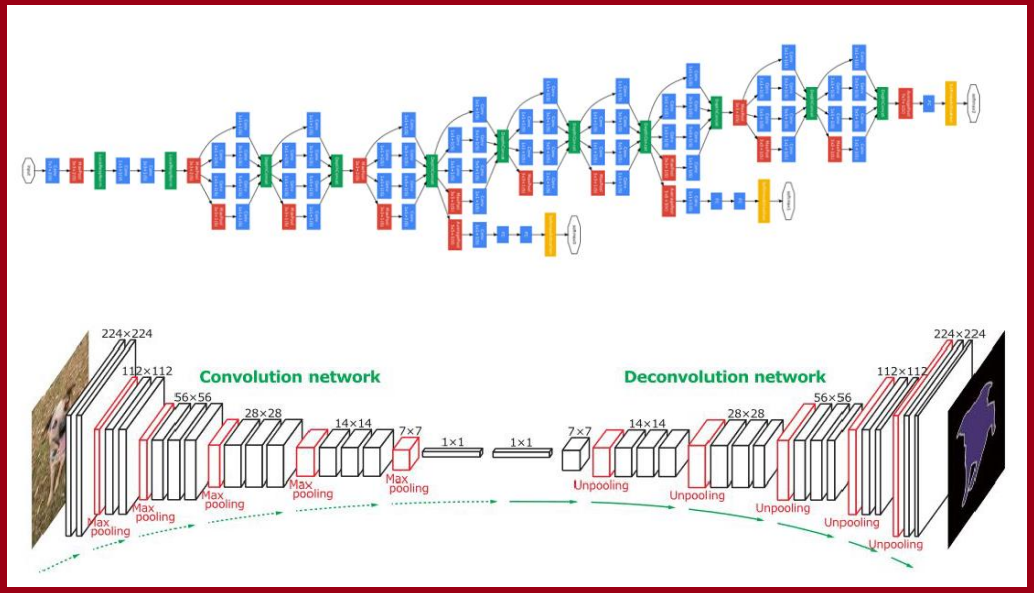
DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

UNIVERSITÀ
DEGLI STUDI
DI PADOVA



Deep Learning: Advanced Approaches

Machine Learning 2023-24

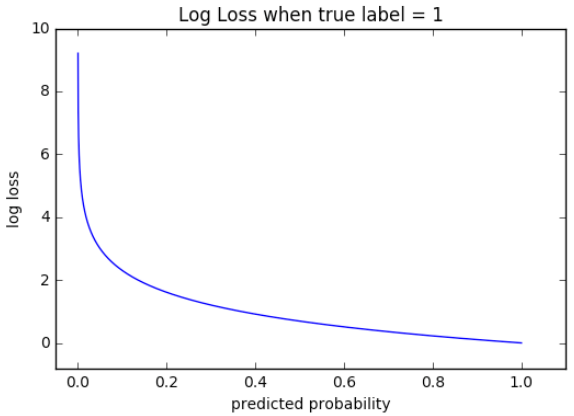
Slides P. Zanuttigh

Some slides from S. Fujimoto, I. Goodfellow and others

Loss Function: Cross Entropy

- For classification tasks the *cross entropy* is commonly used in place of the 0-1 loss
- For binary classification: $L(f(\boldsymbol{x}), y) = -y \log\big(f(\boldsymbol{x})\big) - (1-y)\log(1-f(\boldsymbol{x}))$
- The optimal $f(\boldsymbol{x})$ minimizing this loss function is $f(\boldsymbol{x}) = P(y = 1 \mid \boldsymbol{x})$
  - We are training the neural net output to estimate conditional probabilities
- Note that the expression works if $f(\boldsymbol{x})$ is strictly between 0 and 1
  - An undefined or infinite value would otherwise arise
  - To achieve this, the sigmoid is commonly used as activation for the output layer
- The function is convex

