Introduction to Convolutional Neural Networks

Ing. Zese Riccardo
riccardo.zese@unife.it

Regione Emilia-Romagna

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

FERRARIAE UNIVERSITAS · EX LABORE FRUCTUS · 13 91

Università degli Studi di Ferrara

Outline

- Introduction to Python
- Introduction to Neural Networks
- Convolutional NN
- Recurrent NN
- Autoencoders and self supervised learning

Outline

- Introduction to Python
- Introduction to Neural Networks
- Convolutional NN
- Recurrent NN
- Autoencoders and self supervised learning

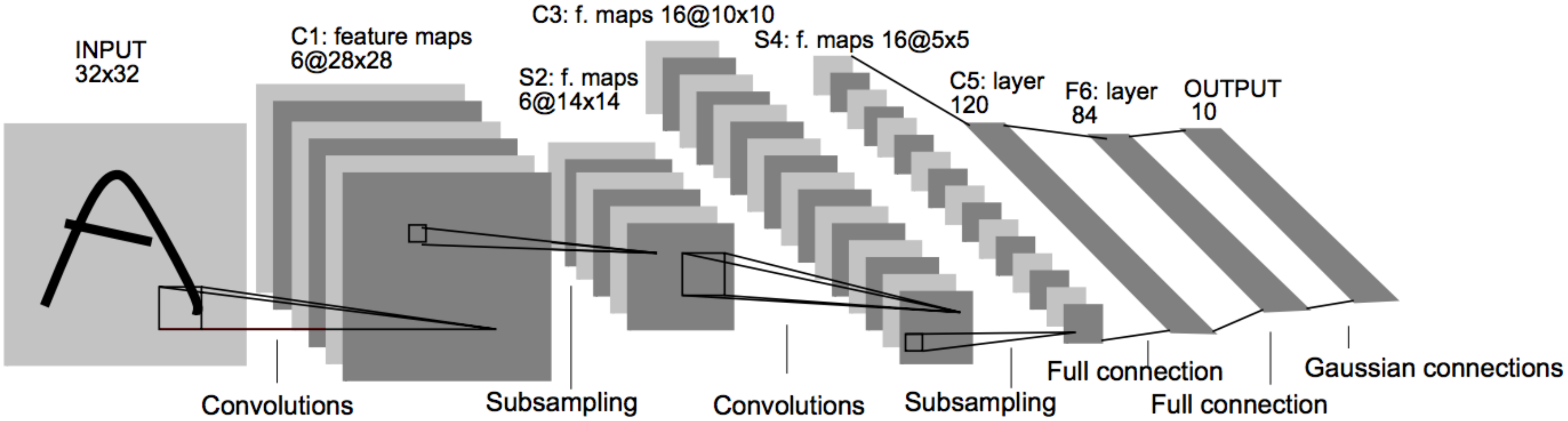
Regione Emilia-Romagna

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Università degli Studi di Ferrara

LeNet 5

- Created by Yann LeCun in the 1990s
- Used on the MNIST data set.
- Novel Idea: Use convolutions to efficiently learn features on data set.



MNIST

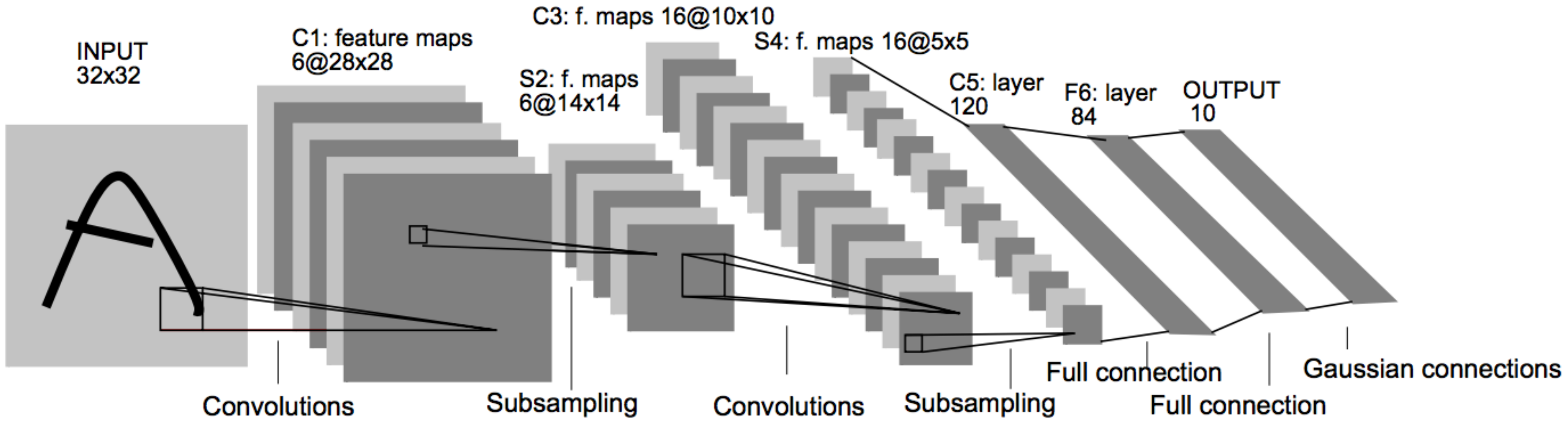
- The MNIST database contains examples of handwritten digits
- Has a training set of 60,000 examples, and a test set of 10,000 examples.
- It is a subset of a larger set available from NIST.
- The original black and white (bilevel) images from NIST were size normalized to fit in a 20x20 pixel box while preserving their aspect ratio. The resulting images contain grey levels as a result of the anti-aliasing technique used by the normalization algorithm. The images were centered in a 28x28 image by computing the center of mass of the pixels, and translating the image so as to position this point at the center of the 28x28 field.

Regione Emilia-Romagna

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

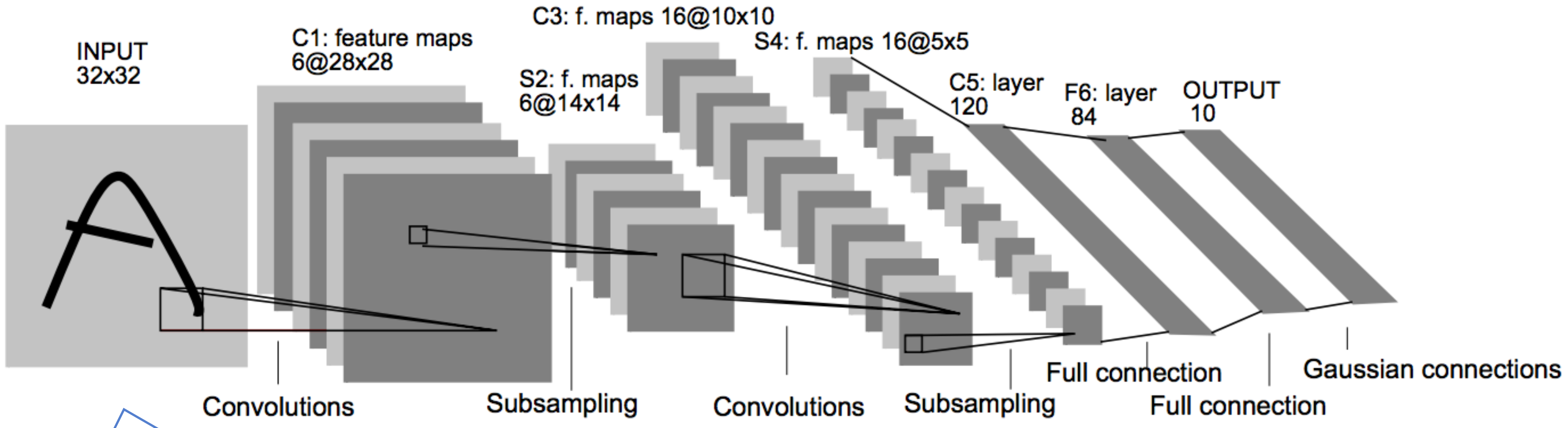
Università degli Studi di Ferrara

LeNet 5



INPUT
32x32
C1: feature maps
6@28x28
S2: f. maps
6@14x14
C3: f. maps 16@10x10
S4: f. maps 16@5x5
C5: layer
120
F6: layer
84
OUTPUT
10
Convolutions
Subsampling
Convolutions
Subsampling
Full connection
Full connection
Gaussian connections

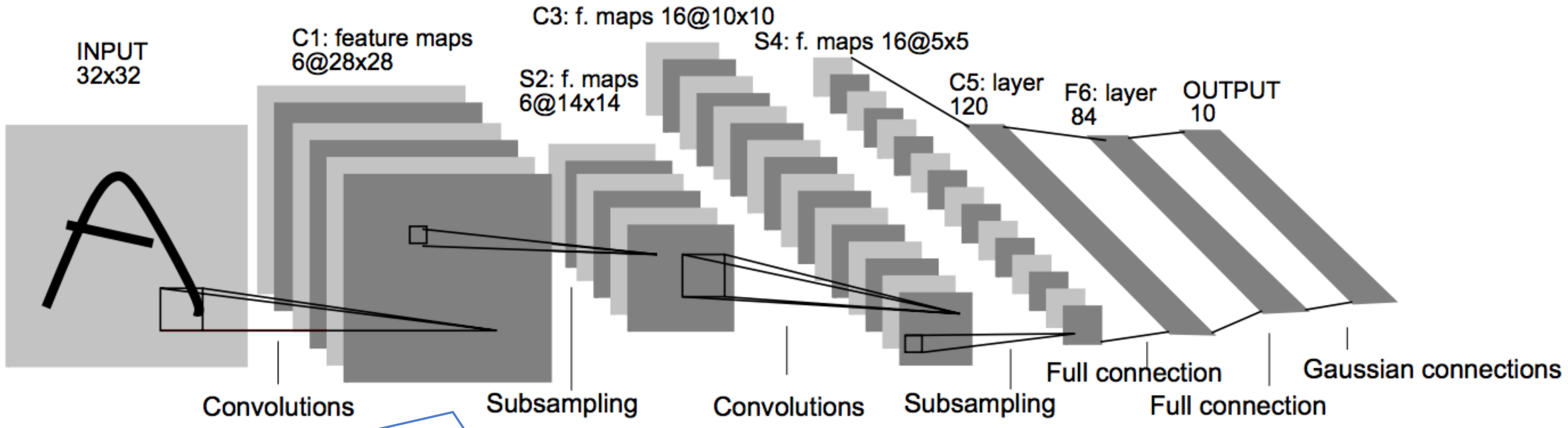
LeNet 5



Input: A 28x28 grayscale image with 2 pixels of padding all around → 32x32

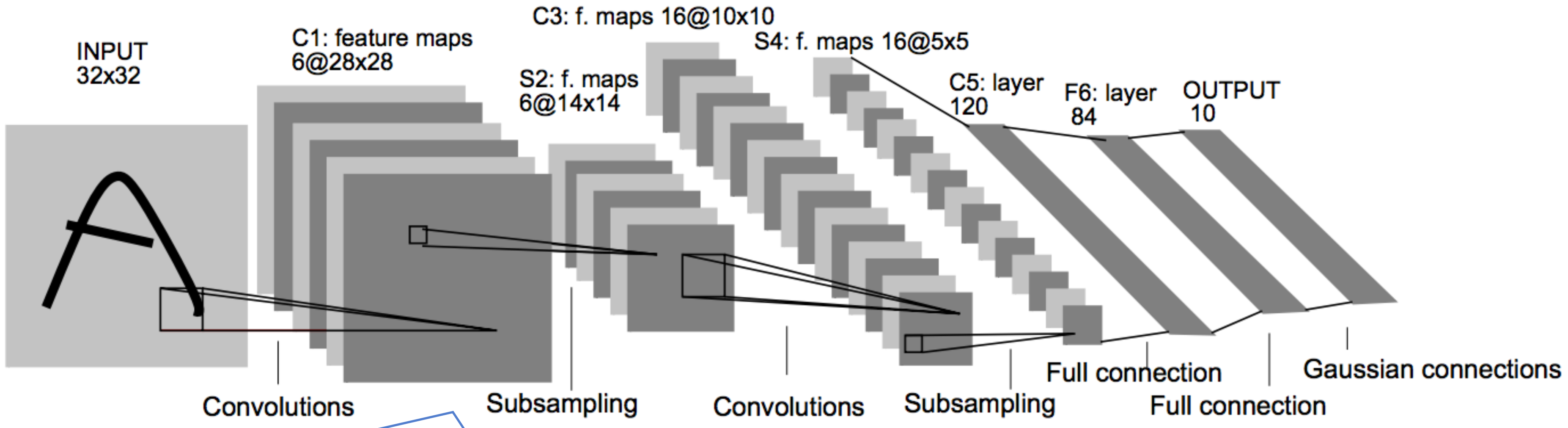
Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

LeNet 5



Convolutional layer with 5x5 kernels and stride 1 → 28x28 (padding removed).
Depth of 6 → 6 different kernels are learned.

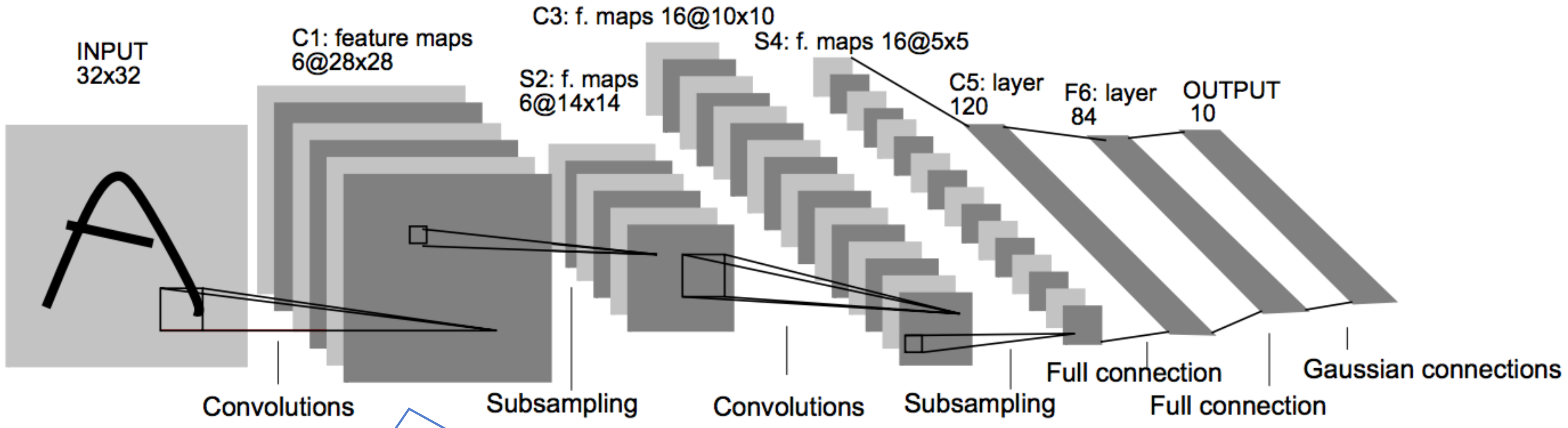
LeNet 5



Convolutional layer with 5x5 kernels and stride 1 → 28x28 (padding removed).
Depth of 6 → 6 different kernels are learned.

