Computer Vision
Architectures of convolution neural networks - lecture 8

Adam Szmigielski
aszmigie@pjwstk.edu.pl
materials: $ftp(public) : //aszmigie/WMAEnglish$

Popular convolutional network architectures

- RNN - fast RNN, faster RNN
- SSD
- YOLO

LeNet-5 1998

Layer	Type	Maps	Size	Kernel Size	Step	F. Activation
Out	Fully connected	—	10	—	—	RBF
F6	Fully connected	—	84	—	—	tanh
C5	convolutions	120	1×1	5×5	1	tanh
S4	Connecting	16	5×5	2×2	2	tanh
C3	convolutions	16	10×10	5×5	1	tanh
S2	connecting	6	14×14	2×2	2	tanh
C1	convolutions	6	28×28	5×5	1	tanh
In	Input	1	32×32	—	—	—

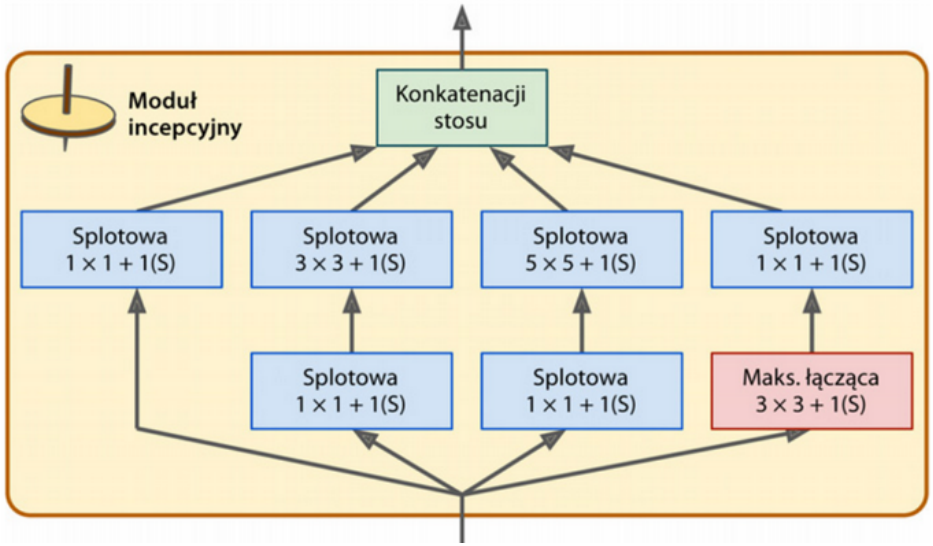
- Została stworzona przez Yanna LeCuna w 1998 roku do rozpoznawania odręcznie pisanych cyfr (MNIST)
- Architektura LeNet-5 stanowi prawdopodobnie najbardziej znany przykład sieci CNN.

AlexNet 2012

Layer	Type	Maps	Size	Kernel Size	Step	Build. with zeros	F. activation
Out	Fully connected	—	1000	—	—	—	Softmax
F9	Fully connected	—	4096	—	—	—	ReLU
F8	Fully connected	—	4096	—	—	—	ReLU
C7	convolutions	256	13×13	3×3	1	SAME	ReLU
C6	convolutions	384	13×13	3×3	1	SAME	ReLU
C5	convolutions	384	13×13	3×3	1	SAME	ReLU
S4	connecting	256	13×13	3×3	2	VALID	—
C3	convolutions	256	27×27	5×5	1	SAME	ReLU
S2	connecting	96	27×27	3×3	2	VALID	—
C1	convolutions	96	55×55	11×11	4	SAME	ReLU
In	Output	3(RGB)	224×224	—	—	—	—

