Advanced Neural Networks for Computer Vision

Dr Amit Sethi, IITB

Module objectives

- Identify problems other than image classification
- Match advanced NN architectures suitable for these problems
- Design training data and methods for training these architectures

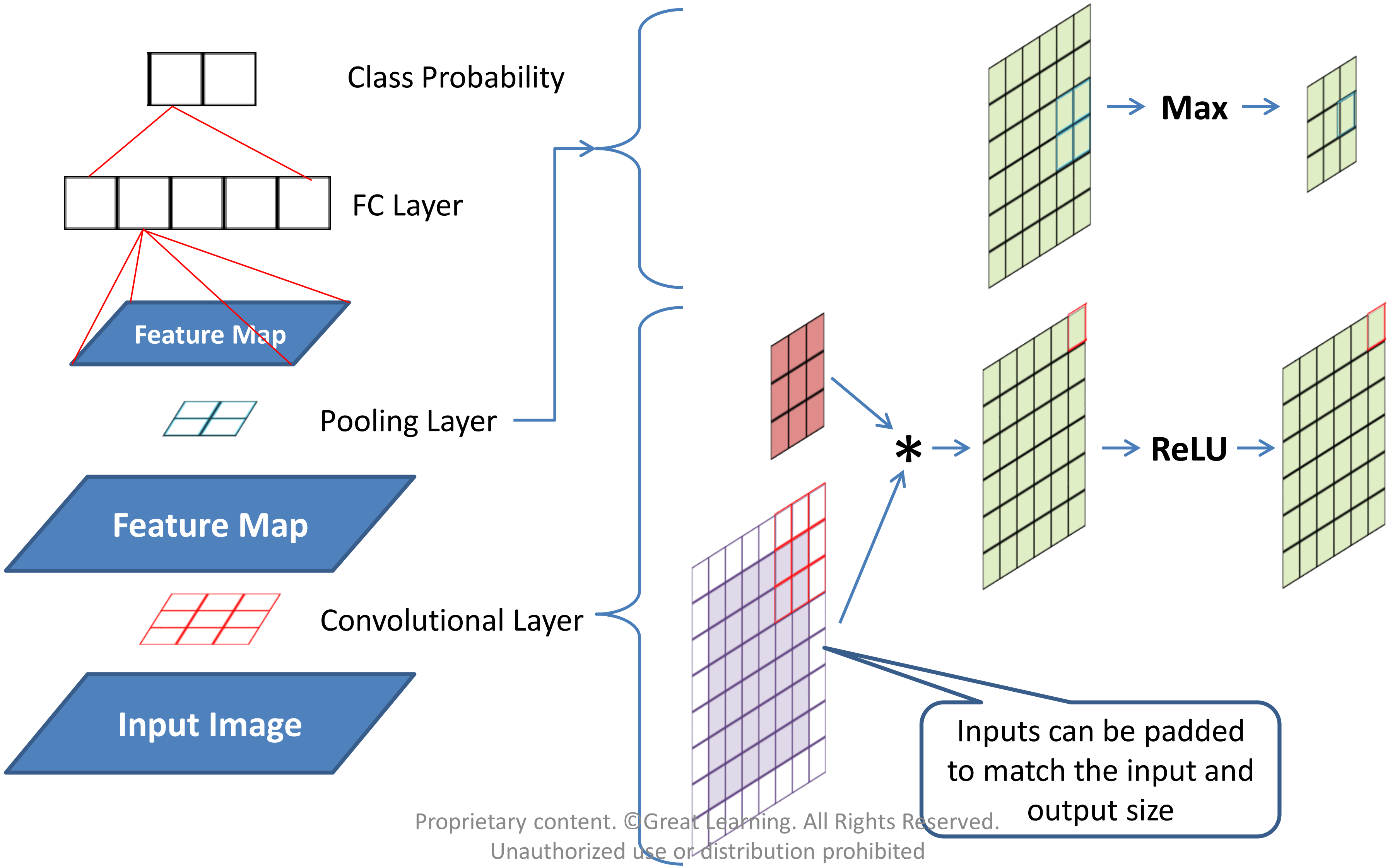
Contents

- FCNs and semantic segmentation
- Other variants of convolution
- Simultaneous localization and recognition
- Siamese network for metric learning

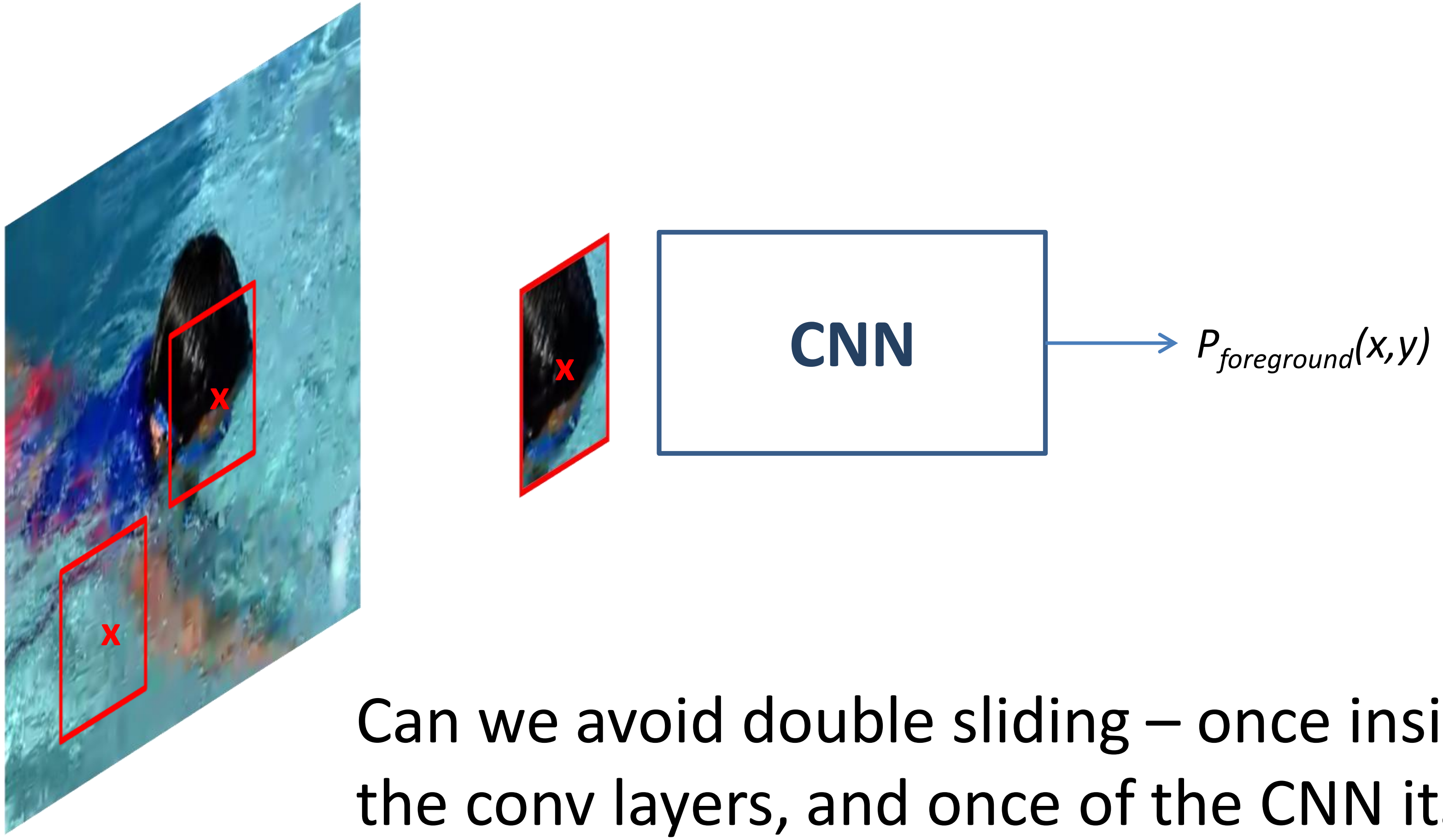
Semantic segmentation is labeling pixels according to their classes

Image Source: "ICNet for Real-Time Semantic Segmentation on High-Resolution Images" Hengshuang Zhao1, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, Jiaya Jia, ECCV'18

CNN Revisited



For segmentation, a pixel class can be predicted using some spatial context



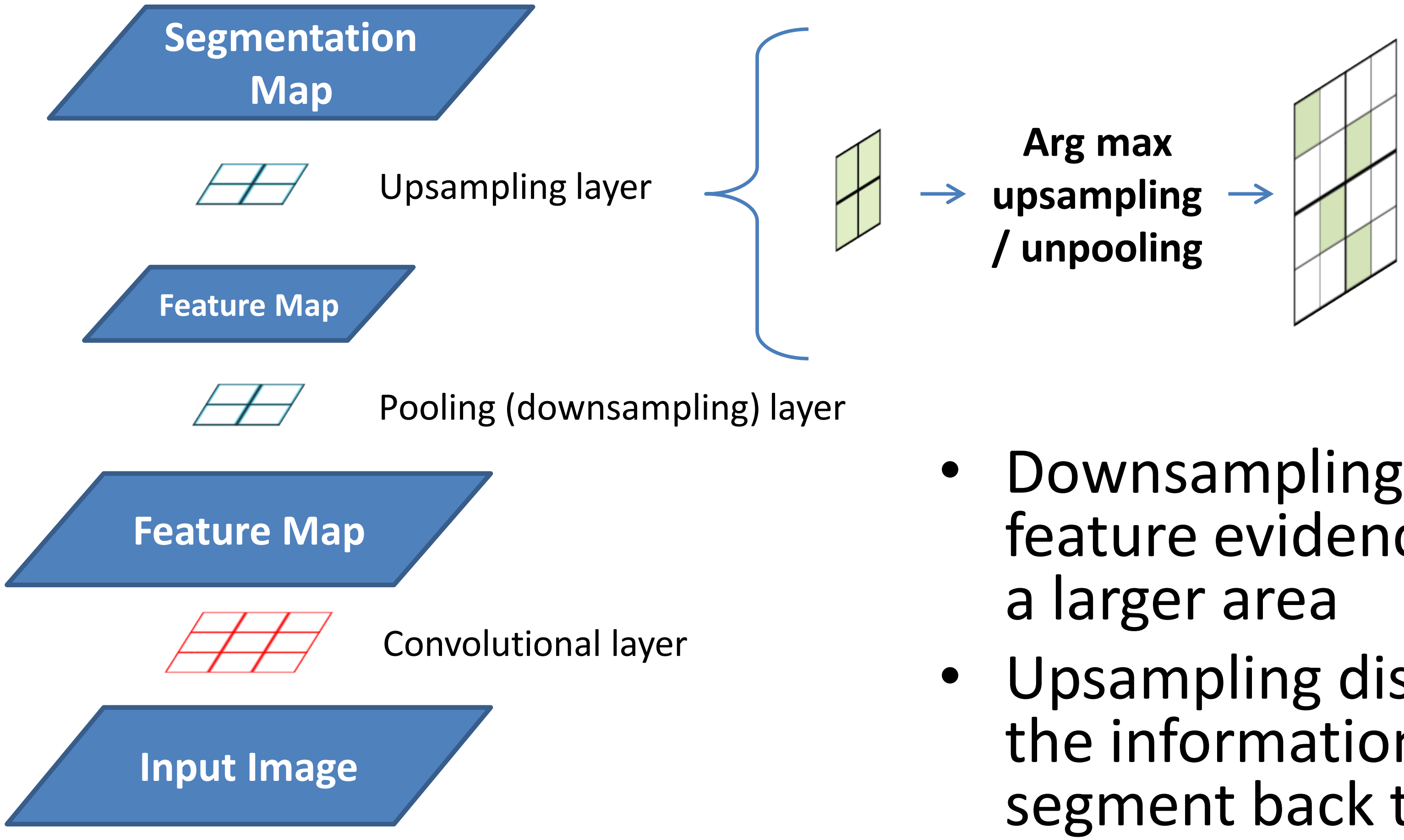
Can we avoid double sliding – once inside the conv layers, and once of the CNN itself?

Pixel labels for training images must be known to train for semantic segmentation

void	road	sidewalk	building	wall
fence	pole	traffic light	traffic sign	vegetation
terrain	sky	person	rider	car
truck	bus	train	motorcycle	bicycle

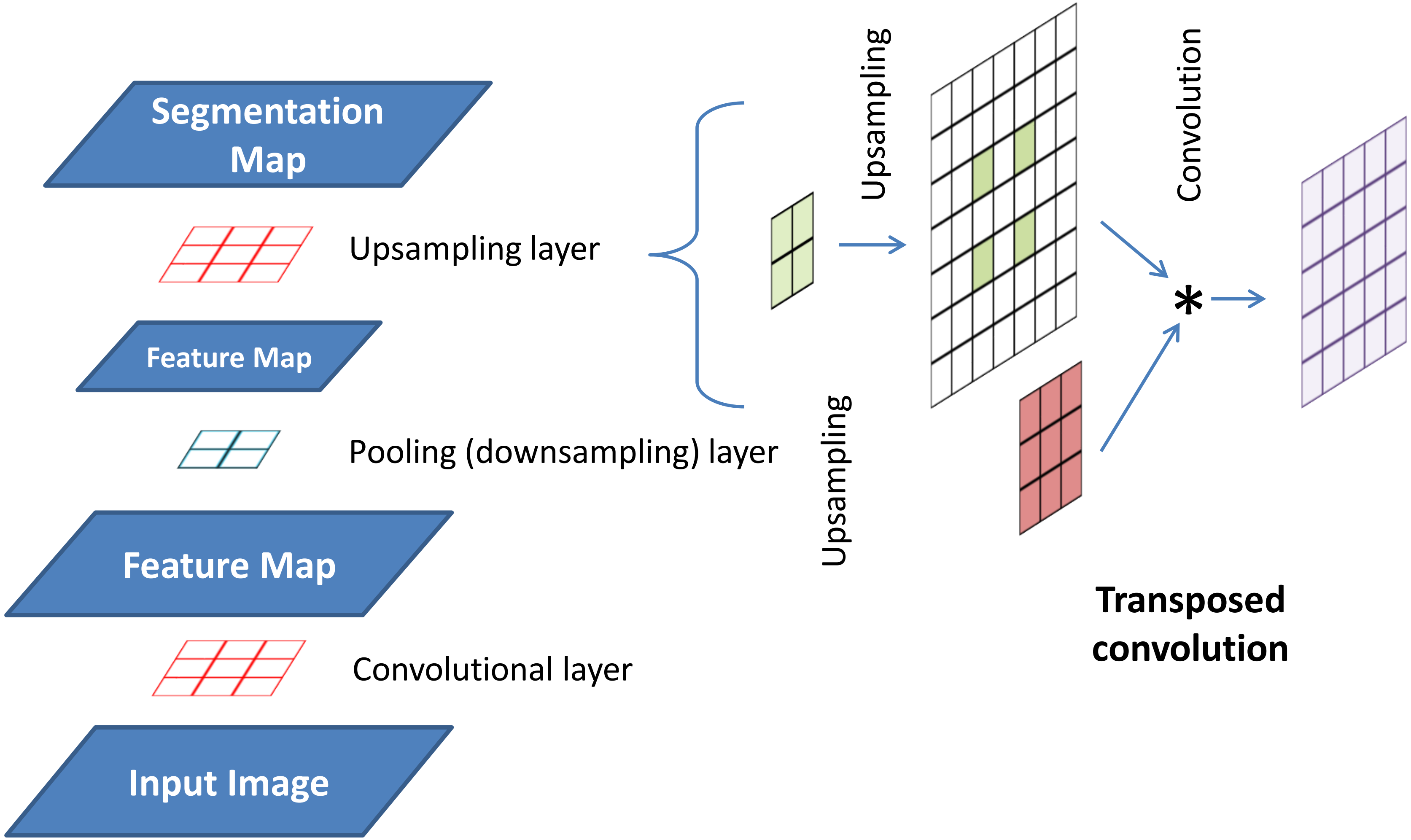
Image Source: "ICNet for Real-Time Semantic Segmentation on High-Resolution Images" Hengshuang Zhao1, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, Jiaya Jia, ECCV'18