Output of 6x28x28

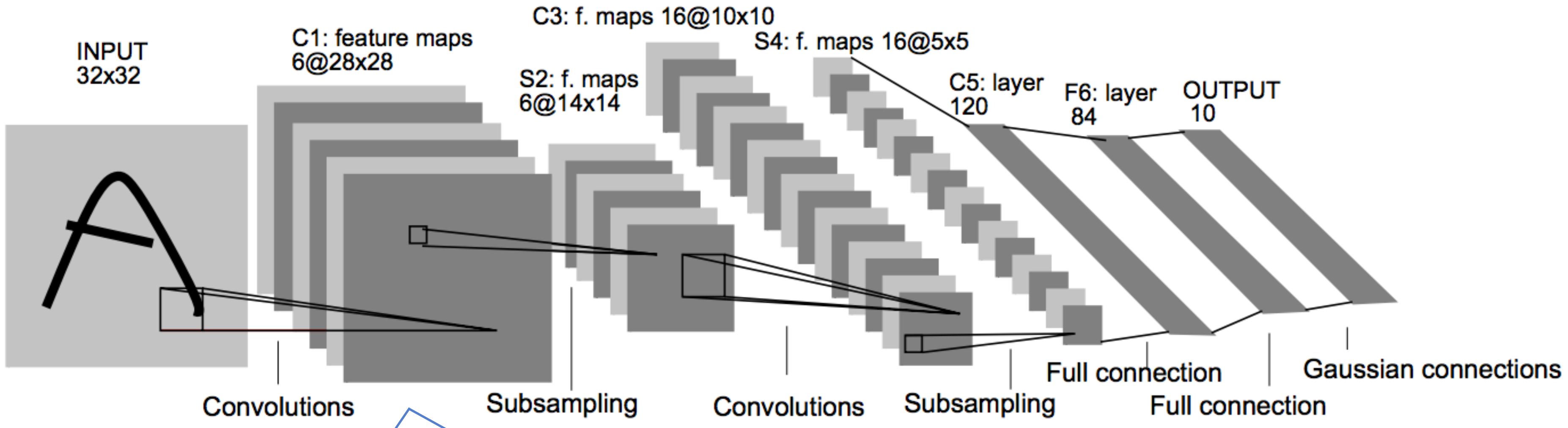
LeNet 5



How many weights to learn in this layer?

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

LeNet 5

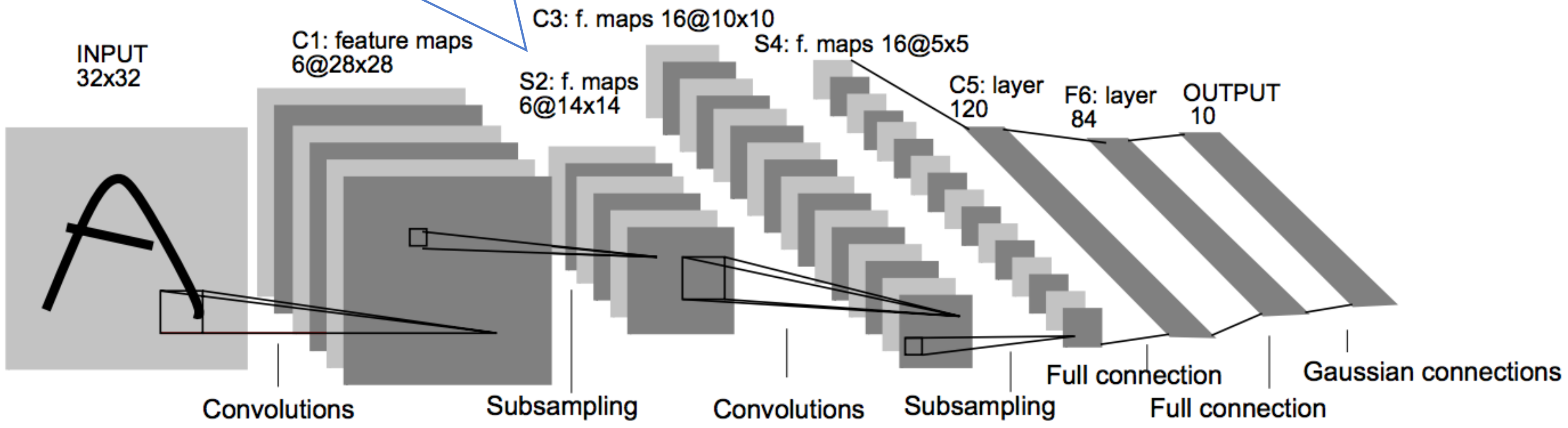


How many weights to learn in this layer?

Answer: Each kernel has 5x5=25 weights (plus a bias term, so actually 26 weights). So total weights = 26*6 = 156.

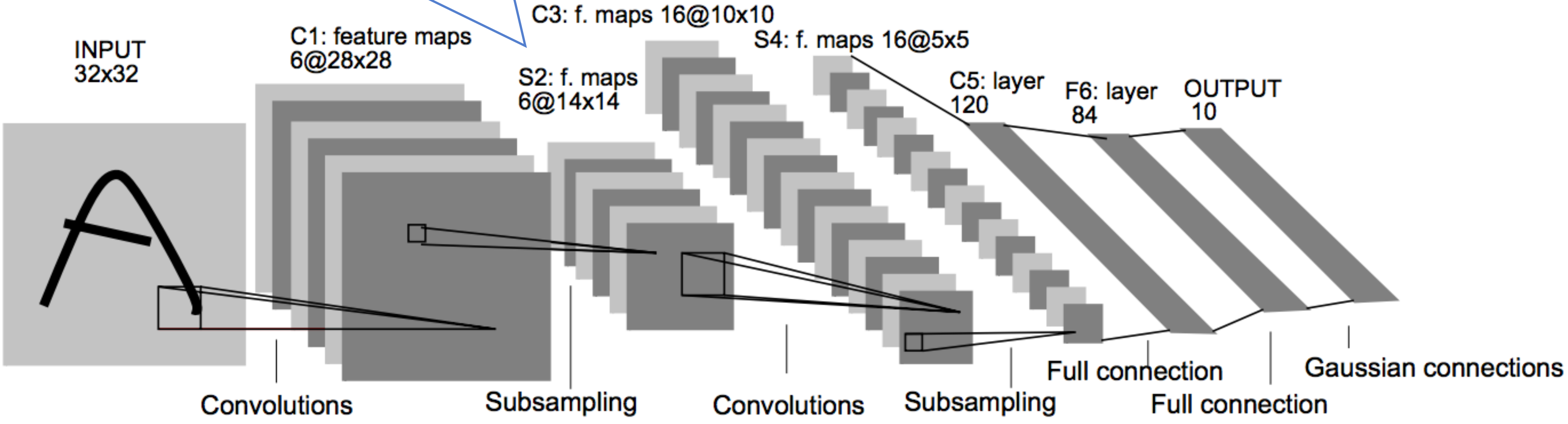
LeNet 5

Could you guess what the next layer does?



LeNet 5

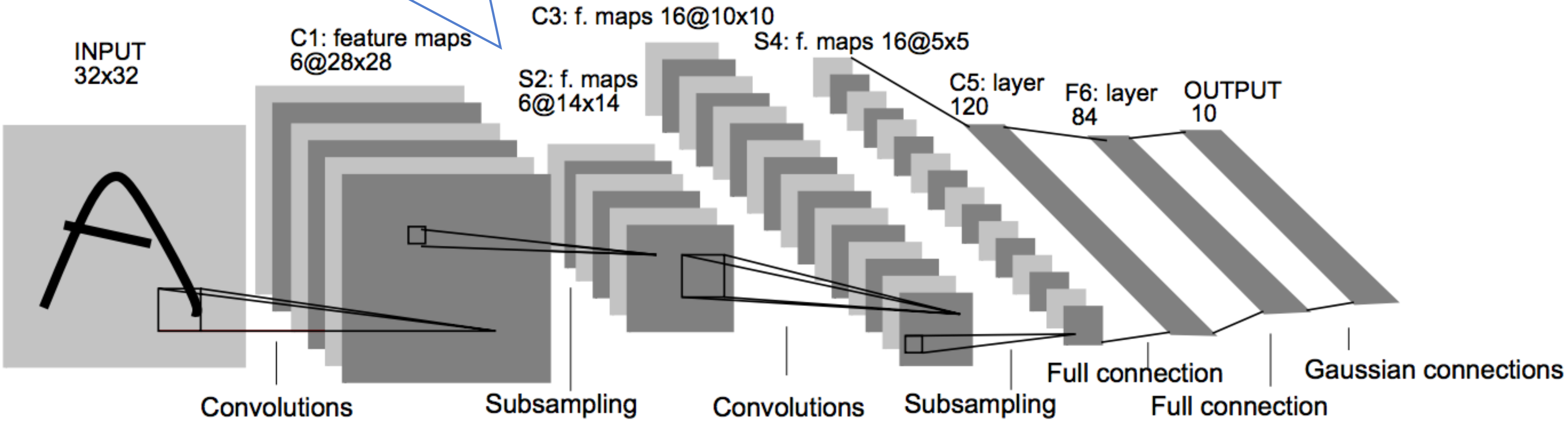
2x2 pooling layer with stride 2



LeNet 5

2x2 pooling layer with stride 2

Output size is 6x14x14 → input is downsampled by a factor of 2

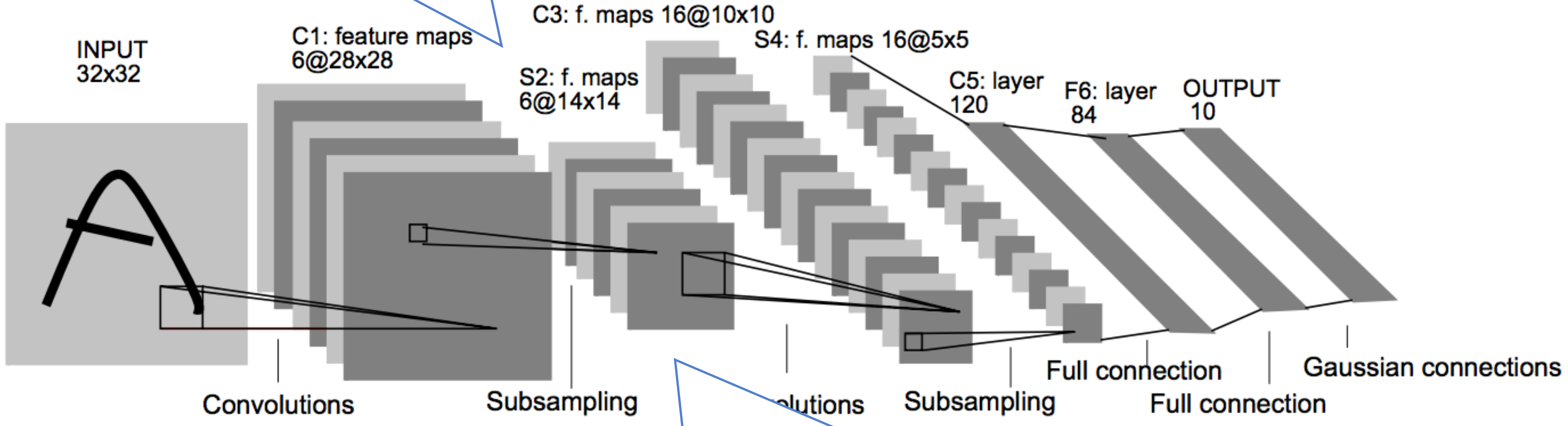


Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

LeNet 5

2x2 pooling layer with stride 2

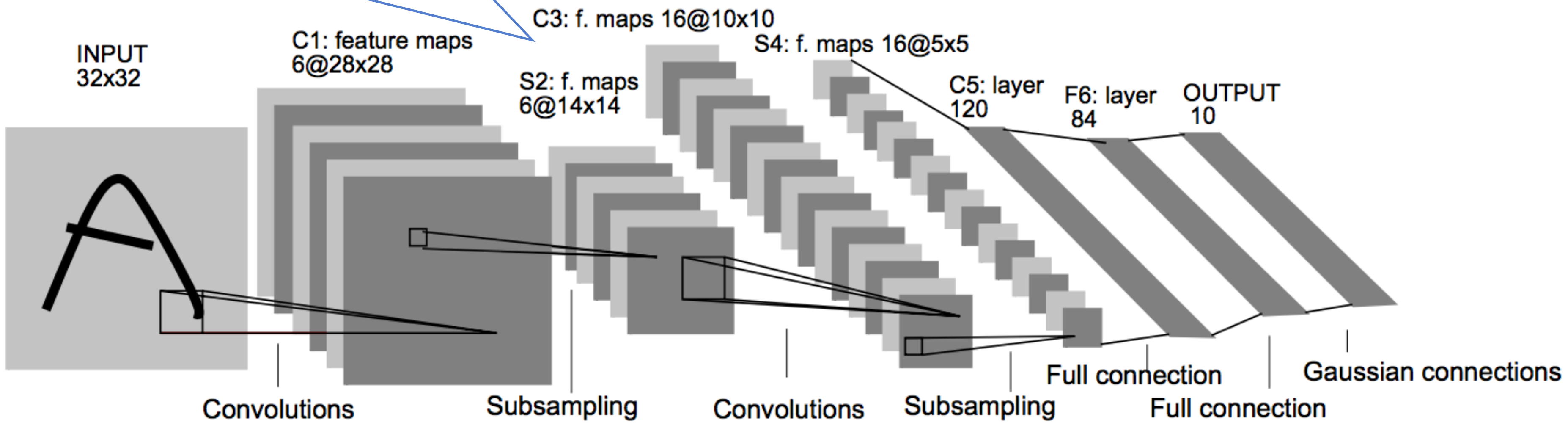
Output size is 6x14x14 → input is downsampled by a factor of 2



Note: The original paper actually does a more complicated pooling then max or avg pooling, but this is considered obsolete now.