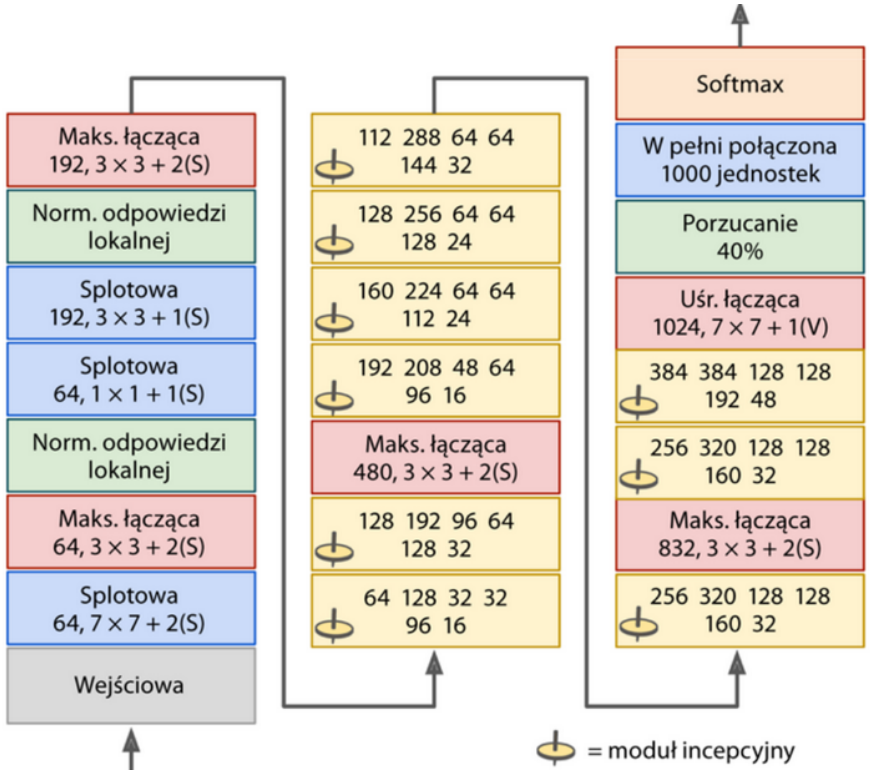
- In order to reduce overtraining, the drop-off method was used on the outputs of the F8 and F9 layers, and they generated data,
- Local response normalization (LRN) was used in layers C1 and C3 - neurons with the highest weights (the most activating) inhibit neurons located in the same position,

GoogLeNet 2014



- The creation of such an architecture was possible due to the introduction of subnets called bf inception modules
- The second set of convolution layers contains kernels of different sizes (1×1, 3×3, and 5×5), which allows them to catch patterns at different scales.

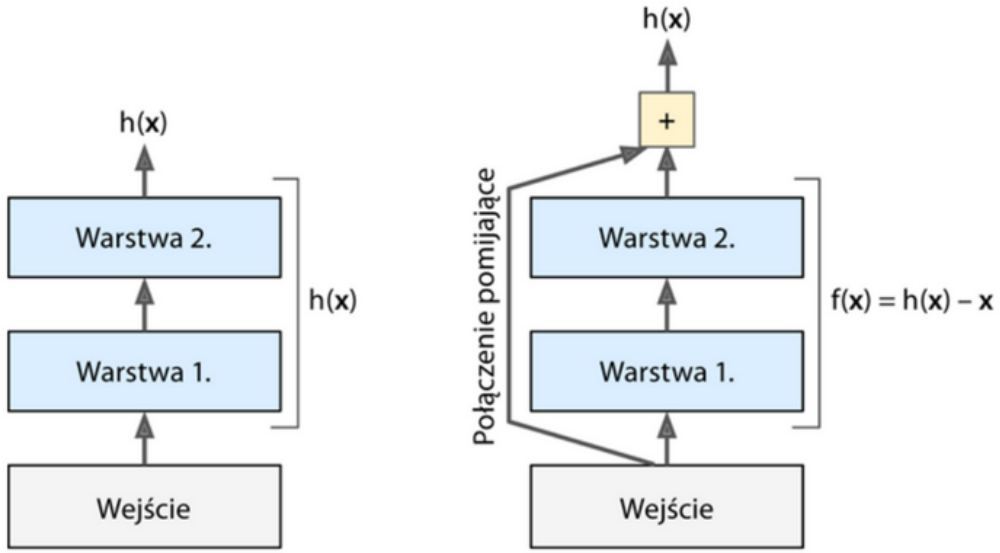
GoogLeNet network architecture



- The first two layers reduce the height and length of the image four times,
- The local response normalization layer instructs previous layers to learn to recognize very different characteristics

