UNIVERSIDADE CATÓLICA PORTUGUESA

BRAGA

Deep Learning

Session 13

Introduction to Computer Vision and Image Classification

Applied Data Science

2024/2025

Computer Vision: Computers that "See"

Self-driving cars

Exploration on Mars

Guided surgery

Visual assistance for people who are blind

Security

Why Discuss Computer Vision with CNNs?

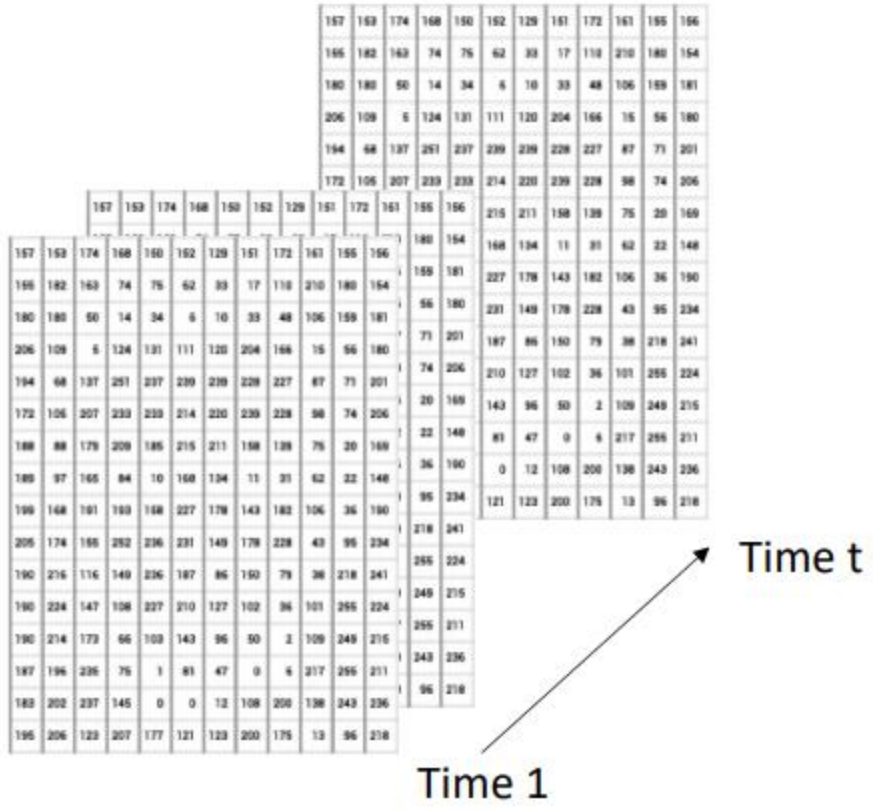
UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- CNNs have a strong track record for vision problems;
- Visual data's representation (i.e., spatial data) is naturally suited for CNNs.

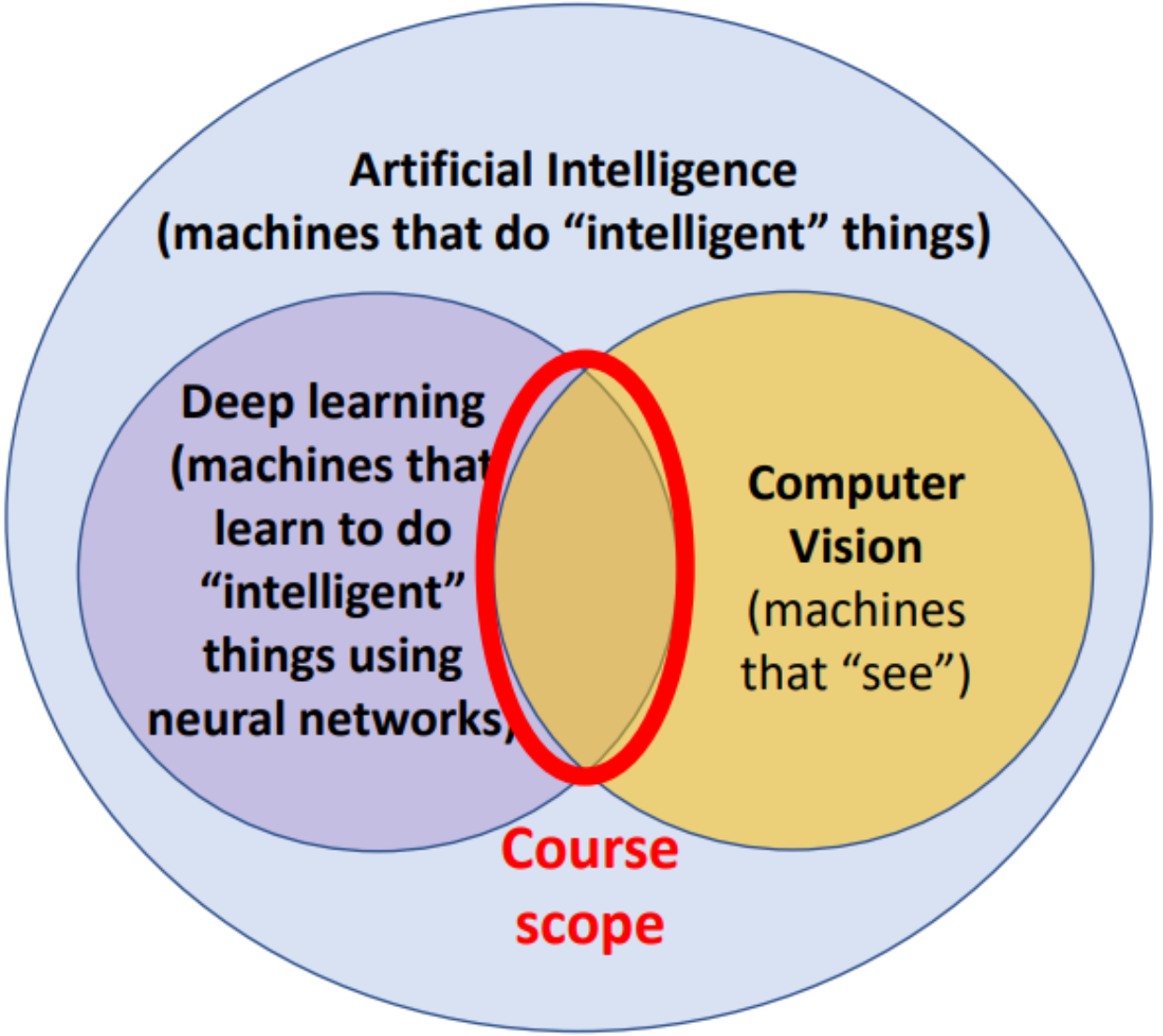
Image:

157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

Video:



Computer Vision



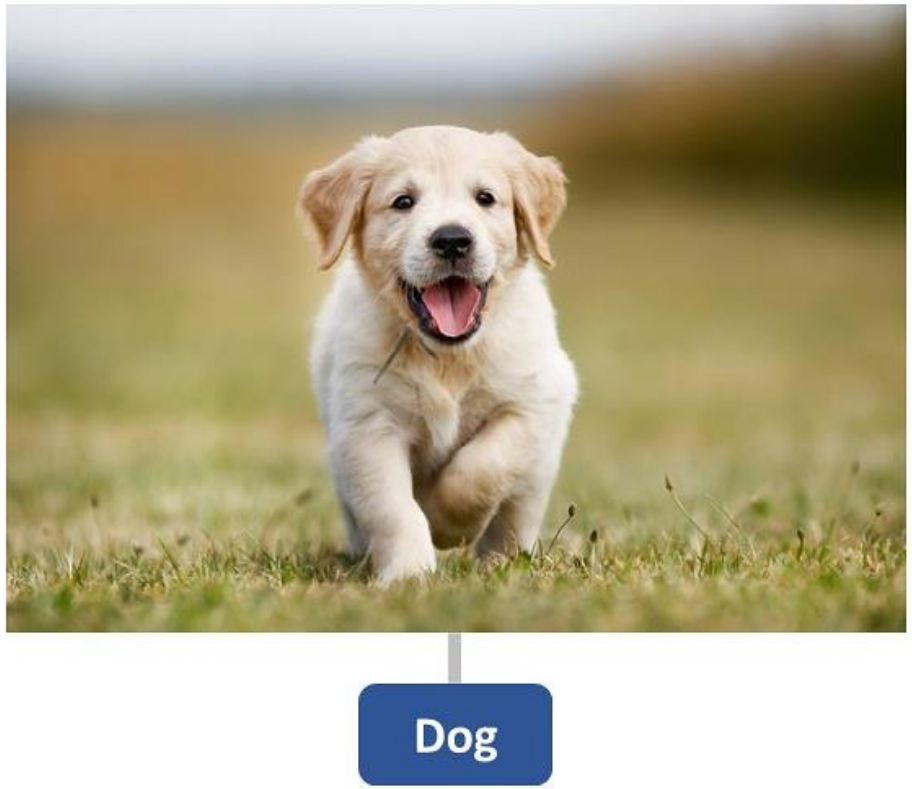
Computer Vision

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- **Key Challenge:** Replicate Human Vision for So Much Variation for **So Many Tasks**!
  - Object Recognition
  - Object Detection
  - Segmentation
  - Image Captioning
  - Visual Question Answering
  - Object Tracking
  - Subjective Problems
  - And many more…

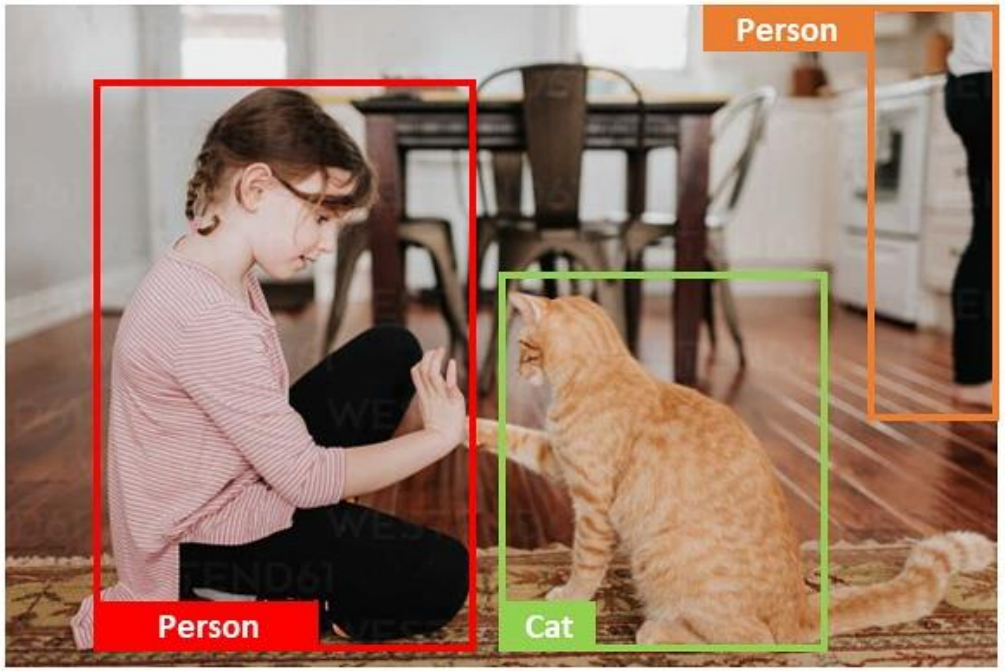
Computer Vision

- **Key Challenge:** Replicate Human Vision for So Much Variation for **So Many Tasks**!
  - **Object Recognition**
  - Object Detection
  - Segmentation
  - Image Captioning
  - Visual Question Answering
  - Object Tracking
  - Subjective Problems
  - And many more…



Computer Vision

- **Key Challenge:** Replicate Human Vision for So Much Variation for **So Many Tasks**!
  - Object Recognition
  - **Object Detection**
  - Segmentation
  - Image Captioning
  - Visual Question Answering
  - Object Tracking
  - Subjective Problems
  - And many more…



Computer Vision

- **Key Challenge:** Replicate Human Vision for So Much Variation for **So Many Tasks**!
  - Object Recognition
  - Object Detection
  - **Segmentation**
  - Image Captioning
  - Visual Question Answering
  - Object Tracking
  - Subjective Problems
  - And many more…

Computer Vision

- **Key Challenge:** Replicate Human Vision for So Much Variation for **So Many Tasks**!
  - Object Recognition
  - Object Detection
  - Segmentation
  - **Image Captioning**
  - Visual Question Answering
  - Object Tracking
  - Subjective Problems
  - And many more…

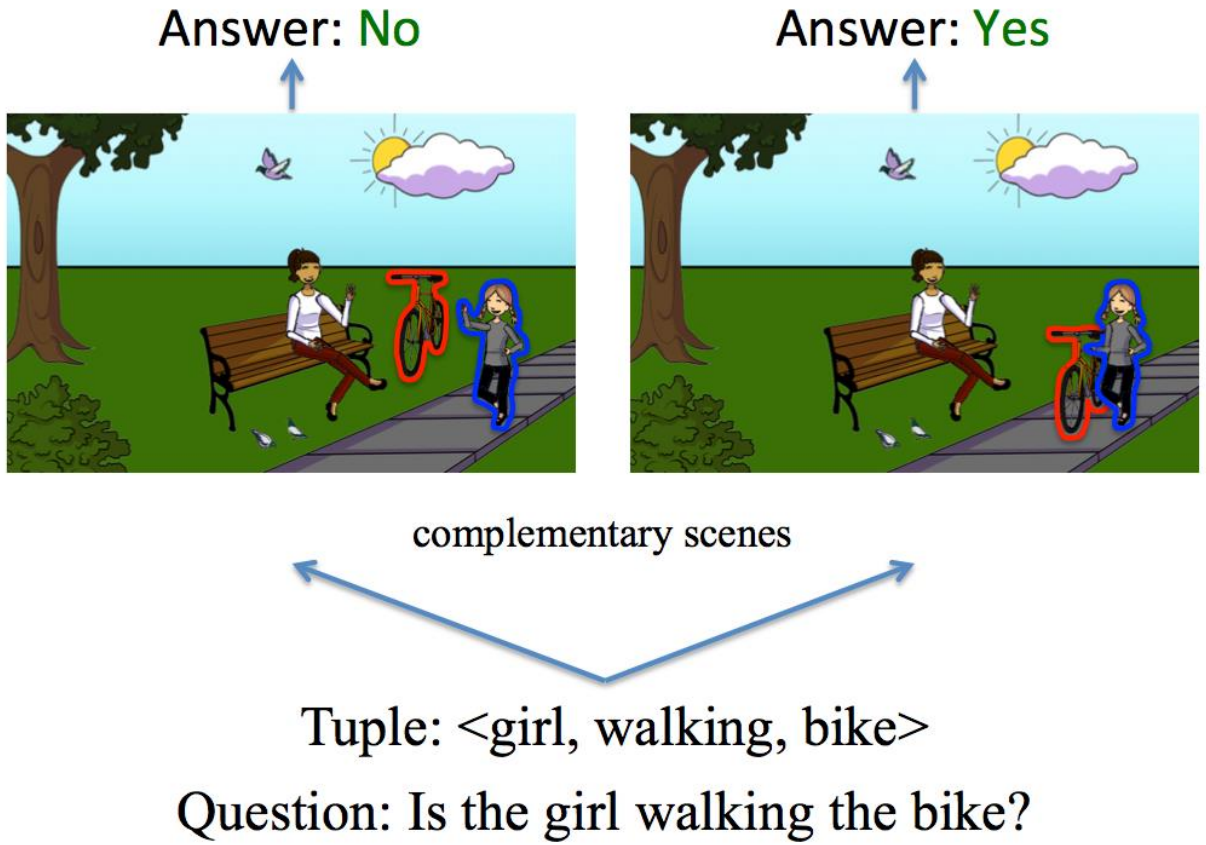
"Man in black shirt is playing guitar."

"Construction worker in orange safety vest is working on road."

"Two young girls are playing with lego toy."

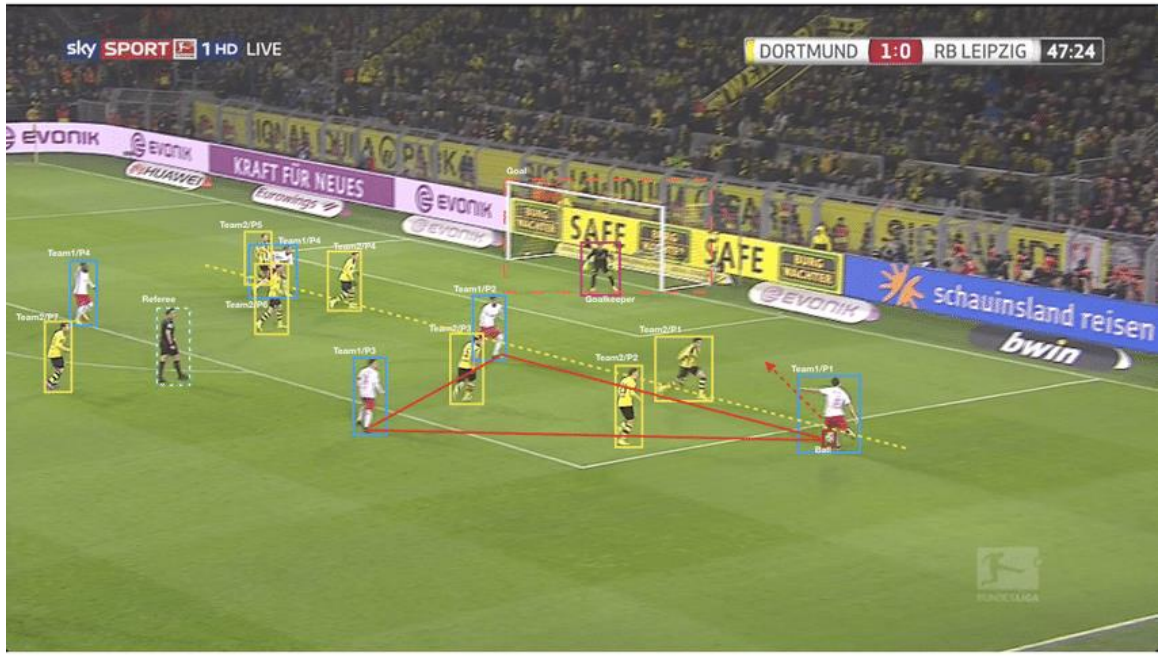
Computer Vision

- **Key Challenge:** Replicate Human Vision for So Much Variation for **So Many Tasks**!
  - Object Recognition
  - Object Detection
  - Segmentation
  - Image Captioning
  - **Visual Question Answering**
  - Object Tracking
  - Subjective Problems
  - And many more…