LeNet 5

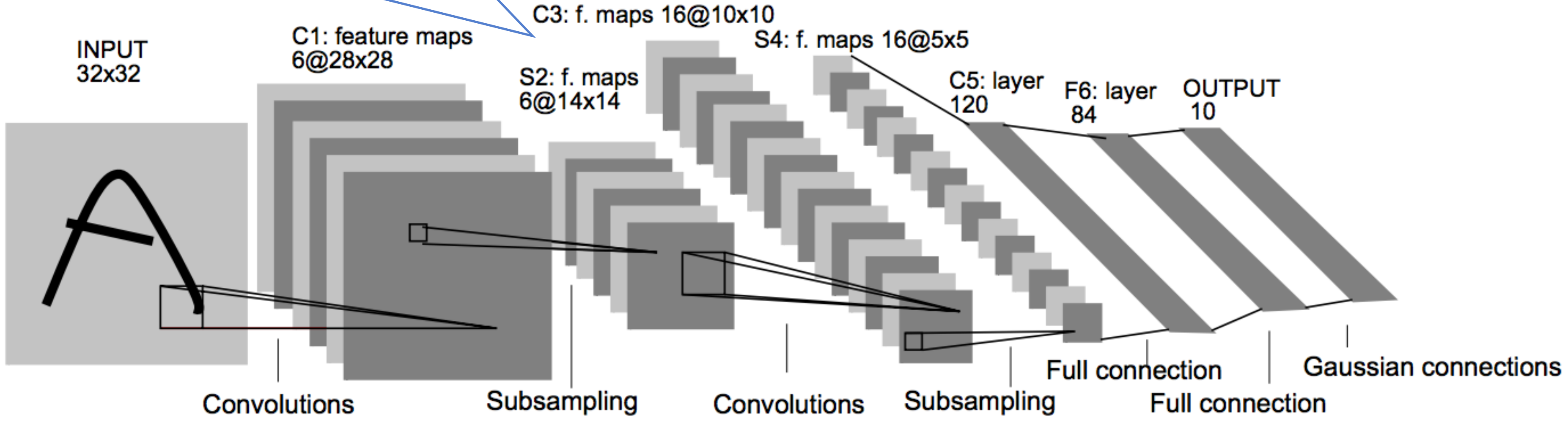
How many weights to learn in this layer?



LeNet 5

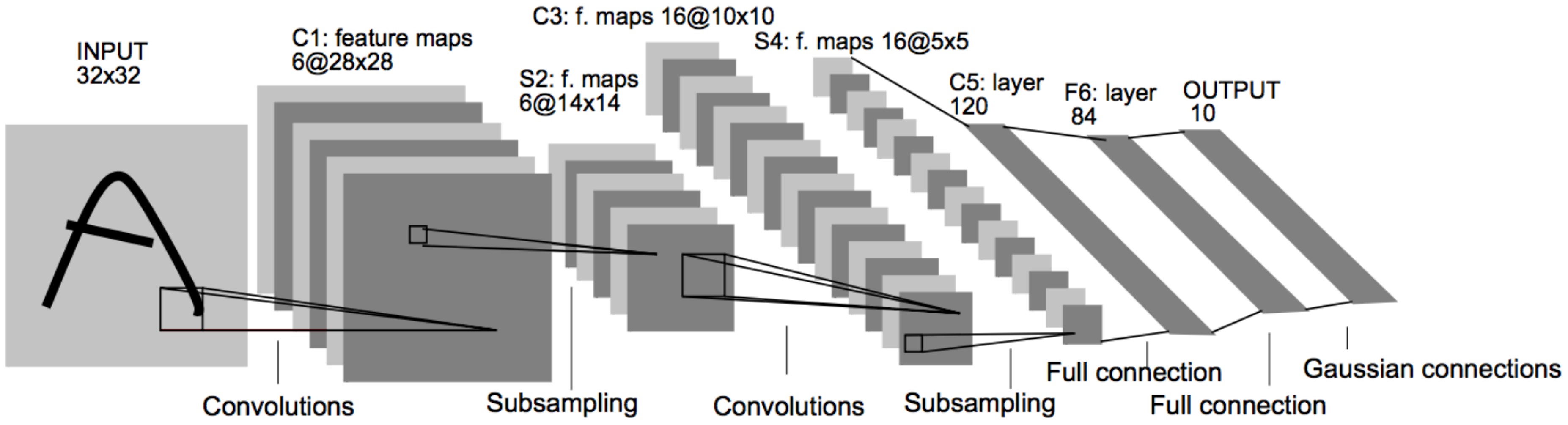
How many weights to learn in this layer?

No weights!
Pooling layers have no weights.



LeNet 5

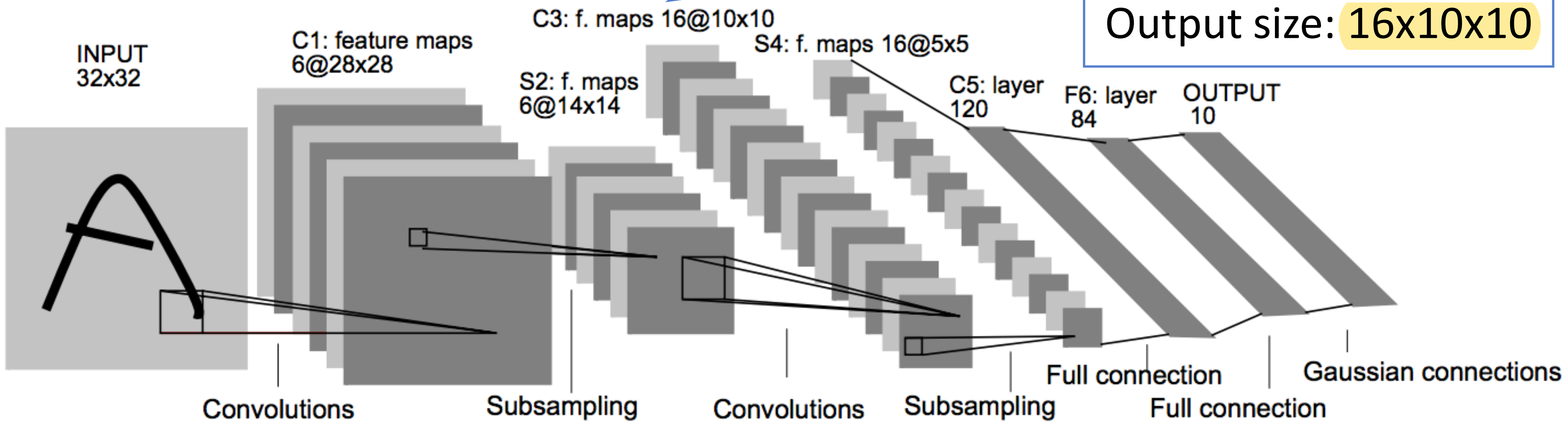
Convolutional layer with 5x5 kernels and stride 2 → 10x10
Depth of 16 → 16 different kernels are learned.



LeNet 5

Convolutional layer with 5x5 kernels and stride 2 → 10x10
Depth of 16 → 16 different kernels are learned.

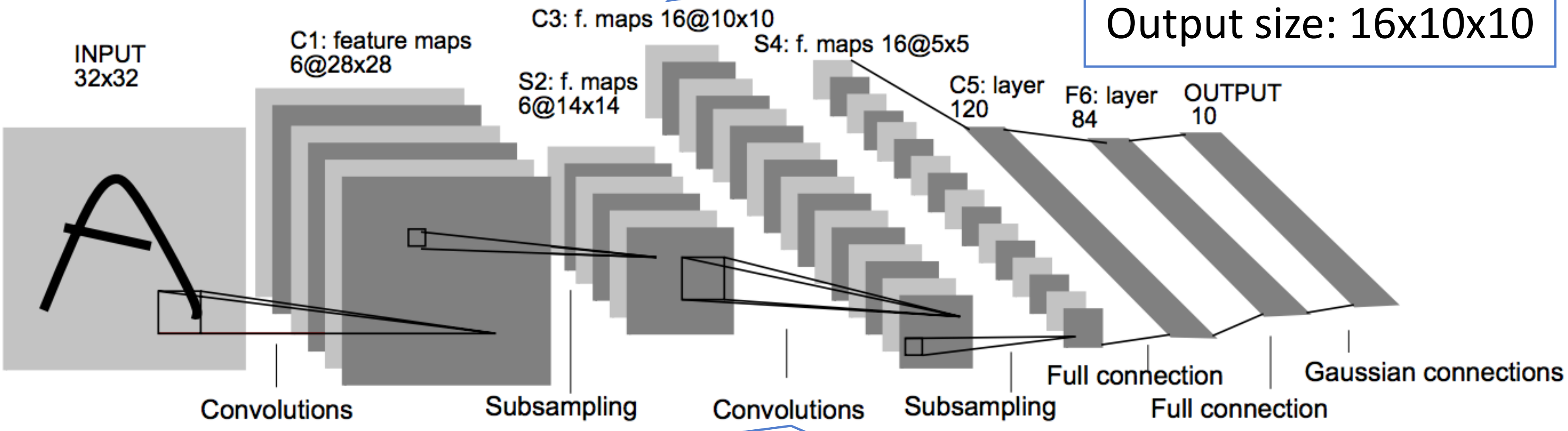
Output size: 16x10x10



LeNet 5

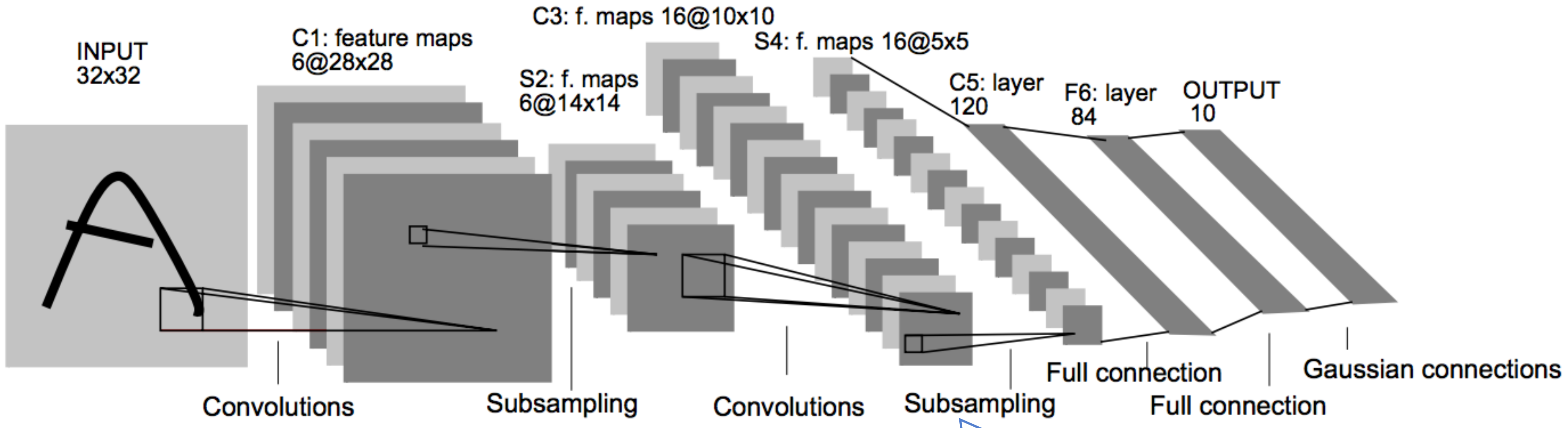
Convolutional layer with 5x5 kernels and stride 2 → 10x10
Depth of 16 → 16 different kernels are learned.

Output size: 16x10x10



In this layer, the kernels "take in" the full depth of the previous layer.
So each 5x5 kernel now "looks at" 6x5x5 pixels.
Each kernel has 6x5x5 = 150 weights + bias term = 151.
Total weights: 151*16= 2416

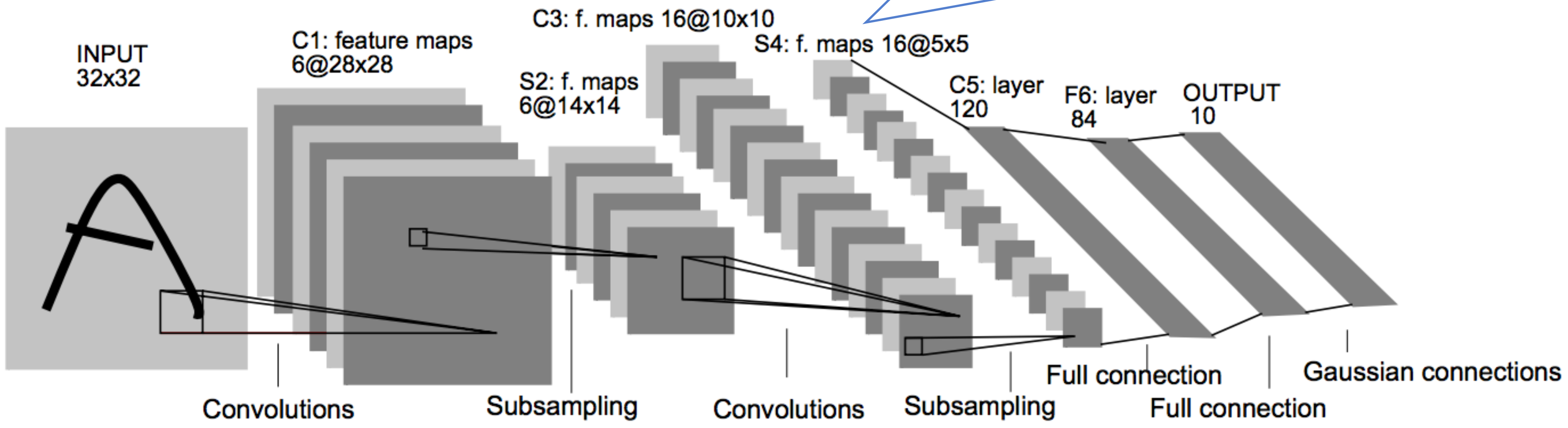
LeNet 5



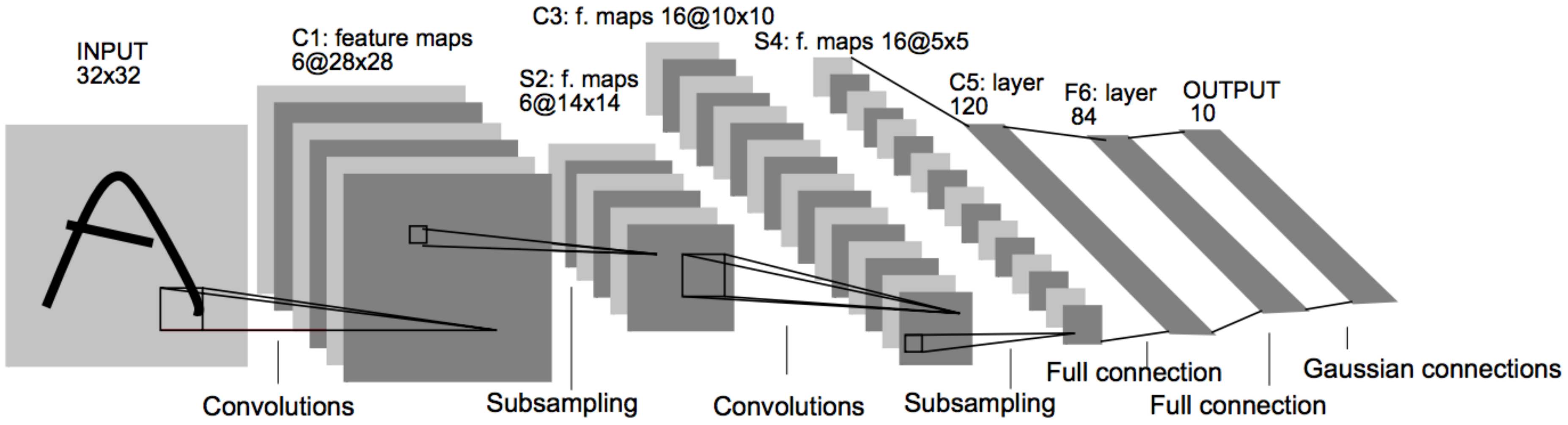
2x2 pooling layer with stride 2.
Output size: 16x5x5