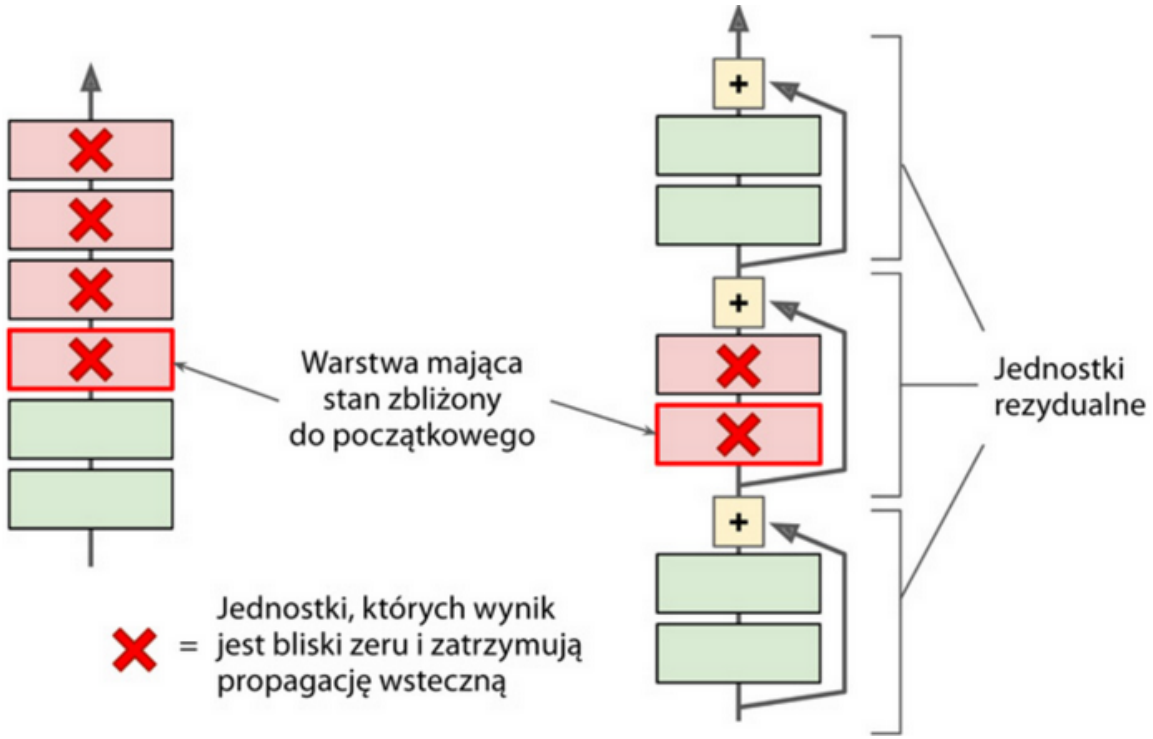
- Two convolution layers, the first of which acts as the boundary layer.
- Local response normalization layer,
- Maximizing blending layer reduces the image dimensions twice as much,
- A tall stack of nine inception modules that
- Dropping is used to regularize, then a fully connected layer using the softmax activation function to display the estimated probabilities of the examples belonging to a given class.

ResNet 2015 - Residual Network



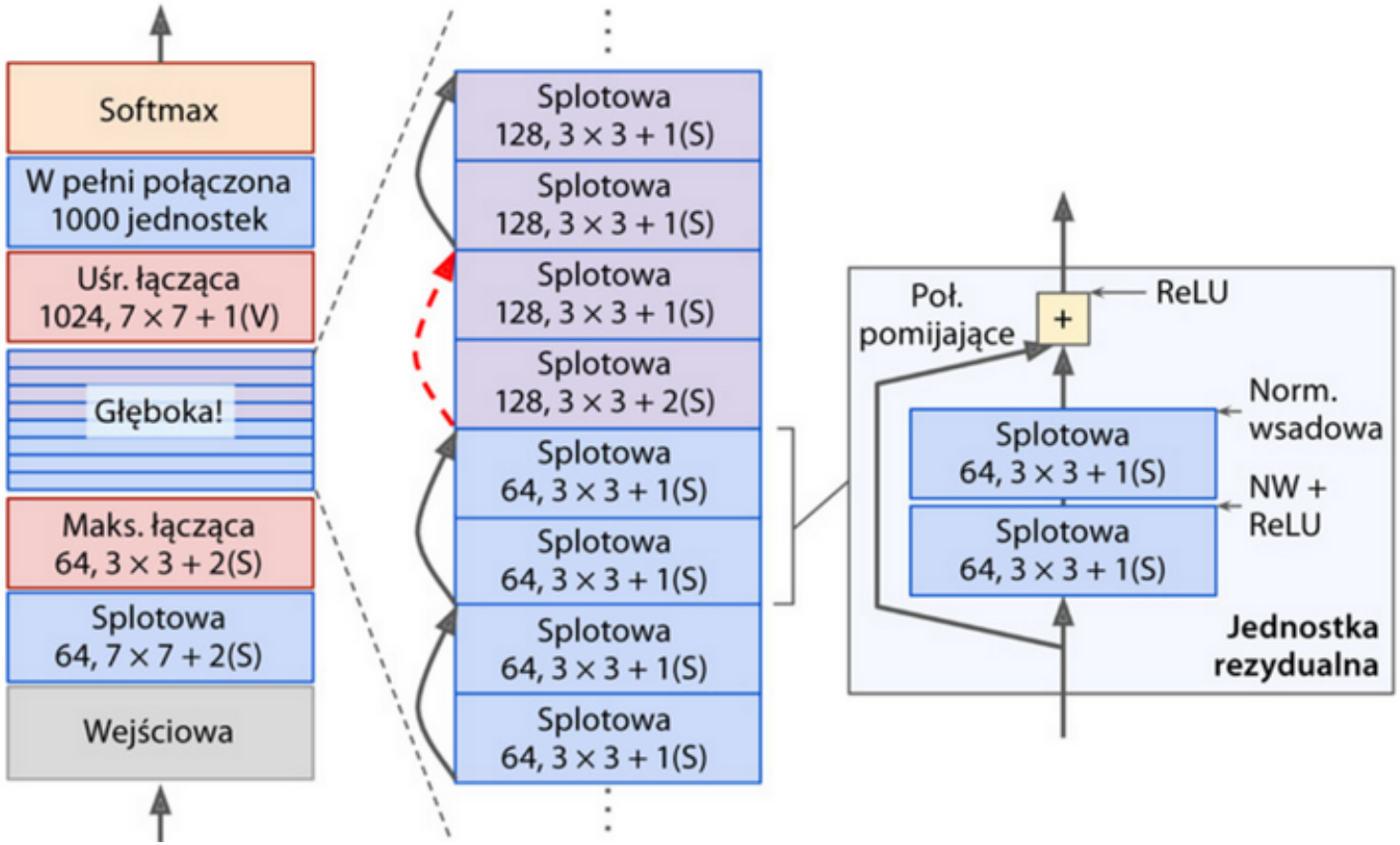
- Deep architecture was used - consisting of 152 layers.
- **connections bypassing** (also called shortcut connections) used
- If we add x input to the network output (skipping connection), the network maps the function $f(x) = h(x) - x$ - **residual learning**.