To produce a segmentation map downsampling is followed by upsampling

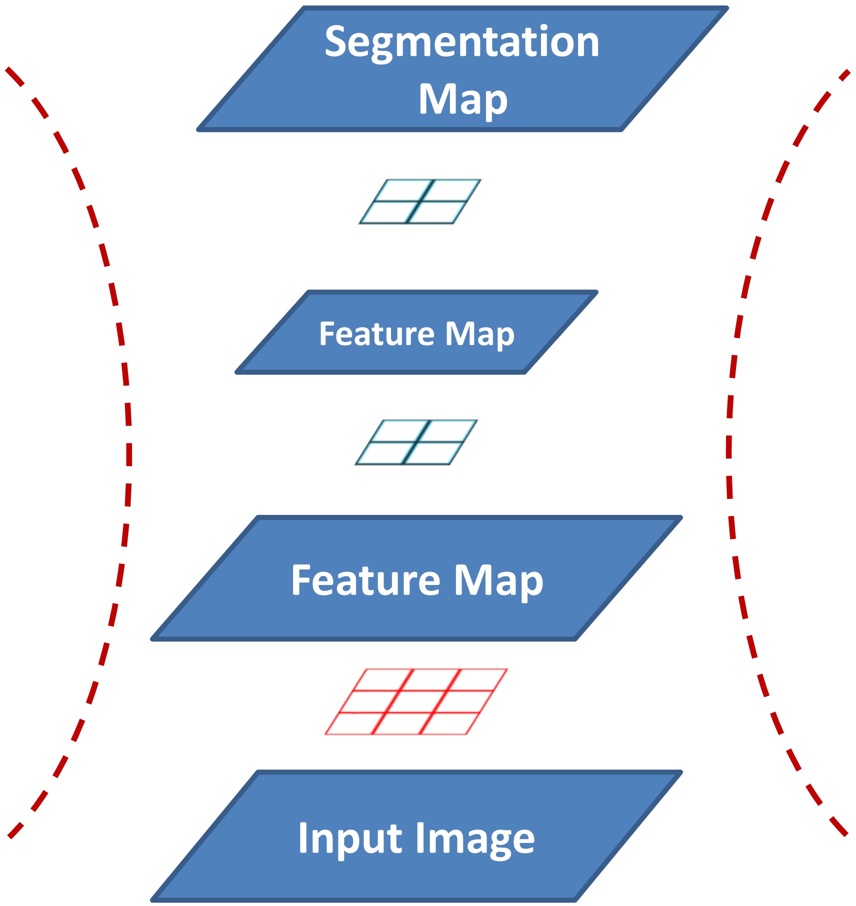


- Downsampling collects feature evidence from a larger area
- Upsampling distributes the information of a segment back to the original pixel domain