LeNet 5

Output flattened to a 400 weights vector



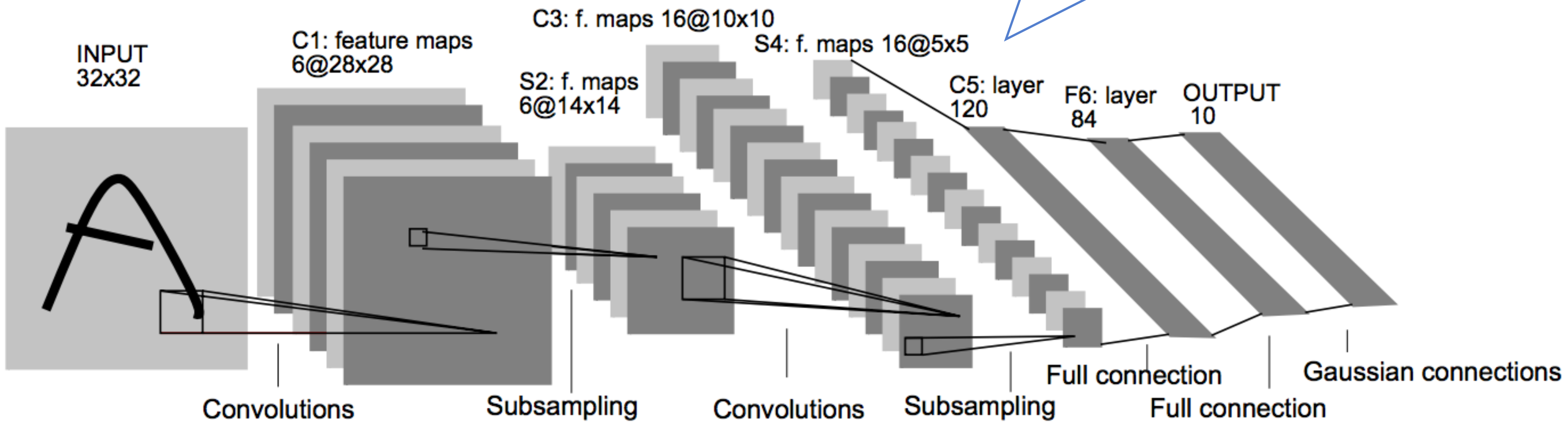
LeNet 5



The following layers are just fully connected layers!

LeNet 5

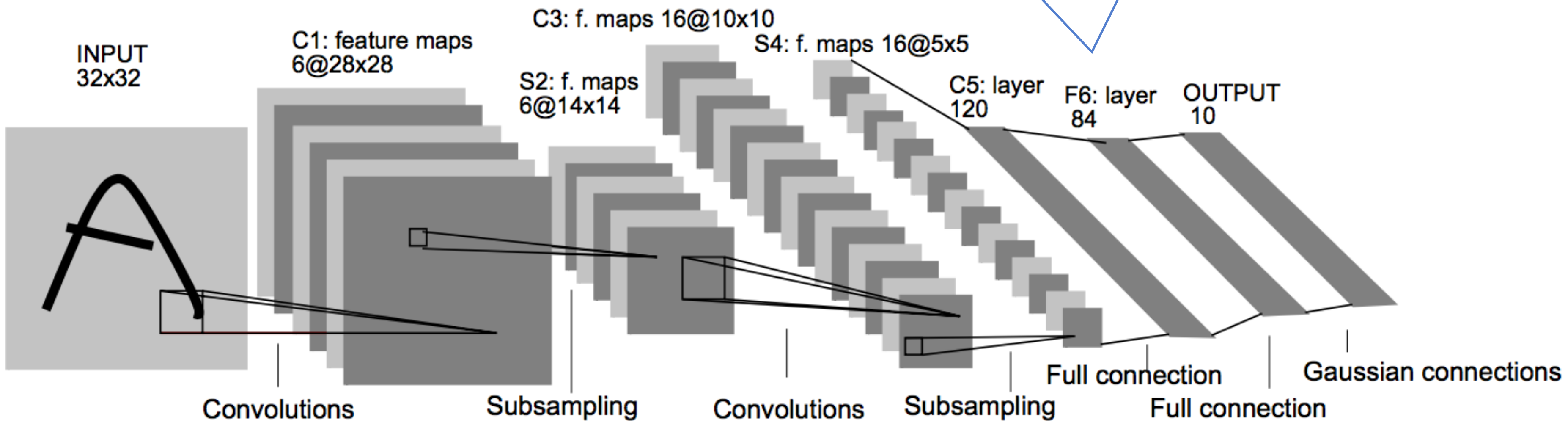
From 400 to 120



The following layers are just fully connected layers!

LeNet 5

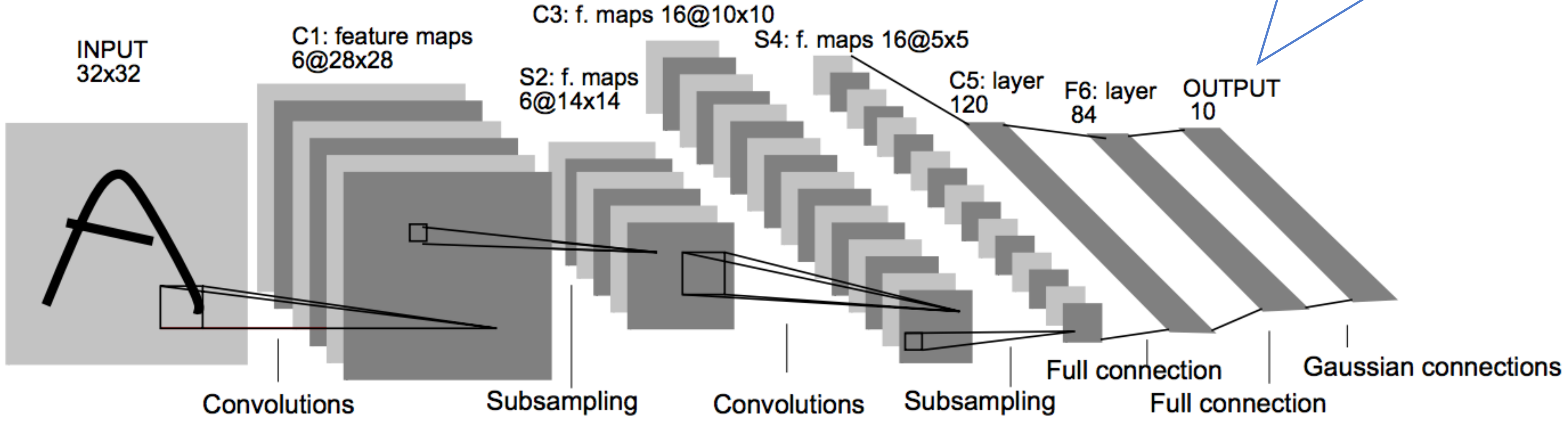
From 120 to 84



The following layers are just fully connected layers!

LeNet 5

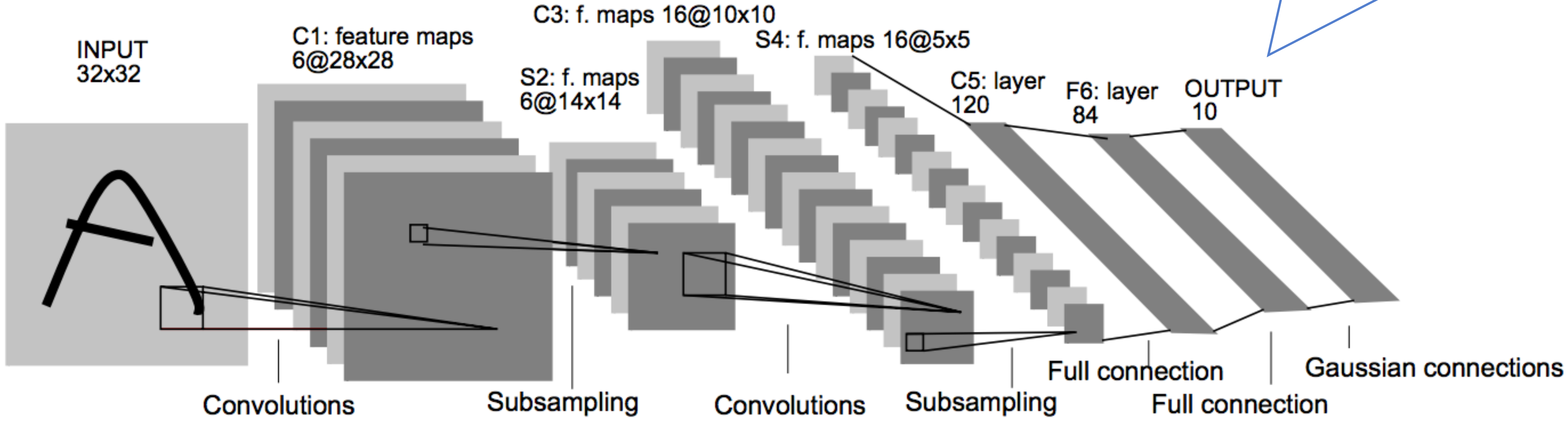
From 84 to 10



The following layers are just fully connected layers!

LeNet 5

Softmax output of size 10 for the 10 digits



The following layers are just fully connected layers!

LeNet 5

- How many total weights in the network?

Conv1: 1*6*5*5 + 6 = 156
Conv3: 6*16*5*5 + 16 = 2416
FC1: 400*120 + 120 = 48120
FC2: 120*84 + 84 = 10164
FC3: 84*10 + 10 = 850
Total = 61706

N. of channels
N. of kernels
Size of kernel
Size of FC weights matrix
Bias

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

LeNet 5

- How many total weights in the network?

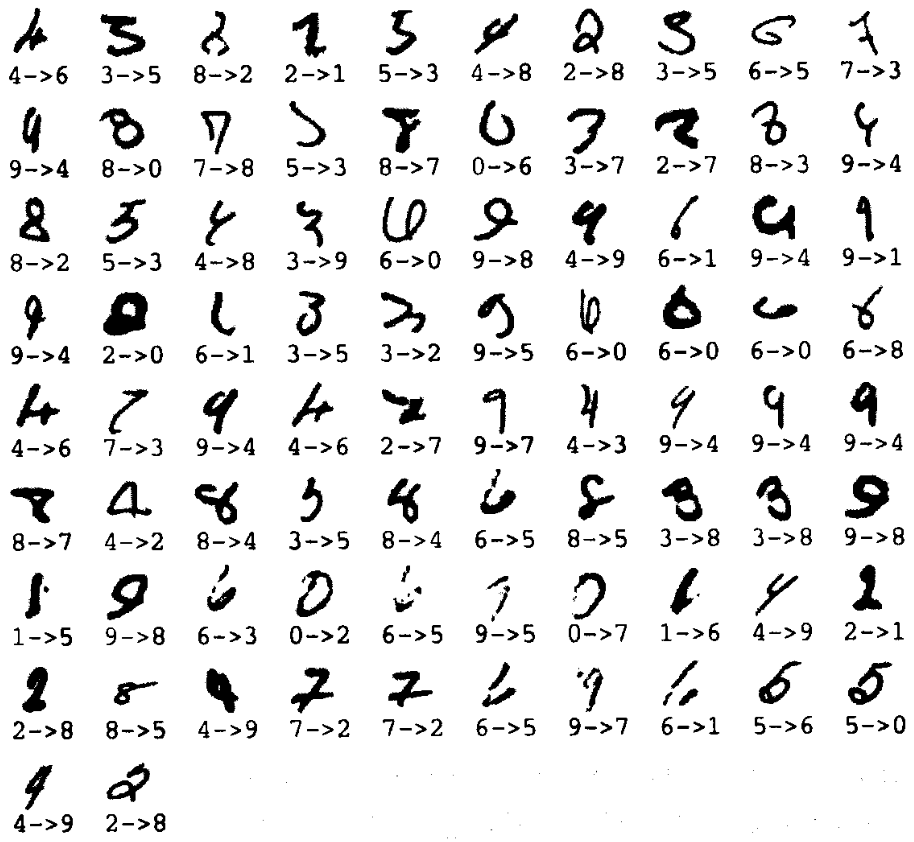
Conv1: 1*6*5*5 + 6 = 156
Conv3: 6*16*5*5 + 16 = 2416
FC1: 400*120 + 120 = 48120
FC2: 120*84 + 84 = 10164
FC3: 84*10 + 10 = 850
Total = 61706

- Less than a single FC layer with 1000x1000 or 1200x1200 weights usually used before the introduction of CNN!

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

LeNet 5

- LeNet misclassified 82 test patterns
- Notice that most of the errors are cases that people find quite easy
- The human error rate is probably 20 to 30 errors but nobody has had the patience to measure it.
- These errors are mostly caused by ambiguous patterns or by digits written in a style that is under-represented in the training set.



AlexNet

- Created in 2012 (Krizhevsky, Sutskever, Hinton —NIPS 2012) for the ImageNet Large Scale Visual Recognition Challenge (ILSVRC)
- Task: predict the correct label from among 1000 classes
- Dataset: around 1.2 million images
- Considered the "flash point" for modern deep learning
- Demolished the competition.
- Top 5 error rate of 15.3%
    - Next best: 26.2%

Results of ILSVRC-2012

Top-1 score computed by checking if the top class (the one having the highest probability) is the same as the target label. Top-5 score computed by checking if the target label is one of your top 5 predictions (the 5 ones with the highest probabilities).

Model	Top-1 (val)	Top-5 (val)	Top-5 (test)
SIFT + FVs [7]	—	—	*26.2%*
1 CNN	40.7%	18.2%	—
5 CNNs	38.1%	16.4%	**16.4%**
1 CNN*	39.0%	16.6%	—
7 CNNs*	36.7%	15.4%	**15.3%**

Table : Comparison of error rates on ILSVRC-2012 validation and test sets. In *italics* are best results achieved by others. Models with an asterisk* were "pre-trained" to classify the entire ImageNet 2011 Fall release.