Deep residual network



- A deep residual network can be viewed as a stack of residual units, or small neural networks containing a bypass connection.

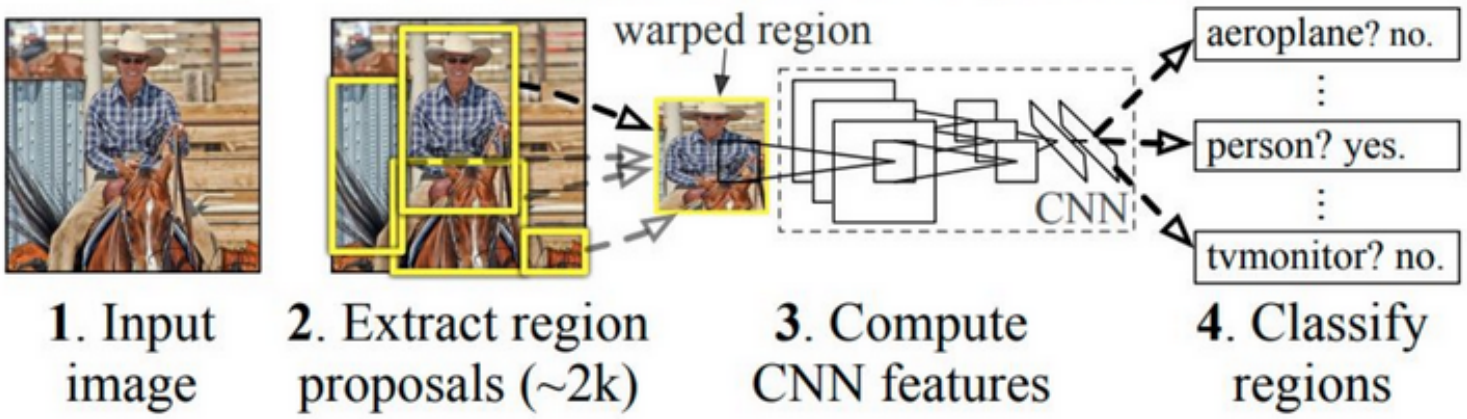
Architecture of the ResNet network



Object detection using convolutional networks

- R-CNN (Regions with CNN features) - fast RNN, faster RNN
- SSD (Single Shot Multibox Detector)
- YOLO (You Only Look Once)