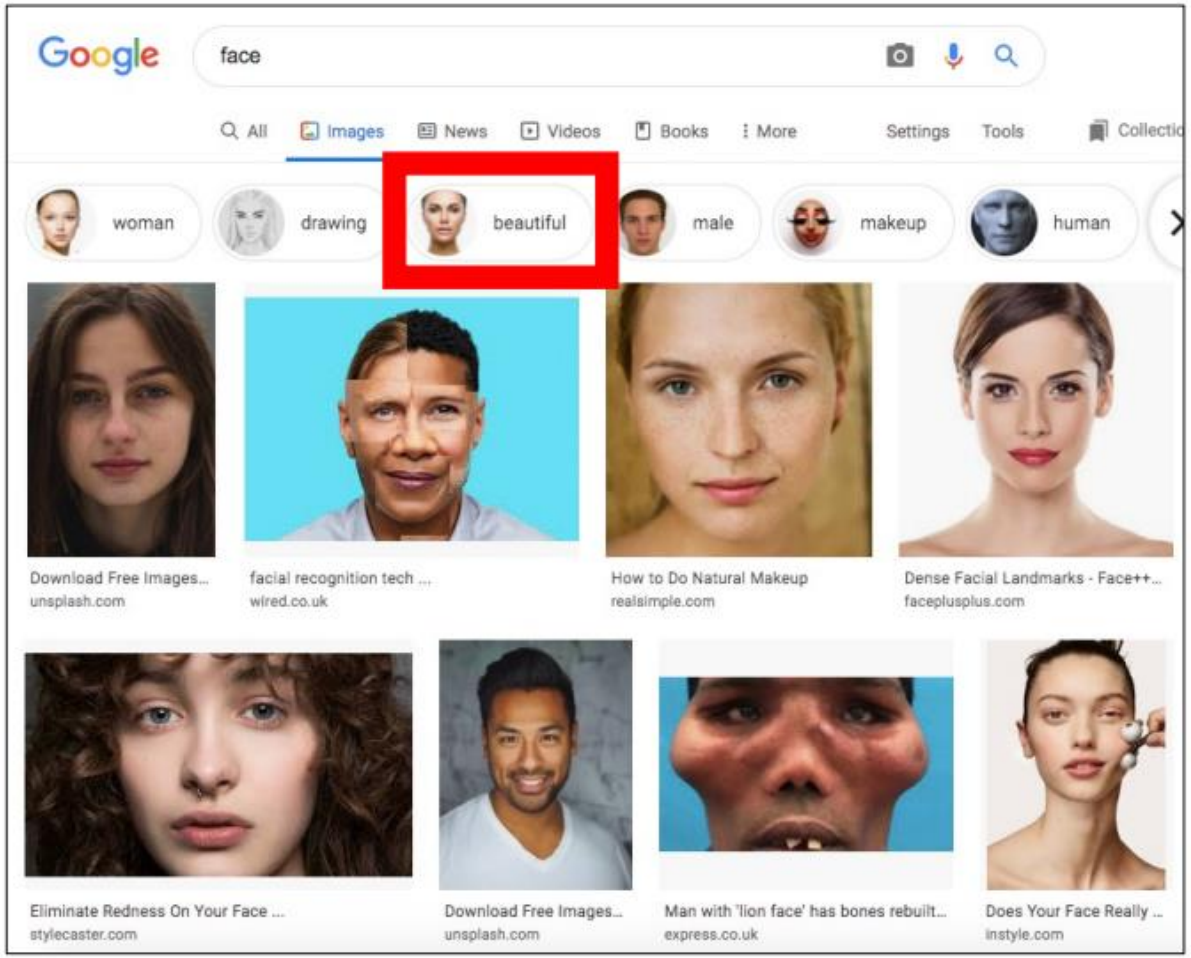
Computer Vision

- **Key Challenge:** Replicate Human Vision for So Much Variation for **So Many Tasks**!
  - Object Recognition
  - Object Detection
  - Segmentation
  - Image Captioning
  - Visual Question Answering
  - **Object Tracking**
  - Subjective Problems
  - And many more…



Computer Vision

- **Key Challenge:** Replicate Human Vision for So Much Variation for **So Many Tasks**!
  - Object Recognition
  - Object Detection
  - Segmentation
  - Image Captioning
  - Visual Question Answering
  - Object Tracking
  - **Subjective Problems**
  - And many more…



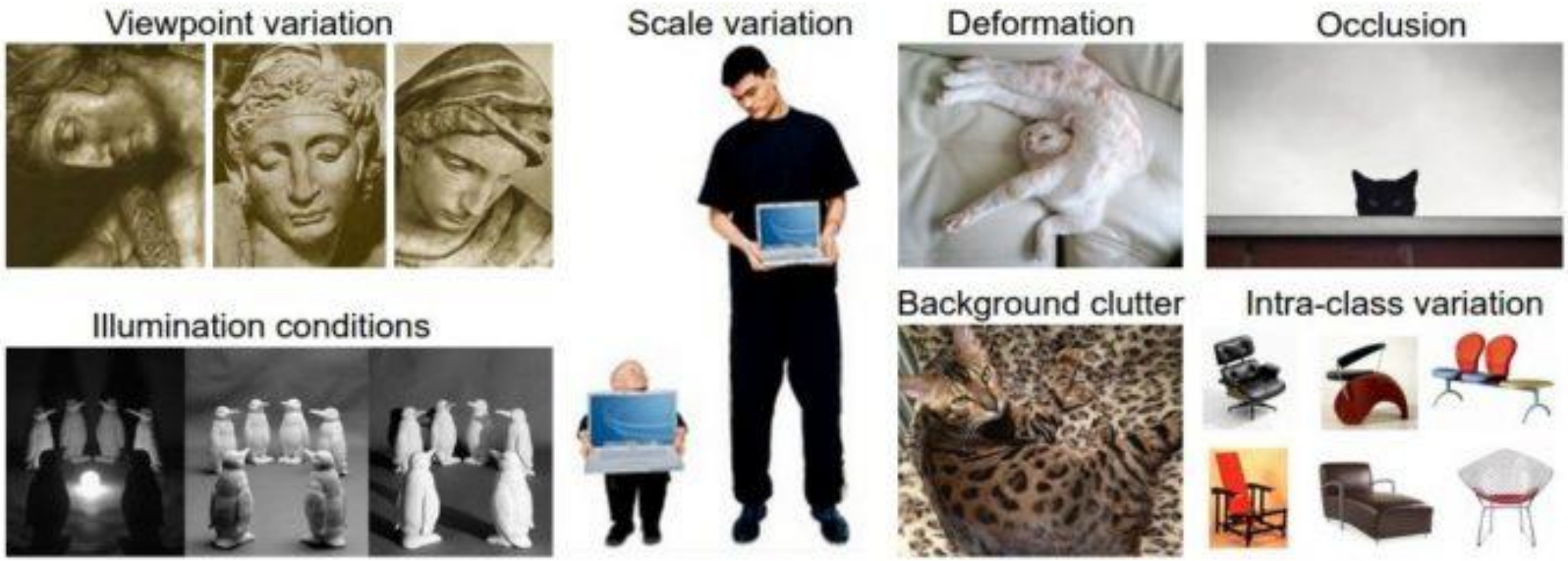
Computer Vision

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- **Key Challenge:** Replicate Human Vision for So Much Variation for **So Many Tasks**!
  - Object Recognition
  - Object Detection
  - Segmentation
  - Image Captioning
  - Visual Question Answering
  - Object Tracking
  - Subjective Problems
  - **And many more…**

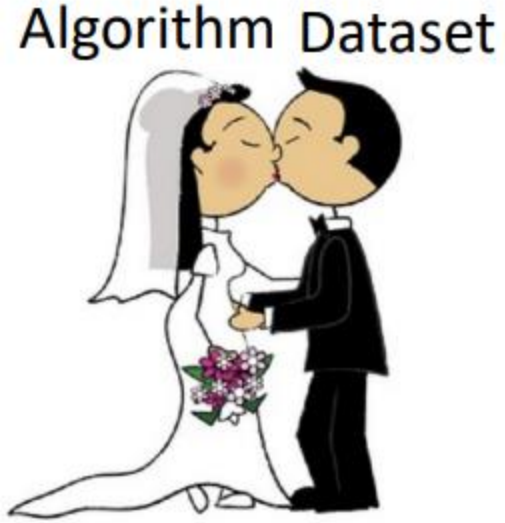
Computer Vision

- **Key Challenge:** Replicate Human Vision for **So Much Variation** for So Many Tasks!

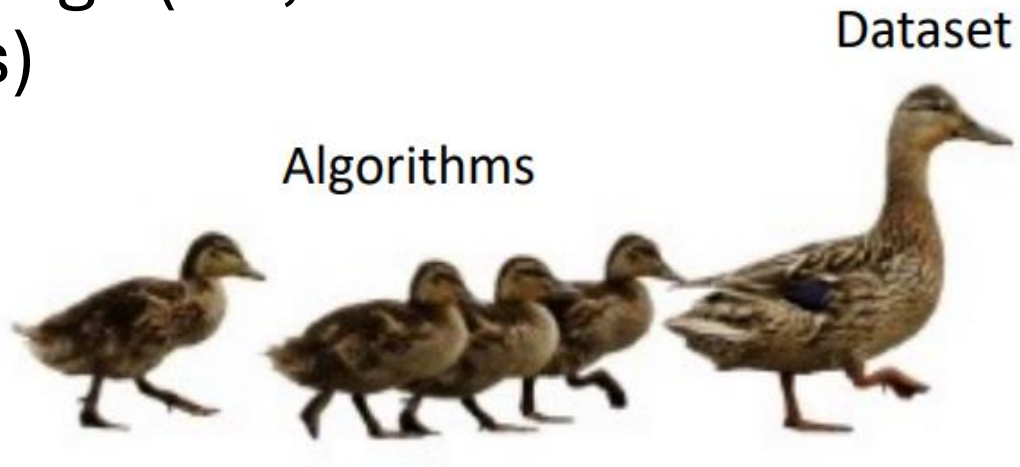


Computer Vision

- Through 1990s, datasets tended to be relatively small (e.g., 10s or 100s of examples)



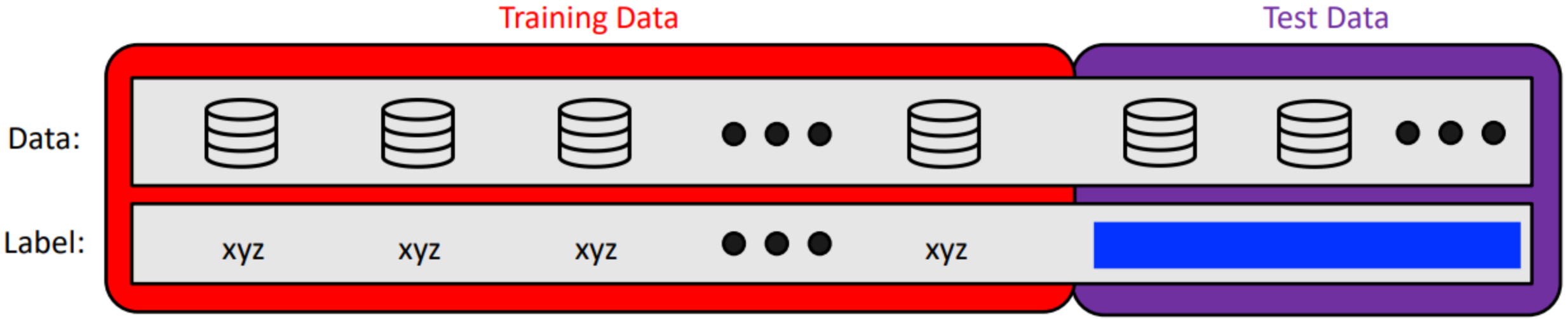
- Since 1990s, datasets tend to be large (i.e., thousands to millions of examples)



Computer Vision: The Era of Dataset Challenges

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- What do you think prompted this shift to large-scale datasets?
  - Progress Charted by Progress on Community Shared Dataset Challenges



  - 1. Dataset split into a "training set" and "test set" with the labels for the "test set" hidden
  - 2. Teams design a model and submit its predictions on the test set to an evaluation server
  - 3. A public leaderboard shows the ranking of performance for all teams' submitted models

Computer Vision: The Era of Dataset Challenges

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- Why challenges?
  - Provide fair comparison between models
  - Create a community around a shared goal

Many Public Dataset Challenges Available

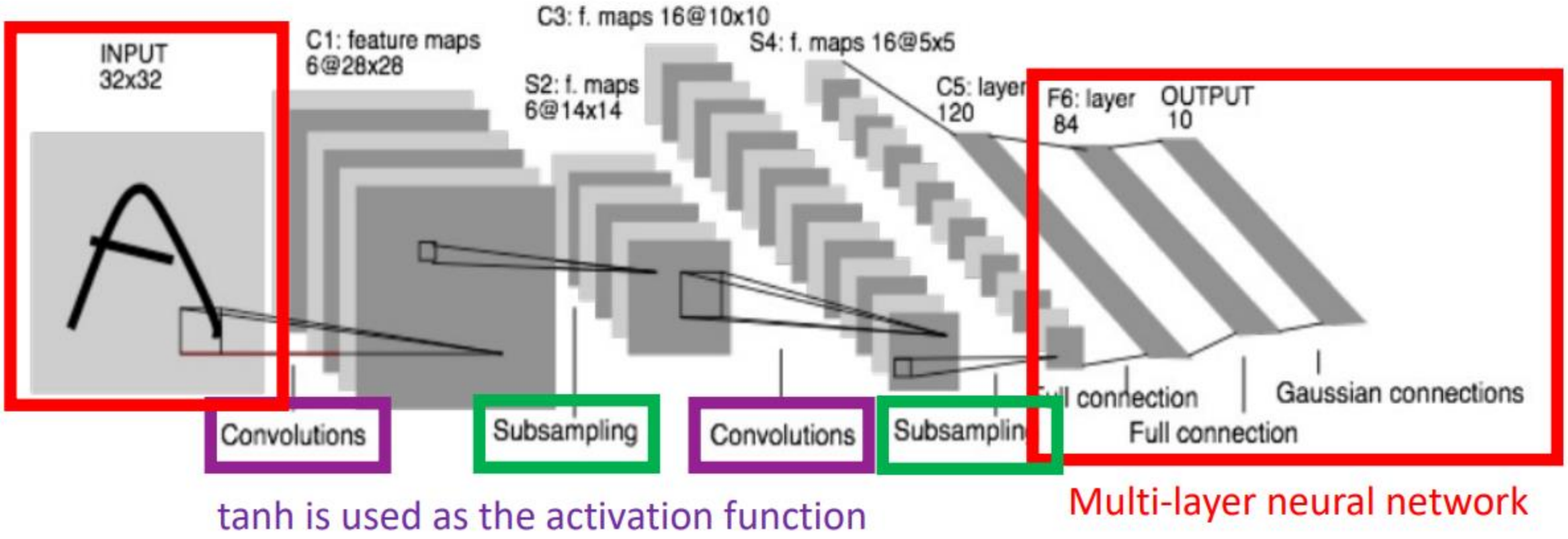
- Google Dataset Search
- Amazon's AWS Datasets
- Kaggle Datasets
- Wikipedia's List
- UC Irvine Machine Learning Repository
- Reddit
- https://dataportals.org/

MNIST Challenge

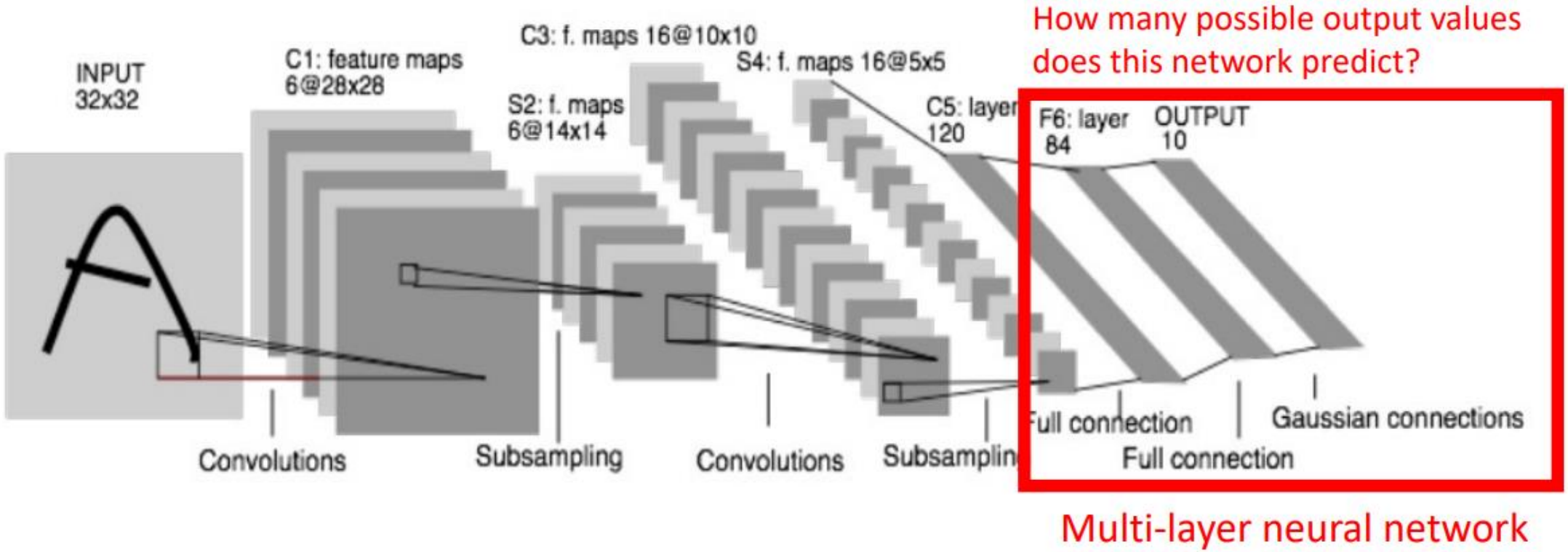
- **Goal:** classify digit as 0, 1, …, or 9
- **Evaluation metric:** accuracy (% correct)
- **Dataset:** 60,000 training and 10,000 test examples, pre-processed to be centered and same dimension; writers were different in the two sets
- **Source:** images collected by NIST from a total of 500 Census Bureau employees and high school students

MNIST Challenge Winner: LeNet

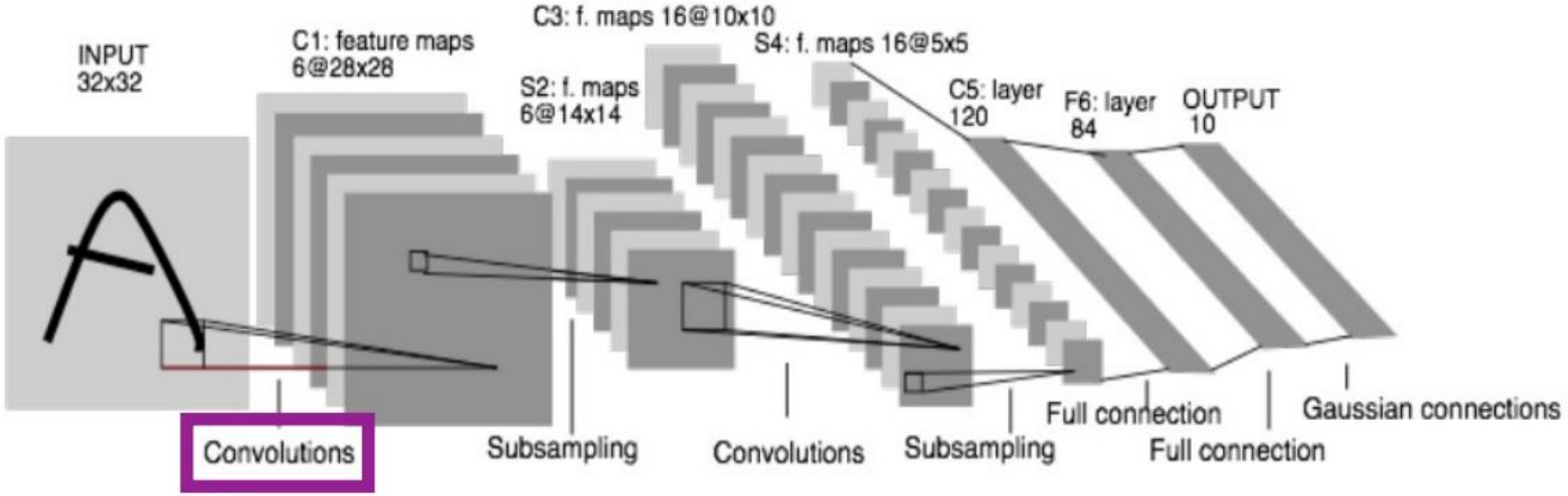
- LeNet: Architecture (like Neocognitron, has convolutional layers and pooling layers)