→ Gradient descent (e.g., SGD) works better



Extension to Multi-Class

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

state
NY
WA
CA

➡

AL	…	CA	…	NY	…	WA	…	WY
0	…	0	…	1	…	0	…	0
0	…	0	…	0	…	1	…	0
0	…	1	…	0	…	0	…	0

- One-hot encoding
  - Output: vector ***y*** with one component for each class
  - $y_i = 1$ if sample in class *i*, $y_i = 0$ otherwise
  - Avoid having some classes "closer" to others as when using class index
  - Increases output data dimensionality
- Extension of cross-entropy to multi-class
  - Labels one-hot encoded, vector function ***f*** to be estimated
  - $f_i(\boldsymbol{x})$ = estimated probability that ***x*** belong to class i

$$L(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) = -\sum_i y_i \log\left(f_i(\boldsymbol{x})\right)$$

In Practice: Many DL Tools.....

- Many deep learning frameworks
- Supported by large research entities and companies
- Optimized for GPU computing


Tensorflow (Google)


Keras: higher level framework for easier implementation


Caffe (University of Berkley)


PyTorch (Meta)

Microsoft Cognitive Toolkit

... and many others

Outline

Deep Learning: Advanced Approaches

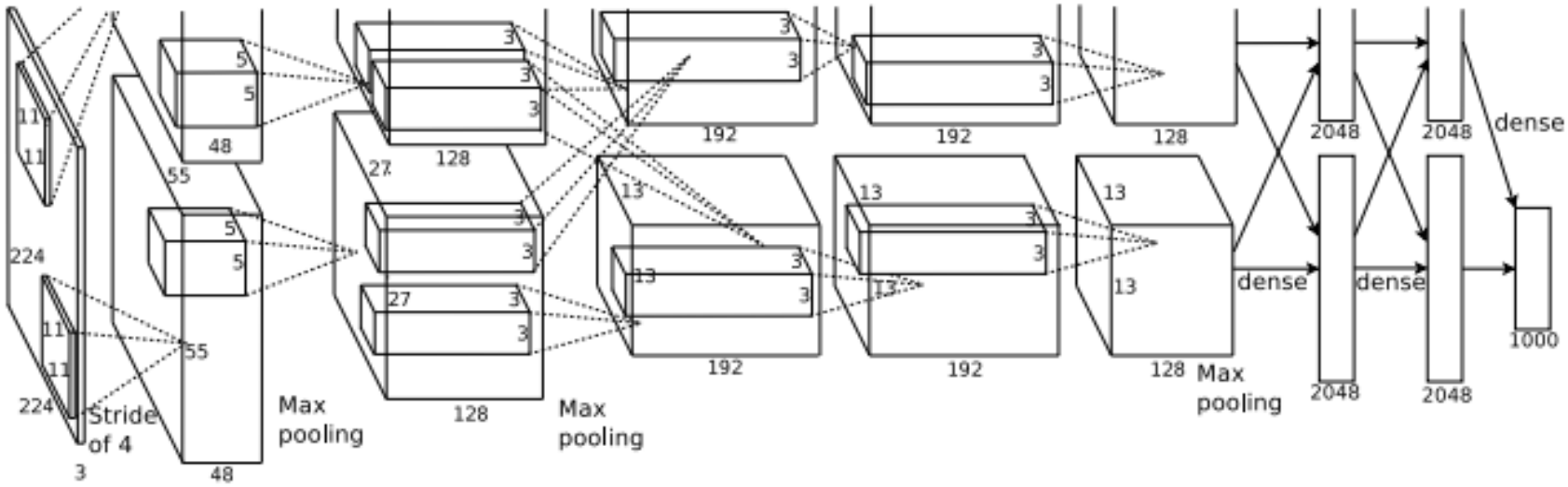
1. *Advanced CNN schemes:* Residual networks, skip connections, auto-encoders
2. *Generative models*: Generative Adversarial Networks (GAN)
3. *Modeling temporal information*: Recurrent Neural Networks (RNN) and Long-Short Term Memory (LSTM) *(not part of the course)*