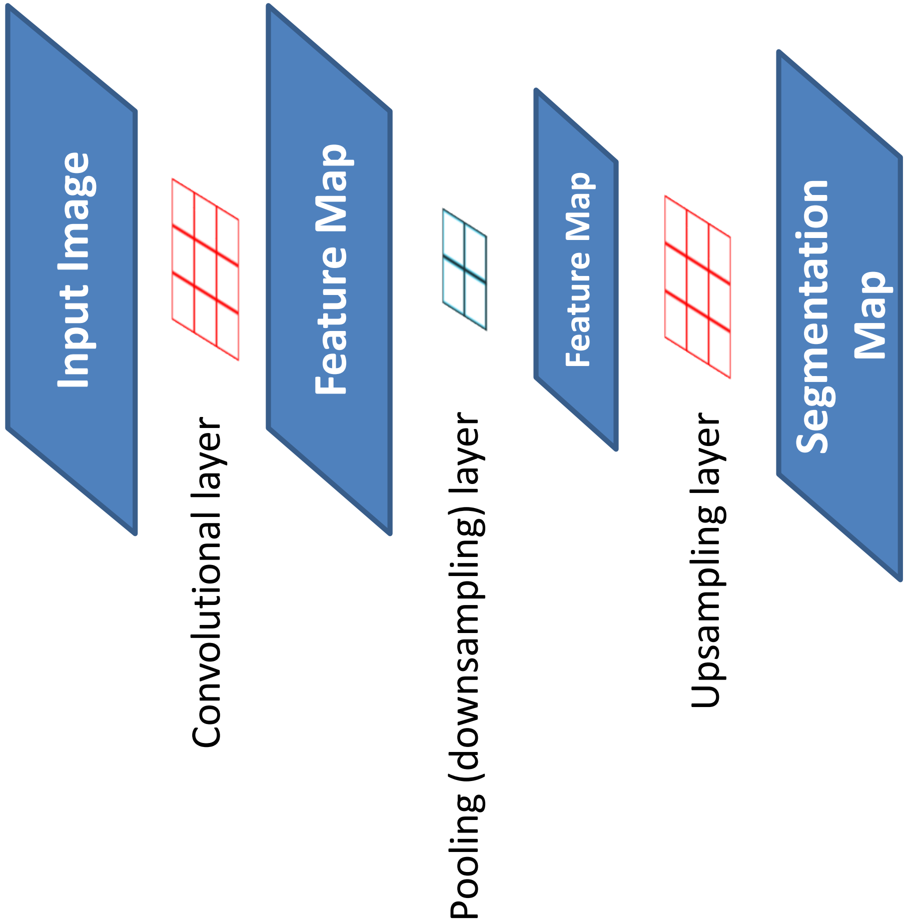
Upsampling can also be learned



Downsampling and upsampling leads to an hour-glass structure

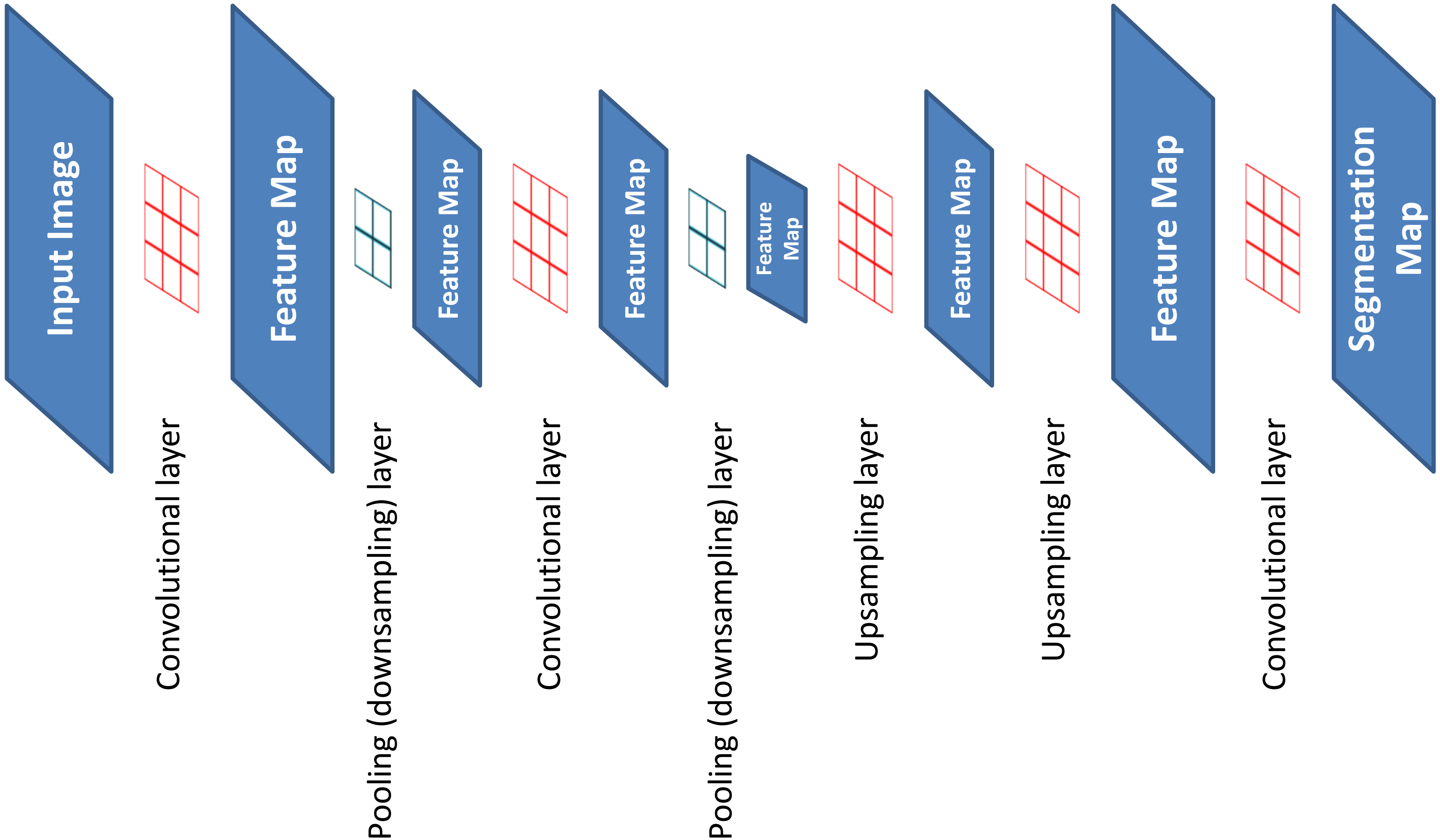


Let us rearrange the layers horizontally

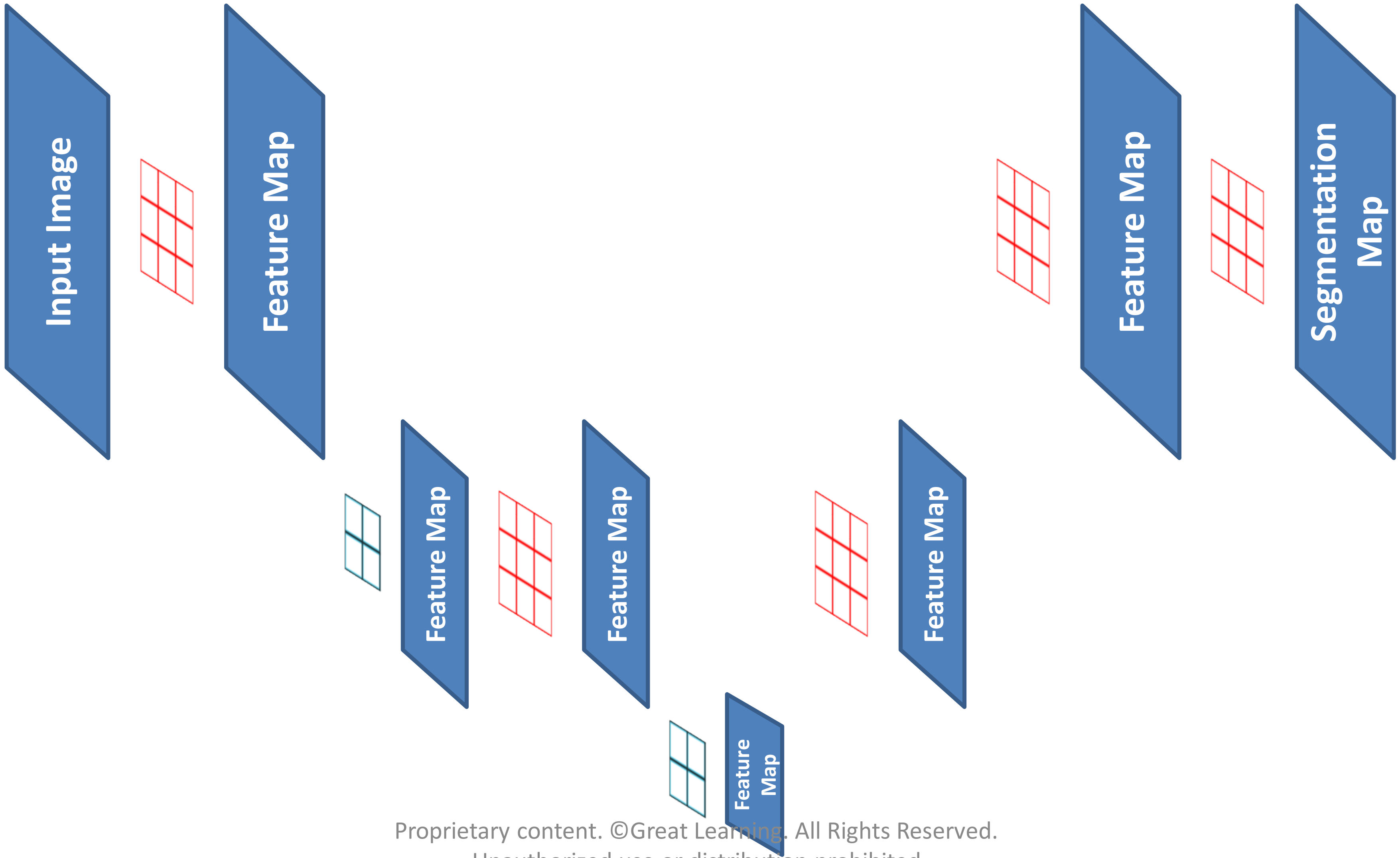


greatlearning
Learning for Life

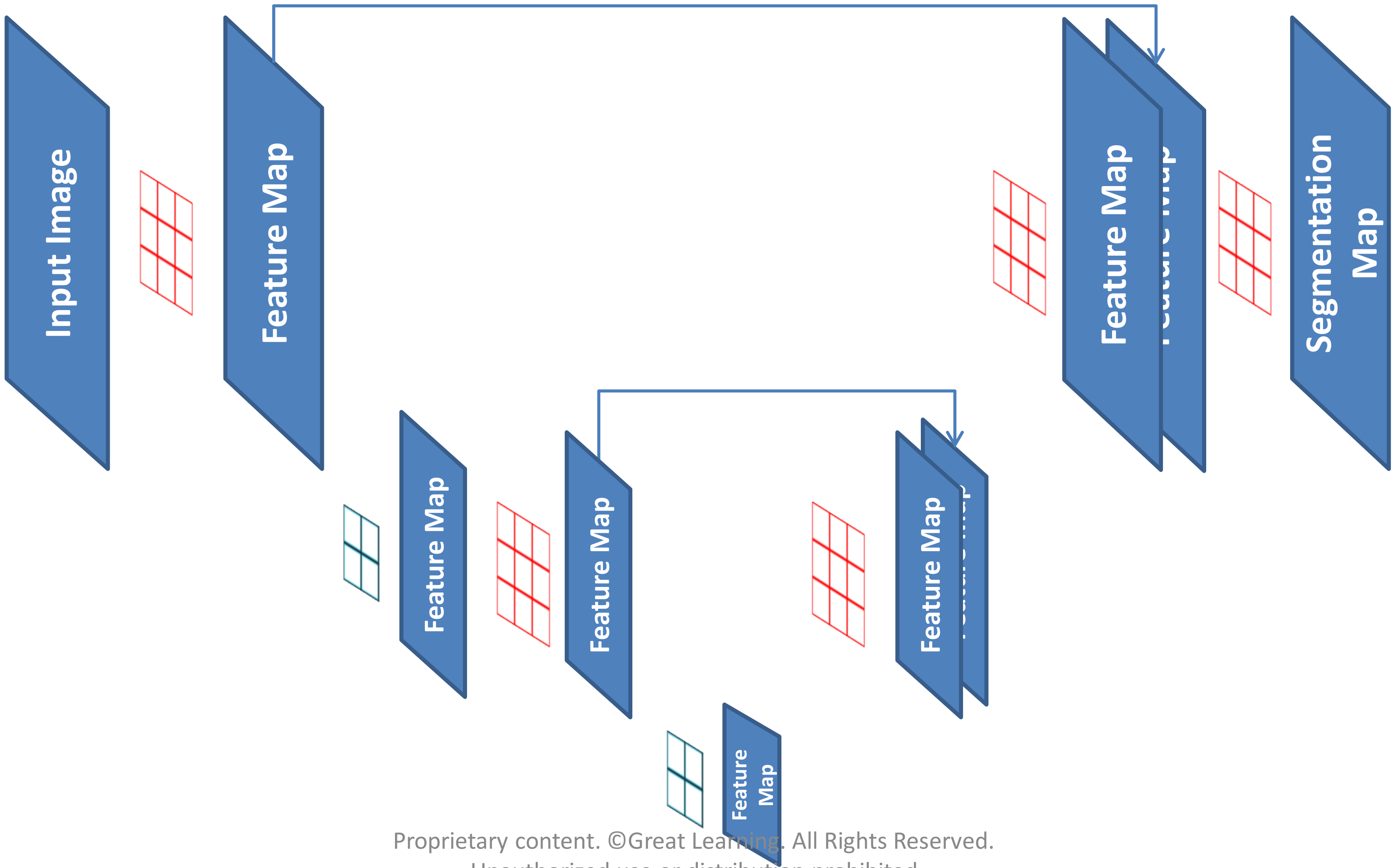
More layers can be added



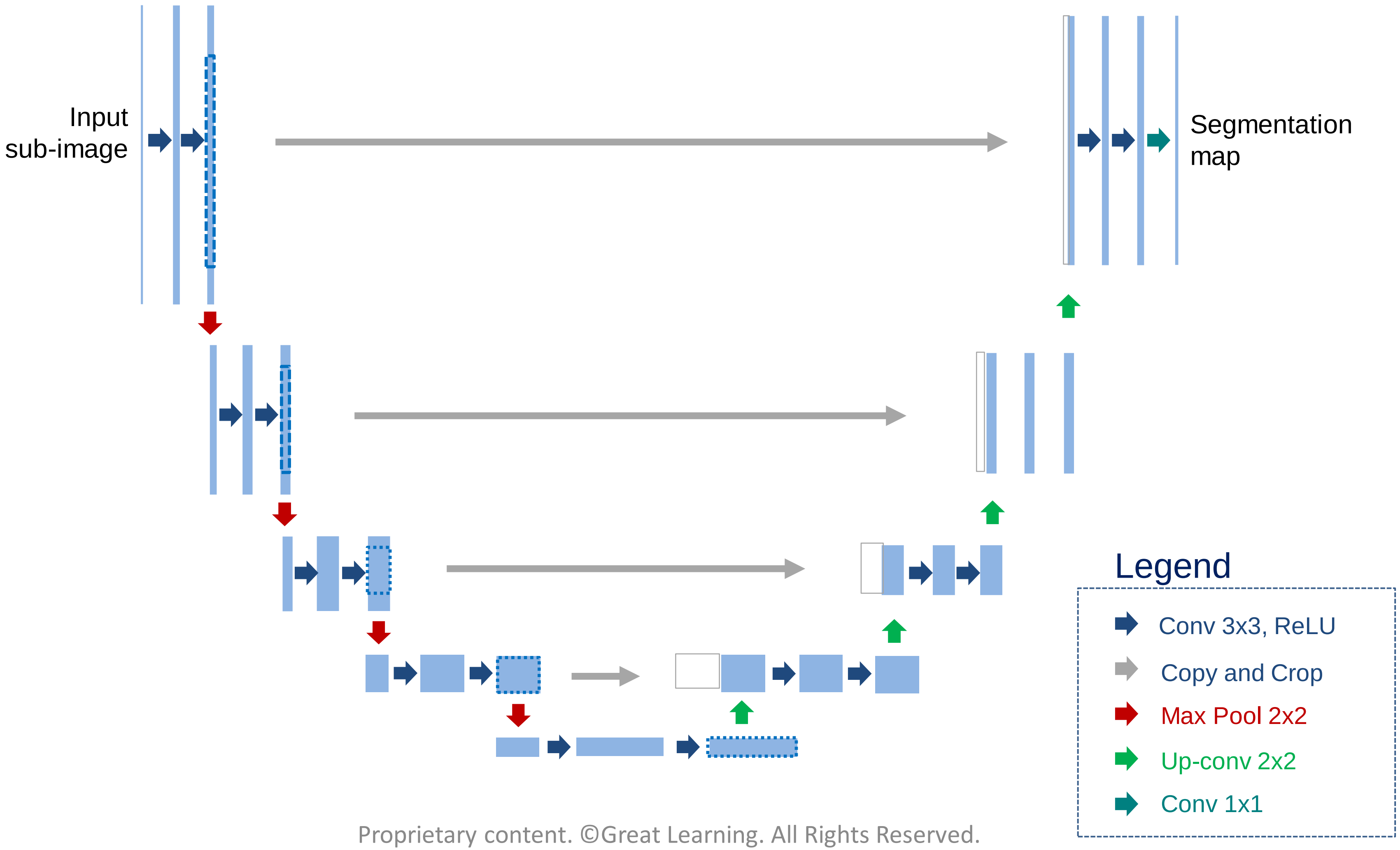
Visually rearrange layers in a big U



Concatenate previous feature maps for finer spatial context

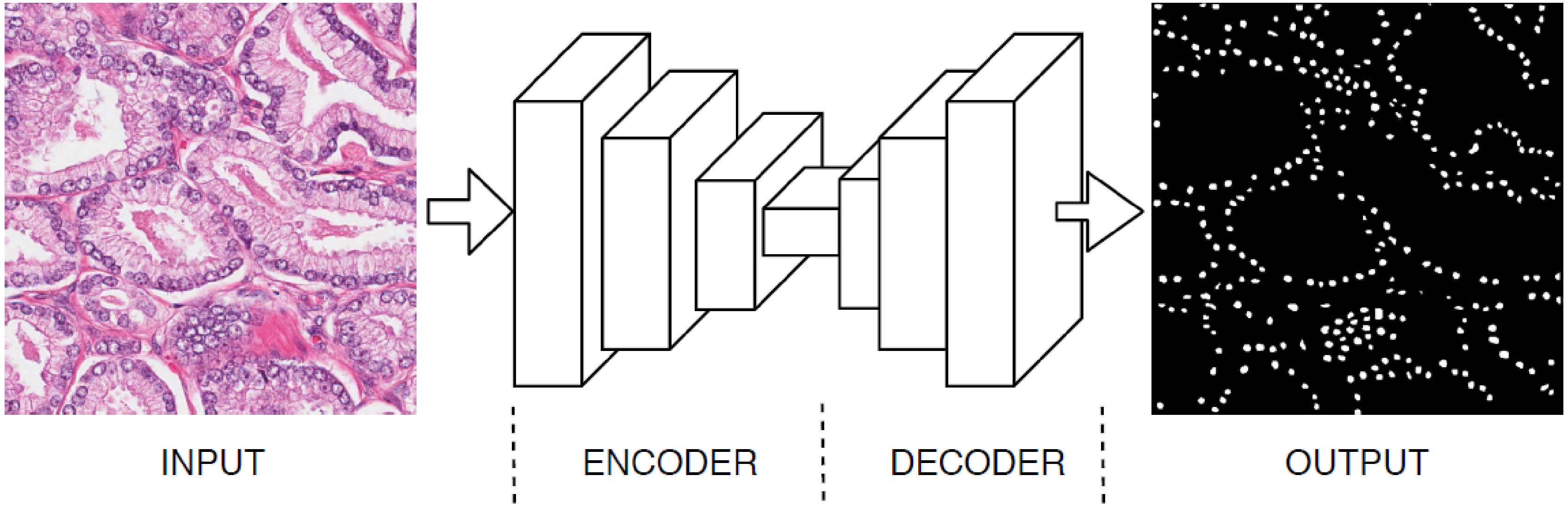


U-Net is based on the ideas described in the previous slides



Source: "U-Net: Convolutional Networks for Biomedical Image Segmentation" Olaf Ronneberger, Philipp Fischer, Thomas Brox, 2015

A sample output for nucleus segmentation in pathology

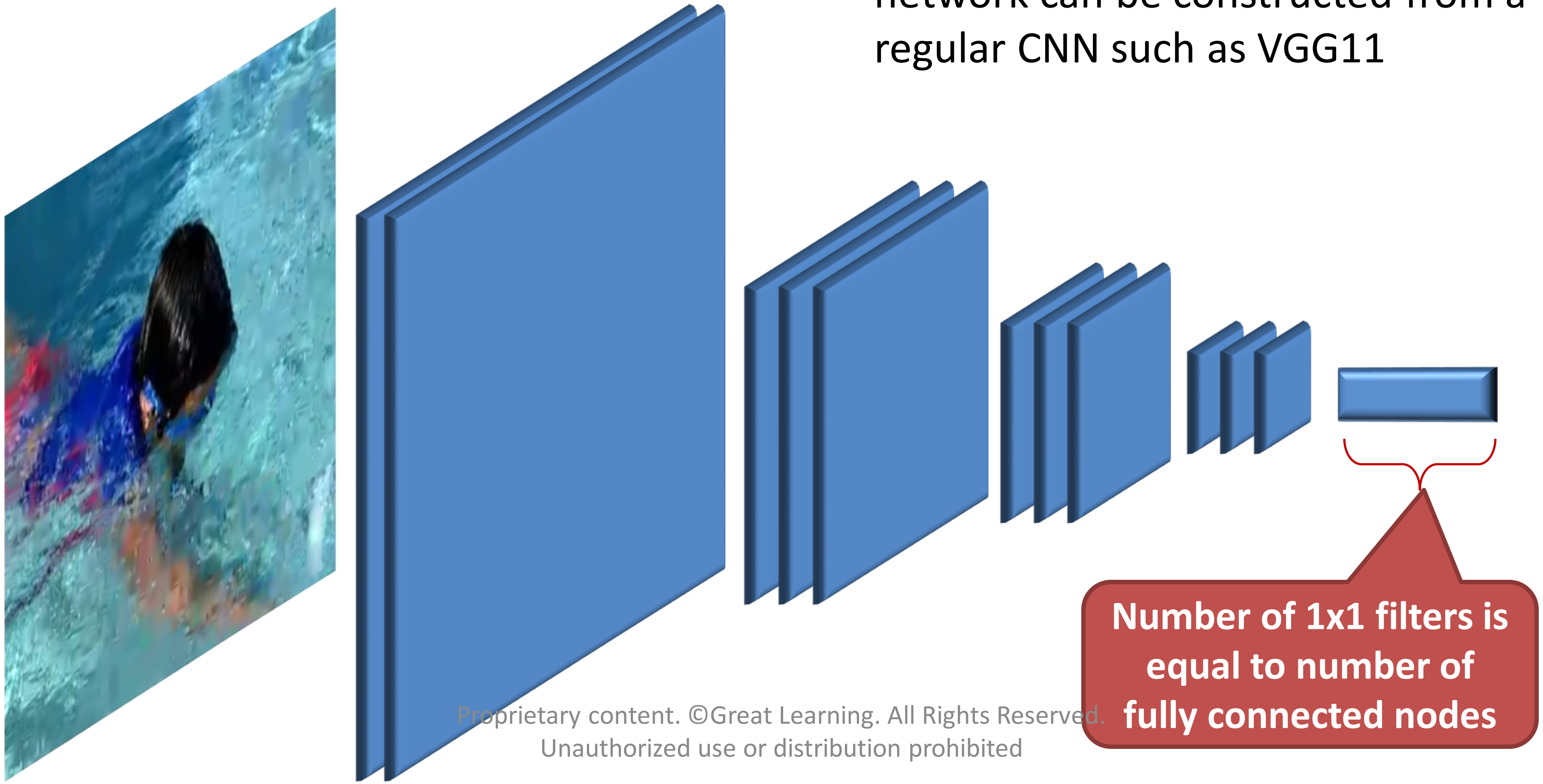


A general representation of fully convolutional networks. The encoder is composed of convolutional and pooling layers for downsampling and the decoder is composed of deconvolutional layers for upsampling.

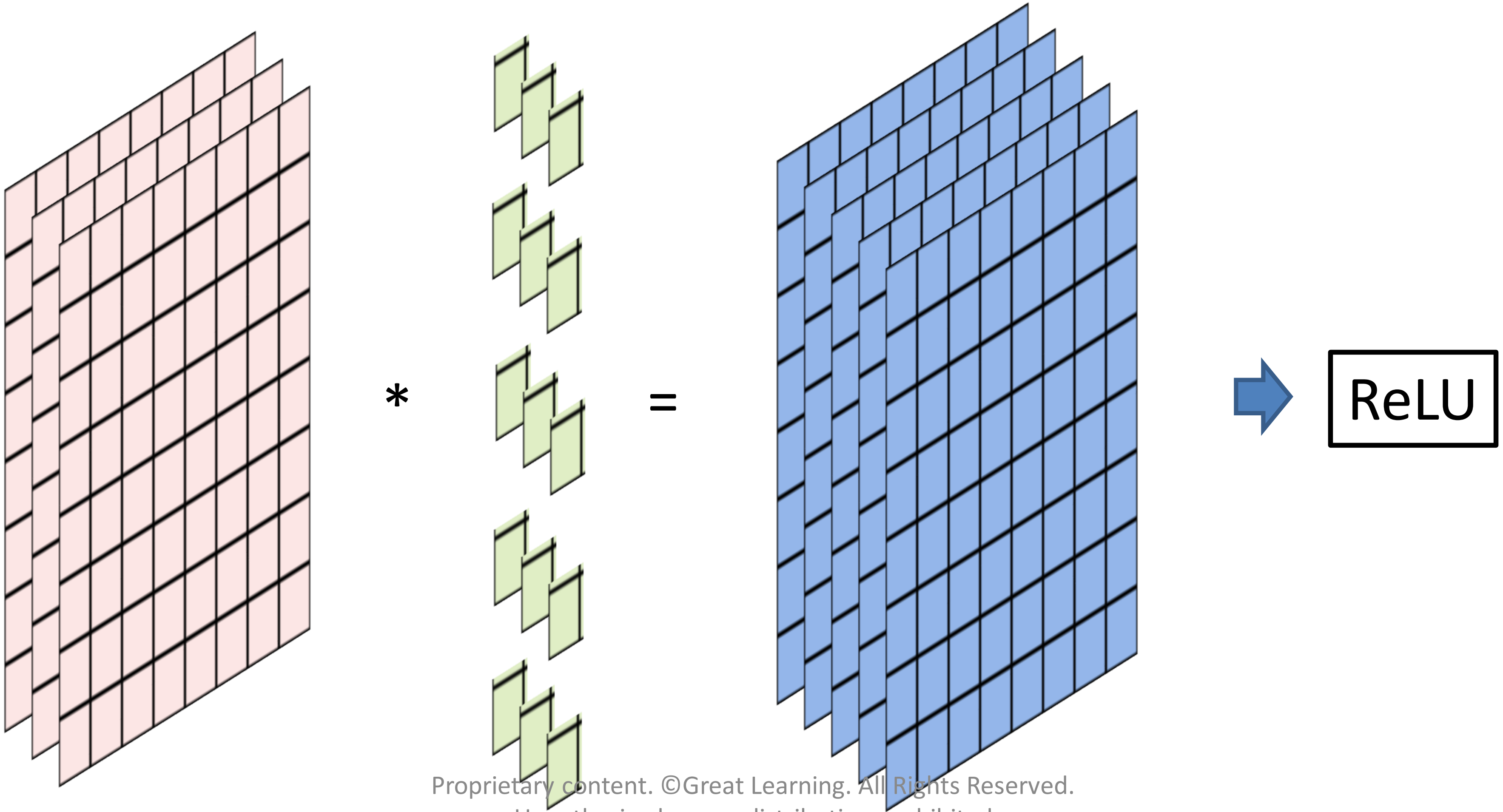
Contents

Using 1x1 convolutions is equivalent to having a fully connected layer

- This way, a fully convolutional network can be constructed from a regular CNN such as VGG11