MNIST Challenge Winner: LeNet

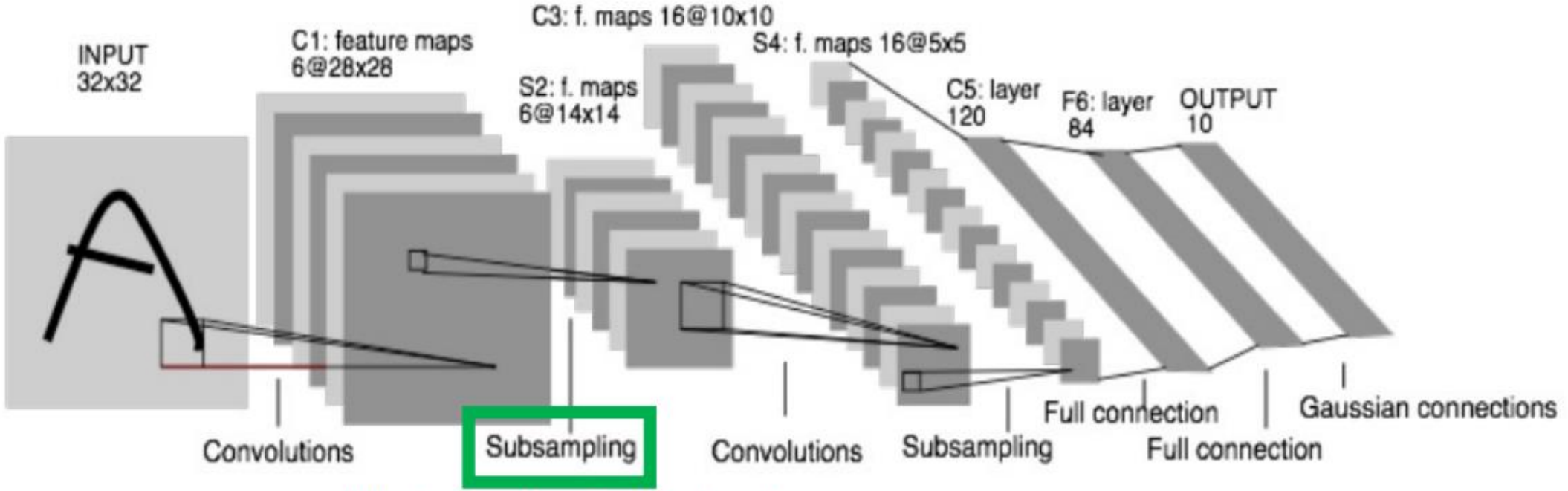


MNIST Challenge Winner: LeNet



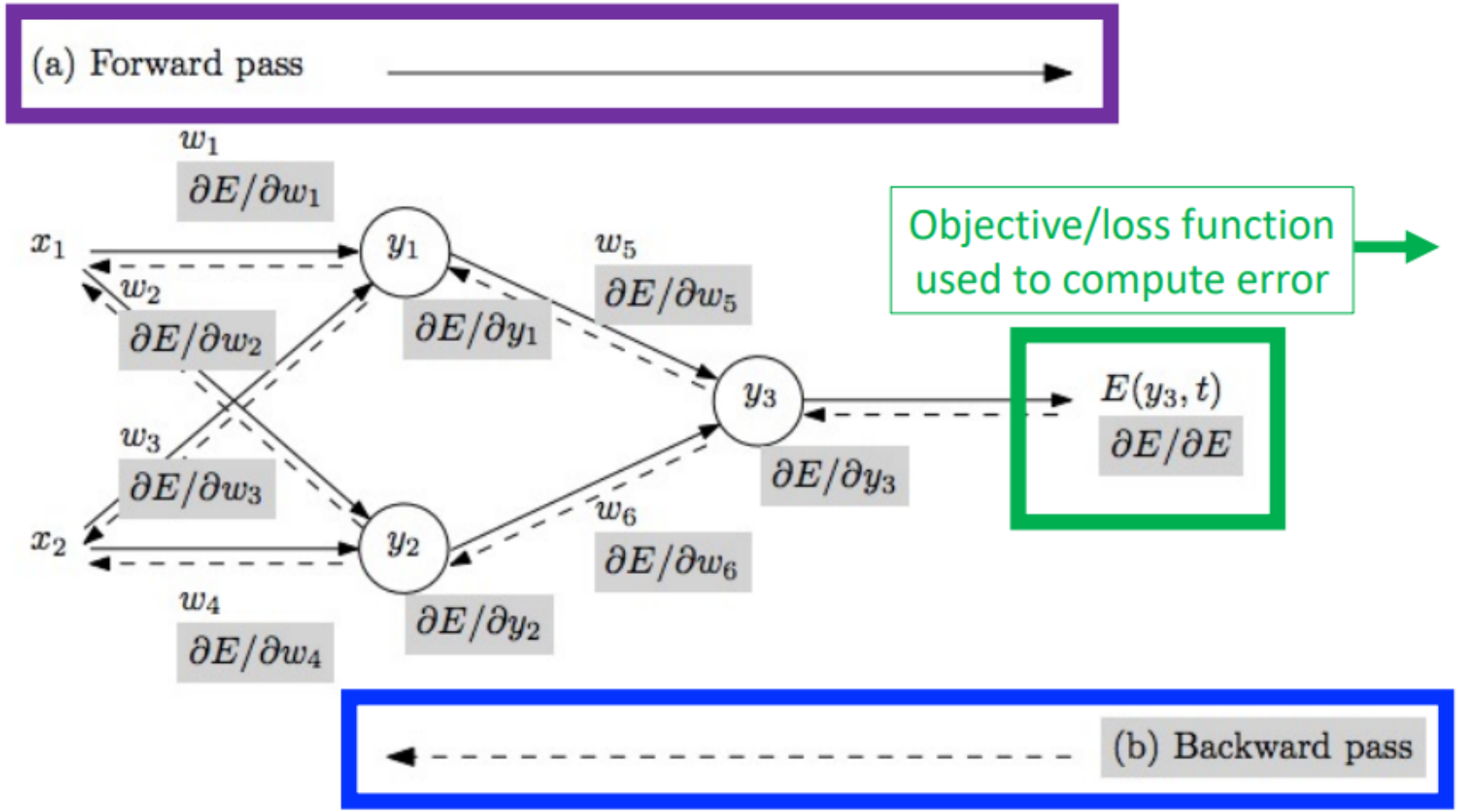
How many filters are between the input and hidden layer 1?

MNIST Challenge Winner: LeNet



What size of a neighborhood is used for this pooling layer?

Training Procedure Approach (Key Novelty)

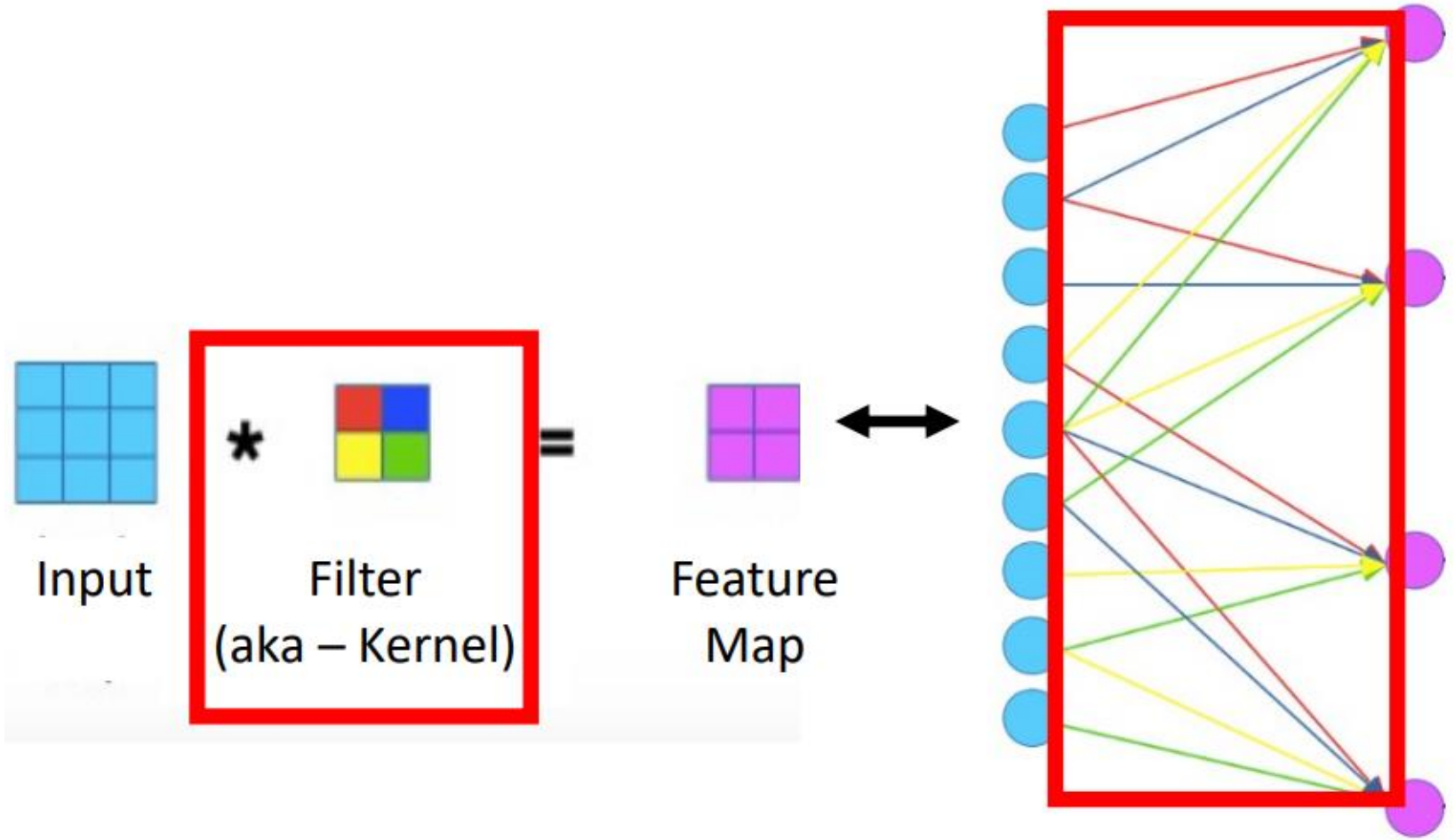


Objective/loss function used to compute error

- Repeat until stopping criterion met:
  1. **Forward pass**: propagate training data through model to make prediction
  2. Quantify the dissatisfaction with a model's results on the training data
  3. **Backward pass**: using predicted output, calculate gradients backward to assign blame to each model parameter
  4. Account for weight sharing by using average of all connections for a parameter
  5. Update each parameter using calculated gradients

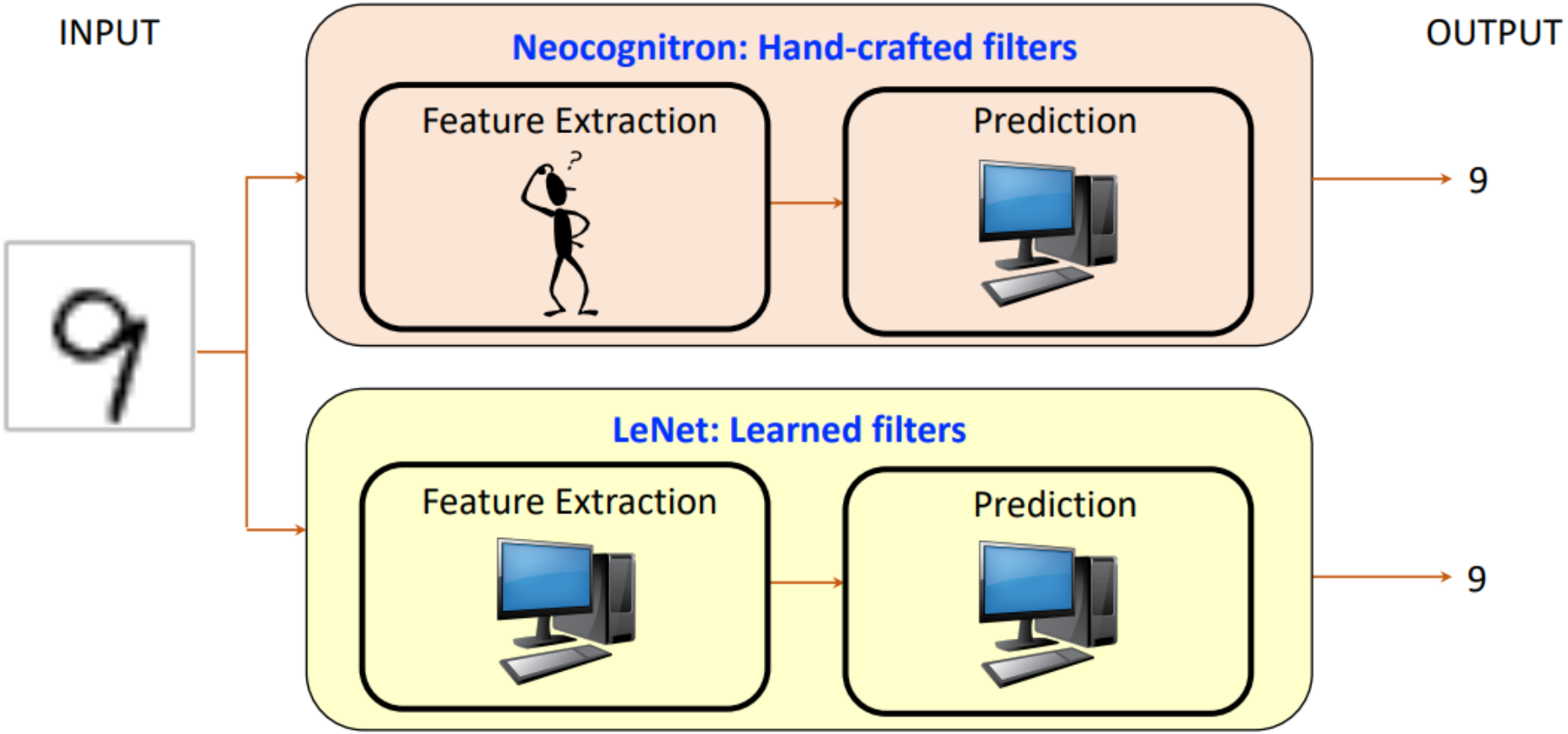
Figure from: Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, Jeffrey Mark Siskind; Automatic Differentiation in Machine Learning: a Survey; 2018

Training Procedure Approach (Key Novelty)



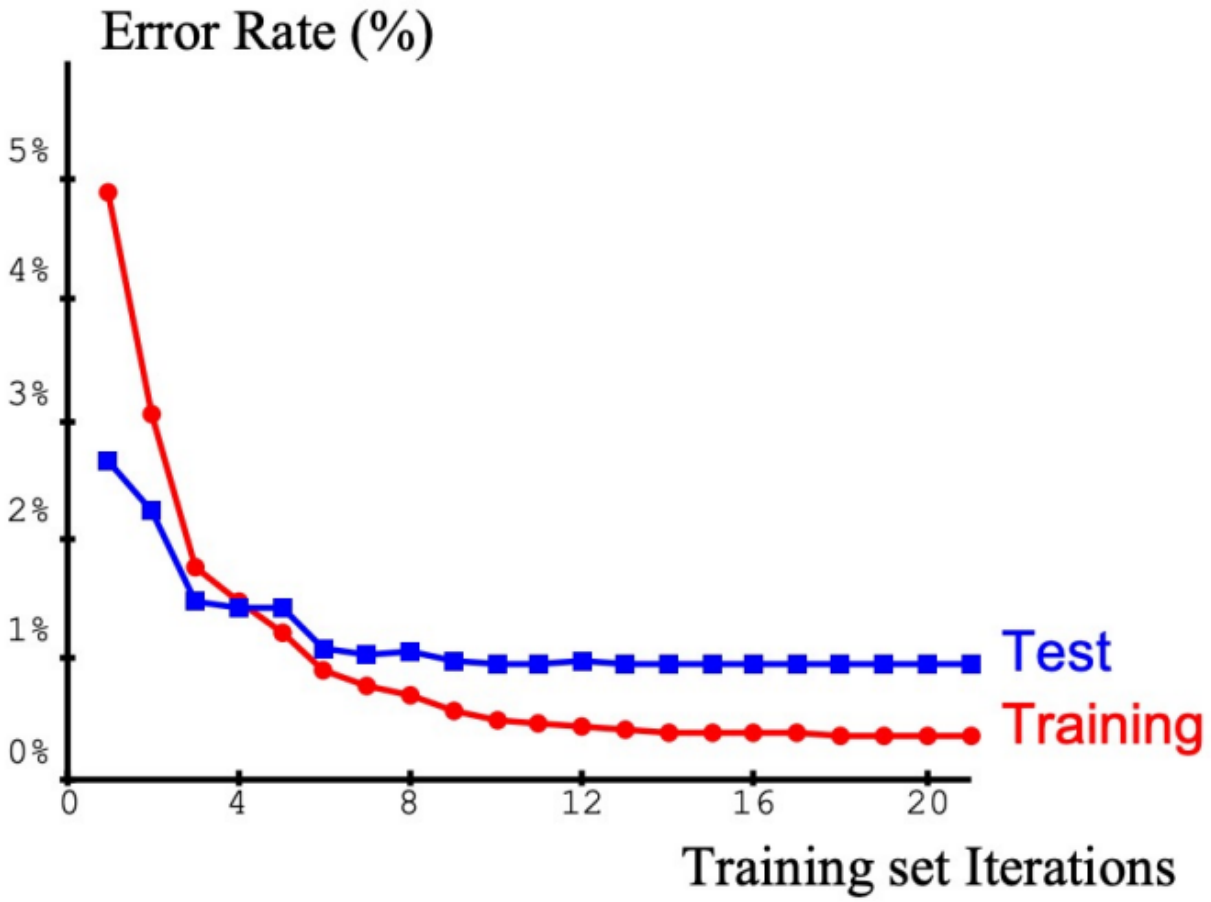
- Repeat until stopping criterion met:
  1. **Forward pass**: propagate training data through model to make prediction
  2. Quantify the dissatisfaction with a model's results on the training data
  3. **Backward pass**: using predicted output, calculate gradients backward to assign blame to each model parameter
  4. Account for weight sharing by using average of all connections for a parameter
  5. Update each parameter using calculated gradients

LeNet vs Neocognitron



LeNet Performance Analysis

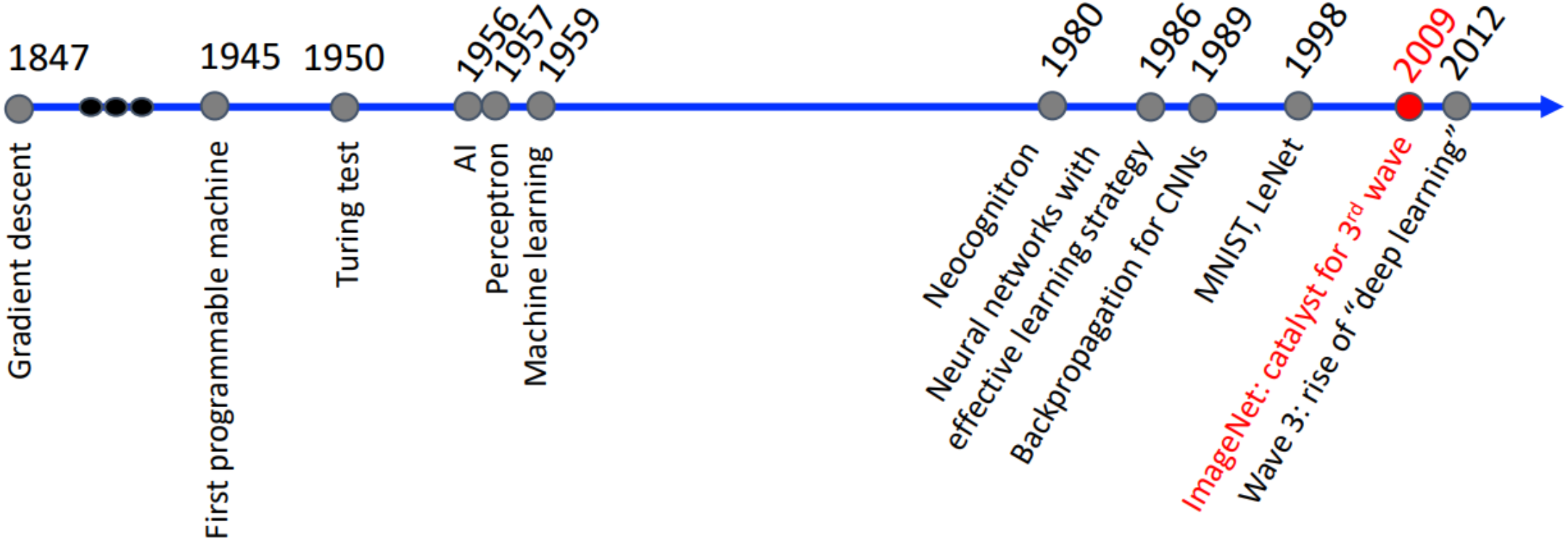
How many epochs are needed for training to converge?