Advanced CNN Models

- We'll see some relatively recent advanced architectures
- Some new concepts will be briefly introduced:
  - Residual Networks
  - Inception Modules
  - Transposed Convolutions

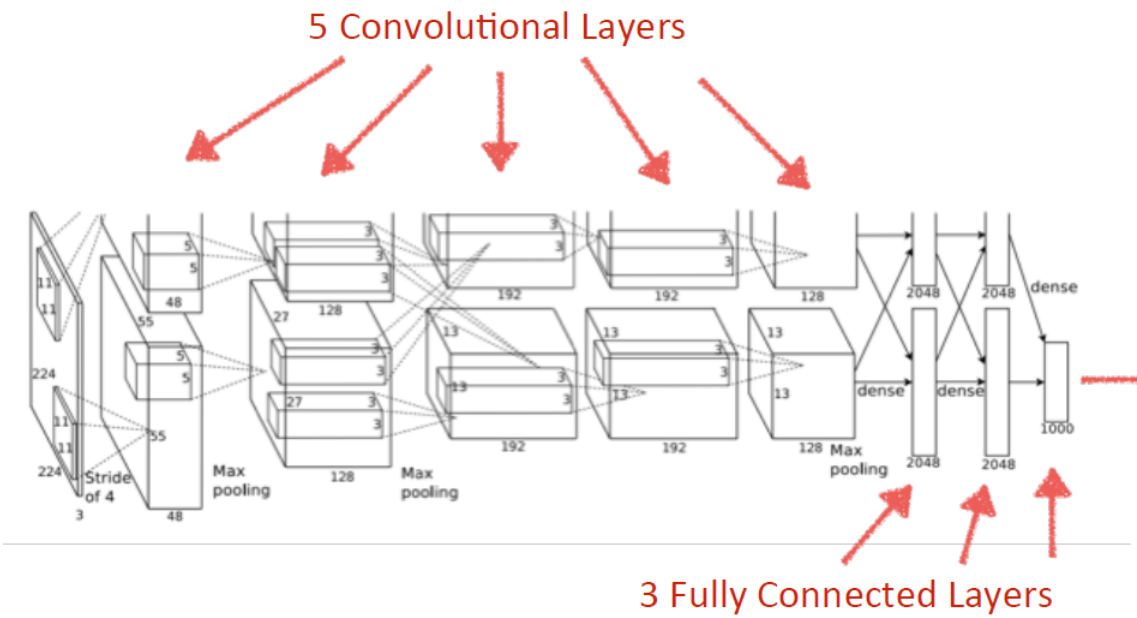
DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

«Historical» Perspective: AlexNet (2012)



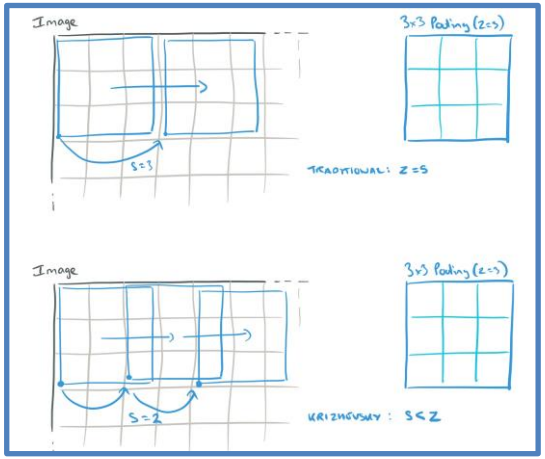
- *AlexNet* [1]: First Deep Learning approach outperforming "classic" ML methods on the image classification task (i.e., outperforming SVM and RF)
- Exploits 11x11, 5x5, 3x3, convolutions, max pooling, dropout, data augmentation, ReLU activations, SGD with momentum
- Split in 2 pipelines since it was trained with 2 GPUs (for 6 days)
  - According to Nvidia the DGX-2 server released in 2018 can train it in 18 mins!!!
- Complex but quite "standard" model