1x1 convolutions can also be used to change the number of feature maps



Inception uses multiple sized convolution filters

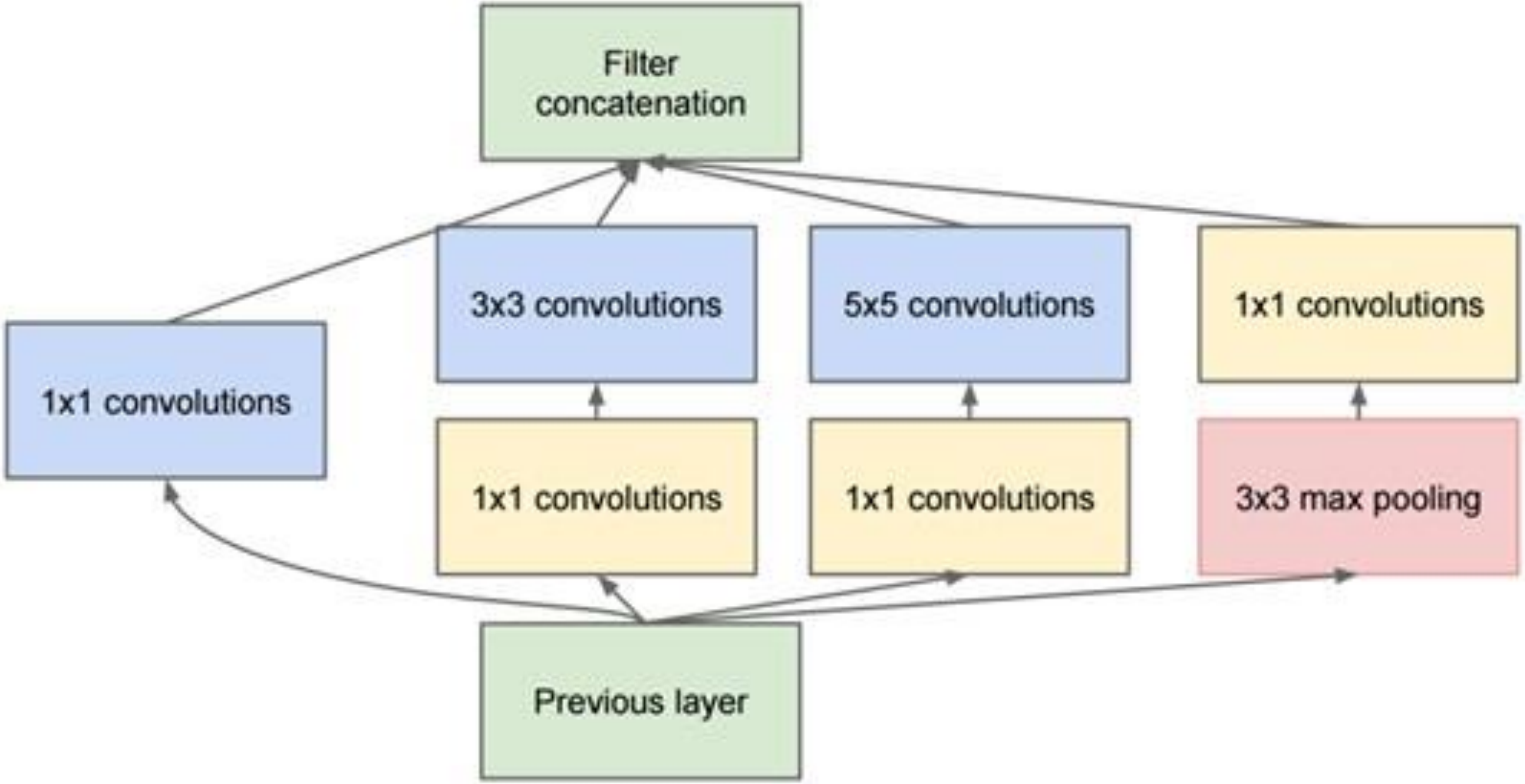
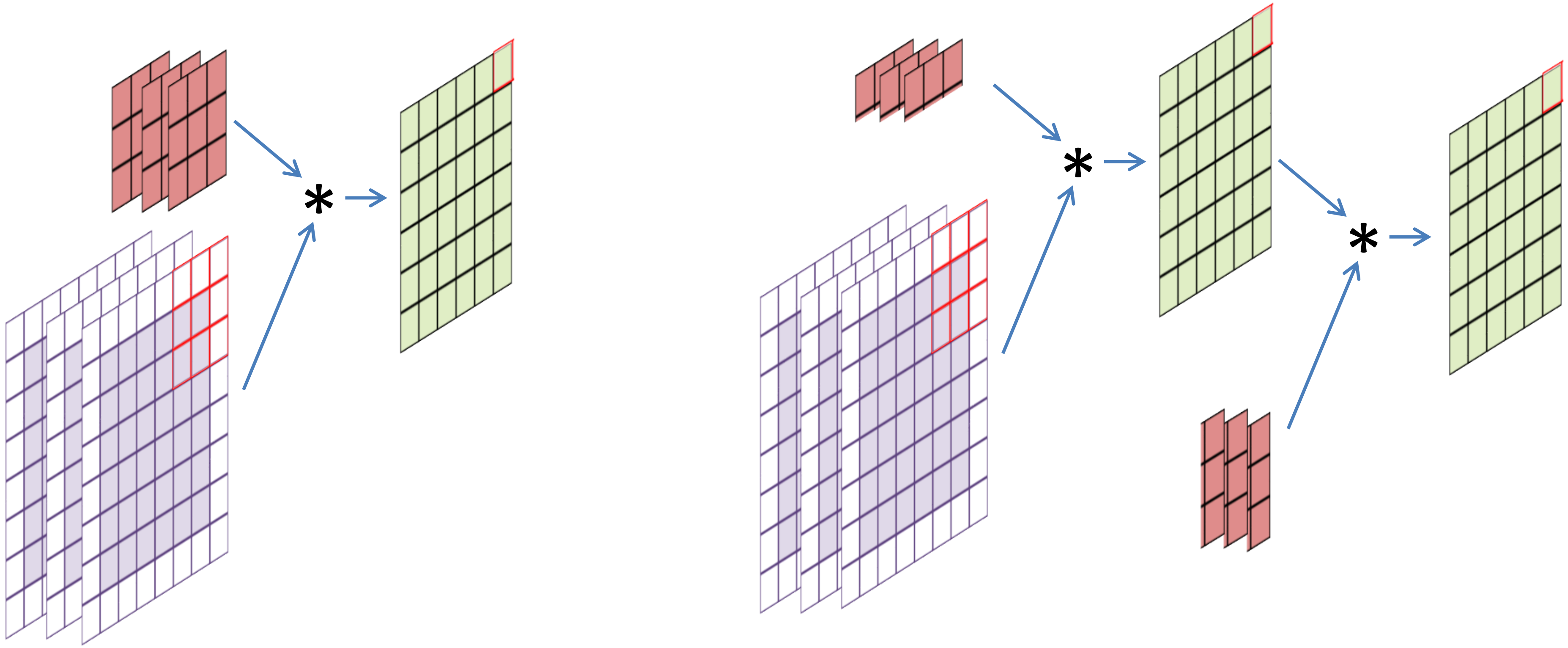
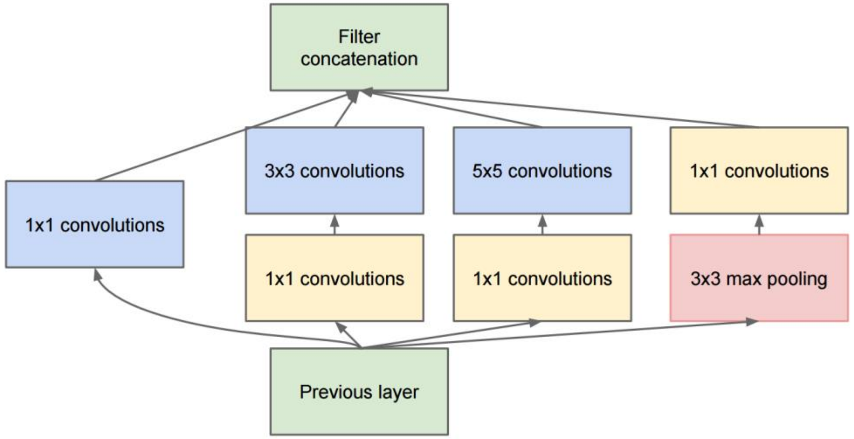


Image source: https://ai.googleblog.com/2016/08/improving-inception-and-image.html

Separable convolutions



Inception uses multiple sized convolution filters



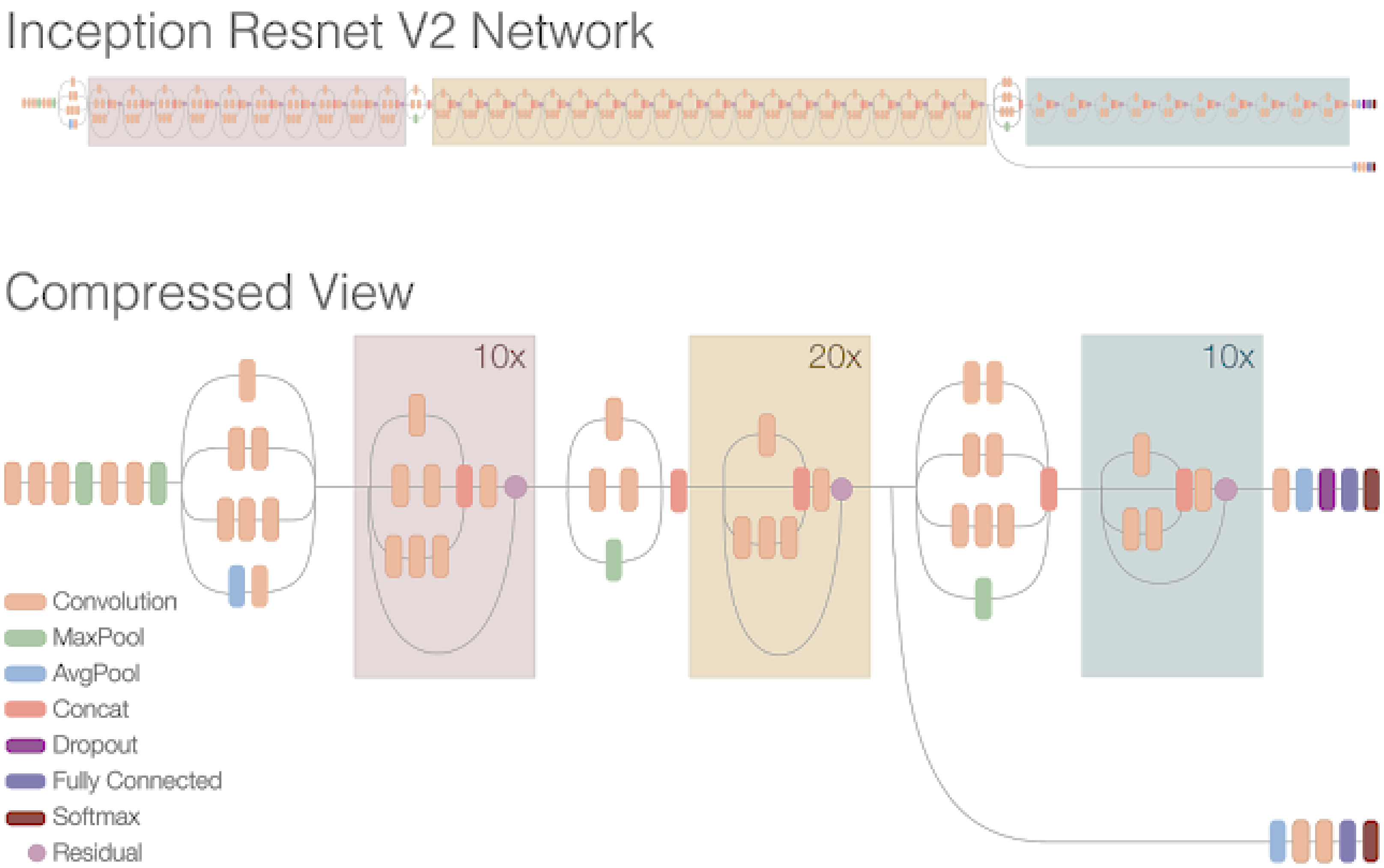
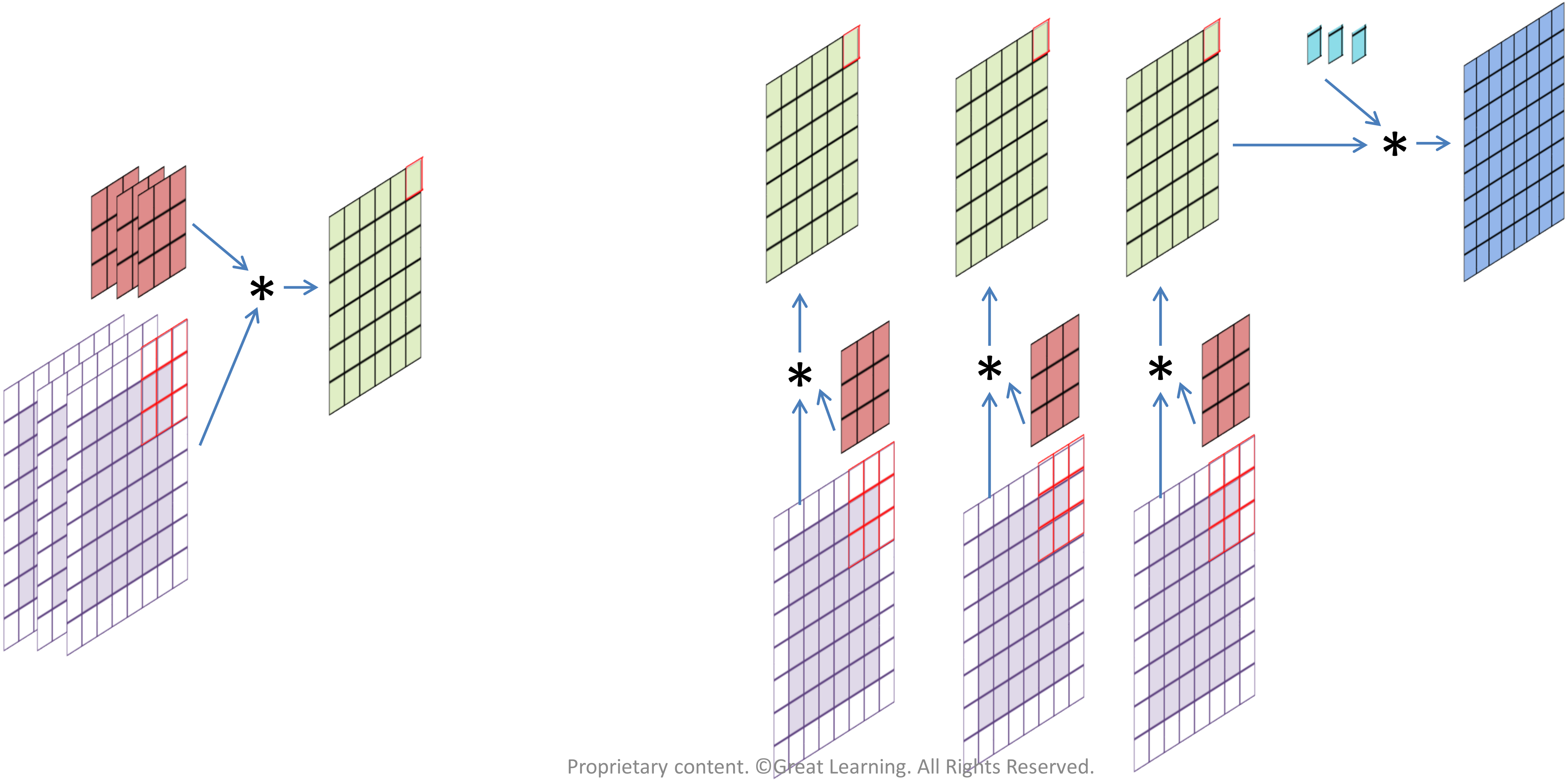