LeNet

- **LeNet, designed on the MNIST Challenge, was used to read over 10% of checks in North America in the 1990s, reading millions of checks every mont!**

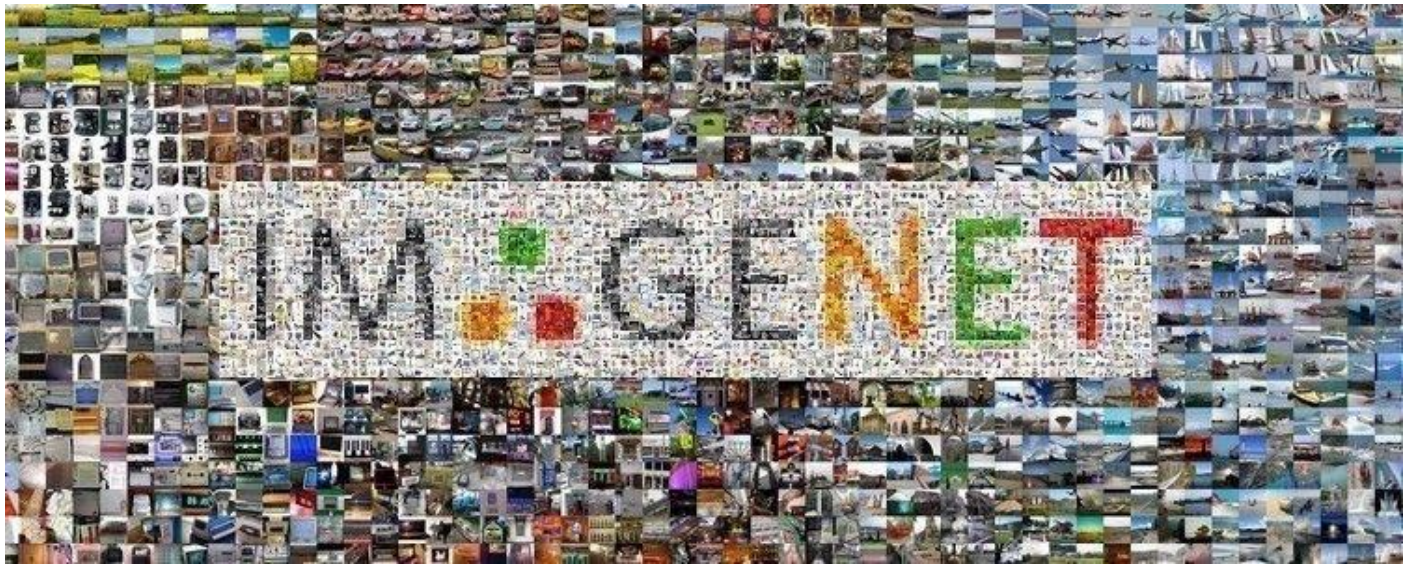
Historical Context



ImageNet: Predict Category from 1000 Options

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- **Evaluation metric:** % correct (top-1 and top-5 predictions)
- **Dataset:** ~1.5 million images
- **Source:** images scraped from search engines, such as Flickr, and labeled by crowdworkers

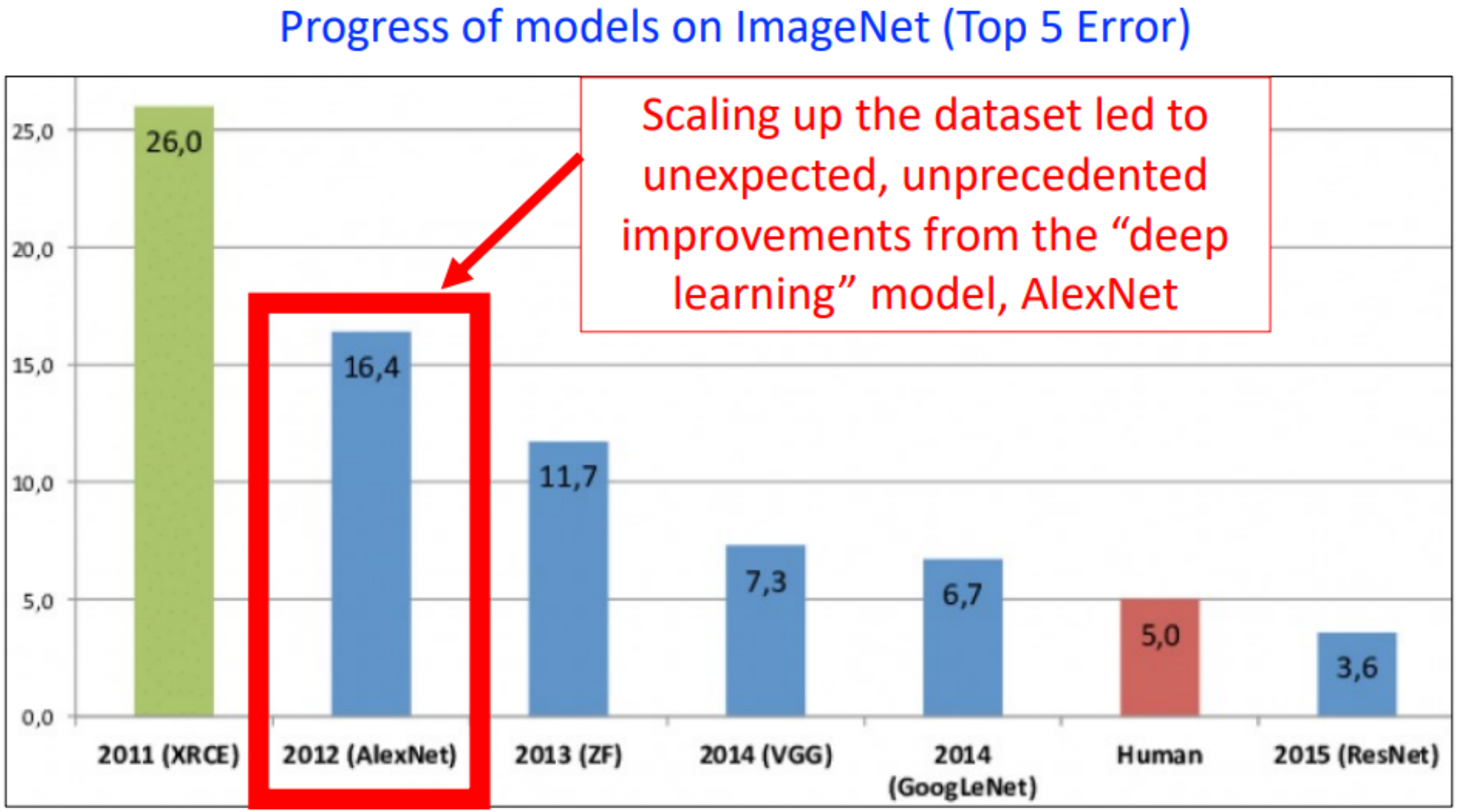


ImageNet vs MNISt

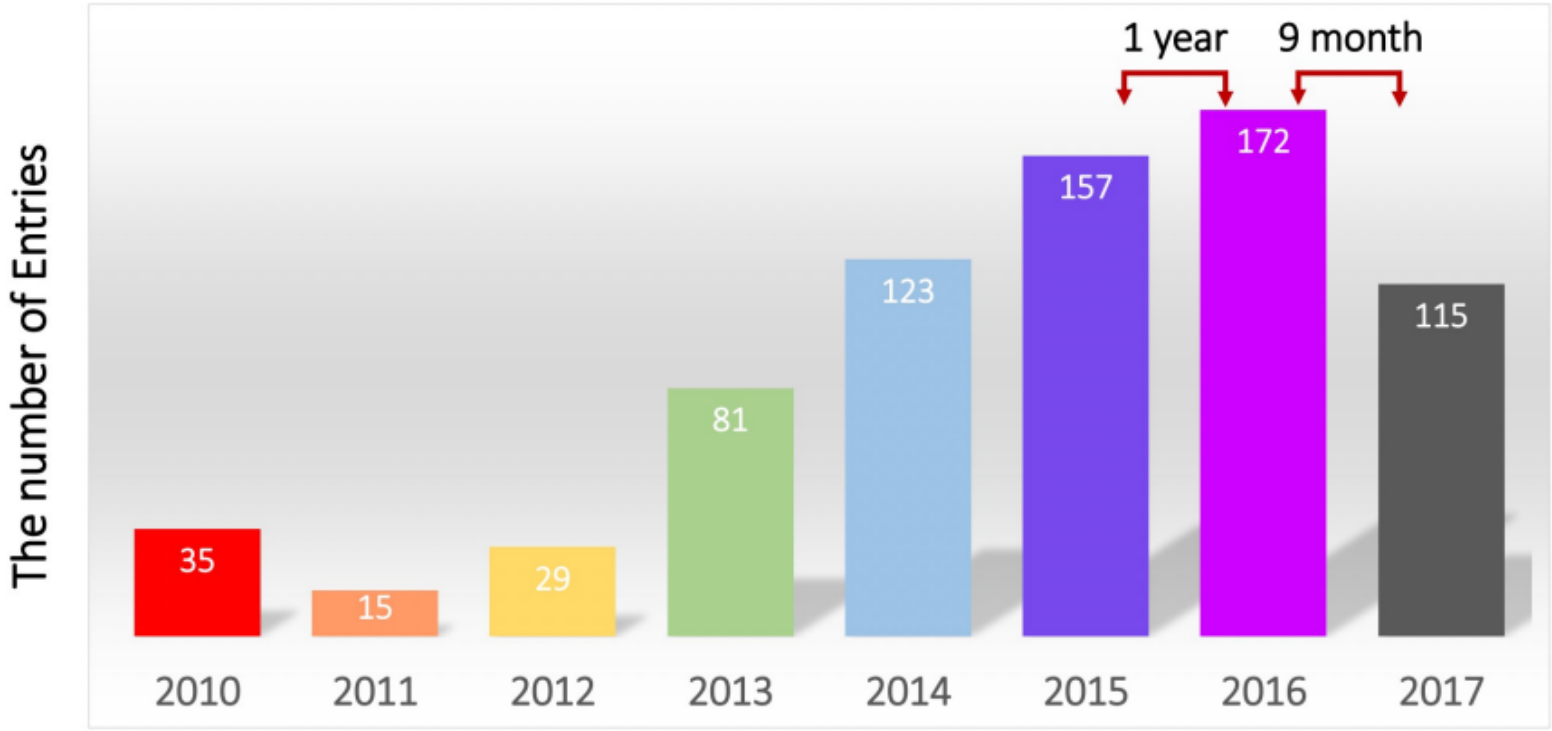
- 3D objects in natural backgrounds

- Many more categories

Rise of Deep Learning

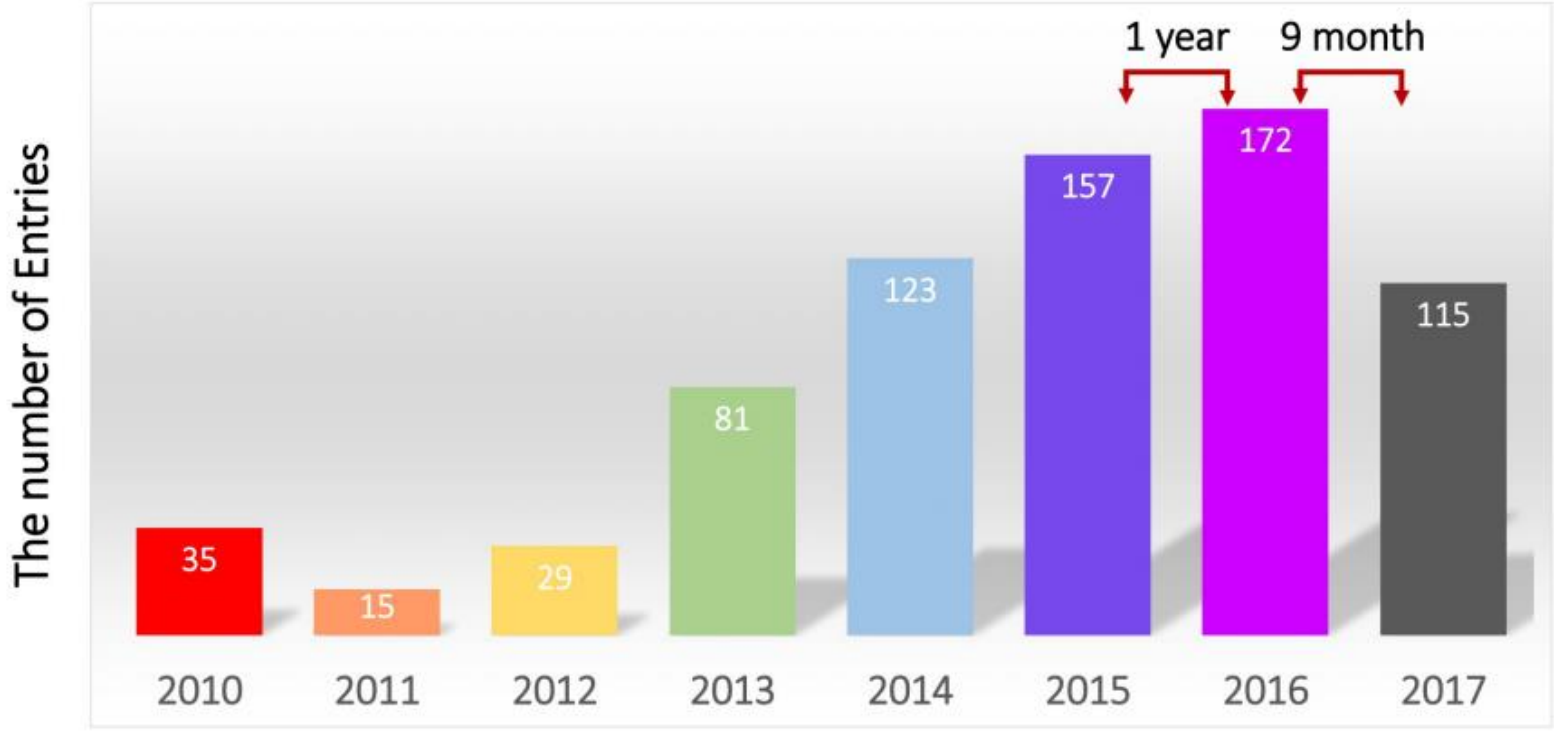


Rise of Deep Learning Following AlexNet



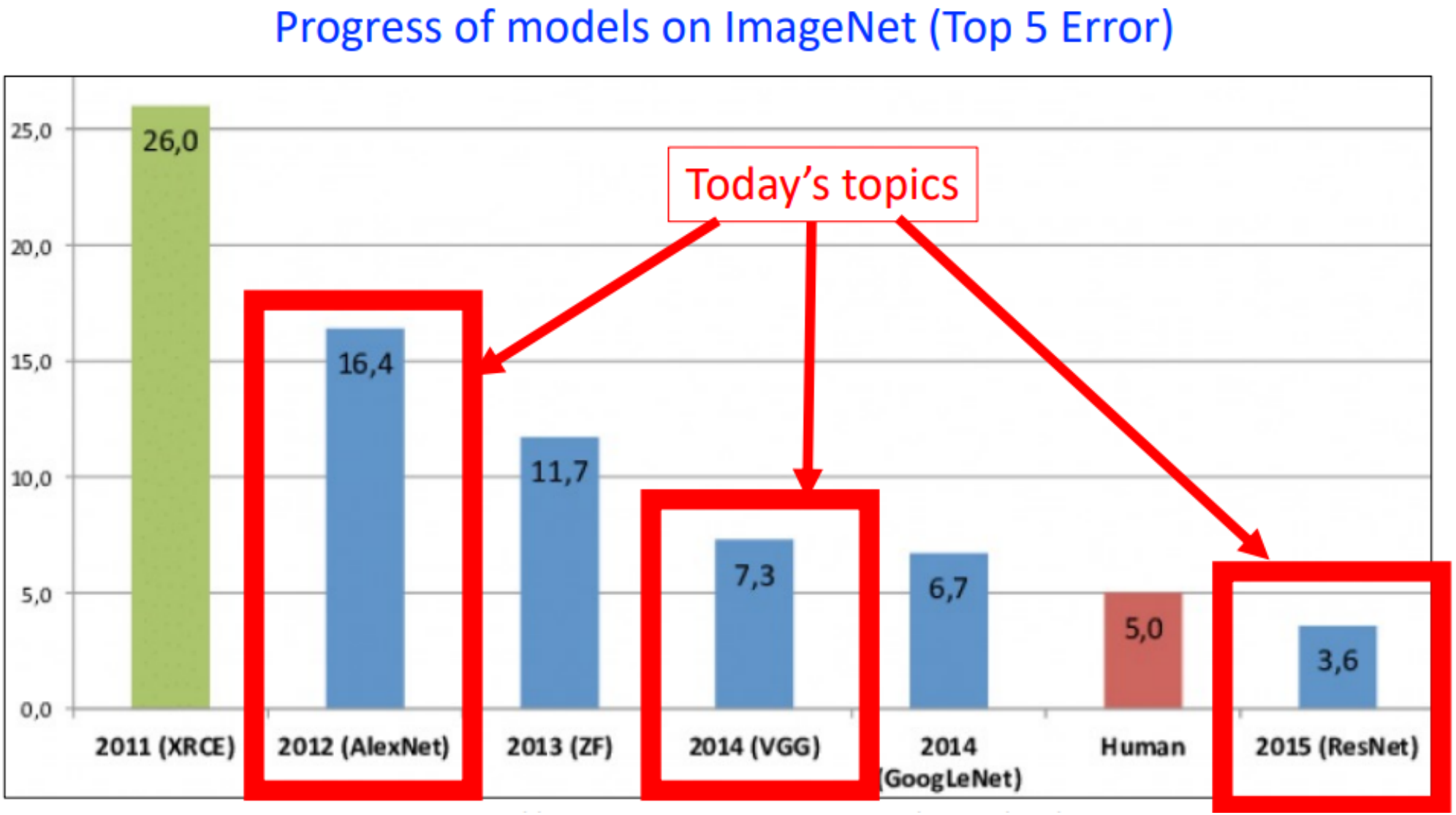
Inspired by AlexNet, many more researchers in the computer vision community proposed neural networks and showed how to make further progress over the years!