AlexNet: the Network



1000-way softmax

overlapping pooling

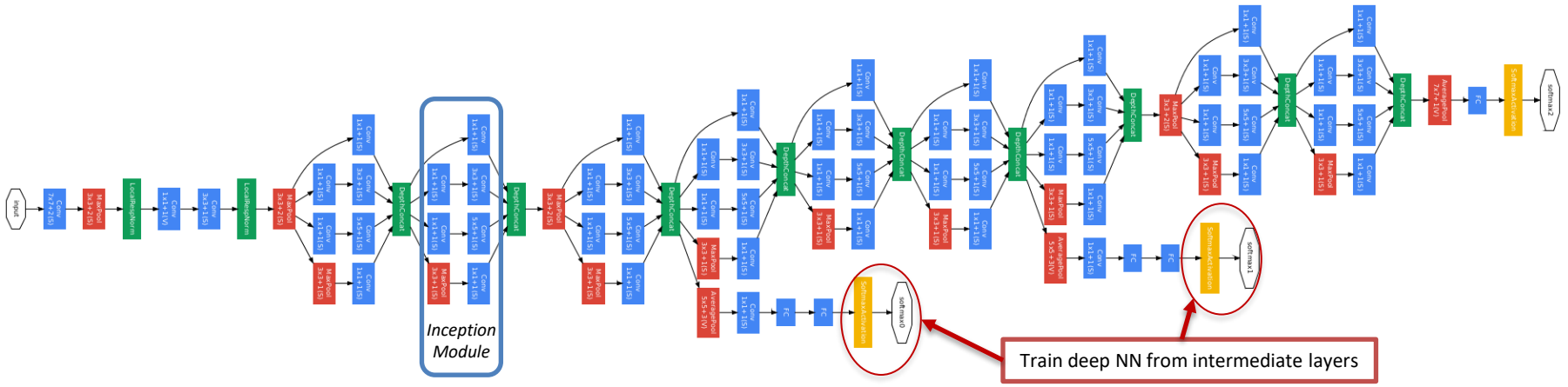


Softmax: maps output values to a set of values in [0,1] range summing up to 1

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{n_c} e^{z_j}}$$

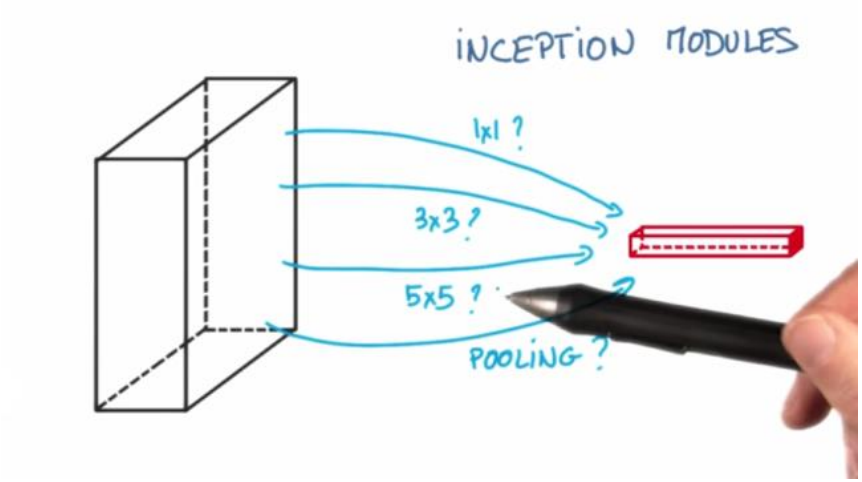
- 5 convolutional layers, 3 fully connected ones
- Many feature maps for each layer
- 650K neurons, 60M parameters
- Rectified Linear Units (ReLU) activations, overlapping pooling, dropout trick
- Training with randomly extracted 224x224 patches for more data