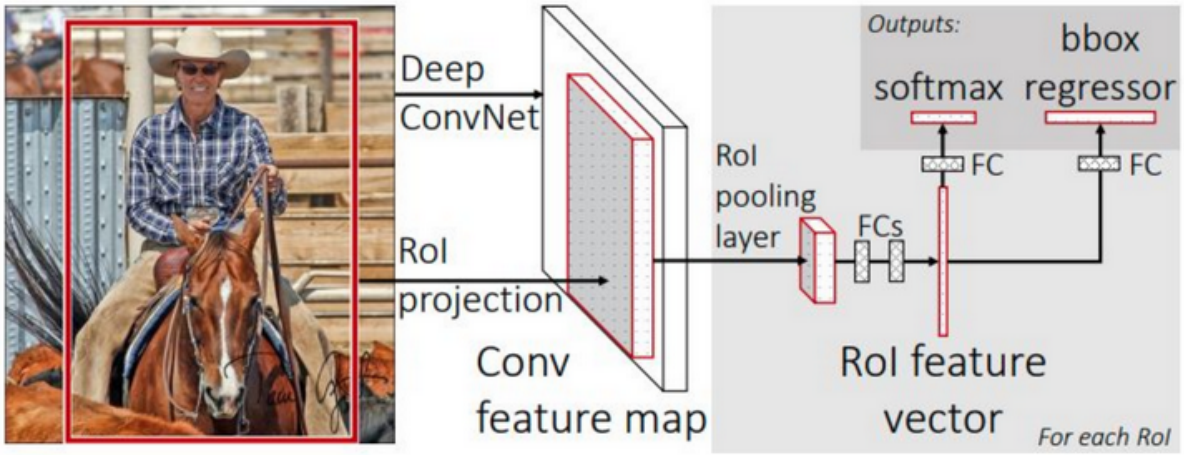
Region convolutional neural network - R-CNN



- **Region proposal** - generating regions where the object can be located,
- **Feature selection module** - feature extraction from each region,
- **Classifier** - classifies features into one of the known classes using a classifier.

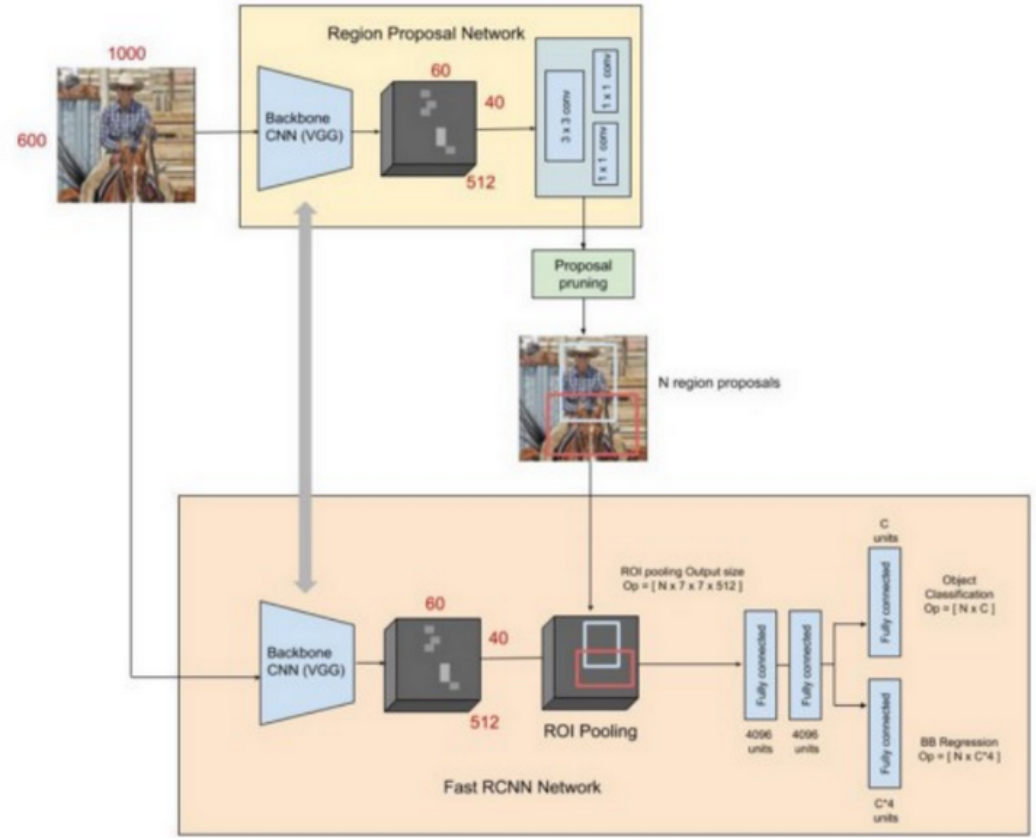
Long training (approx. 2000 region suggestions for each image). Searching for regions is not a learning algorithm - it can lead to wrong propositions and this error will not be corrected.