Rise of Deep Learning Following AlexNet



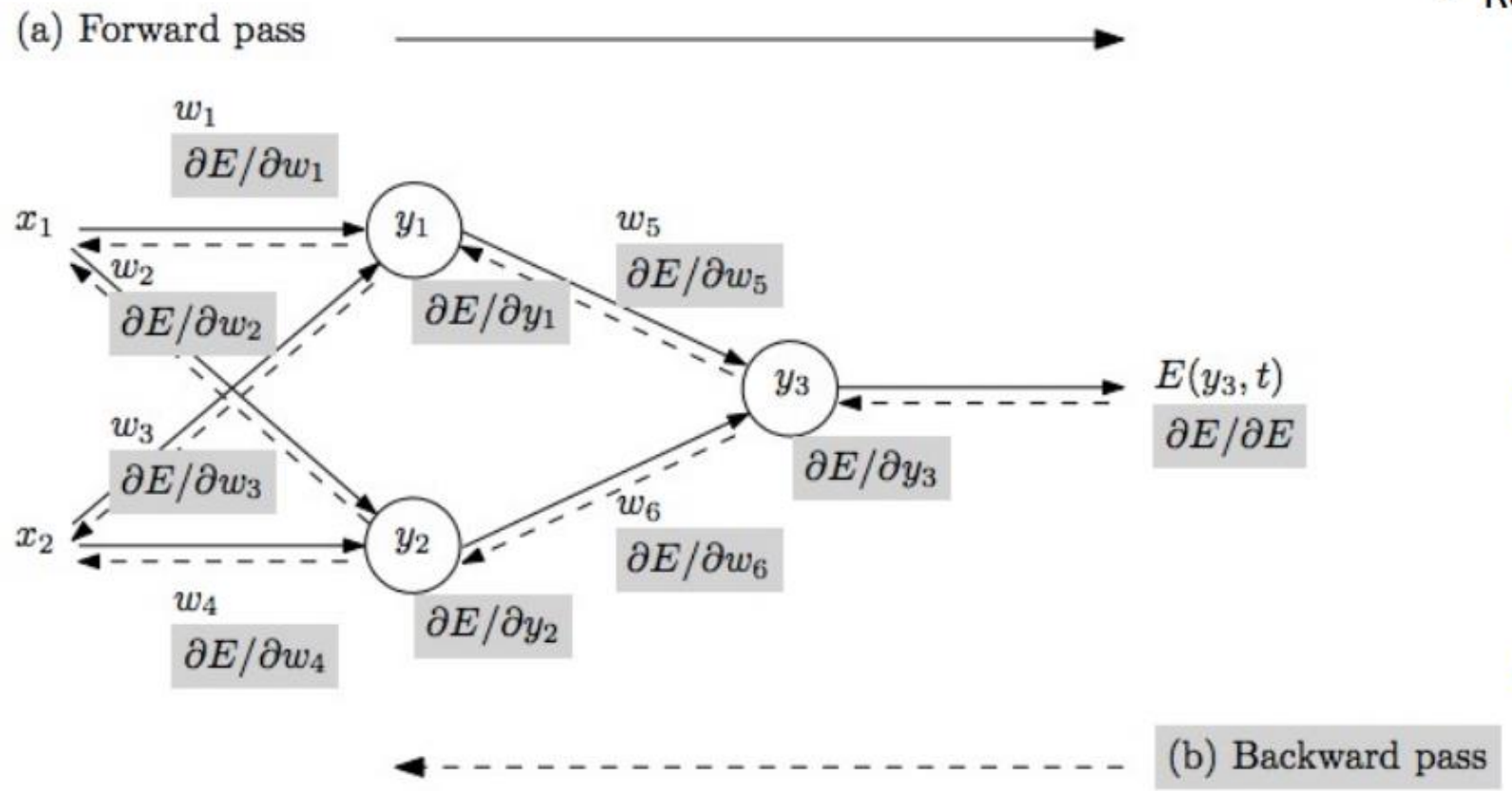
- 727 entries (plus an entry that famously was kicked out in 2015 for cheating from Baidu)
- Labor cost ~$110 million: assuming 3 people contribute to each entry and $50k cost per person

Secret Sauce for State-of-Art: Deeper CNNs



The Problem of the Vanishing Gradients

- Why It Is Difficult to Achieve Better Performance with CNNs That Are Deeper?



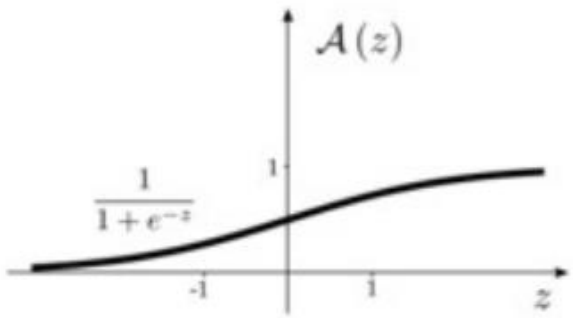
- Repeat until stopping criterion met:
  1. **Forward pass**: propagate training data through model to make prediction
  2. Quantify the dissatisfaction with a model's results on the training data
  3. **Backward pass**: using predicted output, calculate gradients backward to assign blame to each model parameter
  4. Update each parameter using calculated gradients

$$W_x = W_x - \alpha \left(\frac{\partial Error}{\partial W_x}\right)$$

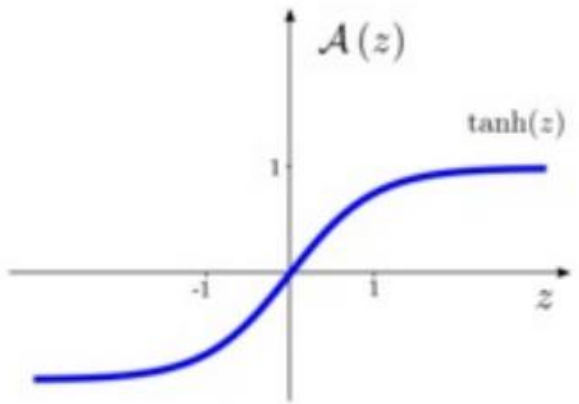
The Problem of the Vanishing Gradients

Recall activation functions and their derivatives:

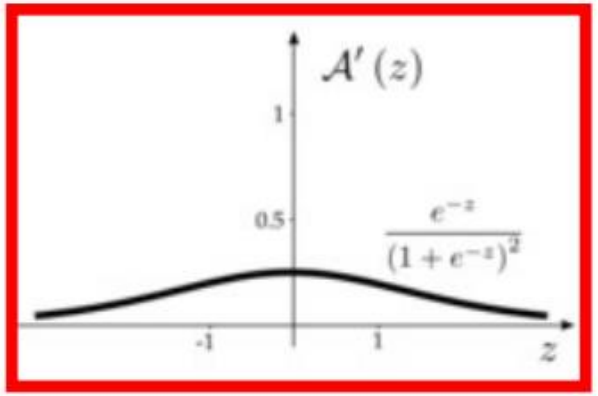
Sigmoid

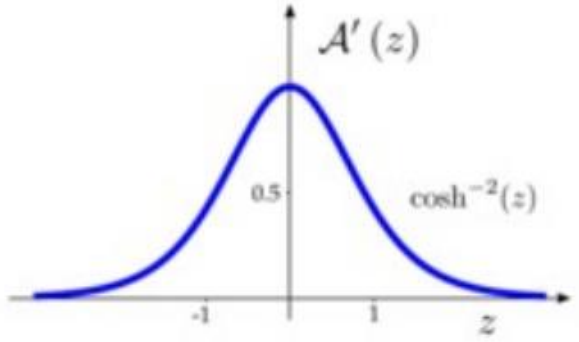


Tanh



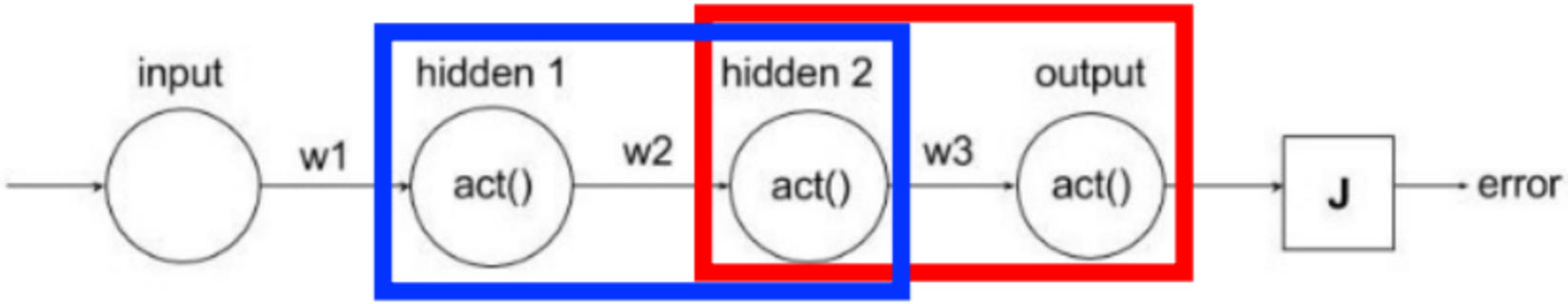
Ranges from 0 to 0.25





The Problem of the Vanishing Gradients

- Sigmoid Example
- Toy example:



- Error Derivative with respect to weight w1:

$$\frac{\partial error}{\partial w1} = \frac{\partial error}{\partial output} * \frac{\partial output}{\partial hidden2} * \frac{\partial hidden2}{\partial hidden1} * \frac{\partial hidden1}{\partial w1}$$

Derivative of sigmoid activation function: (0 to 1/4]

Derivative of sigmoid activation function: (0 to 1/4]

Problem: What happens as you multiply more numbers smaller than 1?

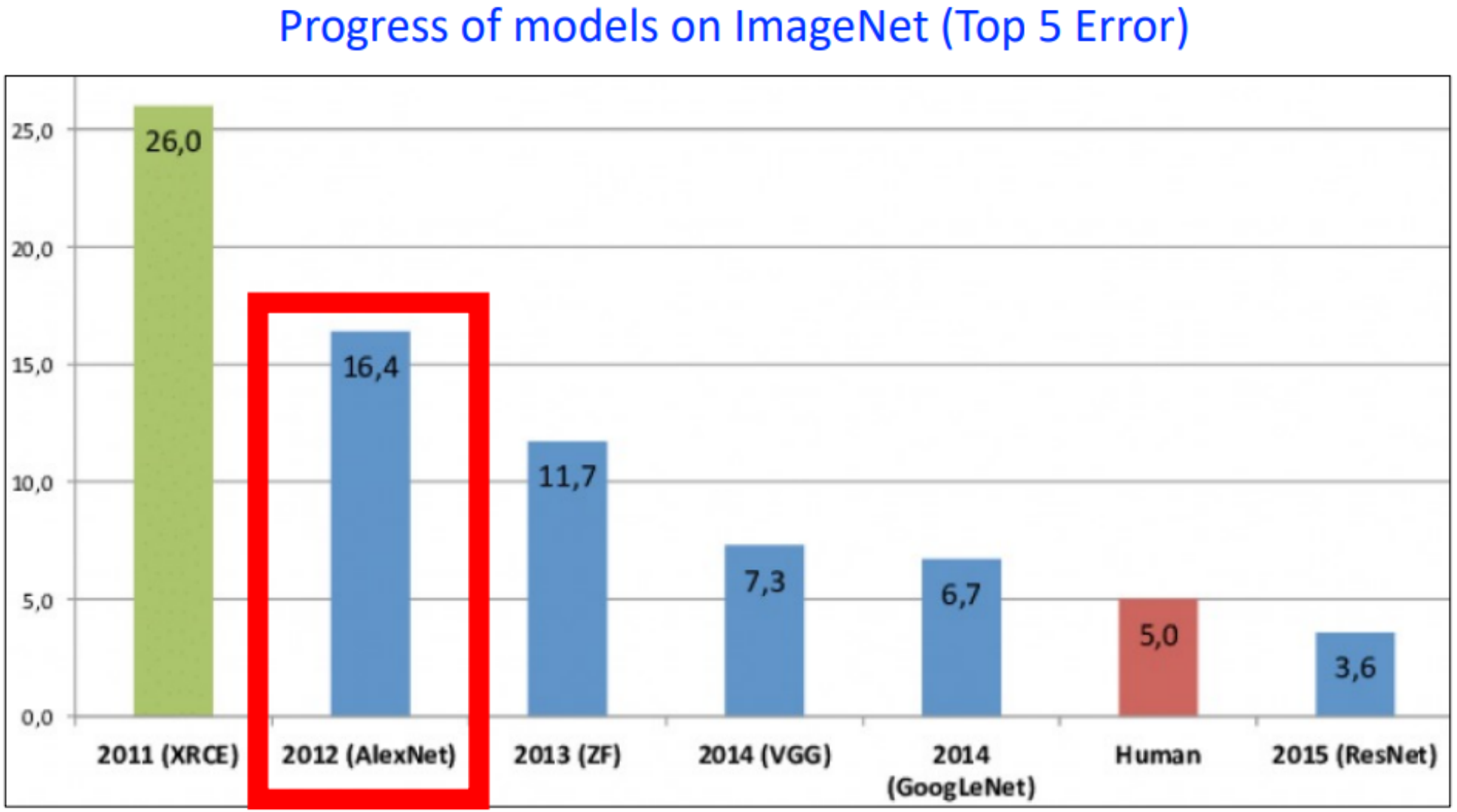
Gradient decreases as further from the last layer… and so weights barely change at training!

The Problem of the Vanishing Gradients

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

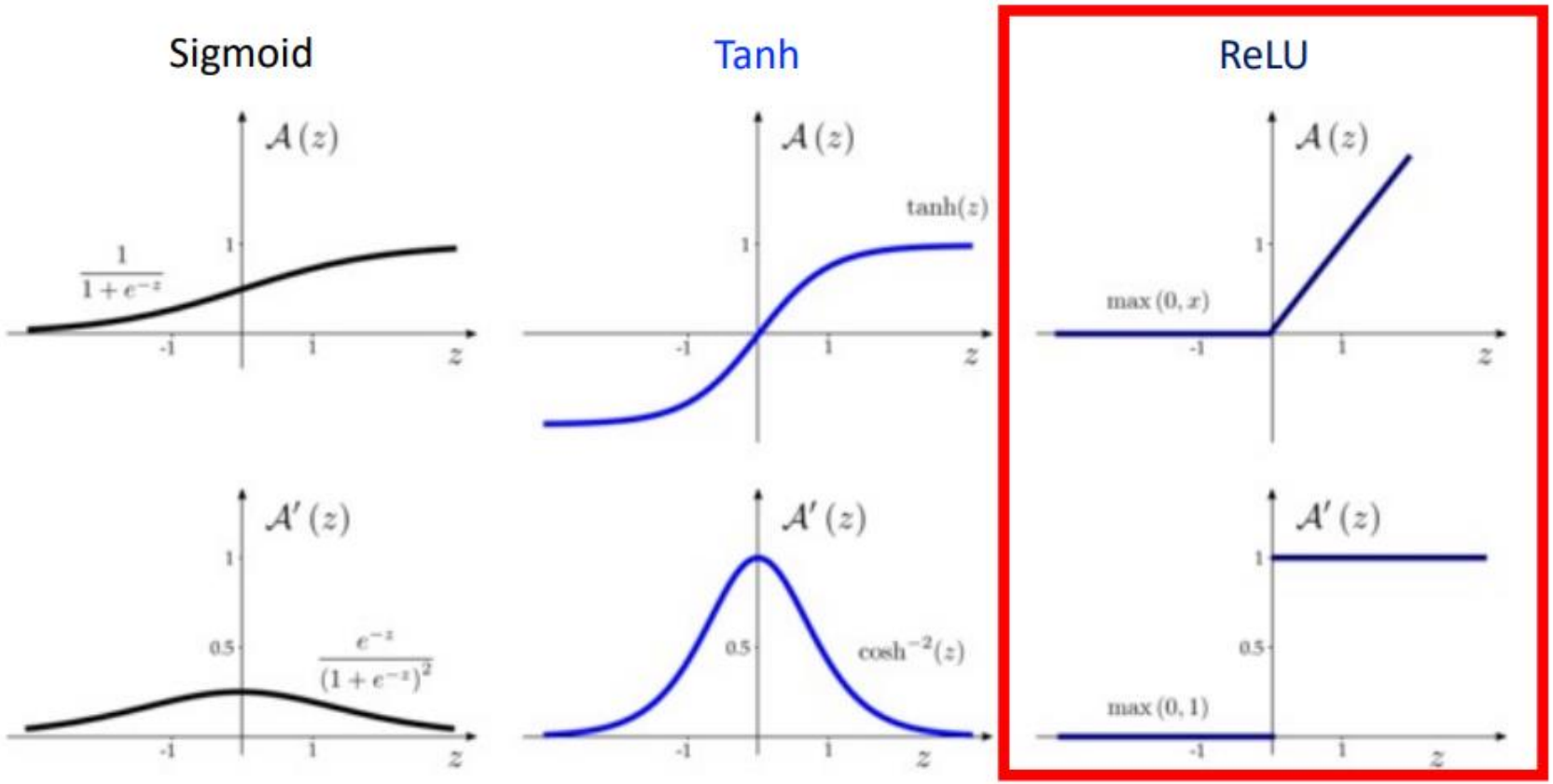
- **How can we avoid the vanishing gradient problem?**

AlexNet: A Deeper CNN



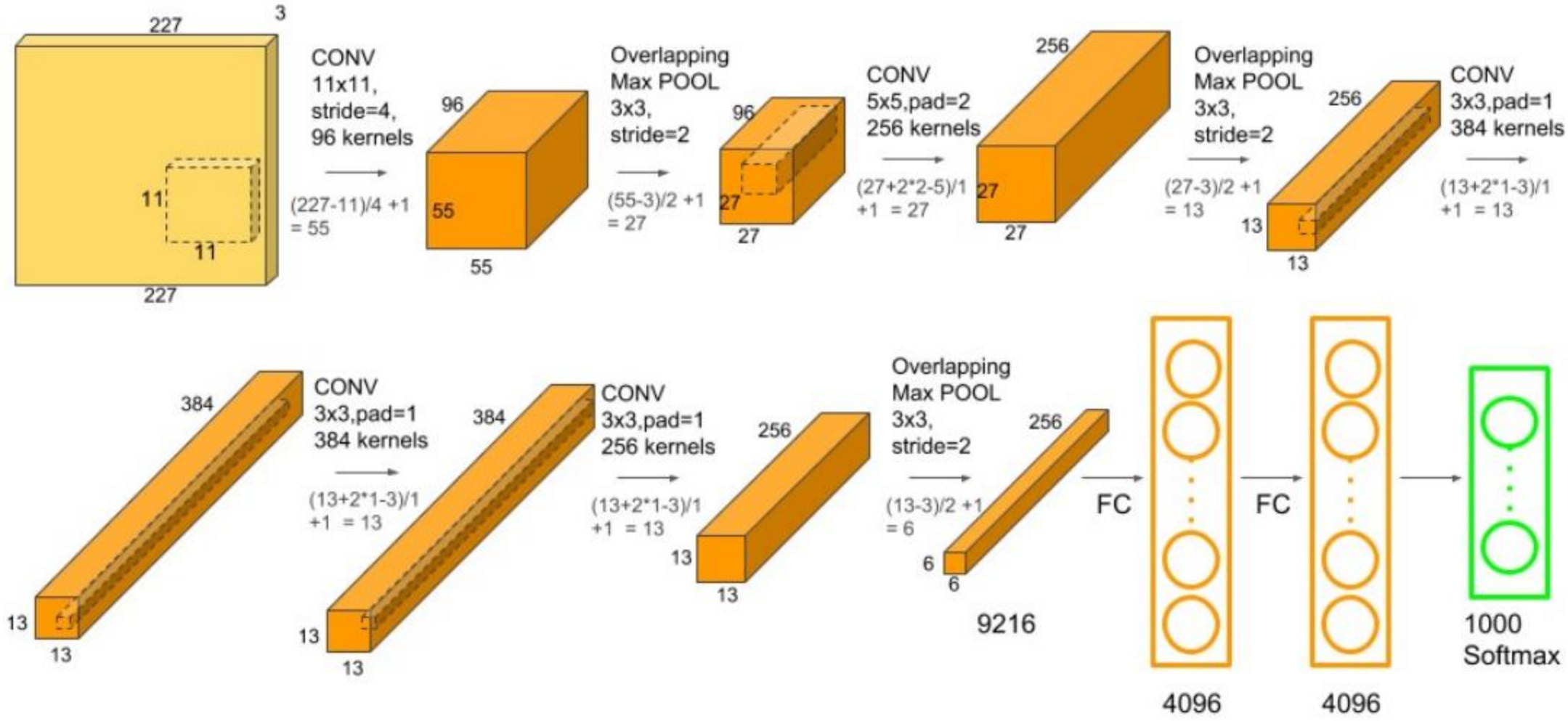
How can we avoid the vanishing gradients problem?