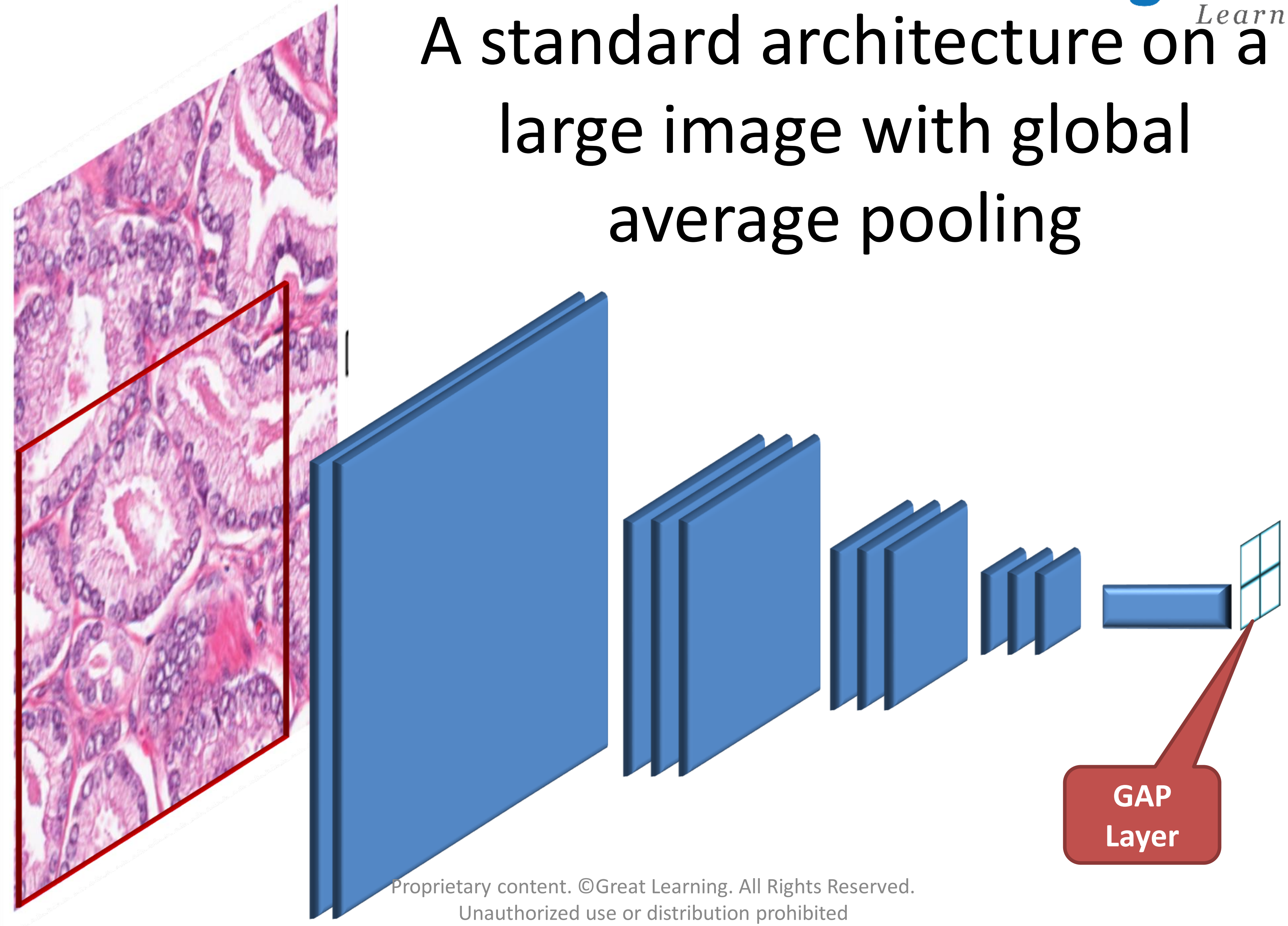
Image source: https://ai.googleblog.com/2016/08/improving-inception-and-image.html

MobileNet filters each feature map separately



"MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications" by Andrew G. Howard Menglong Zhu Bo Chen Dmitry Kalenichenko Weijun Wang Tobias Weyand Marco Andreetto Hartwig Adam, Google , 2017

A standard architecture on a large image with global average pooling



Atrous (dilated) convolutions can increase the receptive field without increasing the number of weights

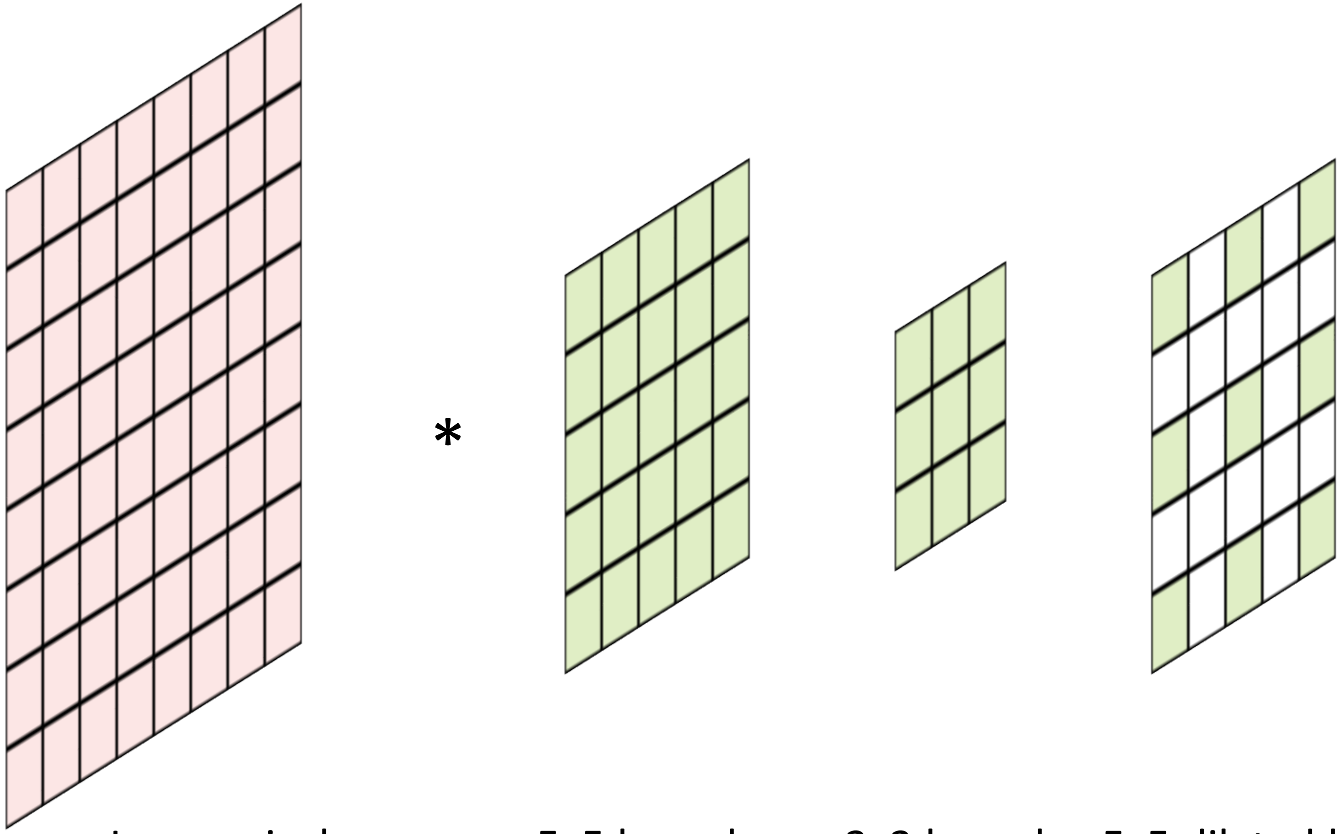


Image pixels

5x5 kernel

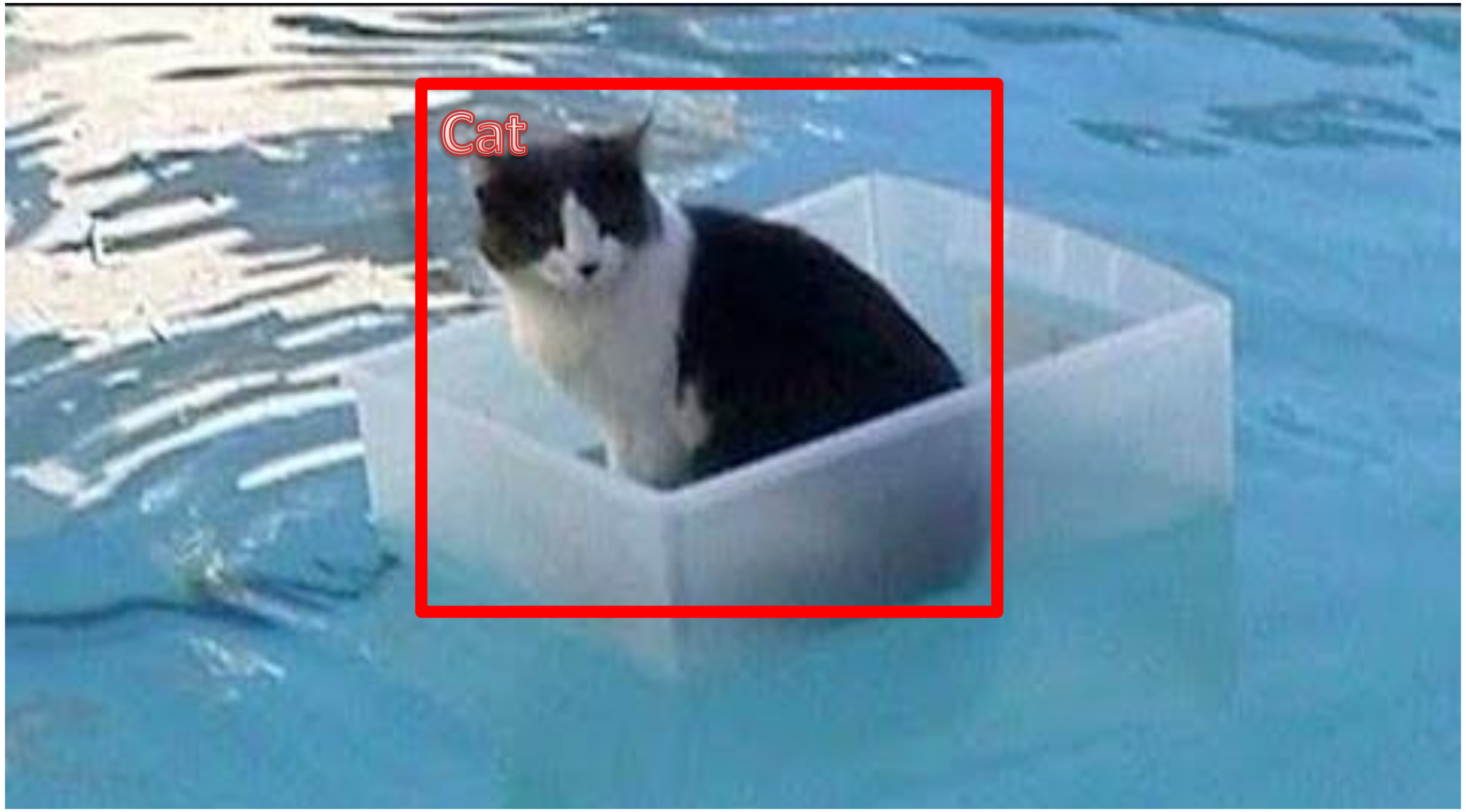
3x3 kernel

5x5 dilated kernel with only 3x3 trainable weights

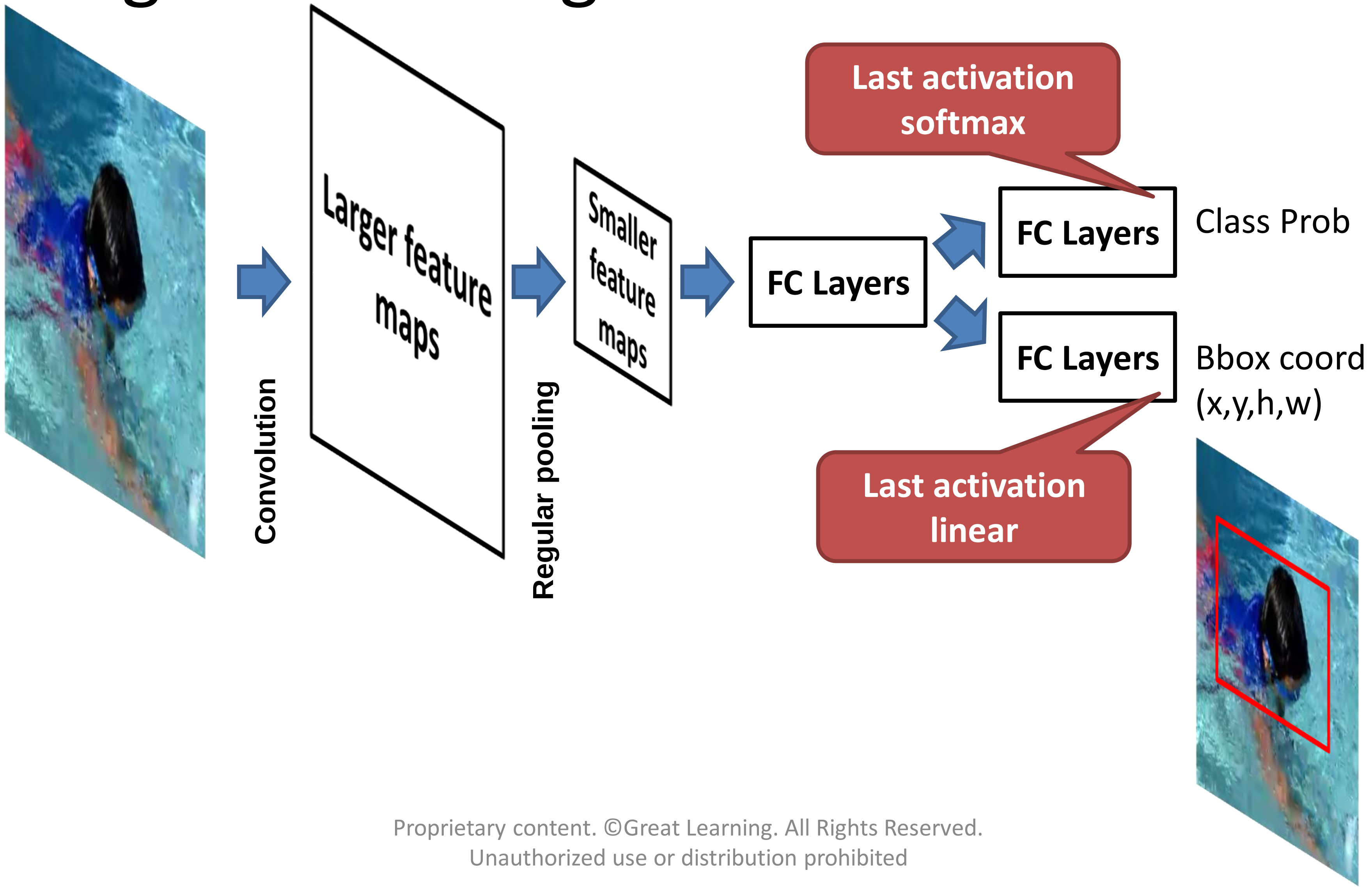
Contents

- FCNs and semantic segmentation
- Other variants of convolution
- Simultaneous localization and recognition
- Siamese network for metric learning