Architecture of the Fast R-CNN



- Single model to learn region and classification proposals at once,
- Uses a pre-trained network to select features.
- At the end of the network is a Region of Interest Pooling Layer (RoI Pooling) that extracts the features for each proposed region.
- Then this data is interpreted in a layer fully connected to two outputs - classifier and **bounding box**

Architecture of the Faster R-CNN model

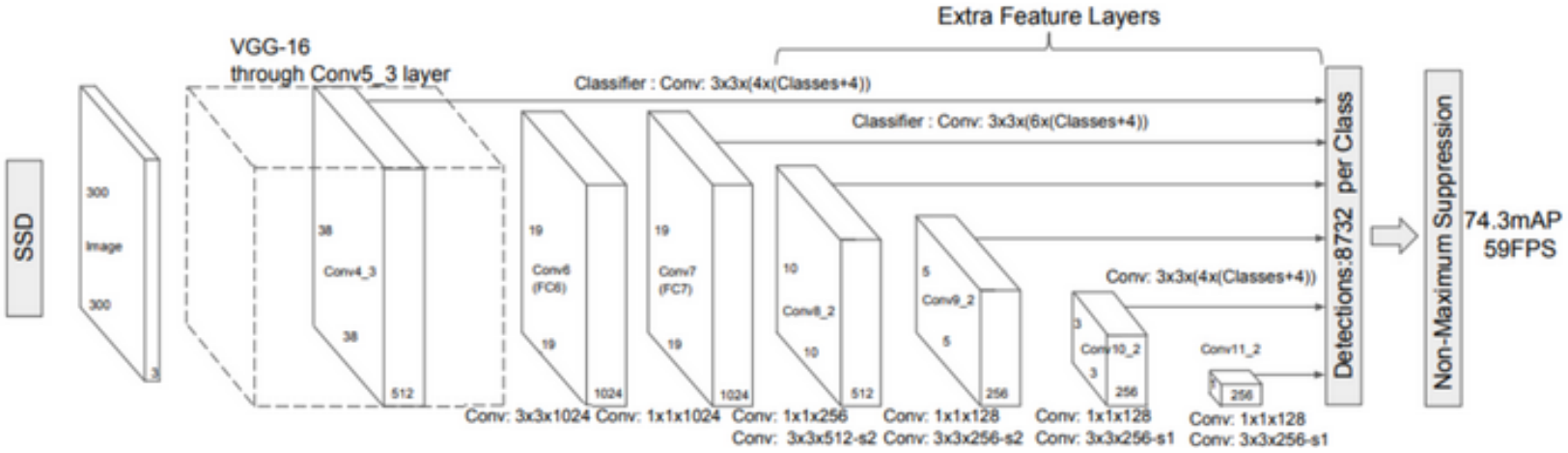


- The main innovation of Faster R-CNN is the application of the Convolutional Network for Region Proposal Network (RPN),
- The RPN process begins with passing the entire image through the

convolutional layers.

- By using RPN instead of selective search, we create a fully learning network that can also share parameters between the convolutional layers in the RPN and the Fast R-CNN detector.
- In Faster R-CNN, the ROI Pooling layer takes the region suggestions not from the selective search results but from the RPN process

Single Shot Multibox Detector SSD 2016

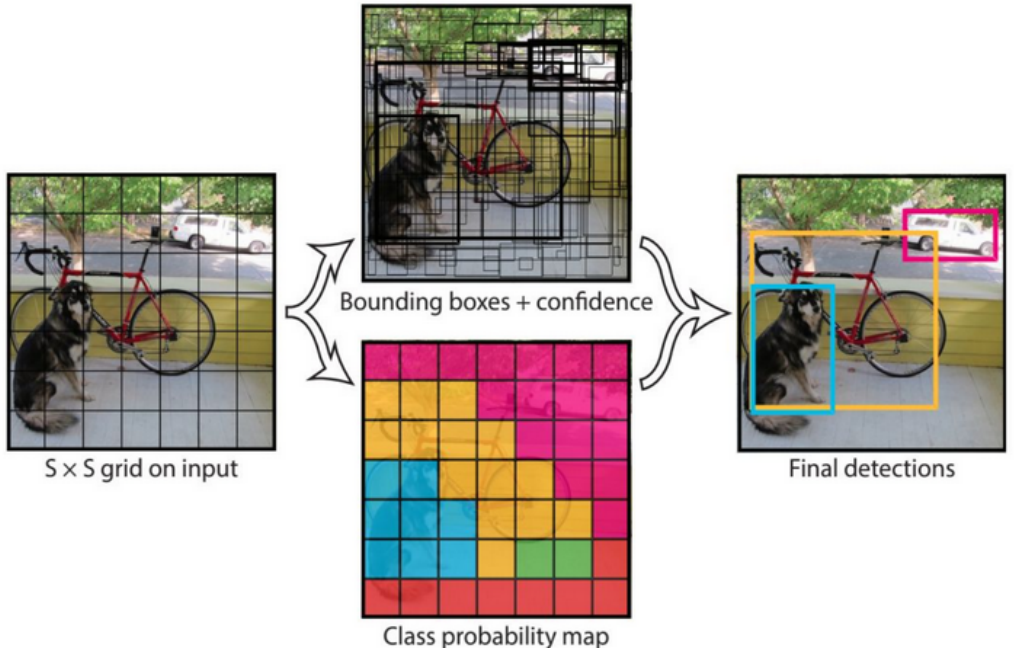


- SSD is an object detection method based on the use of a single deep network,
- At the time of prediction, the network generates results for the presence of each category in the default areas
- The network combines predictions from multiple feature maps with different resolutions,

  SSD has two backbone components and SSD layers - Extra Feature Layers.