- Use non-saturating activation functions:
  - Use activation functions with derivative value equal to 1 (i.e., 1x1x1… doesn't vanish)



AlexNet Architecture

- Similar to LeNet But With More Convolutional and Pooling Layers



AlexNet Architecture

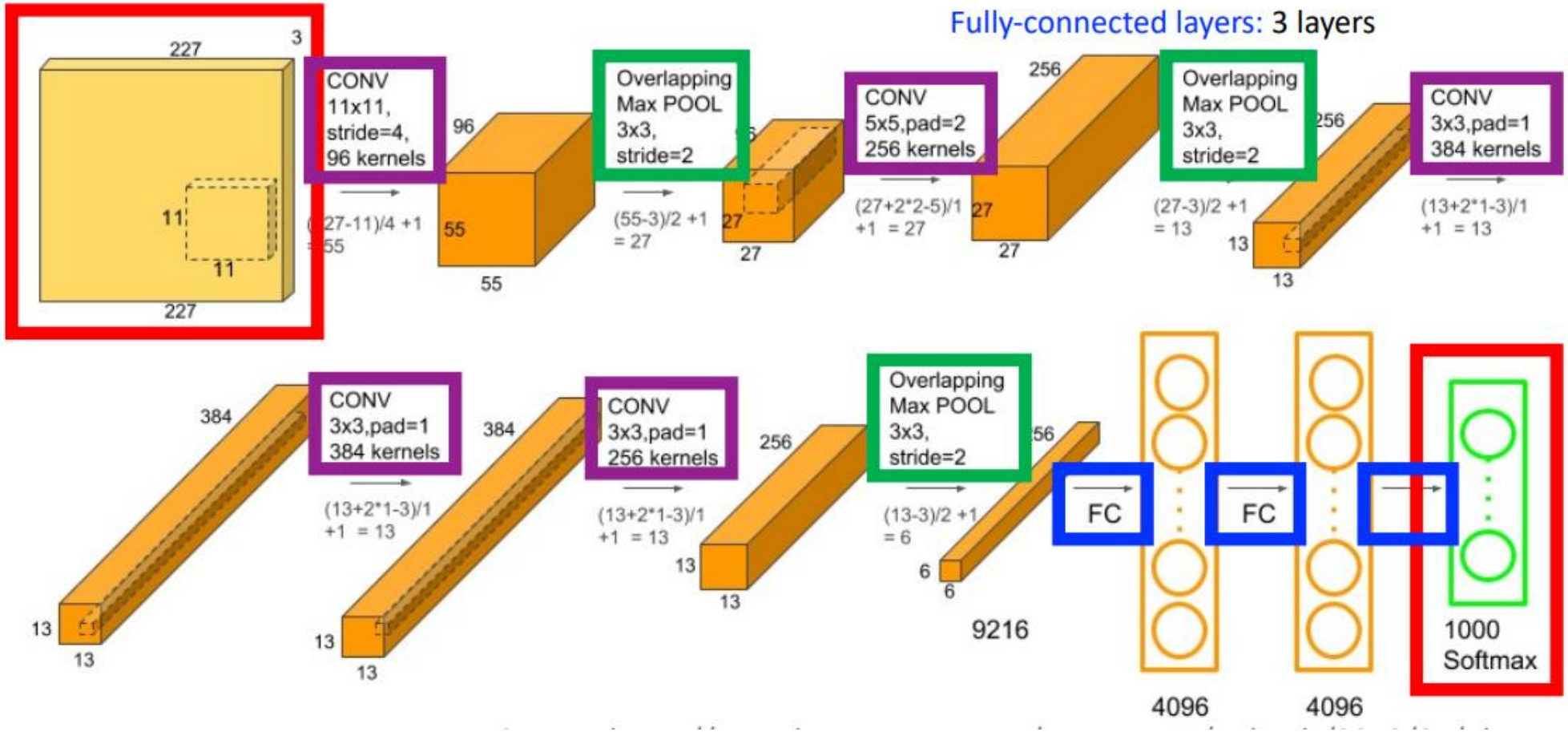
Input: RGB image resized to fixed input size

Output: 1000 class probabilities (sums to 1)

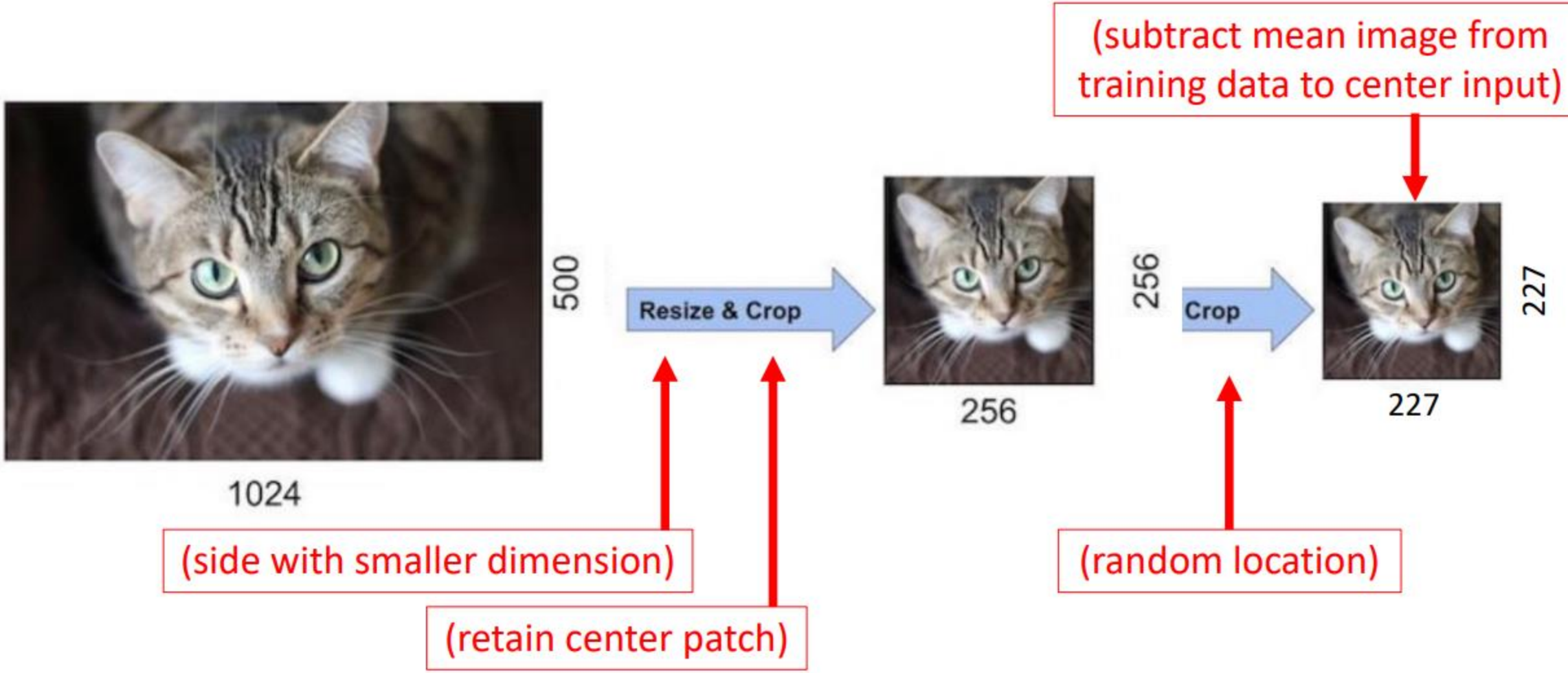
Convolutional layers: 5 layers

Pooling Layers: 3 layers

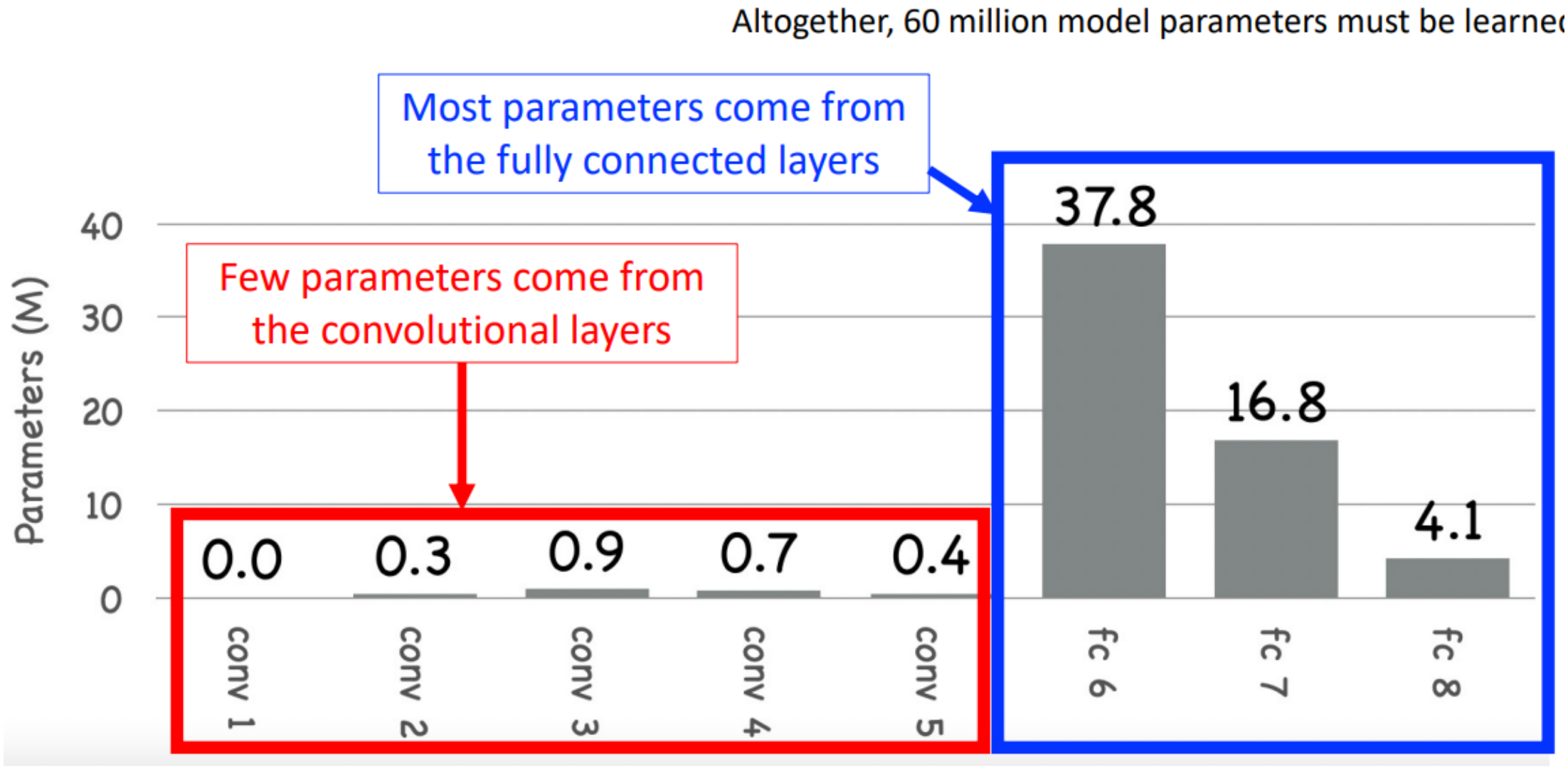
Fully-connected layers: 3 layers



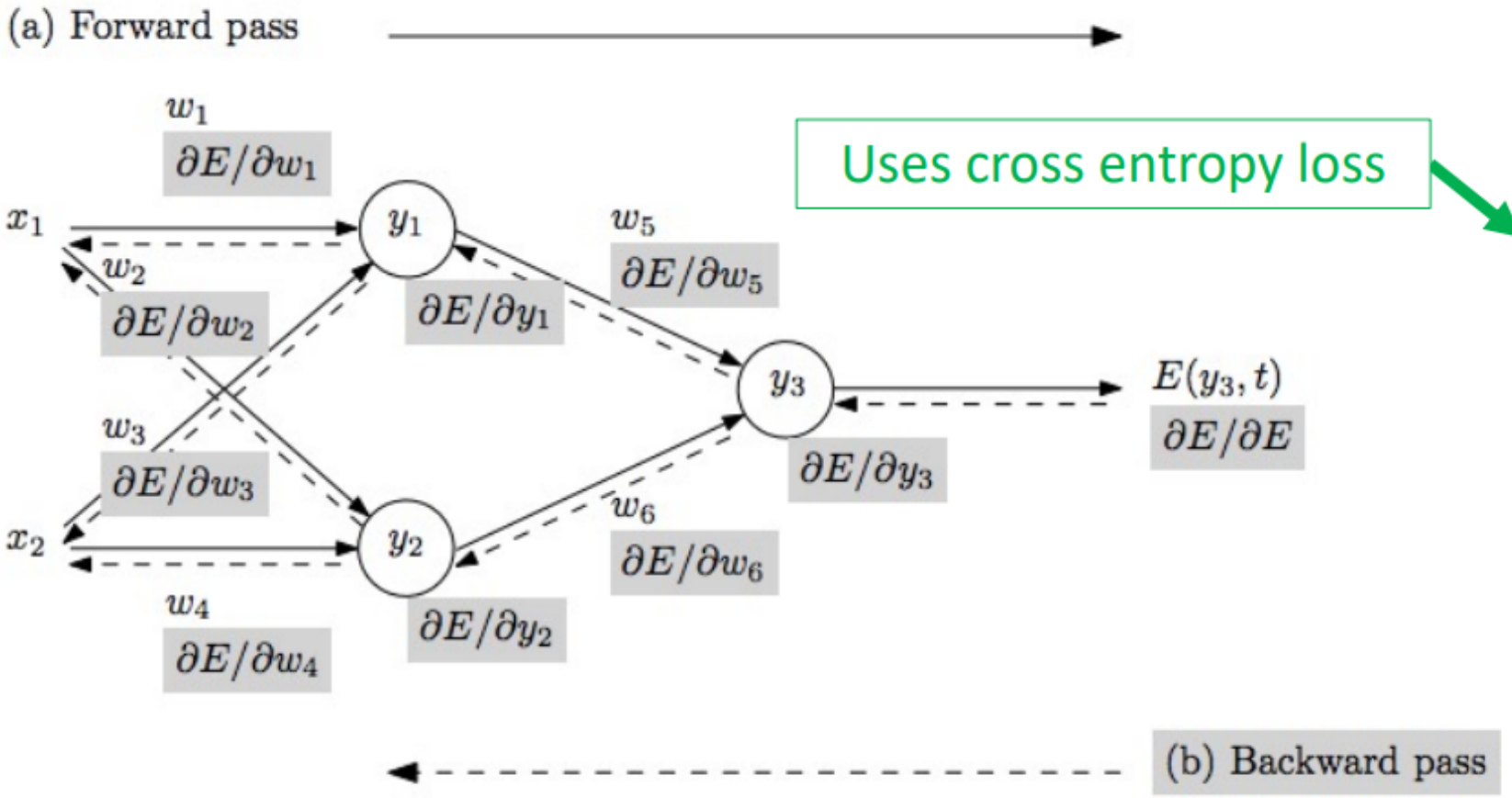
AlexNet: Input Preprocessing



AlexNet Architecture



AlexNet Training: 90 Epochs



Uses cross entropy loss

- Repeat until stopping criterion met:
  1. **Forward pass**: propagate training data through model to make prediction
  2. Quantify the dissatisfaction with a model's results on the training data
  3. **Backward pass**: using predicted output, calculate gradients backward to assign blame to each model parameter
  4. Update each parameter using calculated gradients

AlexNet: Key Tricks for Going Deeper

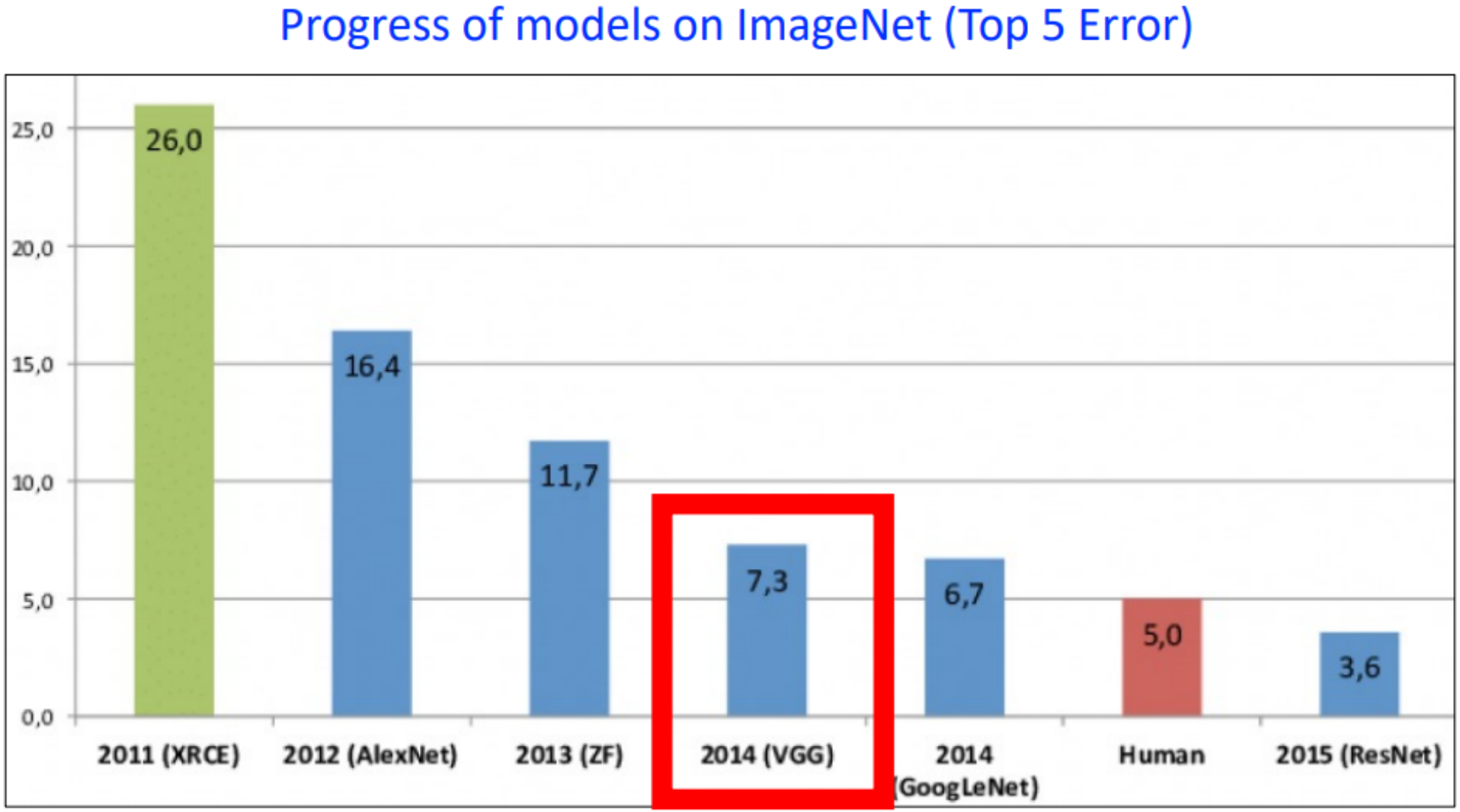
UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- ReLU instead of sigmoid or tanh activation functions

- Regularization techniques
  - Data augmentation
  - Dropout in fully connected layers
  - L2 parameter norm penalty