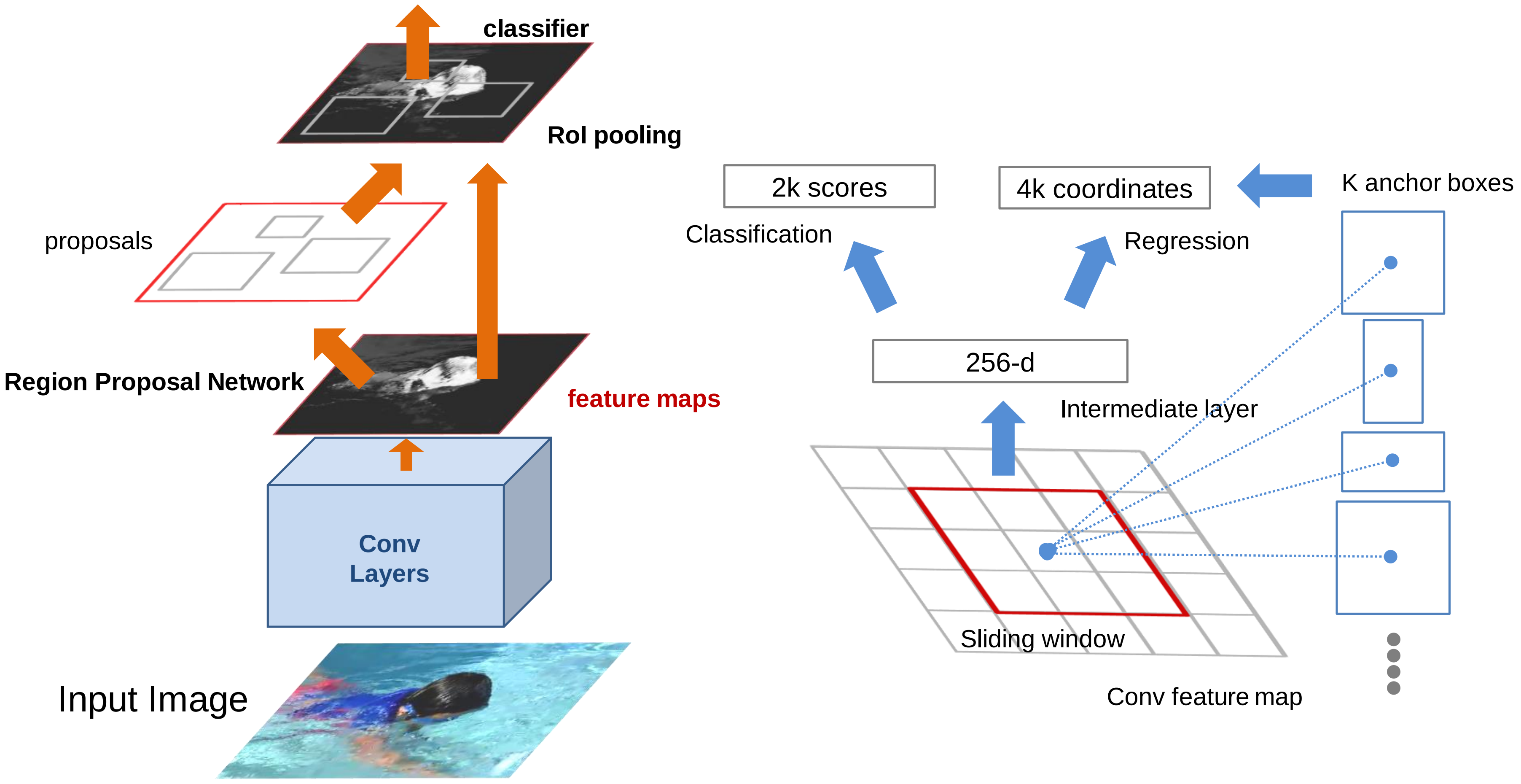
What is localization



We can train a regression network to give bounding box coordinates

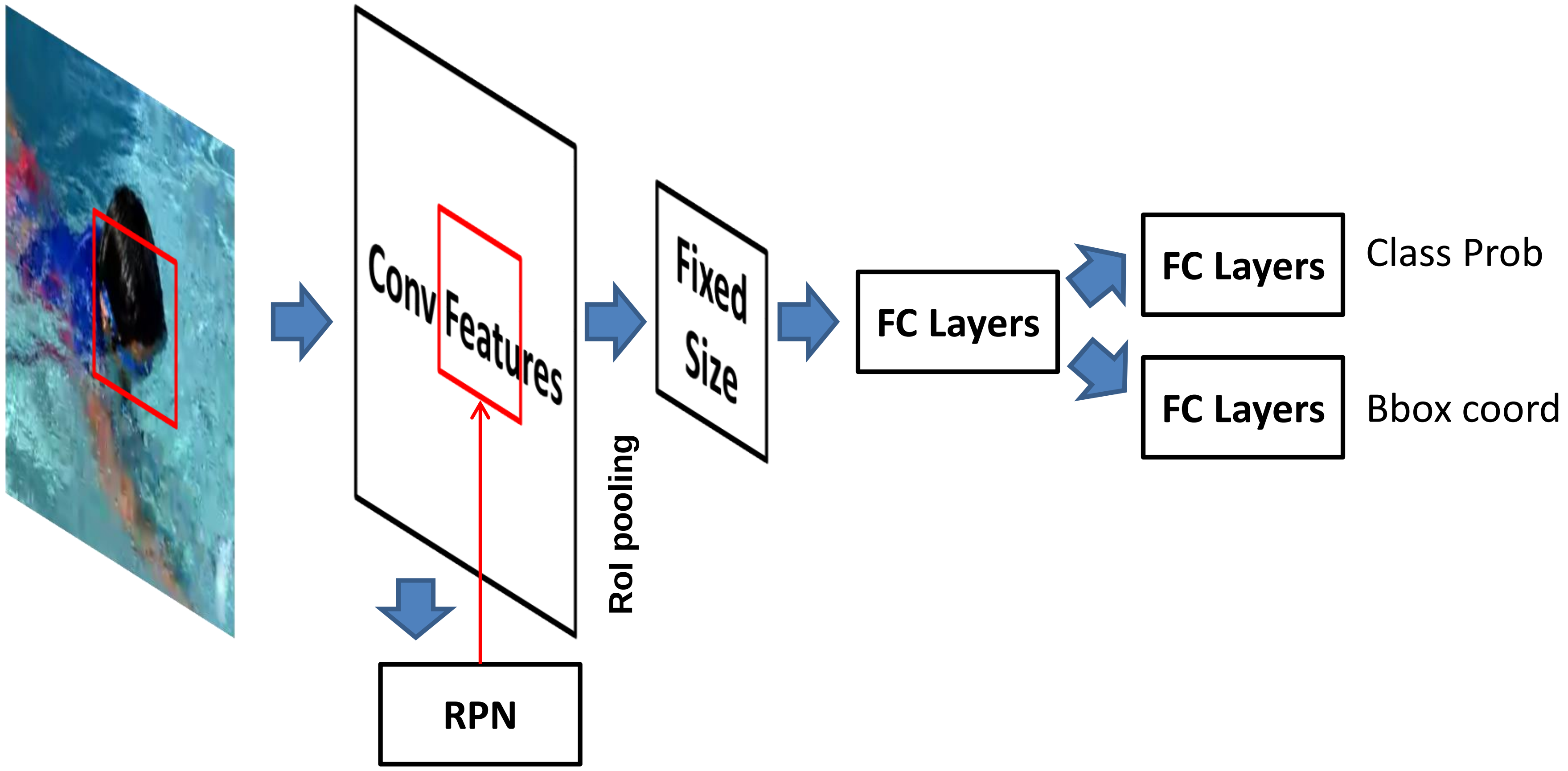


Faster R-CNN architecture

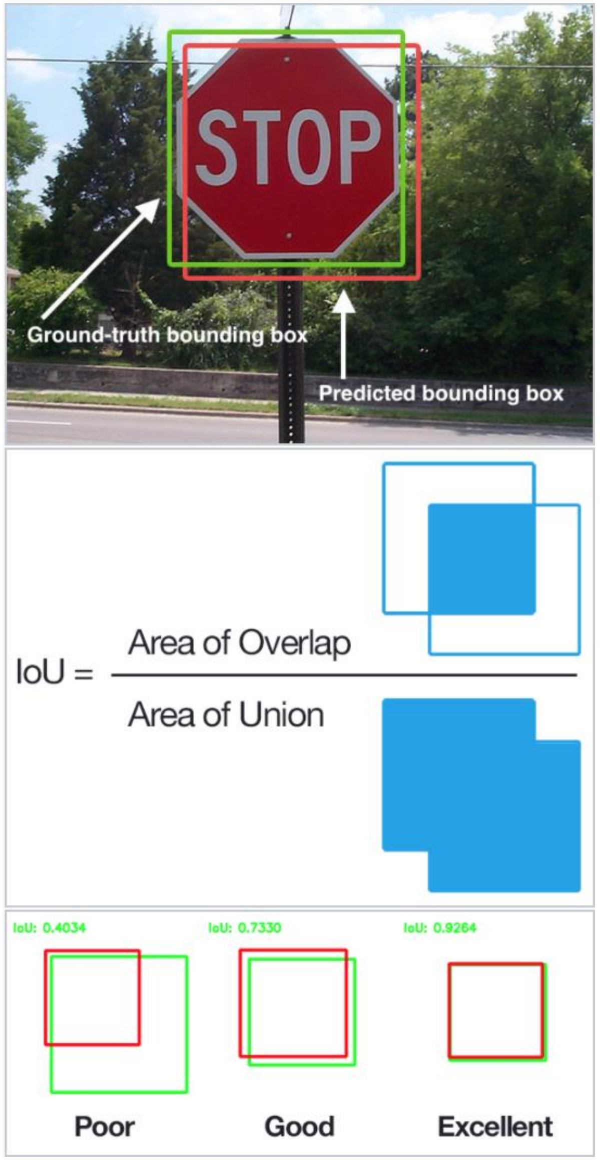


Source: "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks" Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, 2017

Classification and regression on region proposals



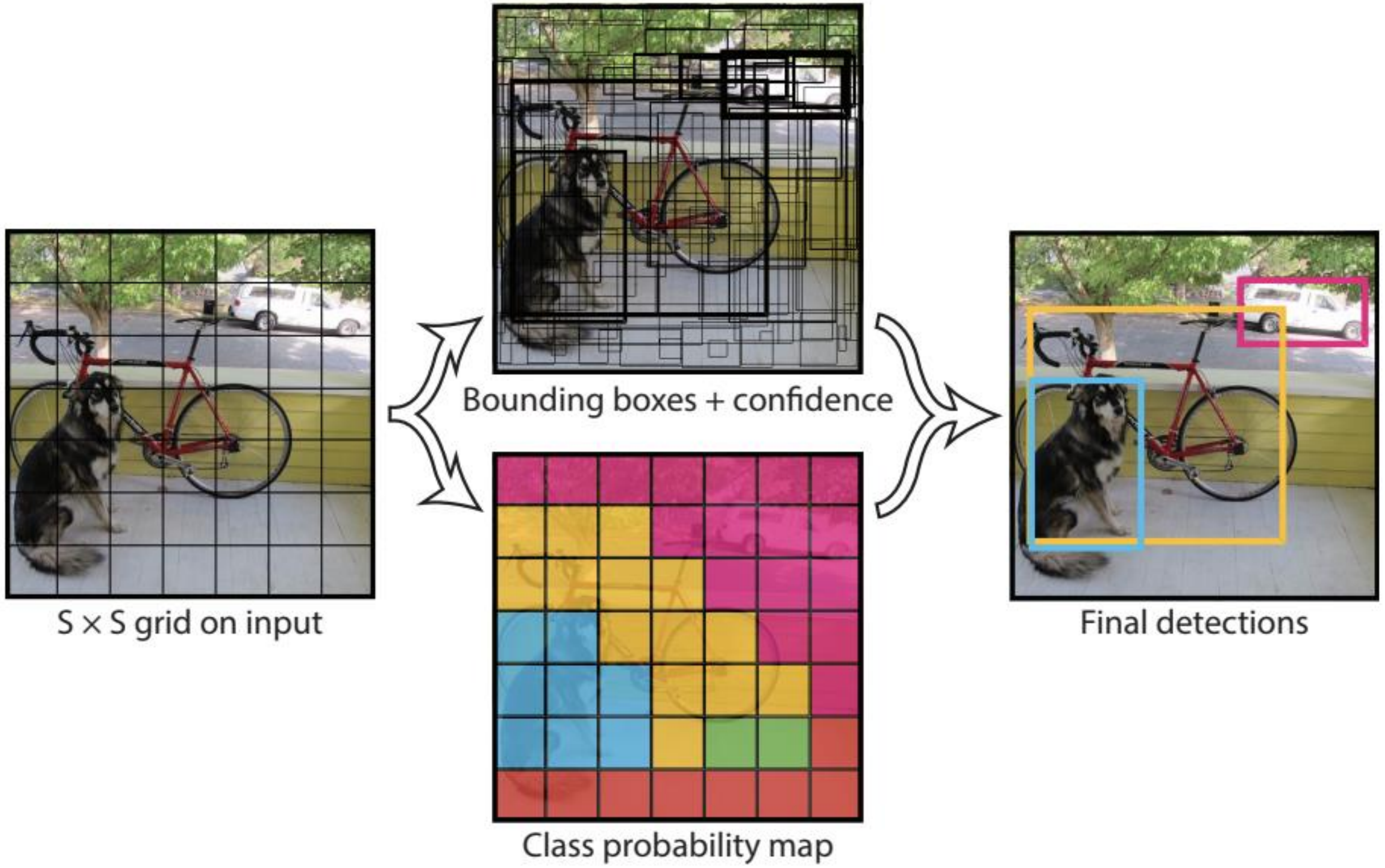
Loss for Simultaneous Classification and Localization



	Classification			Regression
Total_loss =	classification_loss	+	alpha ×	localization_loss
	Cross_Entropy	+	α	Mean_Sq_Error
	Cross_Entropy	+	α	Smooth_L1

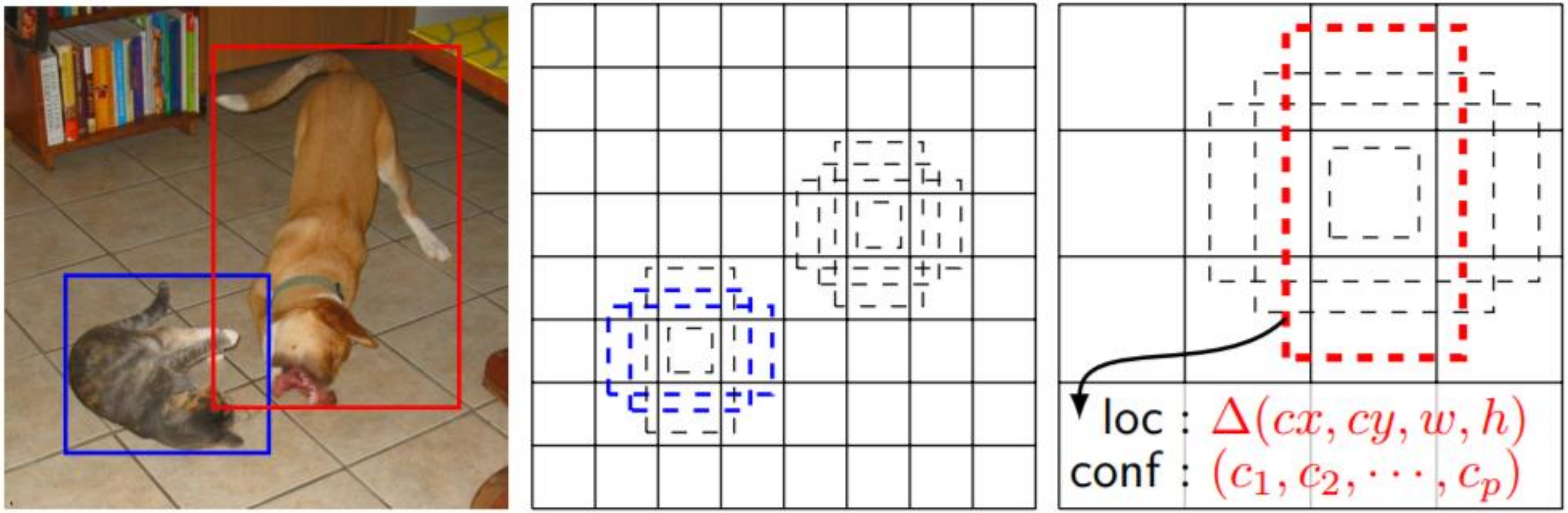
Source: Wikimedia commons

YOLO Approach to Detecting Multiple Objects



“You Only Look Once: Unified, Real-Time Object Detection” Joseph Redmon , Santosh Divvala, Ross Girshick , Ali Farhadi, 2016