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Regione Emilia-Romagna

Università degli Studi di Ferrara

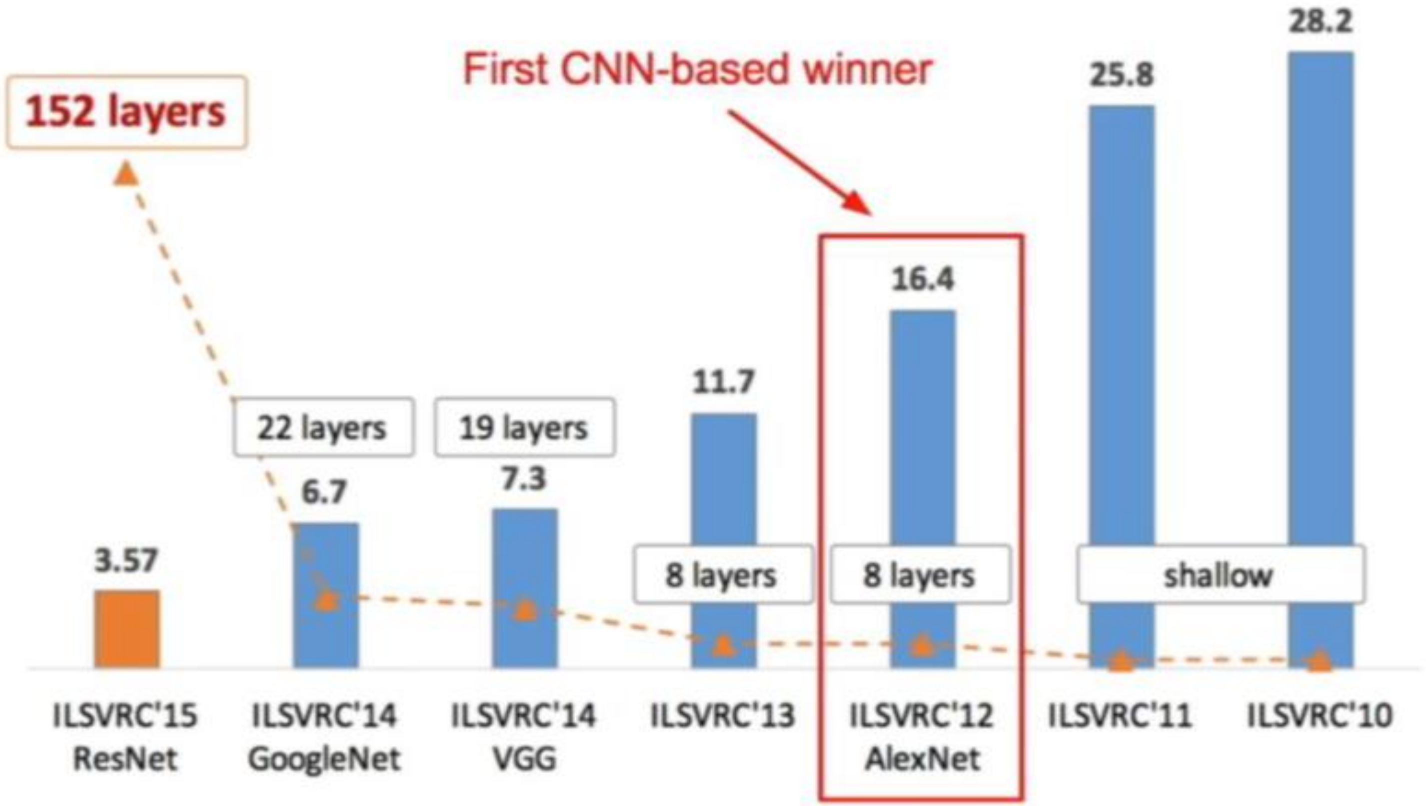
Results of ILSVRC-2012

1 CNN → 1 AlexNet.
5 CNNs → results averaged from prediction of 5 AlexNets. This is a kind of boosting technique already used in LeNet for digit classification.
1 CNN* → addition of one more convolutional layer (pre-trained)
7 CNNs* → results averaged from 2 modified AlexNet and 5 original AlexNet.

Model	Top-1 (val)	Top-5 (val)	Top-5 (test)
SIFT + FVs [7]	—	—	*26.2%*
1 CNN	40.7%	18.2%	—
5 CNNs	38.1%	16.4%	**16.4%**
1 CNN*	39.0%	16.6%	—
7 CNNs*	36.7%	15.4%	**15.3%**

Table : Comparison of error rates on ILSVRC-2012 validation and test sets. In *italics* are best results achieved by others. Models with an asterisk* were "pre-trained" to classify the entire ImageNet 2011 Fall release.

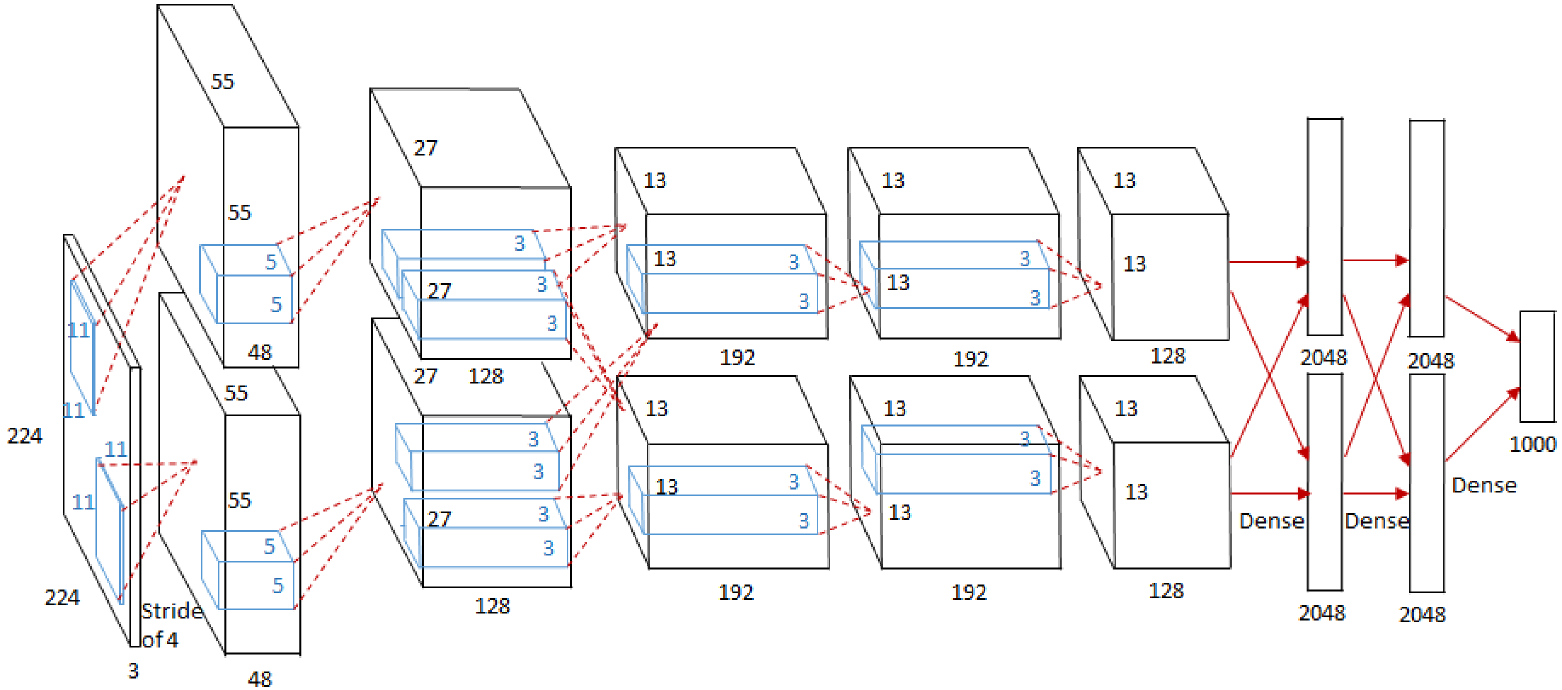
Results of ILSVRC



Img: copyright Kaiming He, 2016.

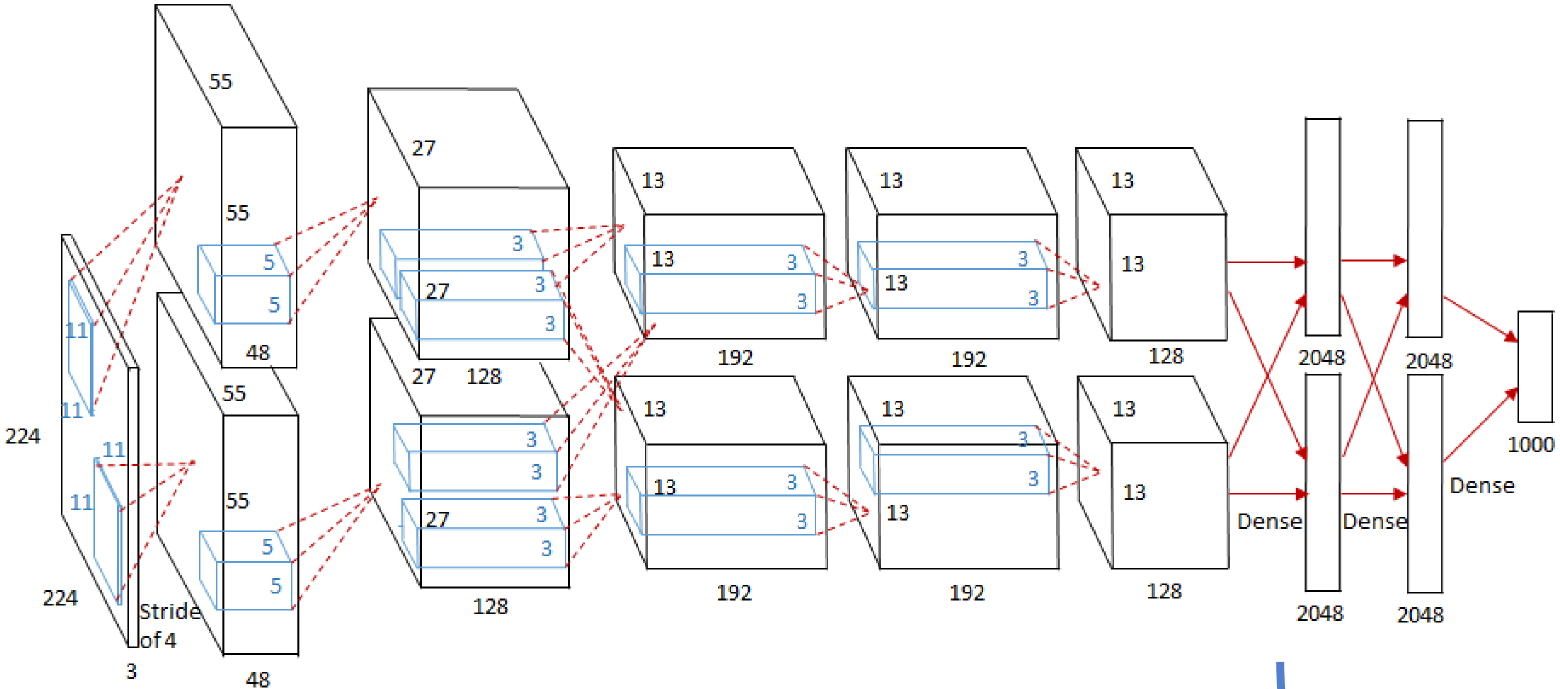
AlexNet

- There are 8 trainable layers: 5 convolutional and 3 fully connected.
- ReLU activations are used for all layers, except for the output layer using softmax.



