-ilsvlc-2014-image-classification-d02355543a11
  - Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
  - Source: Convolutional Neural Networks. https://medium.com/@rajat.k.91/convolutional-neural-networks-why-what-and-how-f8f6dbebb2f9
  - Image Credit: Medium. https://medium.com/@pierre_guillou/understand-how-works-resnet-without-talking-about-residual-64698f157e0c
  - Source: Convolutional Neural Networks. https://medium.com/@rajat.k.91/convolutional-neural-networks-why-what-and-how-f8f6dbebb2f9

Example exam question

What does transfer learning with CNNs involve?

A. Training a model from scratch for each new task.

B. Using a pretrained model and fine-tuning it for a new task.

C. Converting image data to text data to facilitate learning.

D. Combining multiple CNNs into a single model for better performance.