SSD Framework

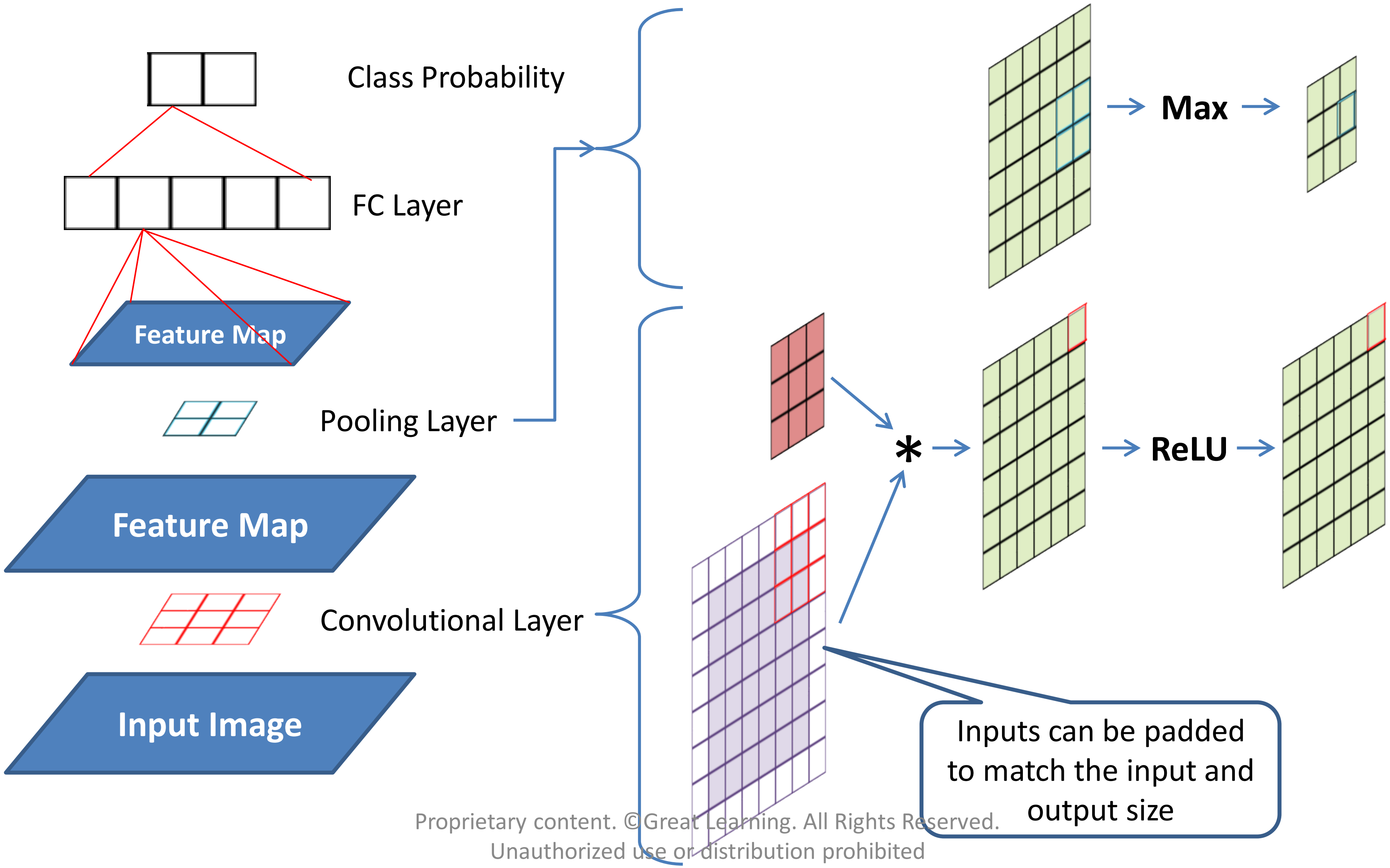


(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

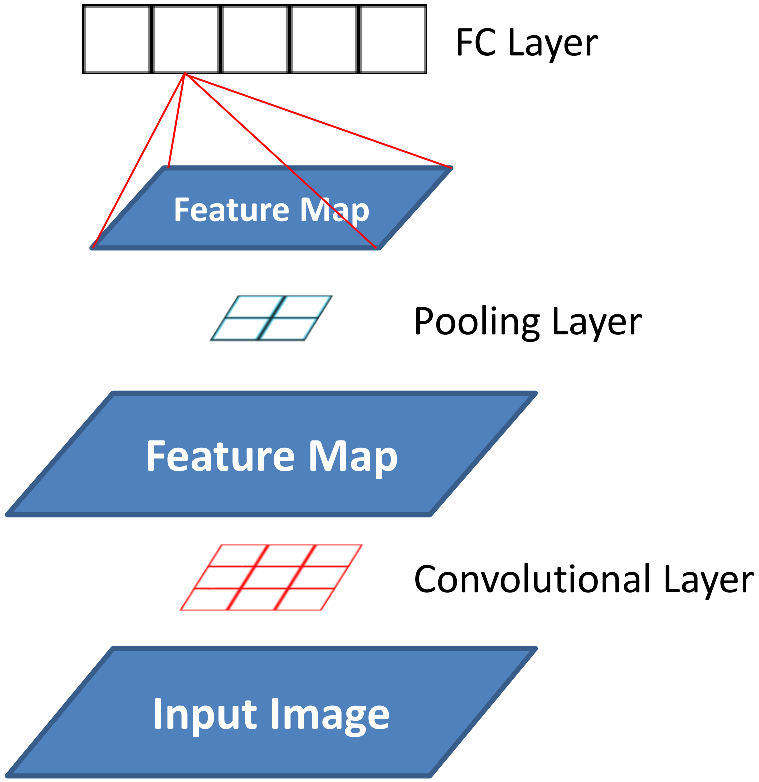
“SSD: Single Shot MultiBox Detector” Wei Liu , Dragomir Anguelov , Dumitru Erhan , Christian Szegedy , Scott Reed , Cheng-Yang Fu , Alexander C. Berg1, Dec 2016

Contents

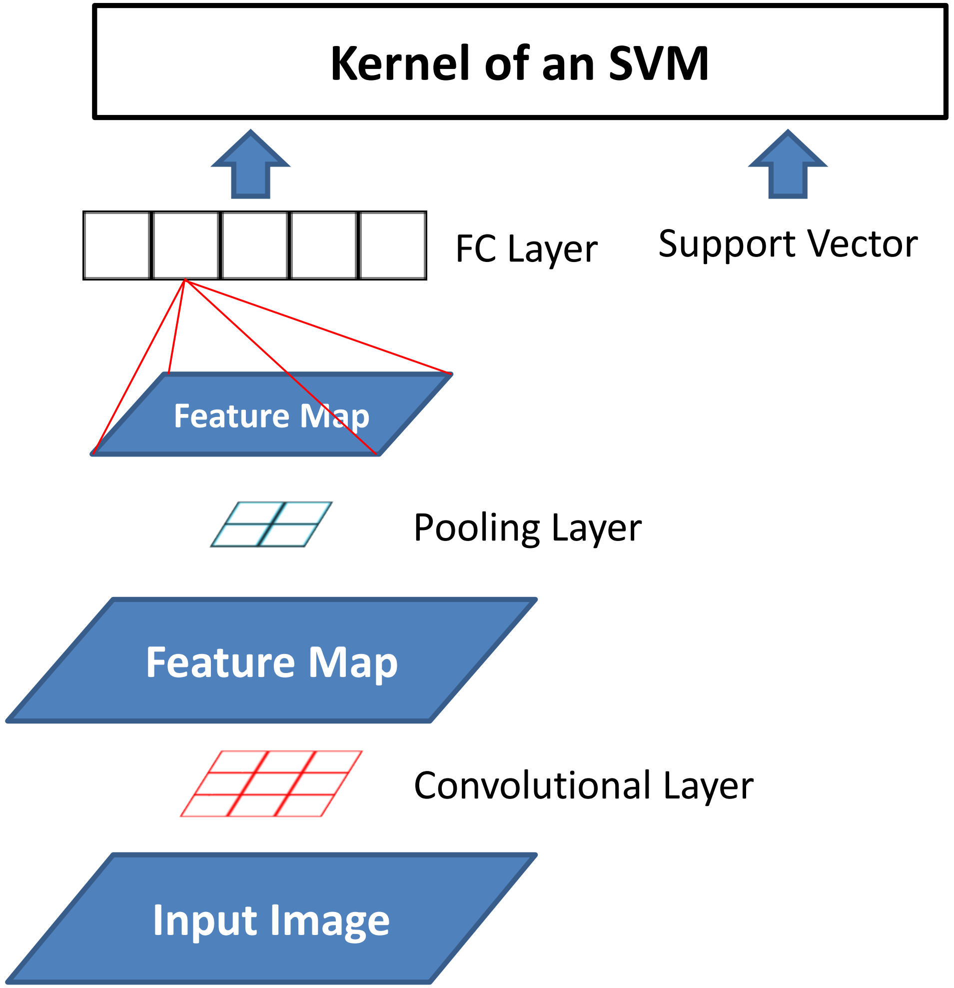
CNN Revisited



The last FC layer gives good features



These features are transferable and can be used in an SVM, for example



$$\exp\left(- ||x_i - x_j||^2/\sigma^2 \right)$$

x_i x_j

Yosinski, Jason, Clune, Jeff, Bengio, Yoshua, and Lipson, Hod. "How transferable are features in deep neural networks?"

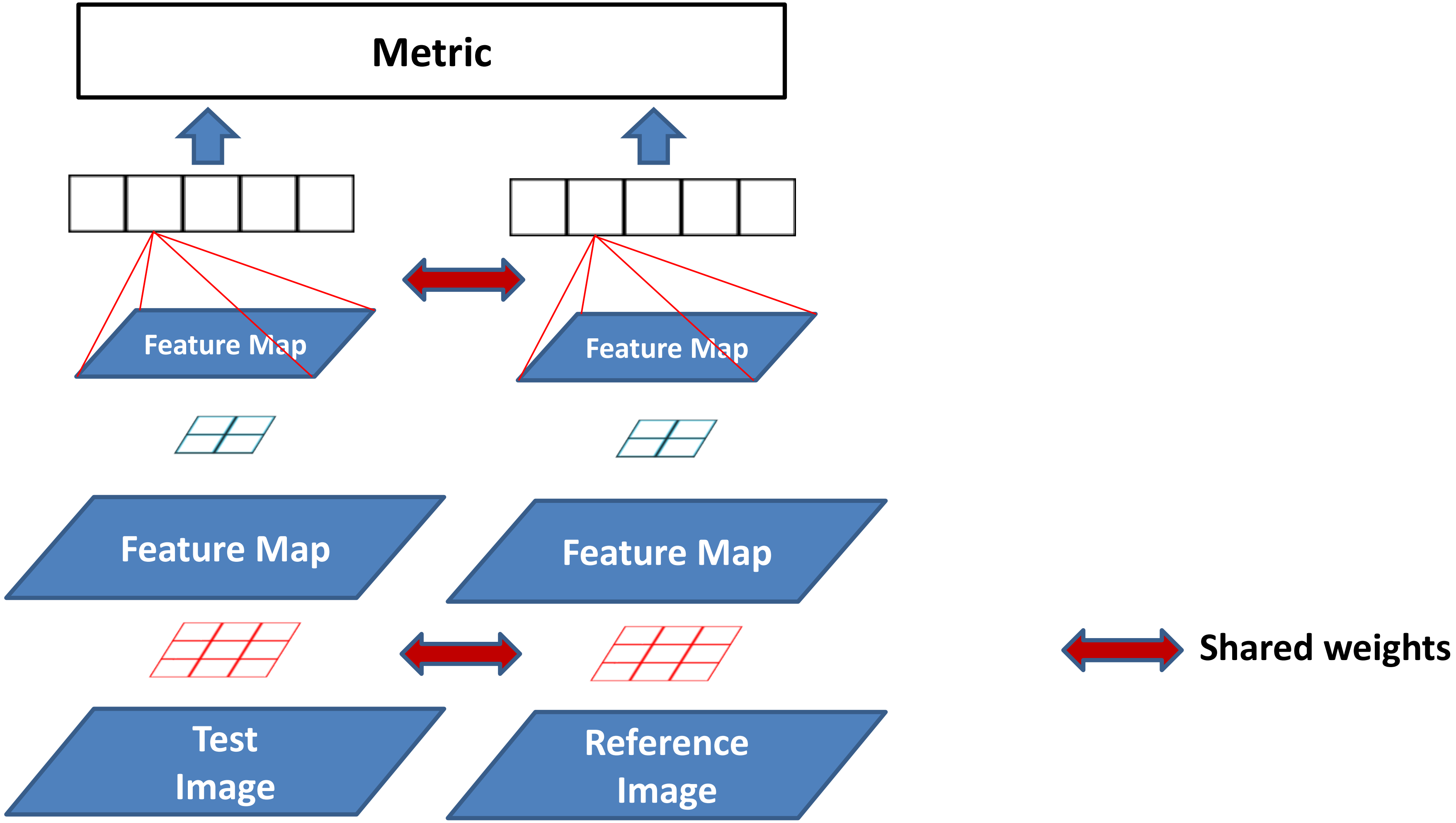
Properties of a kernel

- Similarity metric
- High value for similar pairs of inputs
- Low value for dissimilar inputs
- Positive semi-definite

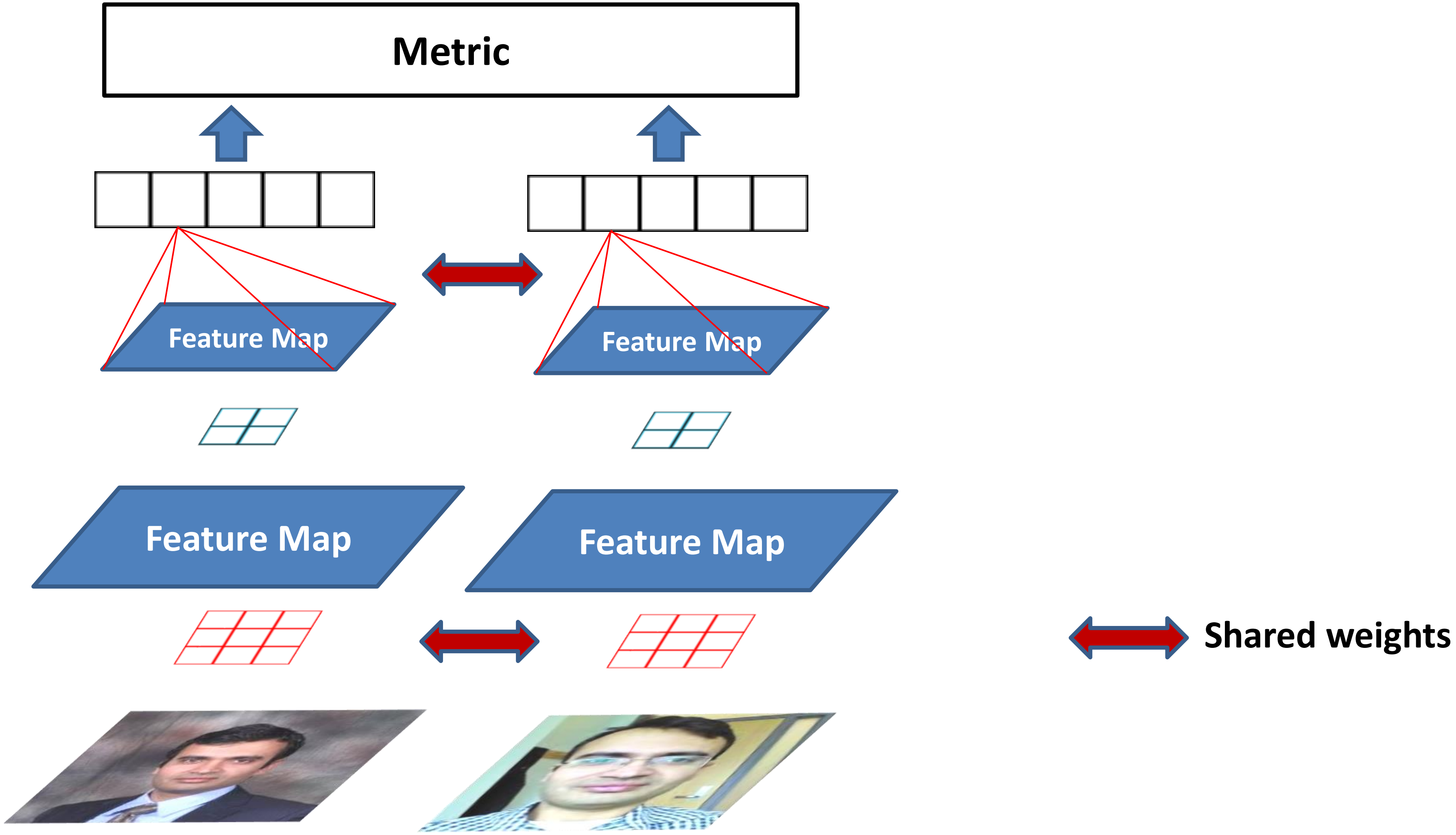
Learning the kernel is called metric learning

- A metric is like a distance
- Inverse of similarity
- It is symmetric
- It follows triangle inequality
- Sometimes, we want to learn a metric