AlexNet

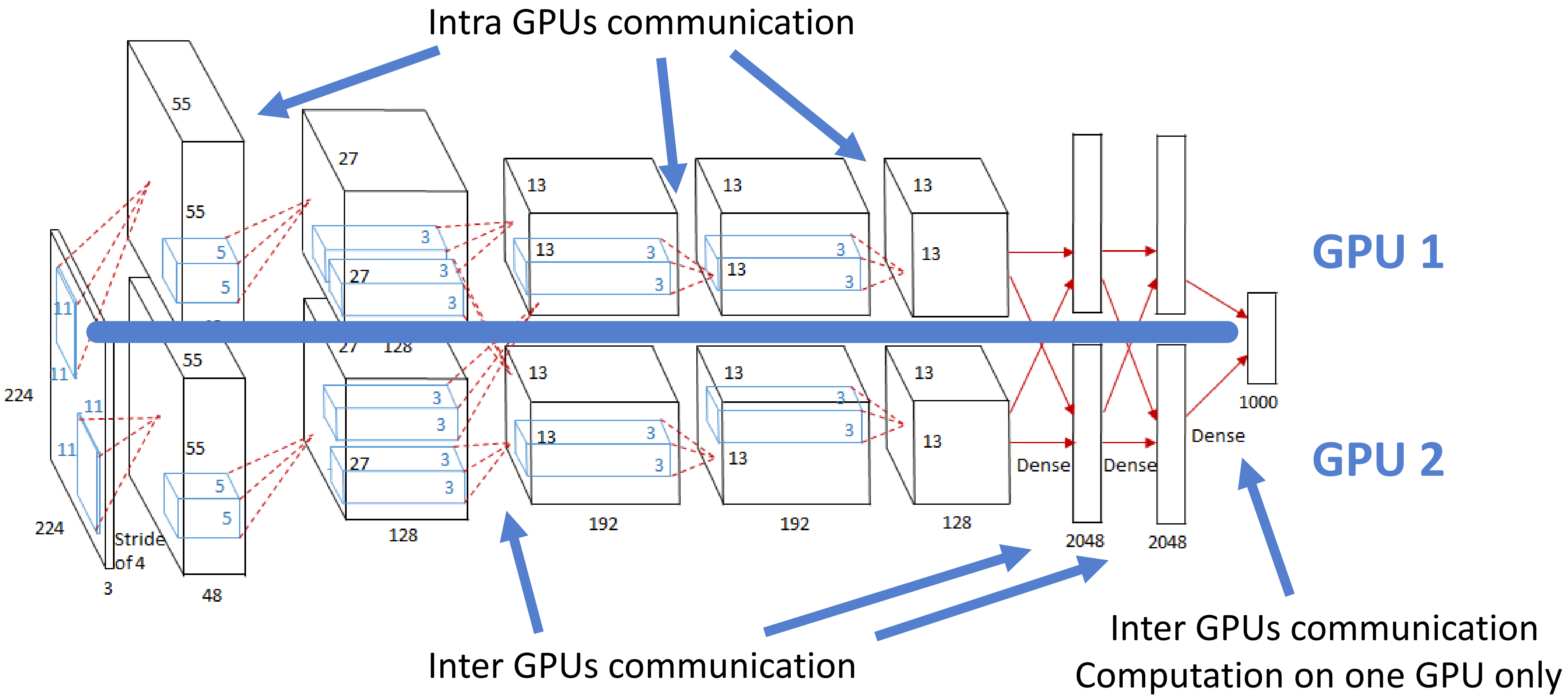
5 Convolutional Layers



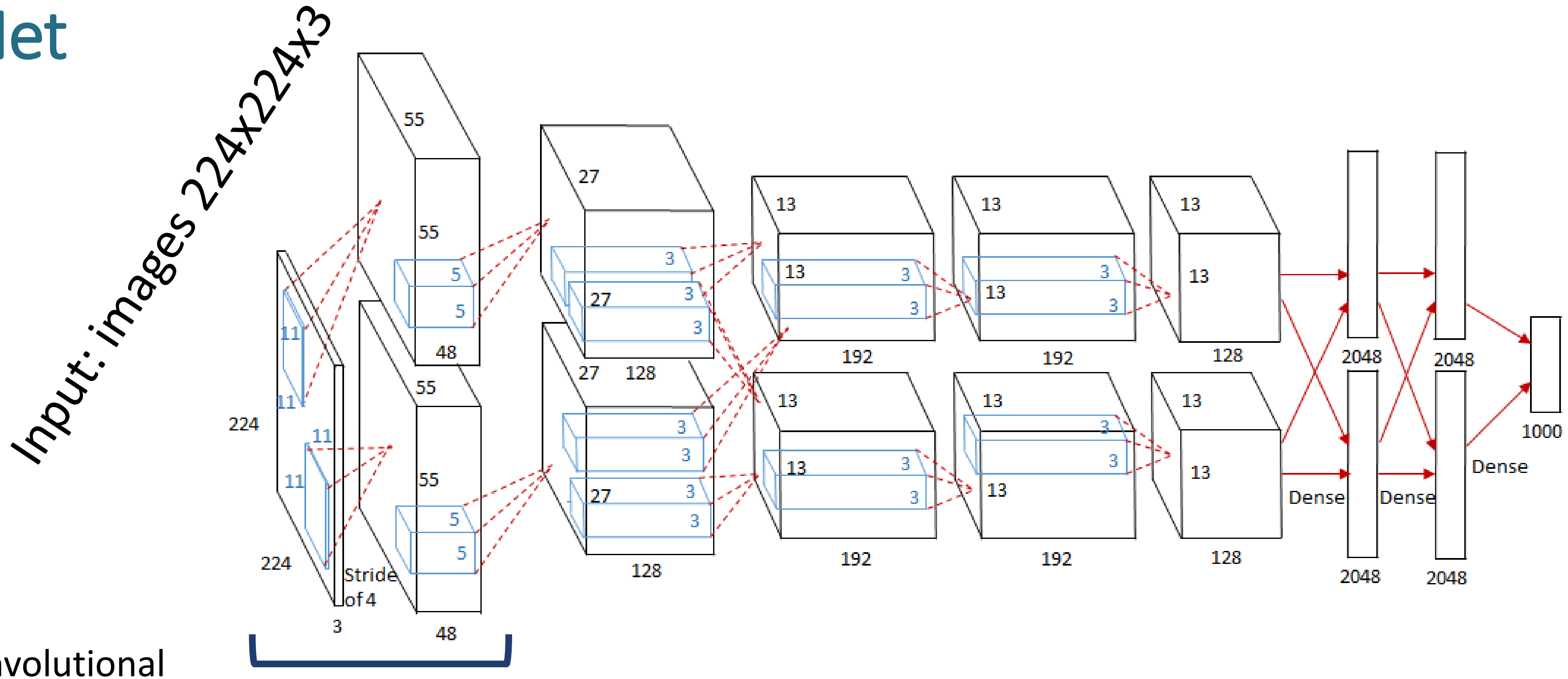
3 Fully Connected Layers

AlexNet

It resides on 2 GPUs → it is split in 2 parts that communicate only partially.



AlexNet



Layer 1: Convolutional

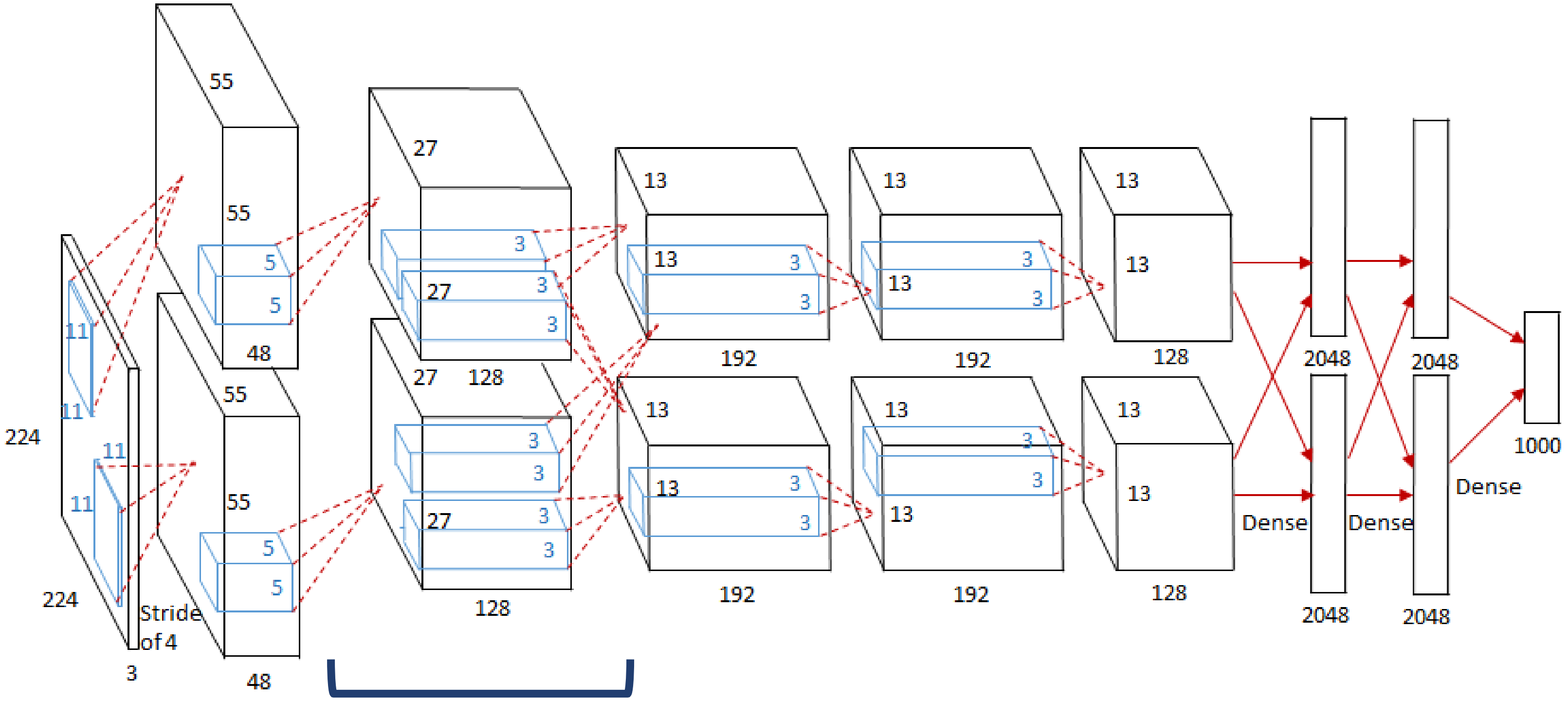
96 kernels of size 11×11×3 (receipt field size: 11) - stride: 4, pad: 0

→ 55×55×96 feature maps (convolution layer output)

3×3 Overlapping Max Pooling - stride: 2 + Local response normalization

→ 27×27×96 feature maps

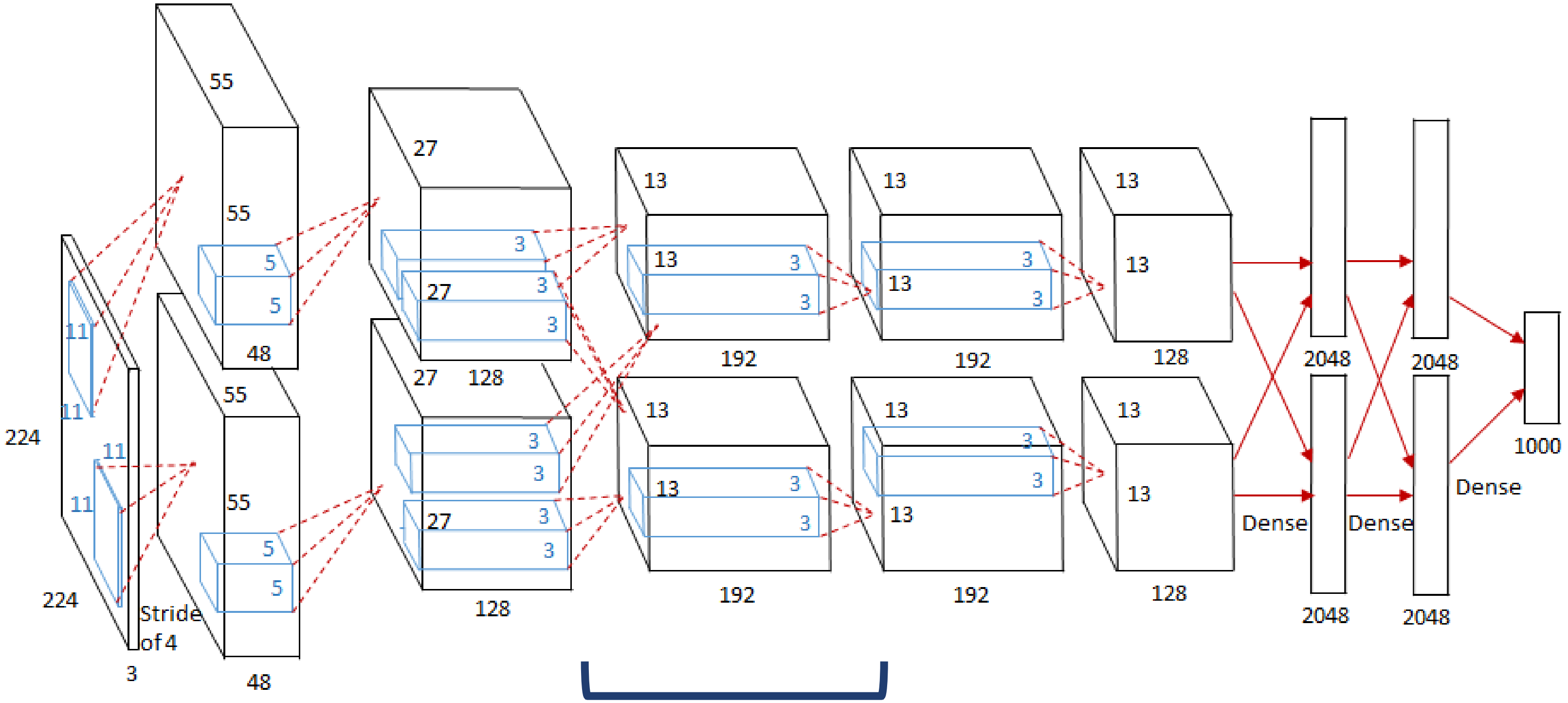
AlexNet



Layer 2: Convolutional
256 kernels of size 5×5×48 - stride: 1, pad: 2
→ 27×27×256 feature maps
3×3 Overlapping Max Pooling - stride: 2 + Local response normalization
→ 13×13×256 feature maps

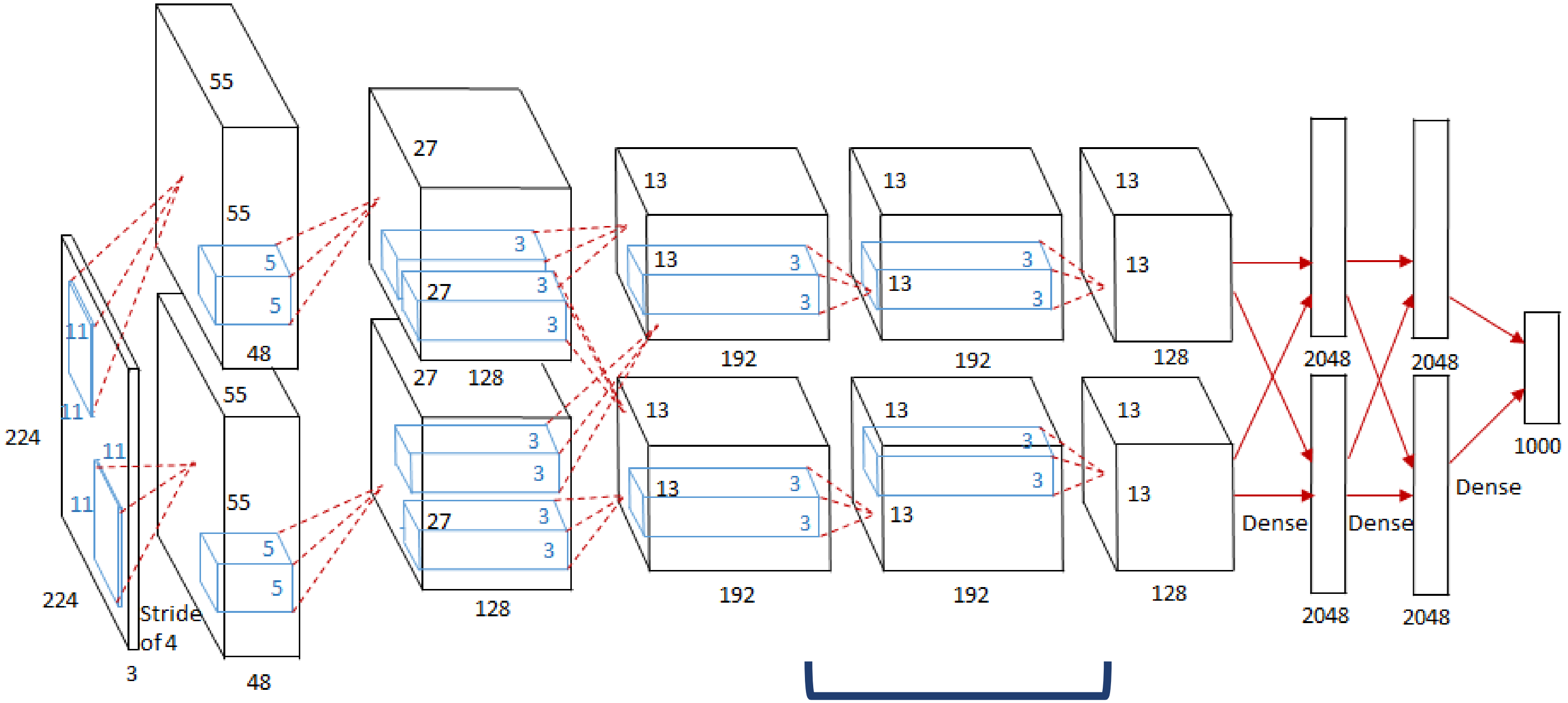
Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