- The base model is usually a pre-trained image classification network,
- SSD does not use the panes moving across the image, it splits the image into a grid and each the cell in this mesh is responsible for detecting the objects
- If no object is detected, this cell becomes a background class cell.
- Tools such as the anchor box and receptive field are used, to easily detect many objects within one cell.

YOLO - You Only Look Once 2016



- `https://storage.googleapis.com/openimages/web/index.html`
- It uses features trained by deep neural networks in the Darknet framework.
- It is a fully convolutional network (FCN) - it is indifferent to the size of the input image.

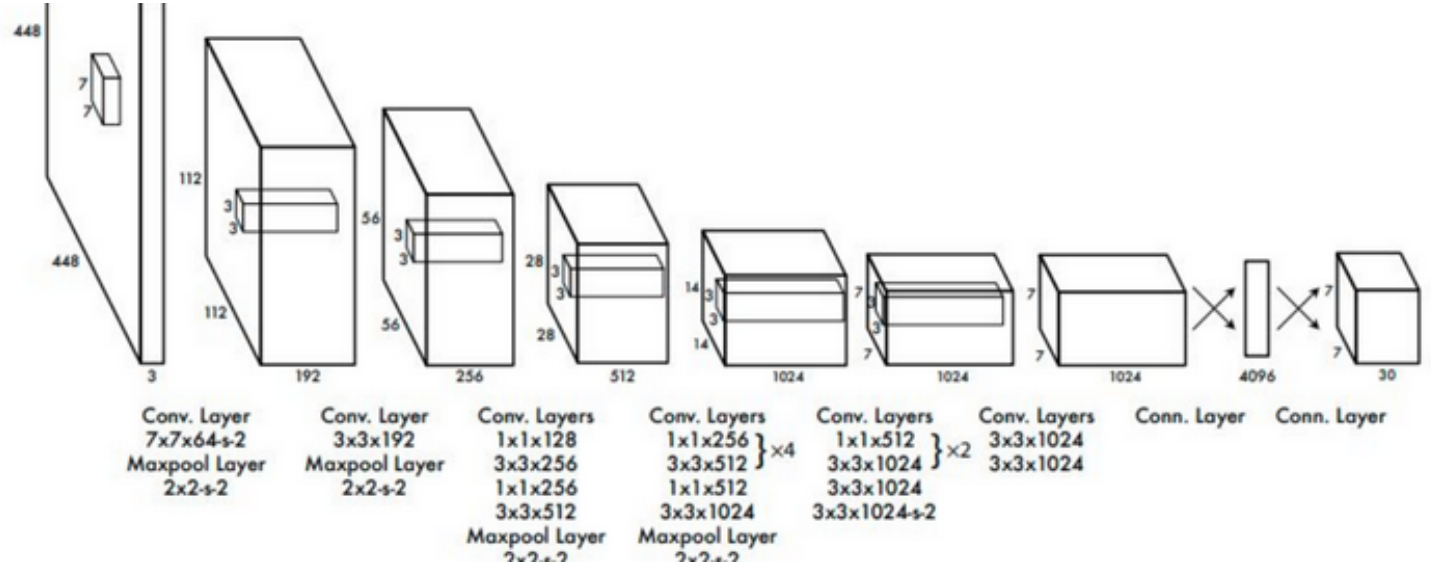
YOLO network performance

- The image is split into cells that form the $S \times S$ grid.
- The input to the YOLO network is the image, and the output is the coordinates x, y of the center of the area containing the object - the bounding box (it bounding box) in this image, its width, height, and the probability for the class detected,
- If the center of an object falls within a grid cell, that cell becomes responsible for detecting that object.
- Each cell detects the B envelope and their confidence scores.
- Confidence intervals define the uncertainty of the detection of the object and the envelope,
- Once you have all possible boundaries, get rid of those that have a low probability of owning an object.