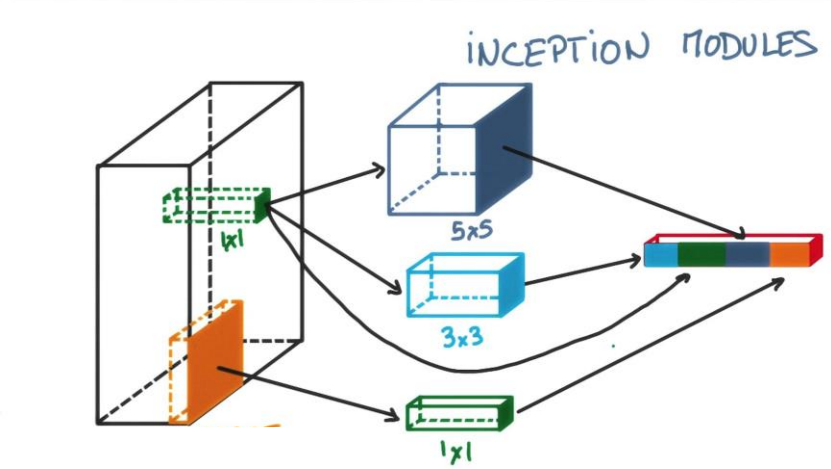
GoogleNet (Inception V1)

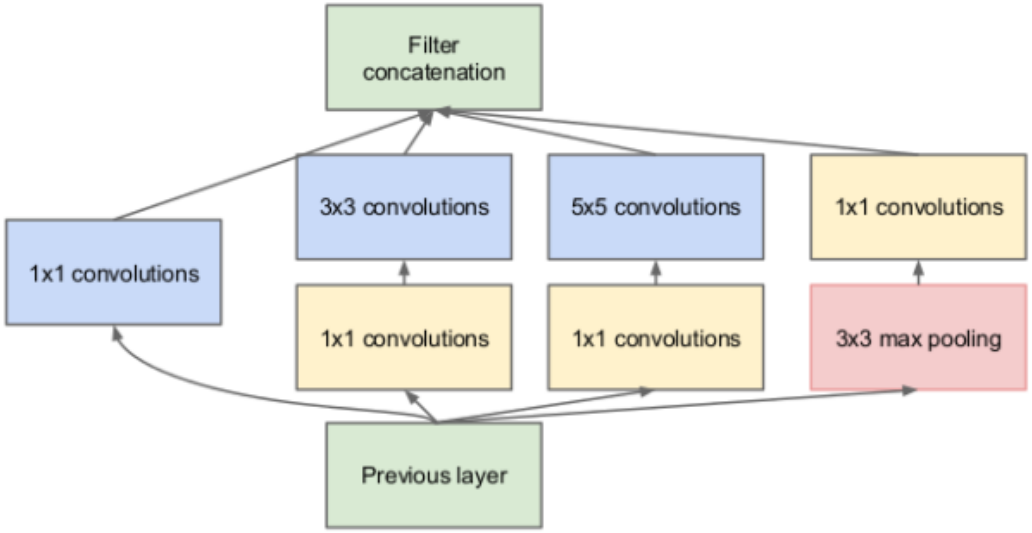


- Released in 2014, 1st method very close to human level performance on image classification
- Implemented a novel element: *the inception module*
  - This module performs multiple small convolutions with different sizes in parallel
- The networks is a 22 layers deep CNN but reduced the number of parameters from 60M of AlexNet to 4M