AlexNet



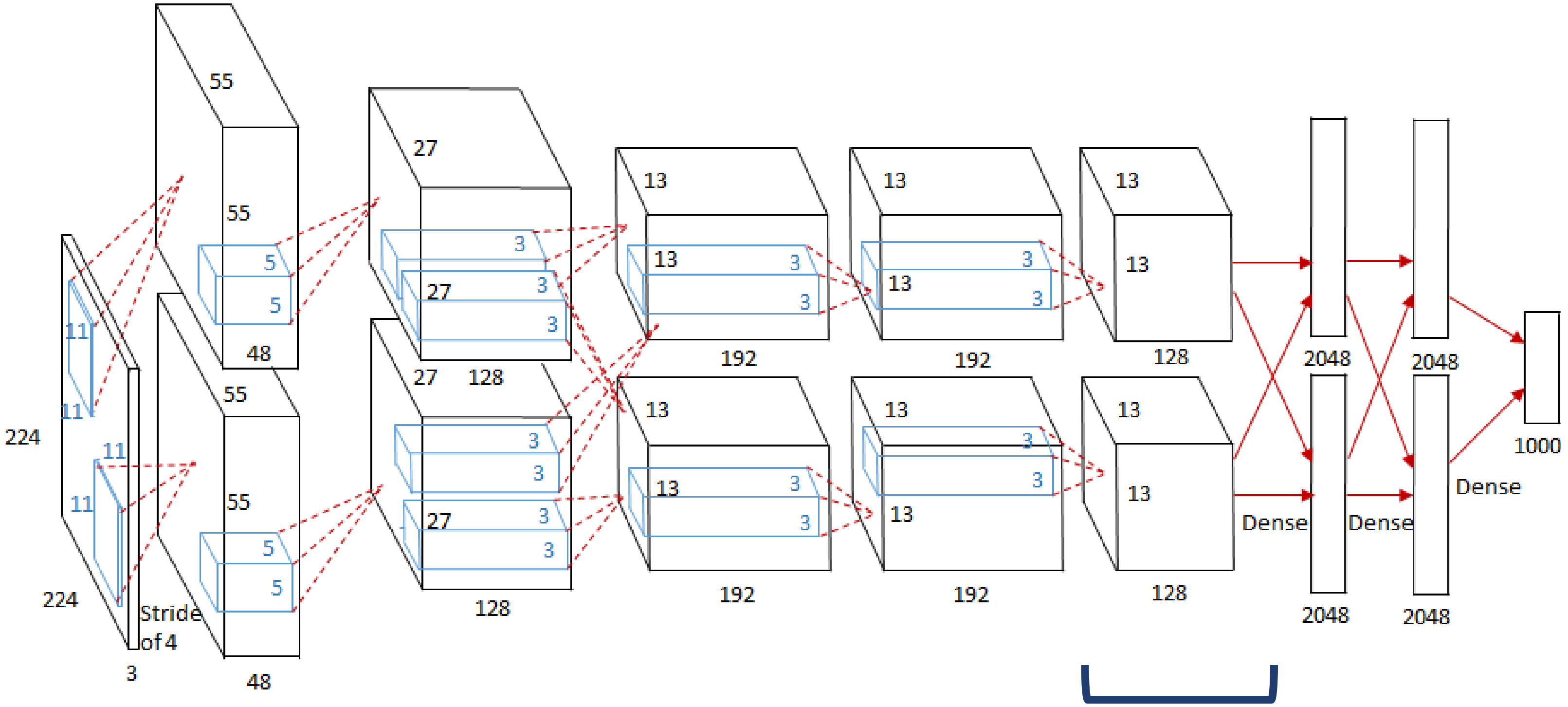
Layer 3: Convolutional
384 kernels of size 3×3×256 - stride: 1, pad: 1
→ 13×13×384 feature maps

AlexNet



Layer 4: Convolutional
384 kernels of size 3×3×192 - stride: 1, pad: 1
→ 13×13×384 feature maps

AlexNet



Layer 5: Convolutional

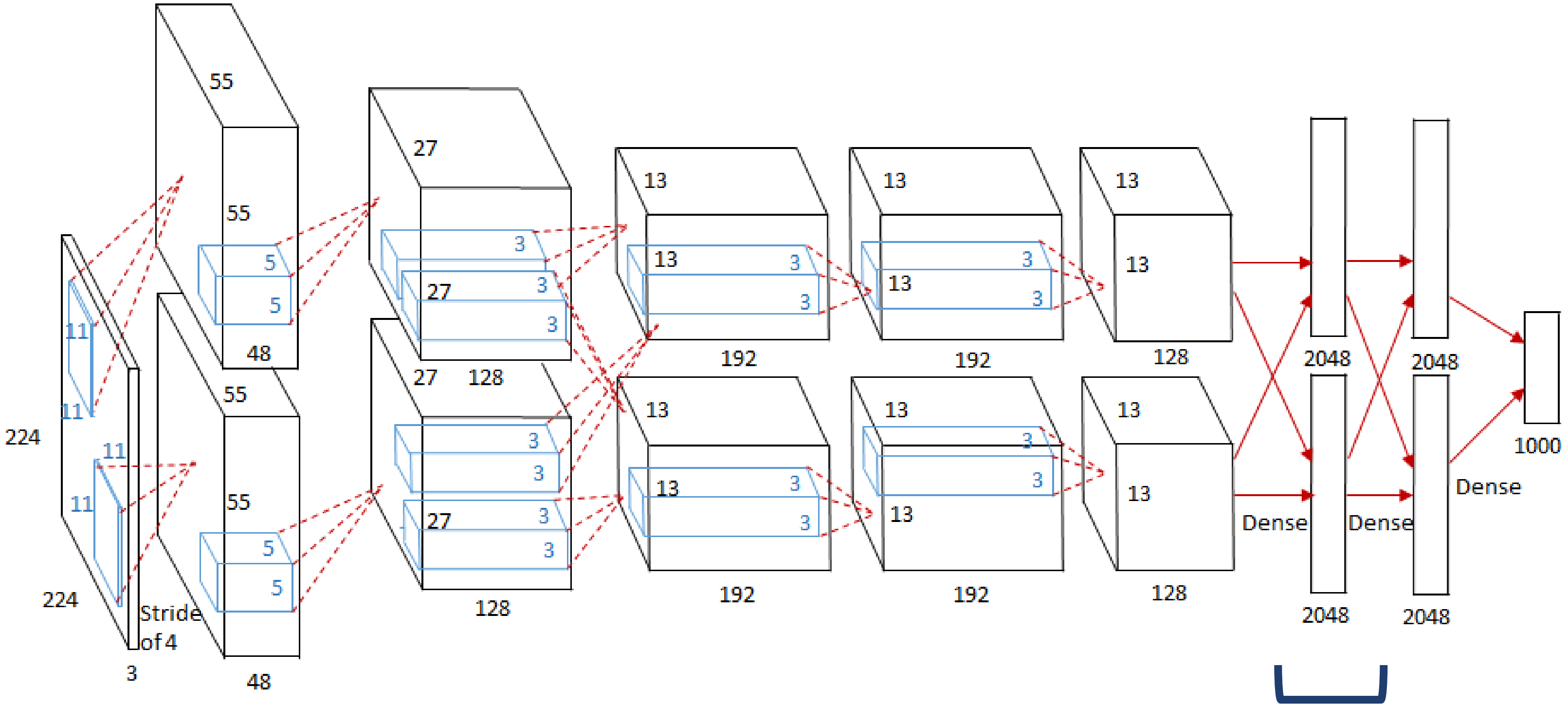
256 kernels of size 3×3×192 - stride: 1, pad: 1

→ 13×13×256 feature maps

3×3 Overlapping Max Pooling - stride: 2

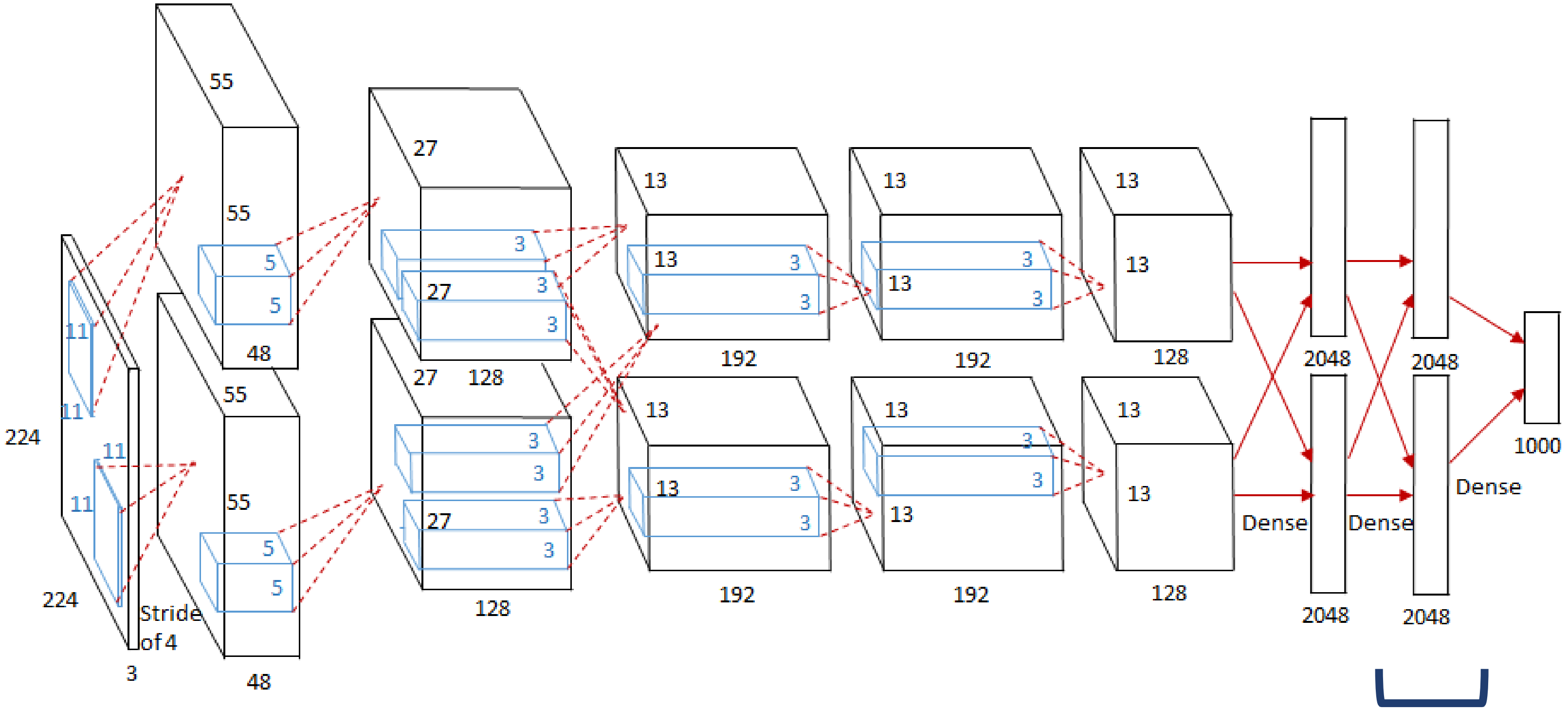
→ 6×6×256 feature maps

AlexNet



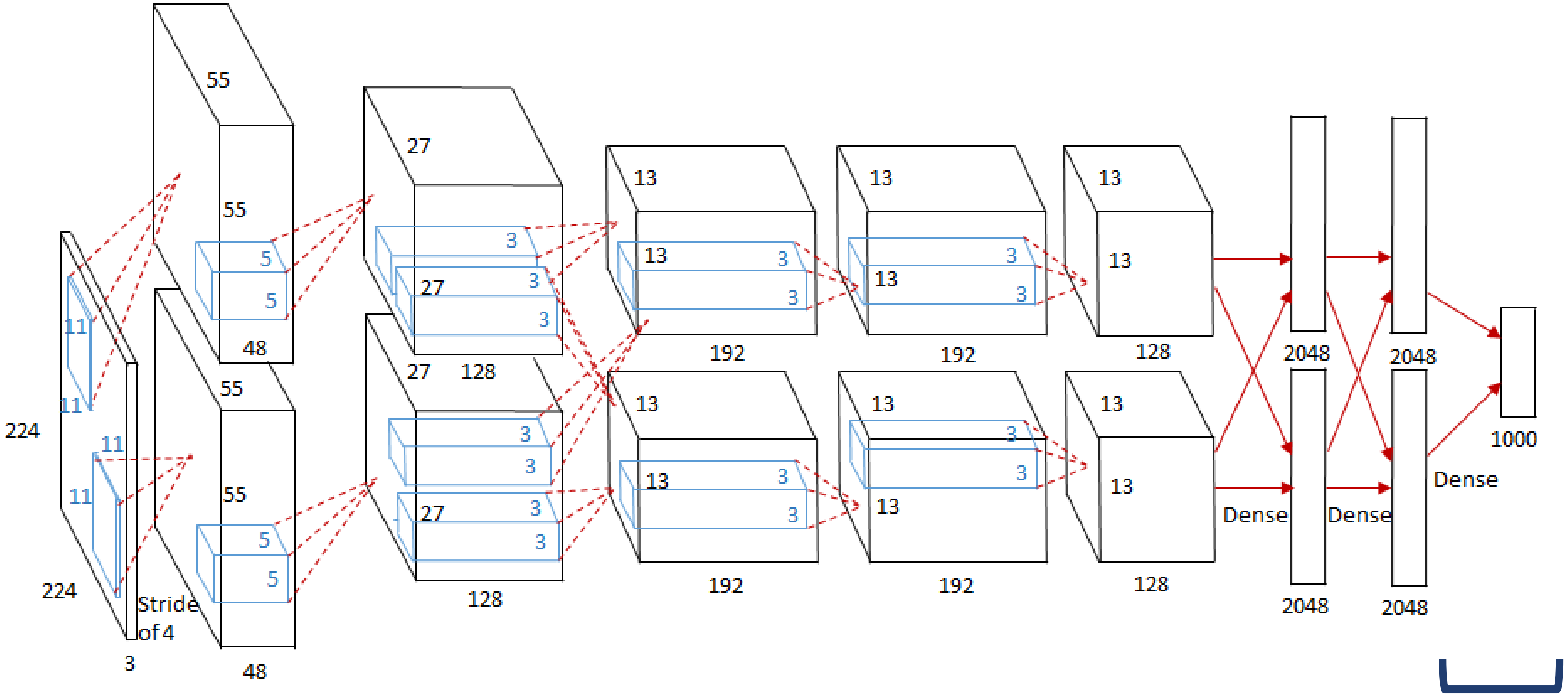
Layer 6: Fully Connected (Dense) Layer
4096 neurons
Dropout: 0.5

AlexNet



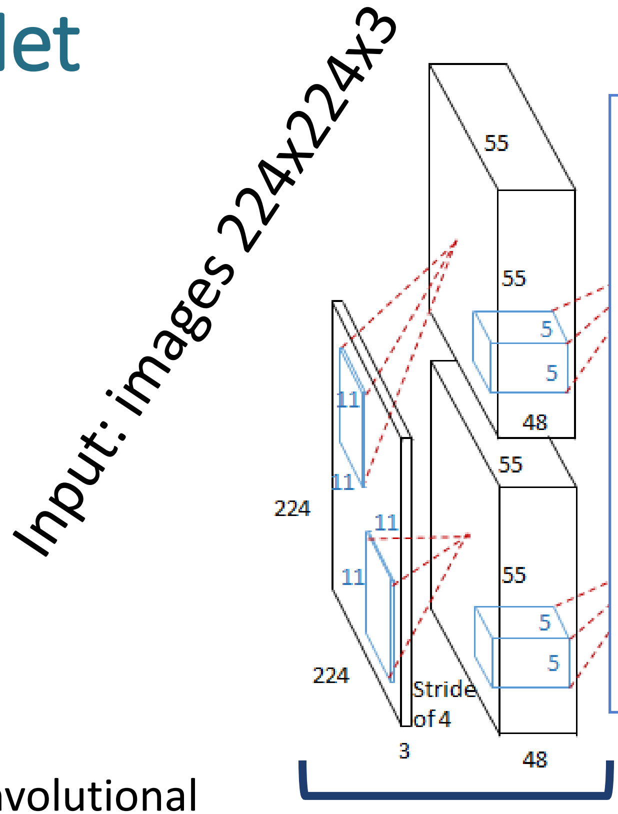
Layer 7: Fully Connected (Dense) Layer
4096 neurons
Dropout: 0.5

AlexNet



Layer 8: Fully Connected (Dense) Layer
Output: 1000 neurons (1000 classes)
Loss: Softmax

AlexNet



If we consider a fully connected network (without using convolution) we would have:
55*55*96 = 290,400 neurons
each has 11*11*3 = 363 weights and 1 bias
290,400 * 364 = 105,705,600 parameters on the first layer of the AlexNet alone!
Use of GPUs is essential.