- For this, the non-maximum suppression algorithm is used to calculate the Jaccard index:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

YOLOv1

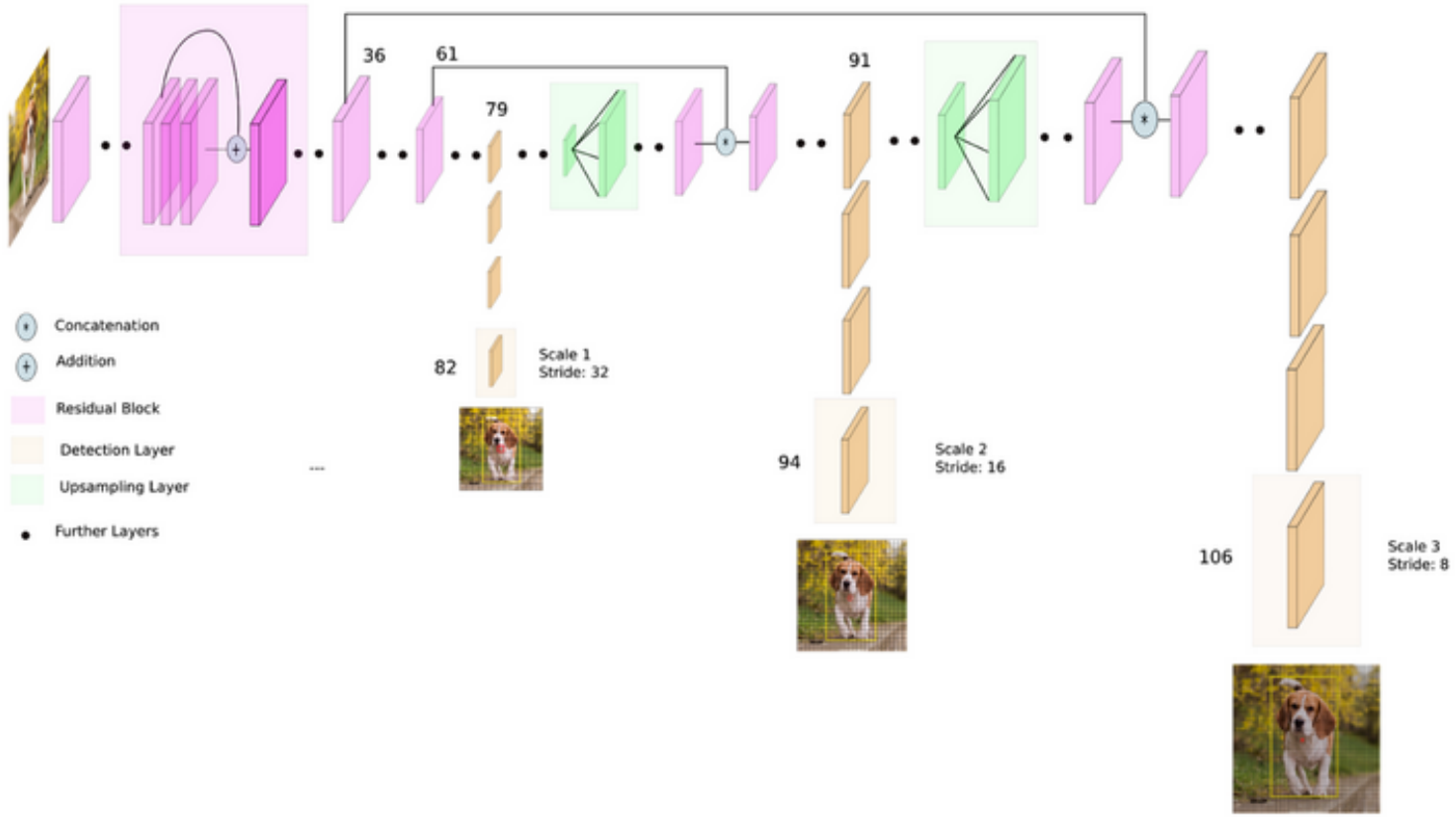


- The network has 24 convolution layers and 2 fully interconnected layers, 2 reducing layers (max-pooling).
- It is a convolutional network that simultaneously predicts multiple areas and class probabilities for those areas.
- YOLOv1 produces poorer results when the image contains many small objects, such as a flock of birds

YOLOv2 - YOLO9000 2016

- In the second version of the algorithm, YOLO detection is based on the also improved, new *Darknet19* model,
- Has fewer convolution layers and 5 reducing layers and a softmax layer,
- Envelope base anchors have been introduced: in YOLOv2 the envelopes have been introduced, as e.g. in Faster R-CNN - the number of envelopes per image increased from 98 to over a thousand,
- The detection of small objects has been improved - it creates detections on the map of features with dimensions of 13×13,
- Training with images at different scales.

YOLOv3 - 2018



- YOLOv3 is based on the improved *Darknet53* architecture - it has 53 convolutional layers,
- Predictions at different scales - the third version network predicts envelopes at three different scales of the image,