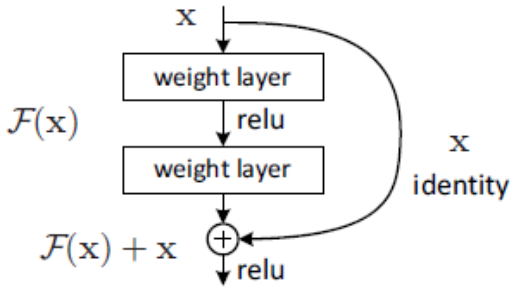
The Inception Module

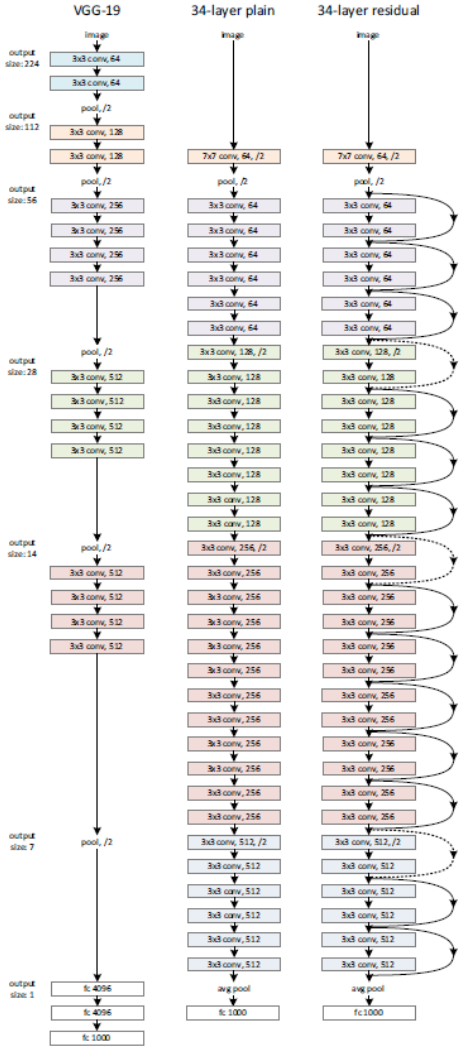






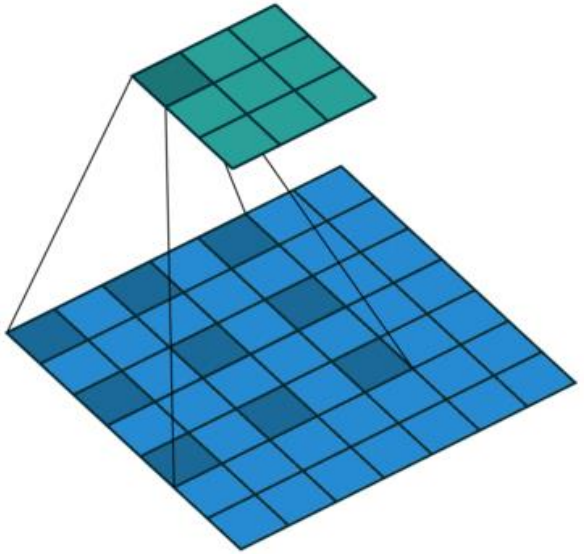
Residual Neural Networks (*ResNet*)





- Residual Neural Network [2] introduced in 2015 a novel architecture with "*skip connections*"
- Idea: try to estimate the residual w.r.t the previous estimation instead of the function itself
- Thanks to this technique they were able to train a NN with 152 layers with reasonable complexity
- Was able to beat human-level performance on image classification tasks

Dilated Convolutions



- Large convolutions have a wide receptive field but requires a lot of parameters
- Use dilated (*atrous*) convolutions, to increase the field of view without increasing the spatial dimensions
- The convolution works on samples spaced apart with a regular step instead of over each single sample in the window.

Multi-Scale Context Aggregation by Dilated Convolutions (2015)

Many Other Approaches….

- This was just a quick overview of some relatively recent results
  - For ICT students more approaches will be presented in computer vision, neural networks and deep learning and many other courses….
- Huge amount of resources is currently spent on Deep Learning research
- Many other schemes exist
- And every month there is a new one outperforming previous results !!!