Layer 1: Convolutional
96 kernels of size 11×11×3 (receipt field size: 11) - stride: 4, pad: 0
→ 55×55×96 feature maps (convolution layer output)
3×3 Overlapping Max Pooling - stride: 2 + Local response normalization
→ 27×27×96 feature maps

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

AlexNet

- Instead, how many total weights in the network?

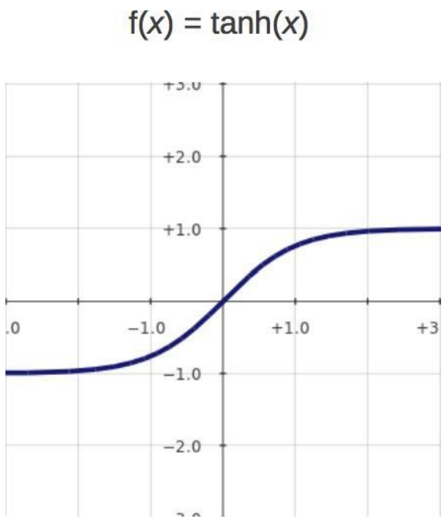
Conv1: 1*3*11*11 + 3	= 366
Conv2: 48*96*5*5 + 96	= 115,296
Conv3: 256*256*3*3 + 256	= 590,080
Conv4: 192*384*3*3 + 384	= 663,936
Conv5: 192*384*3*3 + 384	= 663,936
FC1: (256*6*6)*4096 + 4096	= 37,752,832
FC2: 4096*4096 + 4096	= 16,781,312
FC3: 4096*1000 + 1000	= 4,097,000
Total	= **60,664,758**

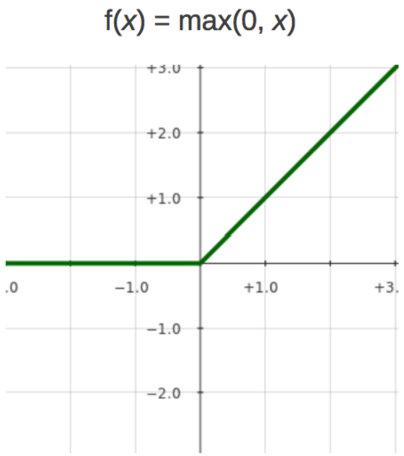
N. of channels N. of kernels Size of kernel
Size of FC weights matrix Bias

Use of ReLU

- Before AlexNet, tanh was used
- AlexNet introduced non-saturating nonlinearity (ReLU)

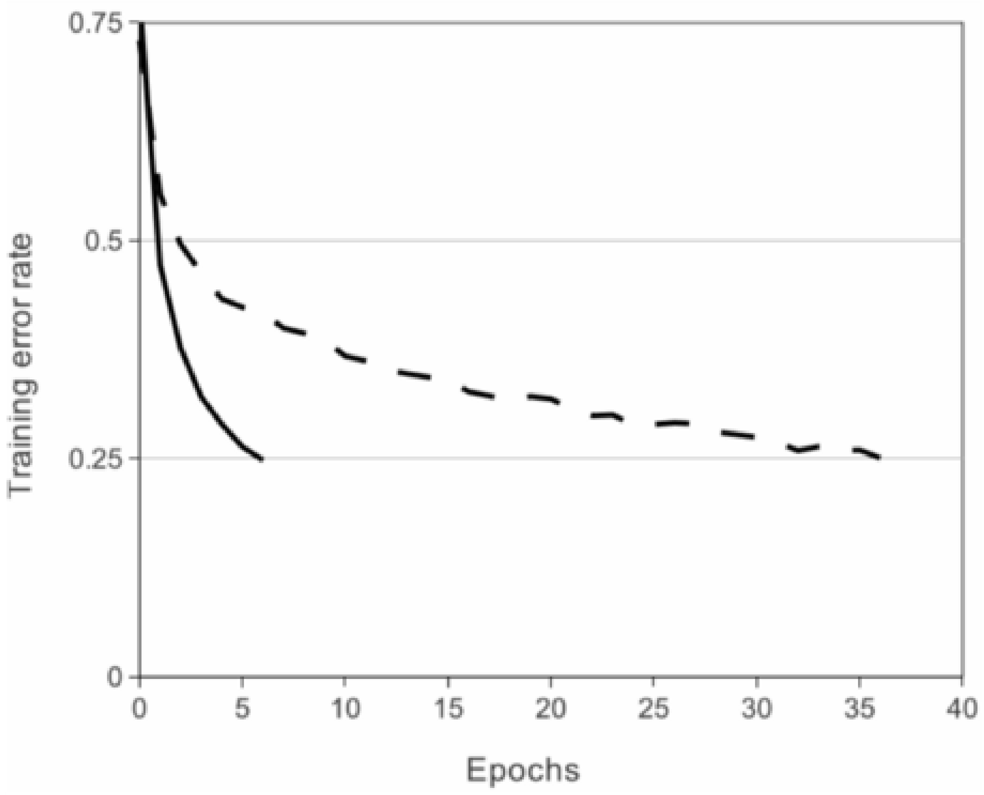
Why?





Use of ReLU

- A 4-layer CNN with ReLUs (solid line) converges **six times faster** than an equivalent network with tanh neurons (dashed line) on CIFAR-10 dataset.
- The learning rate for each network were chosen independently to make training as fast as possible.



AlexNet - Details

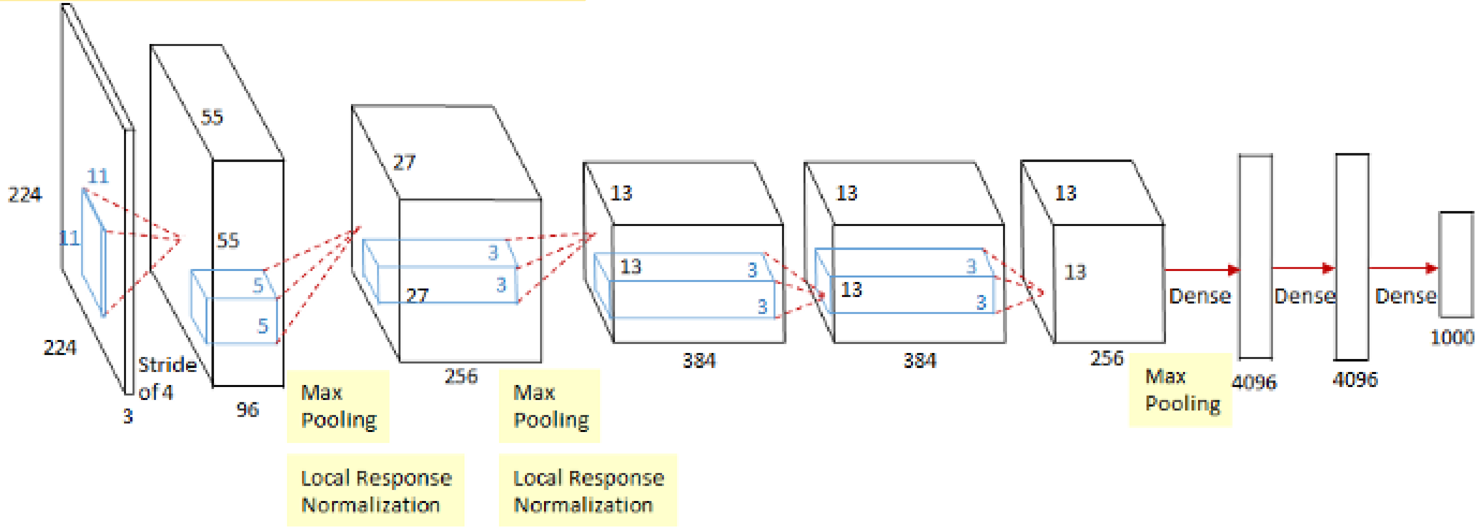
- **Local Response Normalization** is different from the batch normalization.
  - As already seen normalization helps to speed up the convergence.
  - Nowadays, batch normalization is used instead of using local response normalization.
- **Overlapping pooling** reduced tendency to overfit and also reduced test error rates

AlexNet - Details

- **Data augmentation** for training.
  - Cropping, horizontal flipping, and other manipulations
- **Stochastic gradient descent with momentum and learning rate decay** was used for training.
  - Batch size: 128
  - Momentum: 0.9
  - L^2 Weight Decay: 0.0005
  - Learning rate: 0.01, reduced by a factor of 10 manually when validation error rate stopped improving, and reduced by 3 times.
- Ran 5 to 6 days on two NVIDIA GTX 580 3GB GPUs.

CaffeNet

- **CaffeNet** is a 1-GPU version of AlexNet. The architecture is:



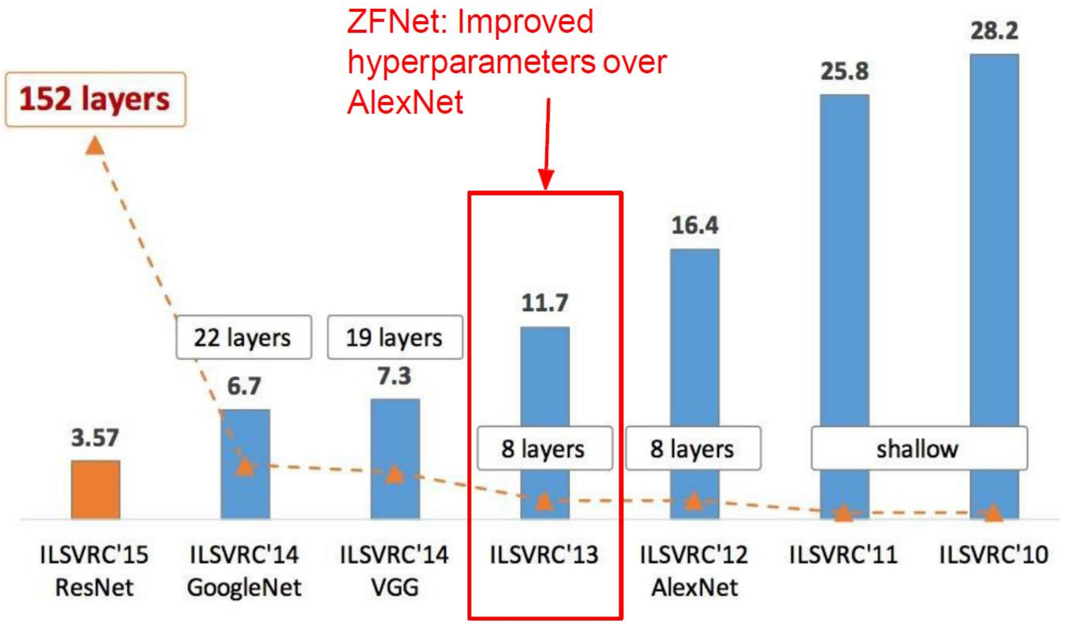
Images by: Sik-Ho Tsang (https://medium.com/coinmonks/paper-review-of-alexnet-caffenet-winner-in-ilsvrc-2012-image-classification-b93598314160)

- By accident, in the early version of CaffeNet the order of pooling and normalization layers was reversed.
- In the current version of CaffeNet provided by Caffe this problem has been fixed.

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

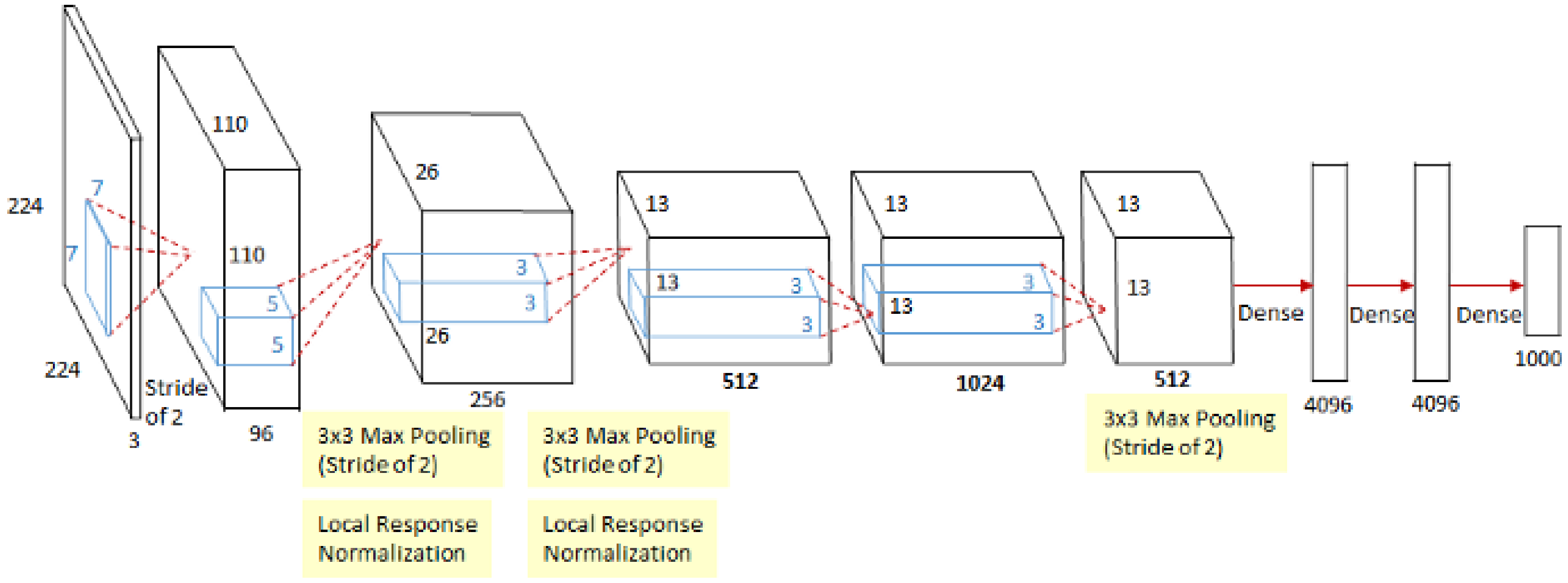
ZFNet

- ZFNet was invented by Dr. Rob Fergus and his PhD student at that moment, Dr. Matthew D. Zeiler in NYU, with the help of Yann LeCun.
- They found improvements to AlexNet, in particular in the first layers.



Img: copyright Kaiming He, 2016.

ZFNet



Images by: Sik-Ho Tsang (https://medium.com/coinmonks/paper-review-of-zfnet-the-winner-of-ilsvlc-2013-image-classification-d1a5a0c45103)

Conv1: change from (11x11 stride 4) to (7x7 stride 2)
Conv3,4,5: instead of 384, 384, 256 filters use 512, 1024,512