- Trained across 2 GPUs

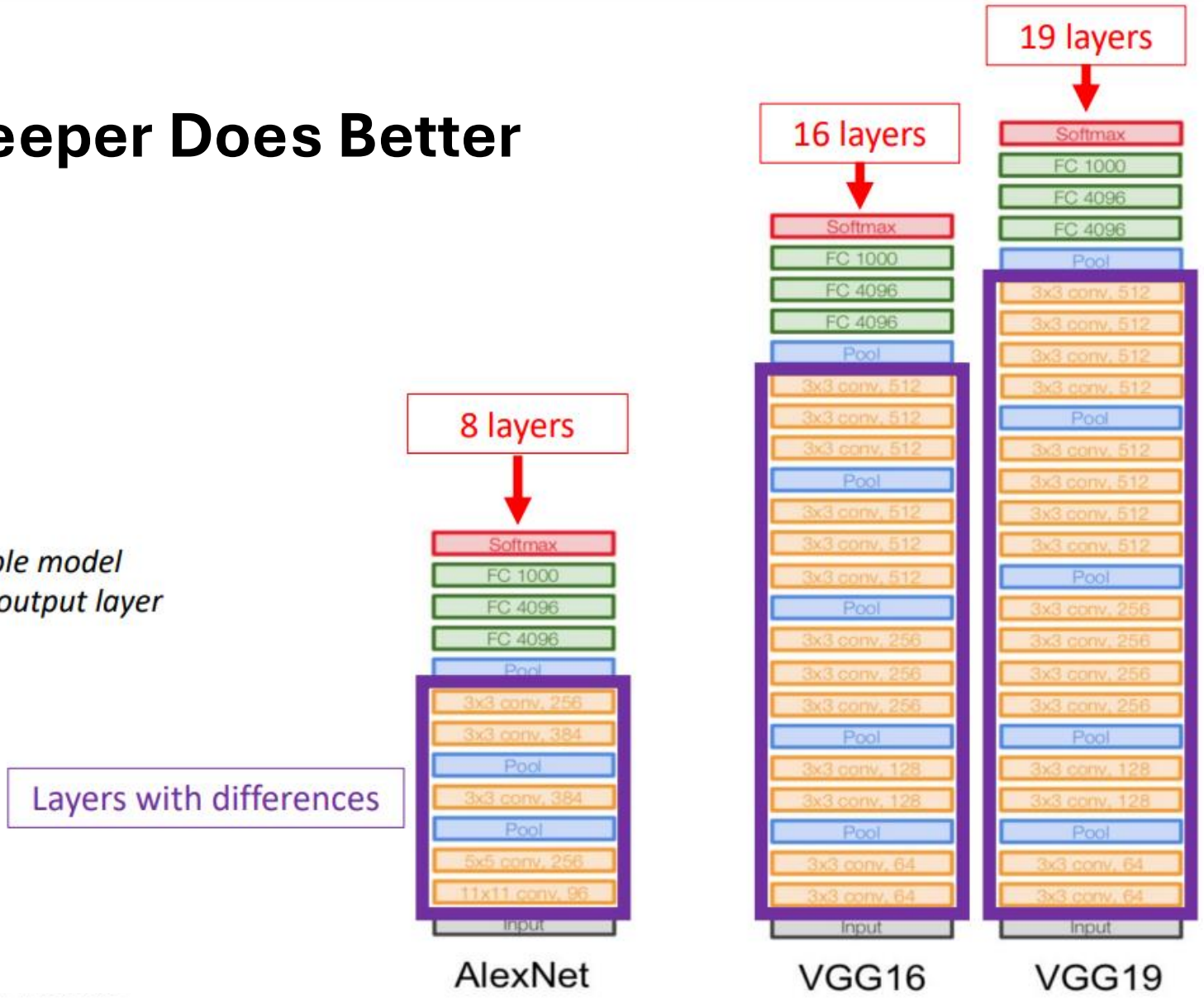
VGG: A Even Deeper CNN



VGG

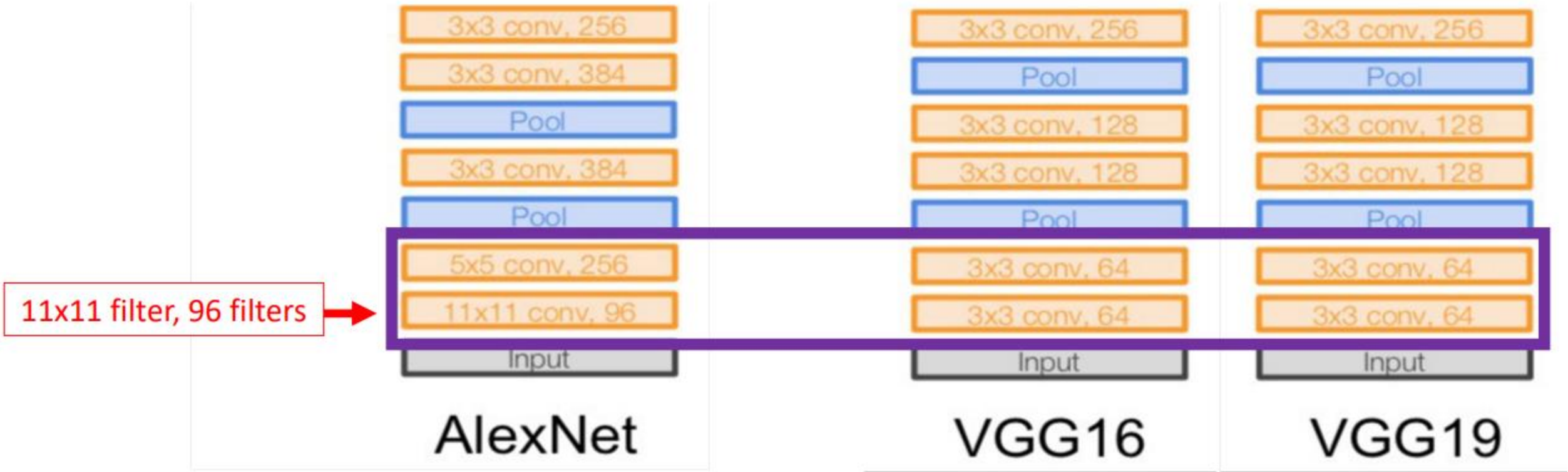
- Key Novelty: **Deeper Does Better**

** Number of layers with learnable model parameters between input and output layer (i.e., exclude pooling layers)*



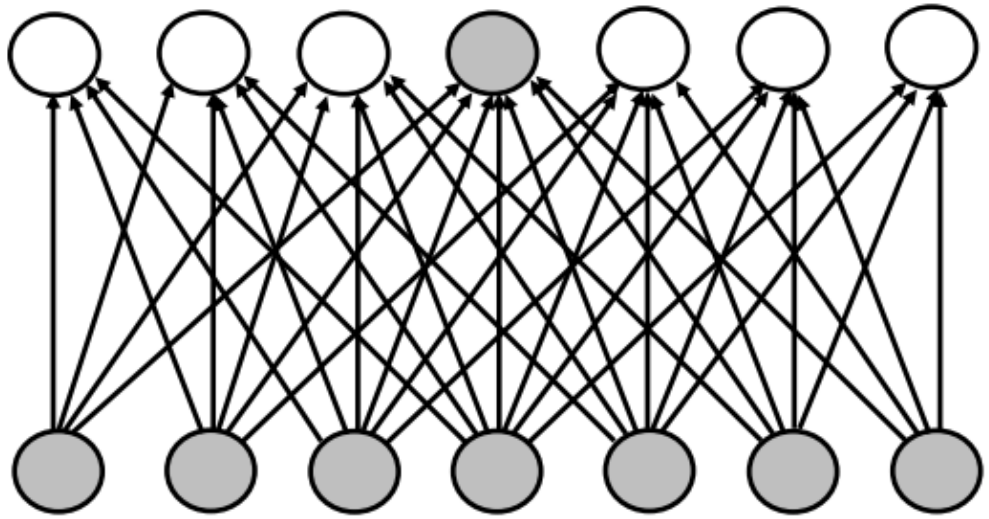
VGG

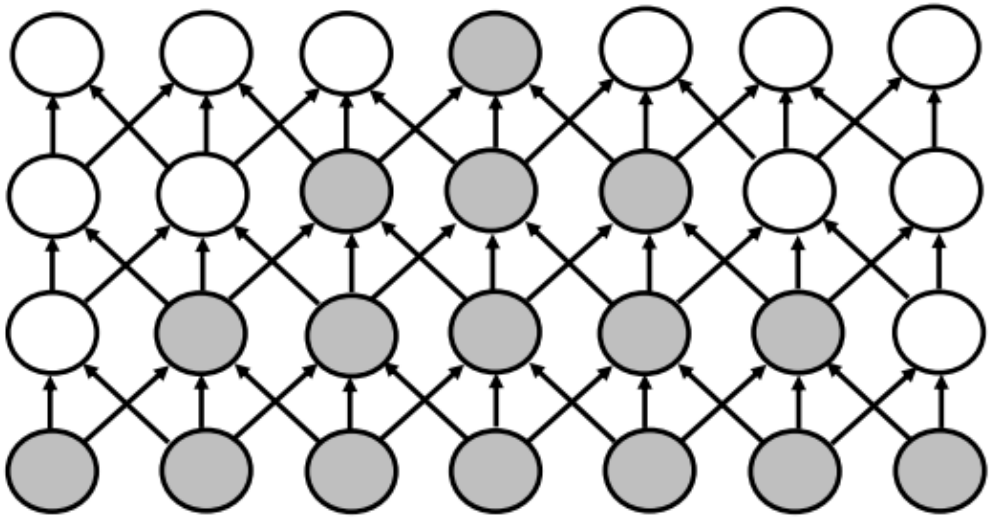
- Replace larger filter with stack of smaller filters



VGG

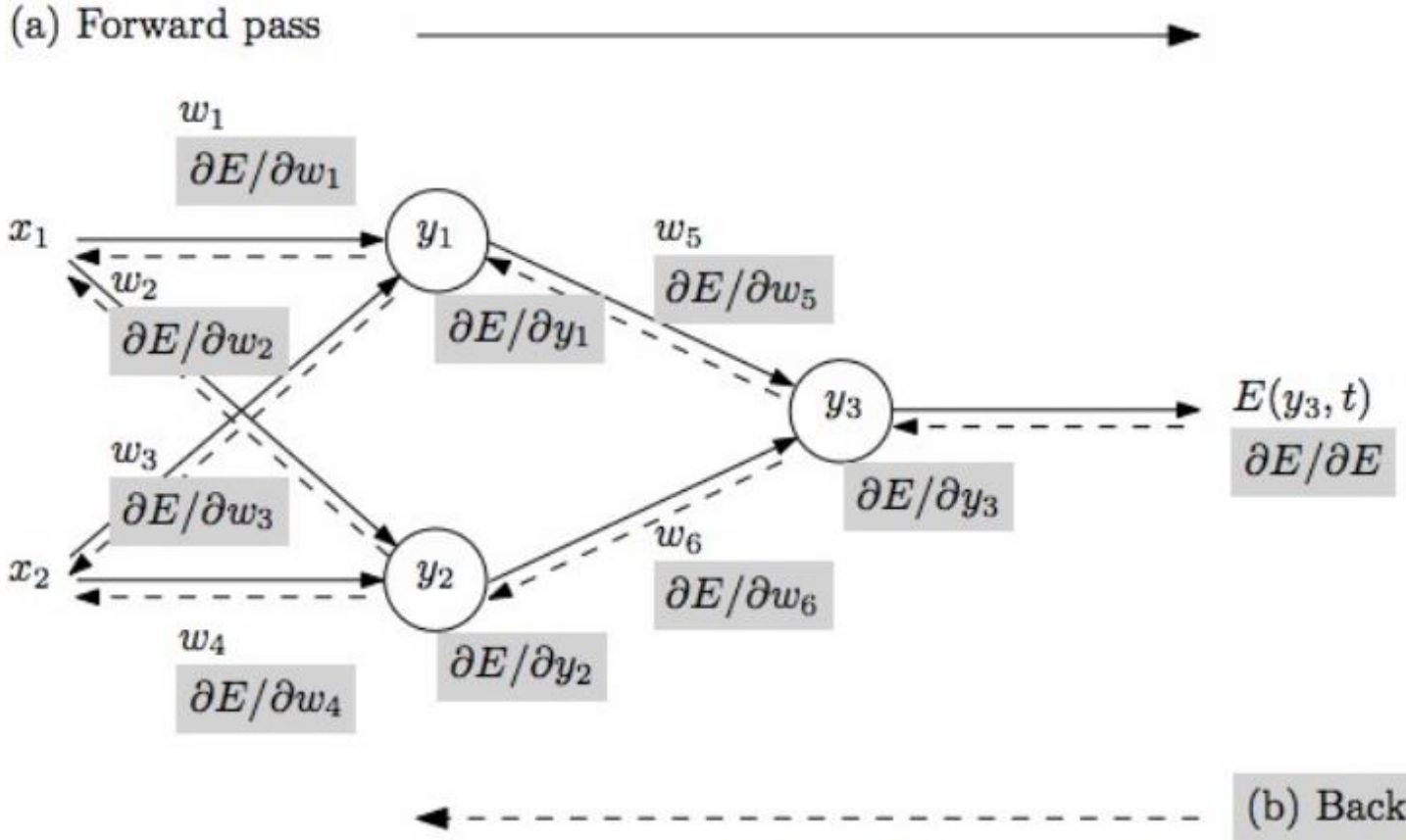
- Replace larger filter with stack of smaller filters; e.g., replace 7x7 with three 3x3s





- Benefits:
  - More discriminative classifier since more non-linear rectifications: 3 vs 1
  - Reduces # of parameters: multiple of 27 (3 x 32) parameters vs 49 (7 x 7) parameters

VGG Training (Similar to AlexNet): 74 Epochs



- Repeat until stopping criterion met:
  1. **Forward pass**: propagate training data through model to make prediction
  2. Quantify the dissatisfaction with a model's results on the training data
  3. **Backward pass**: using predicted output, calculate gradients backward to assign blame to each model parameter
  4. Update each parameter using calculated gradients

Can still encounter vanishing gradients

$$W_x = W_x - \alpha \left(\frac{\partial Error}{\partial W_x}\right)$$

Weight Initialization

- **Problem:** vanishing gradients can occur since derivative expression contains weight multiplication, and weights are generally initialized to < 1

- **Solution:** Strategic initialization of weights
  - Methods like **Xavier** or **He initialization** scale weights based on layer size.
  - This keeps activations and gradients stable across layers.

Xavier Initialization

- For a given layer with n_{in} input units and n_{out} output units:

- Weights are drawn from a uniform distribution between:

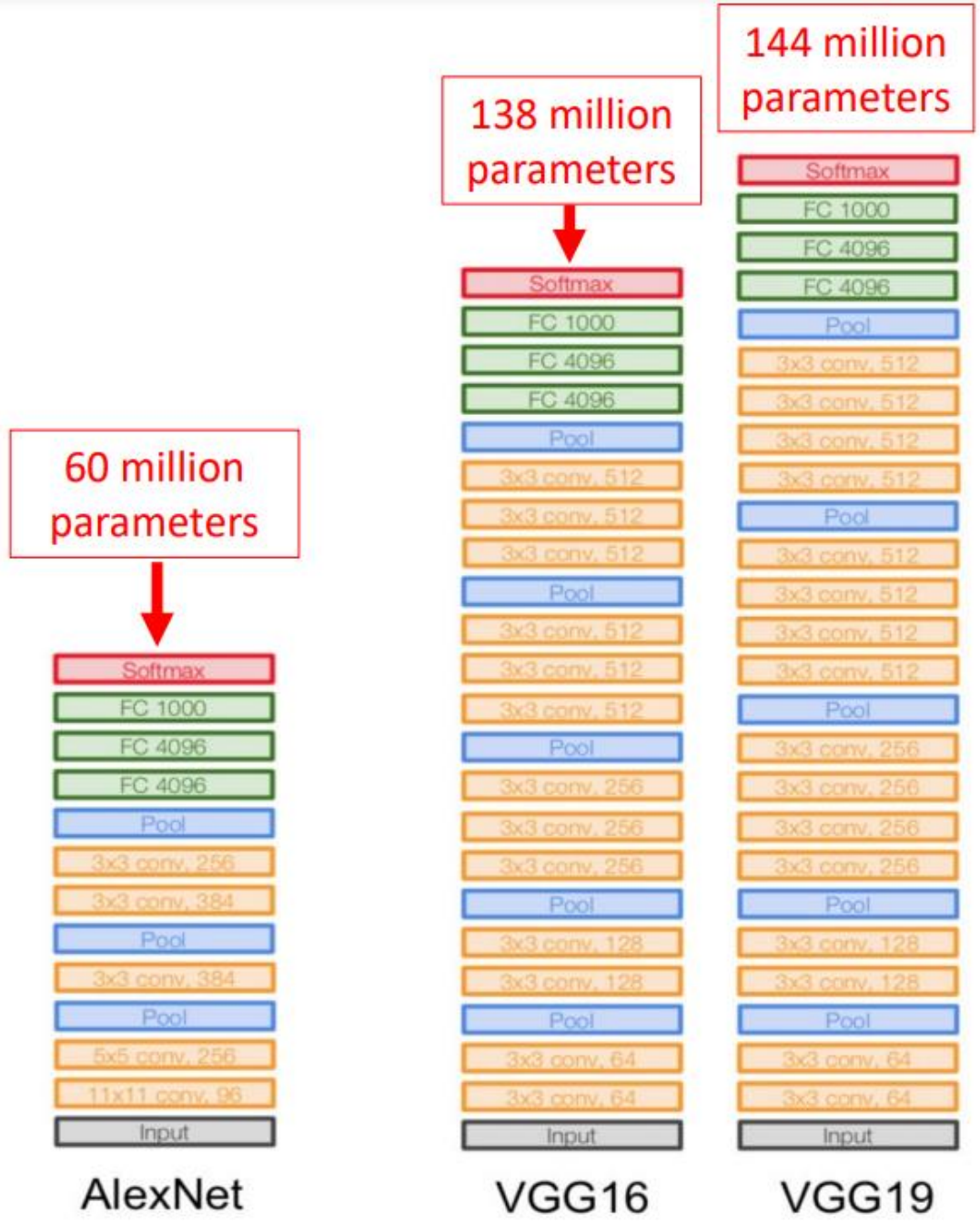
$$W \sim \text{Uniform}\left(-\sqrt{\frac{6}{n_{\text{in}} + n_{\text{out}}}}, \sqrt{\frac{6}{n_{\text{in}} + n_{\text{out}}}}\right)$$

- Alternatively, weights can be drawn from a normal distribution with:

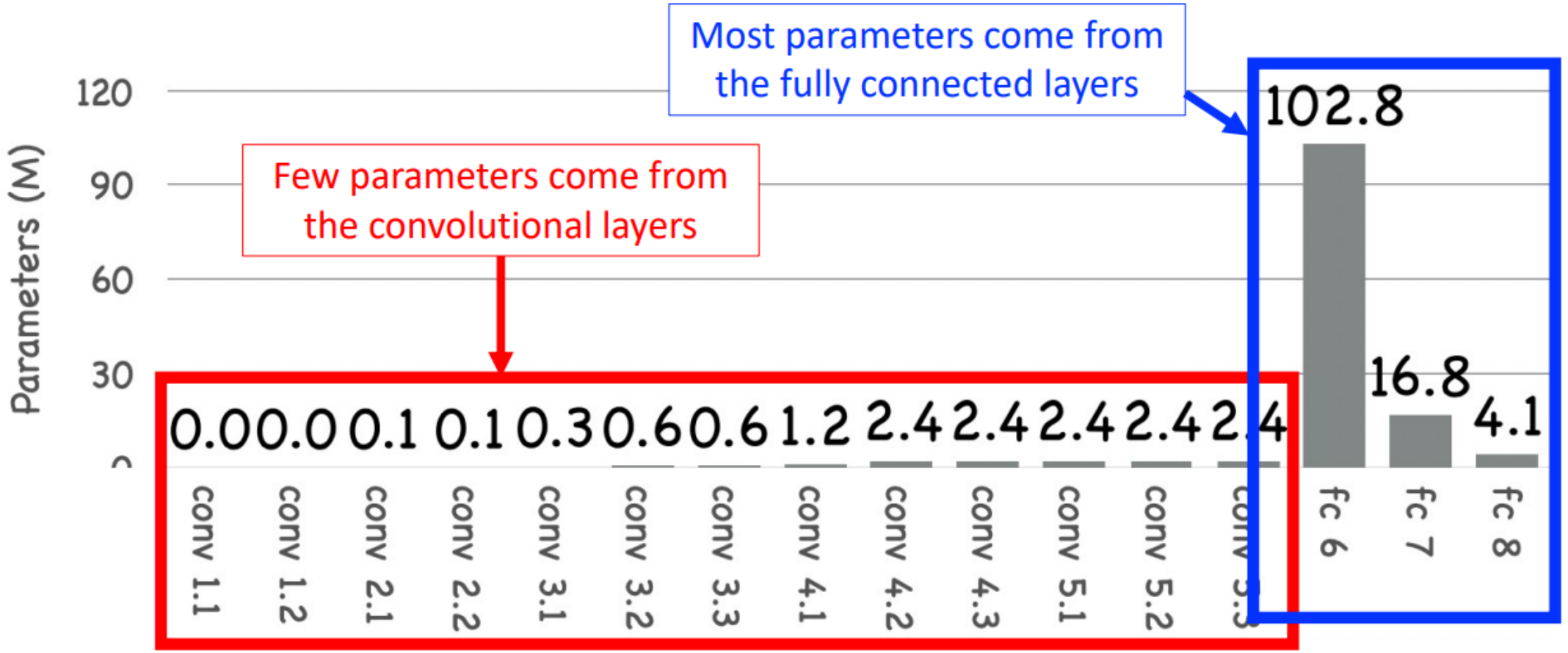
$$W \sim \text{Normal}\left(0, \sqrt{\frac{2}{n_{\text{in}} + n_{\text{out}}}}\right)$$

VGG Limitation

- Models are large!



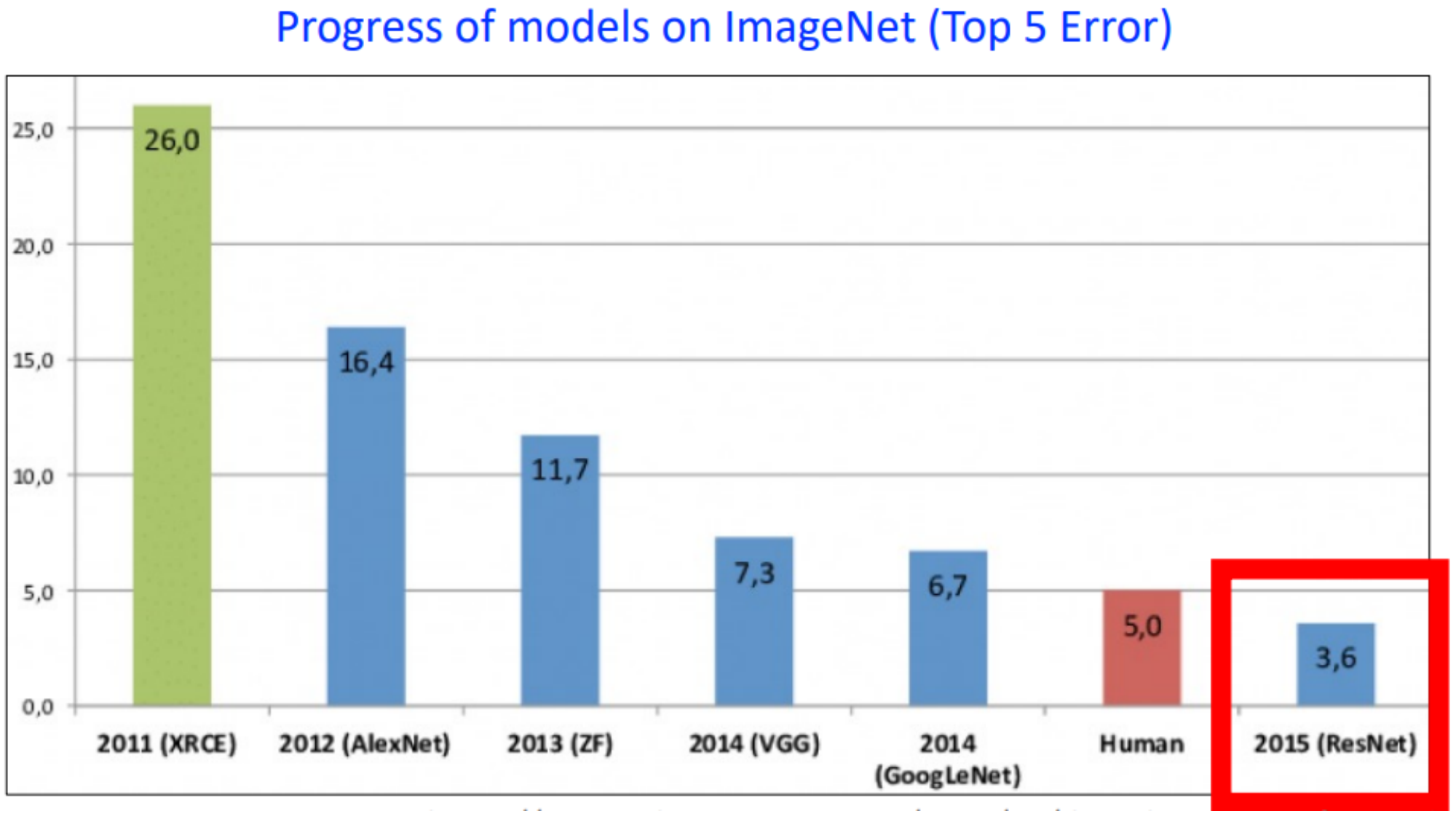
VGG Limitation



VGG: Key Tricks for Going Deeper

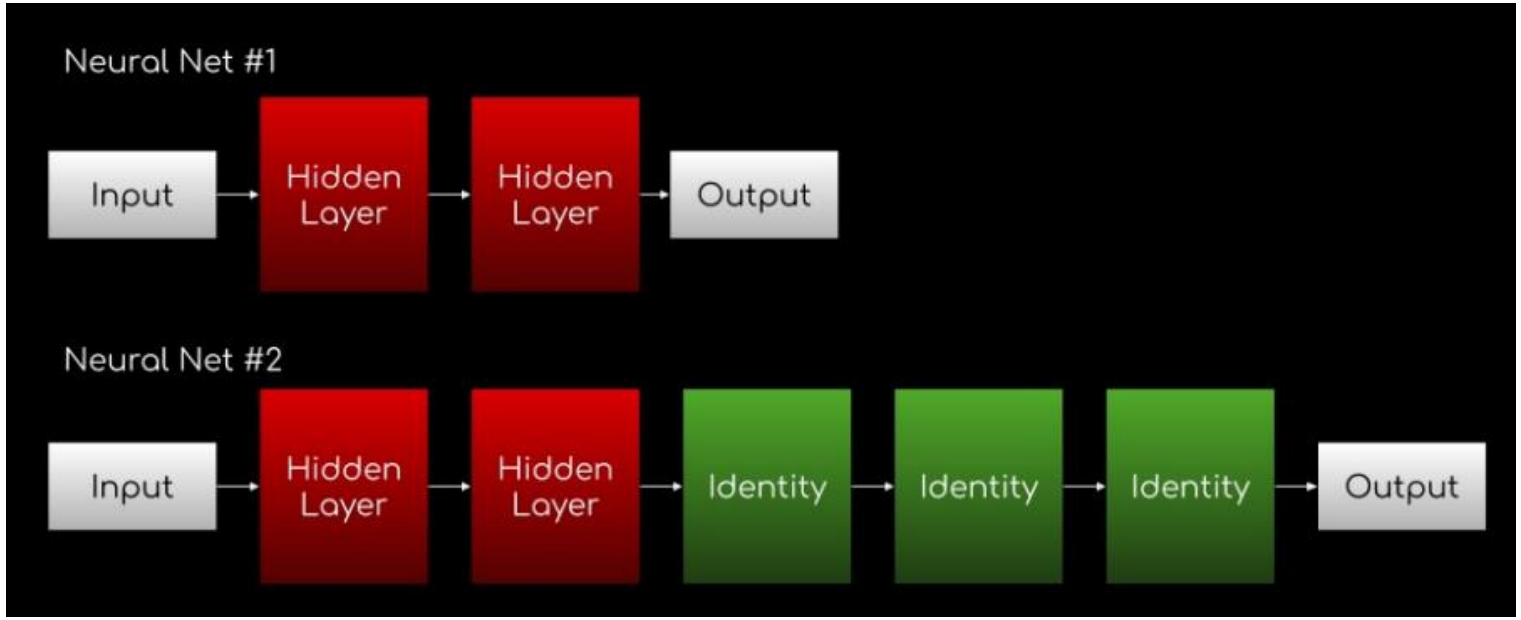
- 3x3 filters instead of larger filters
- Weight initialization with Xavier Glorot procedure
- Regularization techniques
  - Data augmentation
  - Dropout in fully connected layers
  - L2 parameter norm penalty
- Trained across multiple GPUs

ResNet: A Even Even Deeper CNN



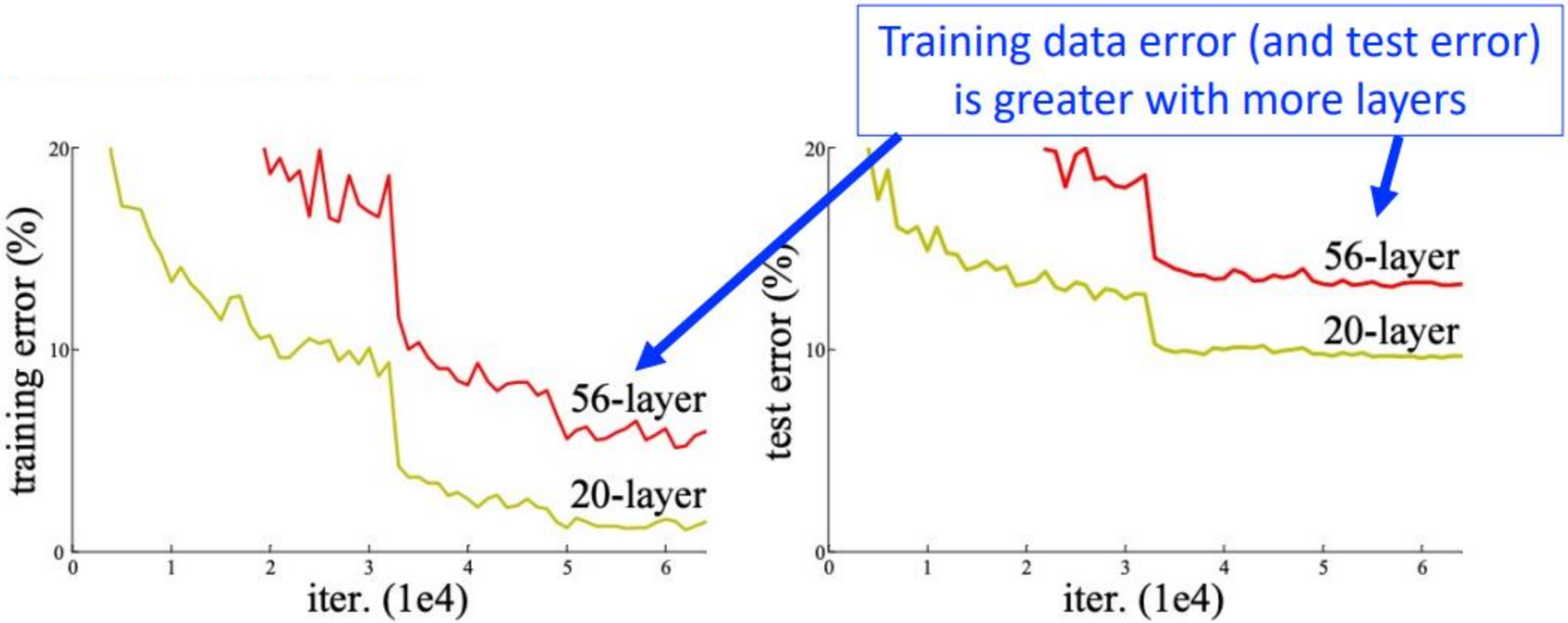
Motivation for ResNet

- **Idea:** a deeper network should perform as good if not better than shallower networks since they can learn the shallower function by simply learning "identity" functions for later layers
- **Observation:** adding more layers leads to WORSE results!
- Is the problem overfitting?



Motivation for ResNet

- Is the problem overfitting? **NO**



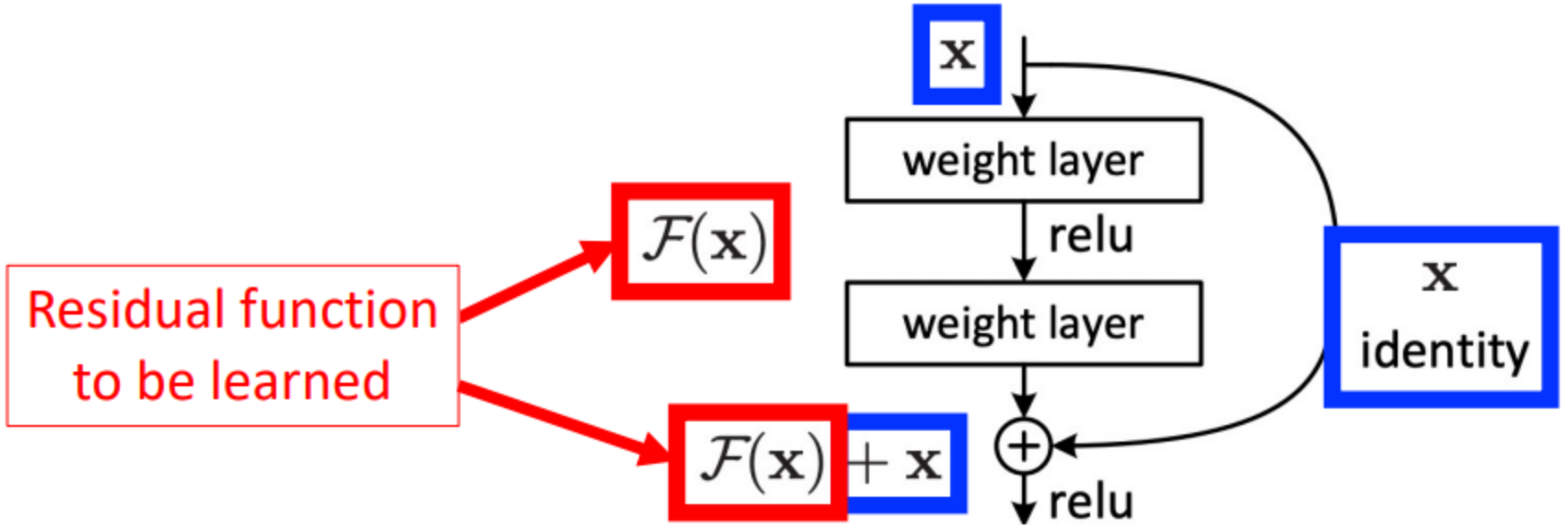
Motivation for ResNet

- **Idea:** a deeper network should perform as good if not better than shallower networks since they can learn the shallower function by simply learning "identity" functions for later layers
- **Observation:** adding more layers leads to WORSE results!
- Is the problem overfitting?

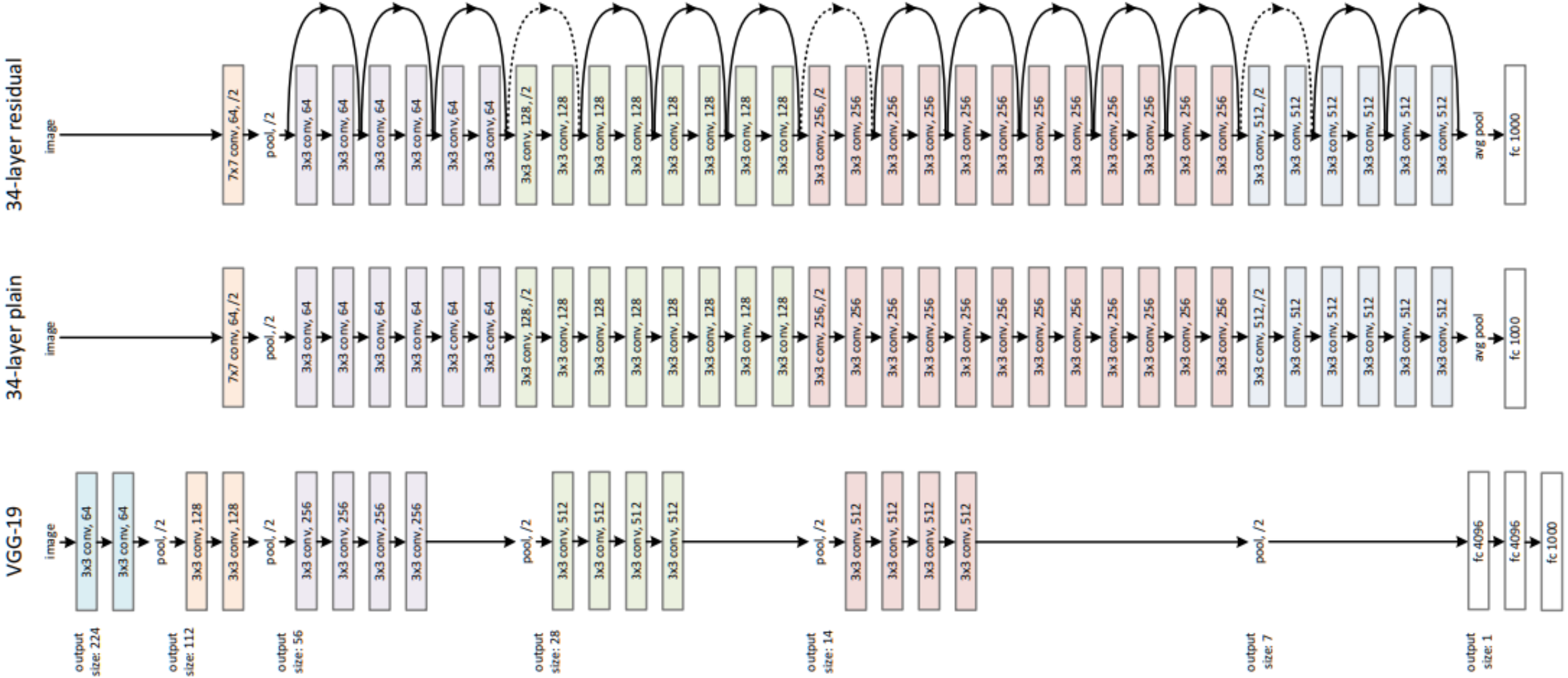
- **Problem: It is difficult to learn for the algorithm to learn layers of identity mappings**

ResNet: Skip Connections

- Skip Connection allows the model to learn the identity functions which ensures the higher layer will perform at least as good as the lower layer, and not worse.



ResNet


34-layer residual
image
7x7 conv, 64, /2
pool, /2
3x3 conv, 64
3x3 conv, 128
3x3 conv, 256
3x3 conv, 512
avg pool
fc 1000
34-layer plain
image
7x7 conv, 64, /2
pool, /2
3x3 conv, 64
3x3 conv, 128
3x3 conv, 256
3x3 conv, 512
avg pool
fc 1000
VGG-19
image
3x3 conv, 64
pool, /2
3x3 conv, 128
3x3 conv, 256
3x3 conv, 512
fc 4096
fc 1000
output size: 224
output size: 112
output size: 56
output size: 28
output size: 14
output size: 7
output size: 1

Experimental Results on Validation Set

model	top-1 err.	top-5 err.
VGG-16 [40]	28.07	9.33
GoogLeNet [43]	-	9.15
PReLU-net [12]	24.27	7.38
ResNet-50	22.85	6.71
ResNet-101	21.75	6.05
ResNet-152	**21.43**	**5.71**

Performance improves with more layers

ResNet: Key Tricks for Going Deeper

- **Skip Connections**

Deeper Model Perform Better