Siamese network as metric learning



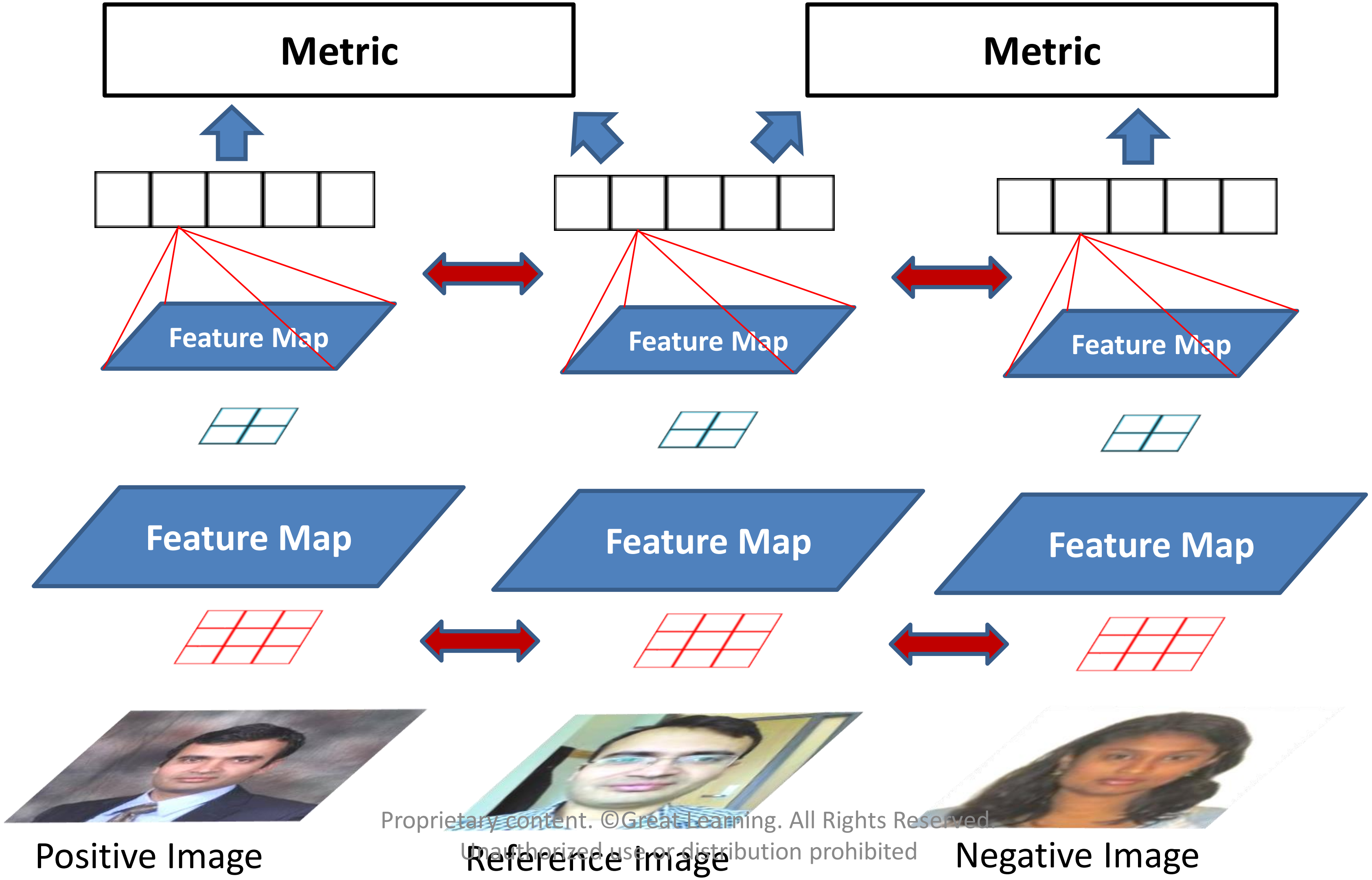
For example, face verification



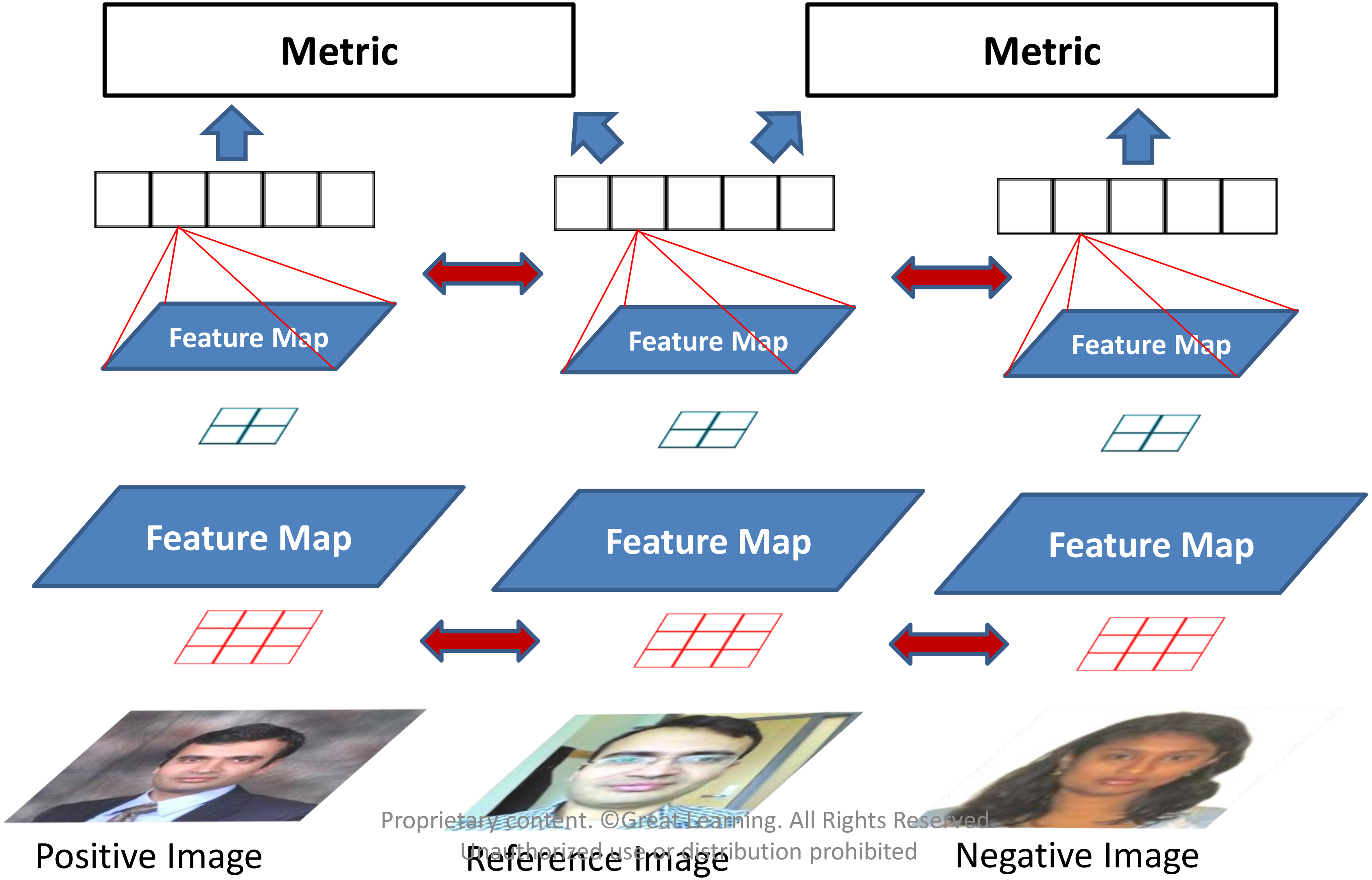
Test Image

Reference Image

Target values differ for similar and dissimilar pairs



Or, the relative values are different



Two ways of viewing a metric

- Absolute terms (Regular Siamese training)
  - Distance (x_{ref}, x_+) = Low; Distance (x_{ref}, x_-) = High
  - Similarity (x_{ref}, x_+) = High; Similarity (x_{ref}, x_-) = Low
- Relative terms (Triplet Siamese training)
  - Distance (x_{ref}, x_-) – Distance (x_{ref}, x_+) > Margin
  - Similarity (x_{ref}, x_+) – Similarity (x_{ref}, x_-) > Margin

- Class probability was based on a single input
  - ClassProb (x,c) = High when $x \in c$; otherwise low

Some distance and similarity measures

- Distances examples
  - L2 norm of difference (Euclidean distance)
  - L1 norm of difference (City-block/Manhattan dist.)

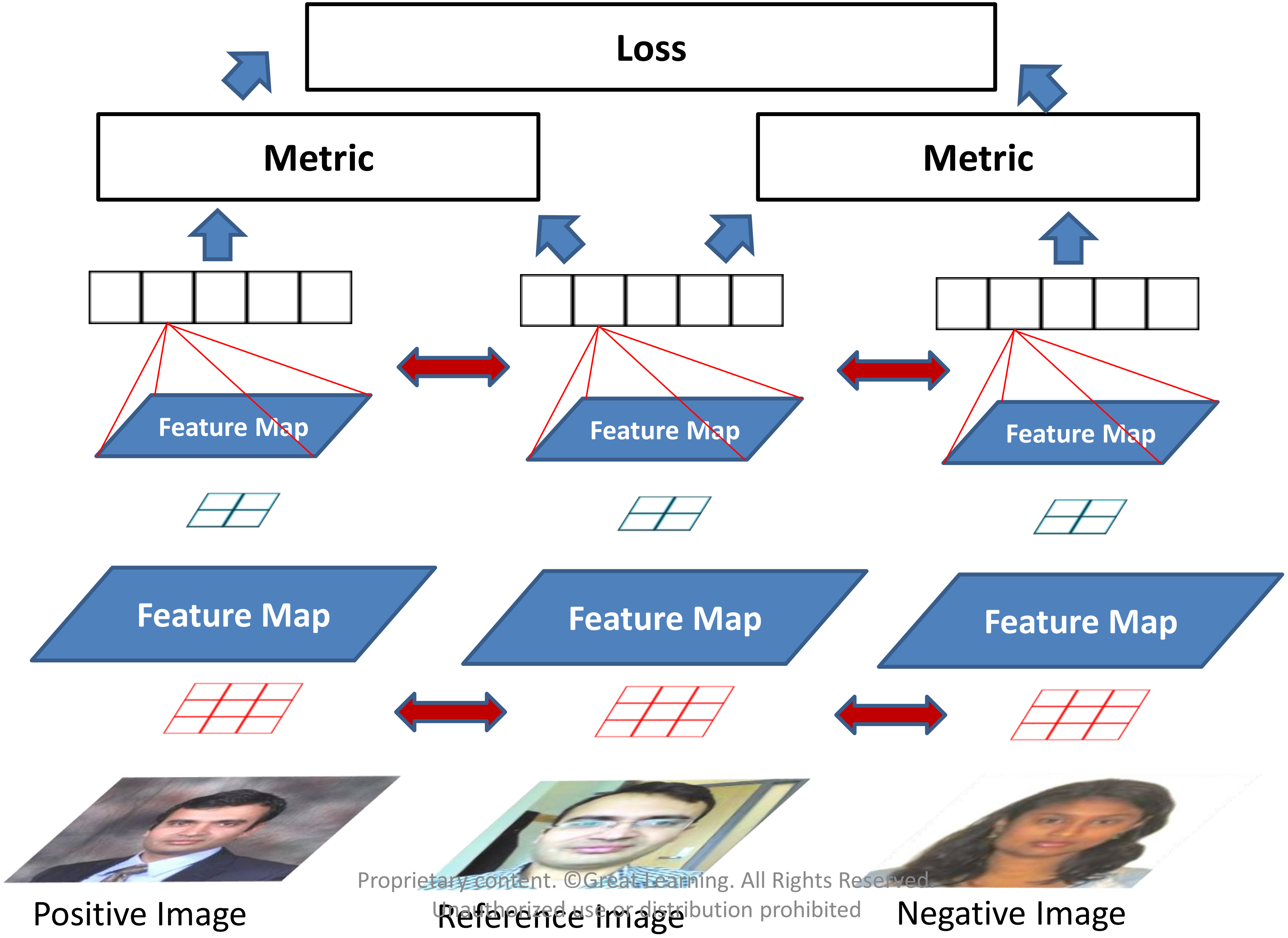
- Similarity examples
  - Dot product
  - Arc cosine
  - Radial basis function (RBF)

Some distance and similarity measures

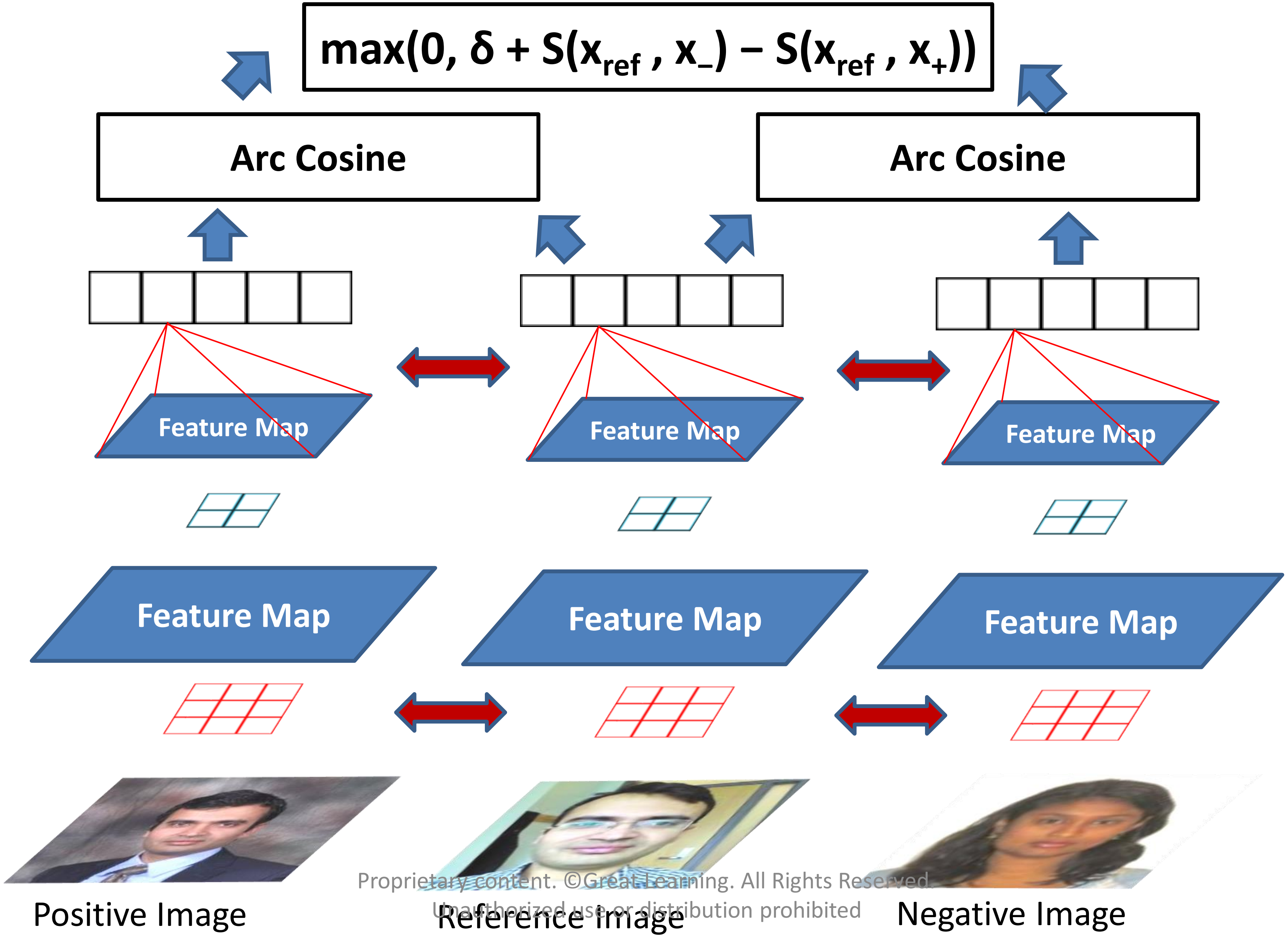
- Distances examples
  - $||(f(x_i) - f(x_j)||_2^2$
  - $|(f(x_i) - f(x_j)|_1$

- Similarity examples
  - $f(x_i)^T f(x_j)$ or $f(x_i) \cdot f(x_j)$
  - $f(x_i) \cdot f(x_j)$ / $(||f(x_i)|| \; ||f(x_j)||)$
  - $\exp(- ||x_i - x_j||^2/\sigma^2)$

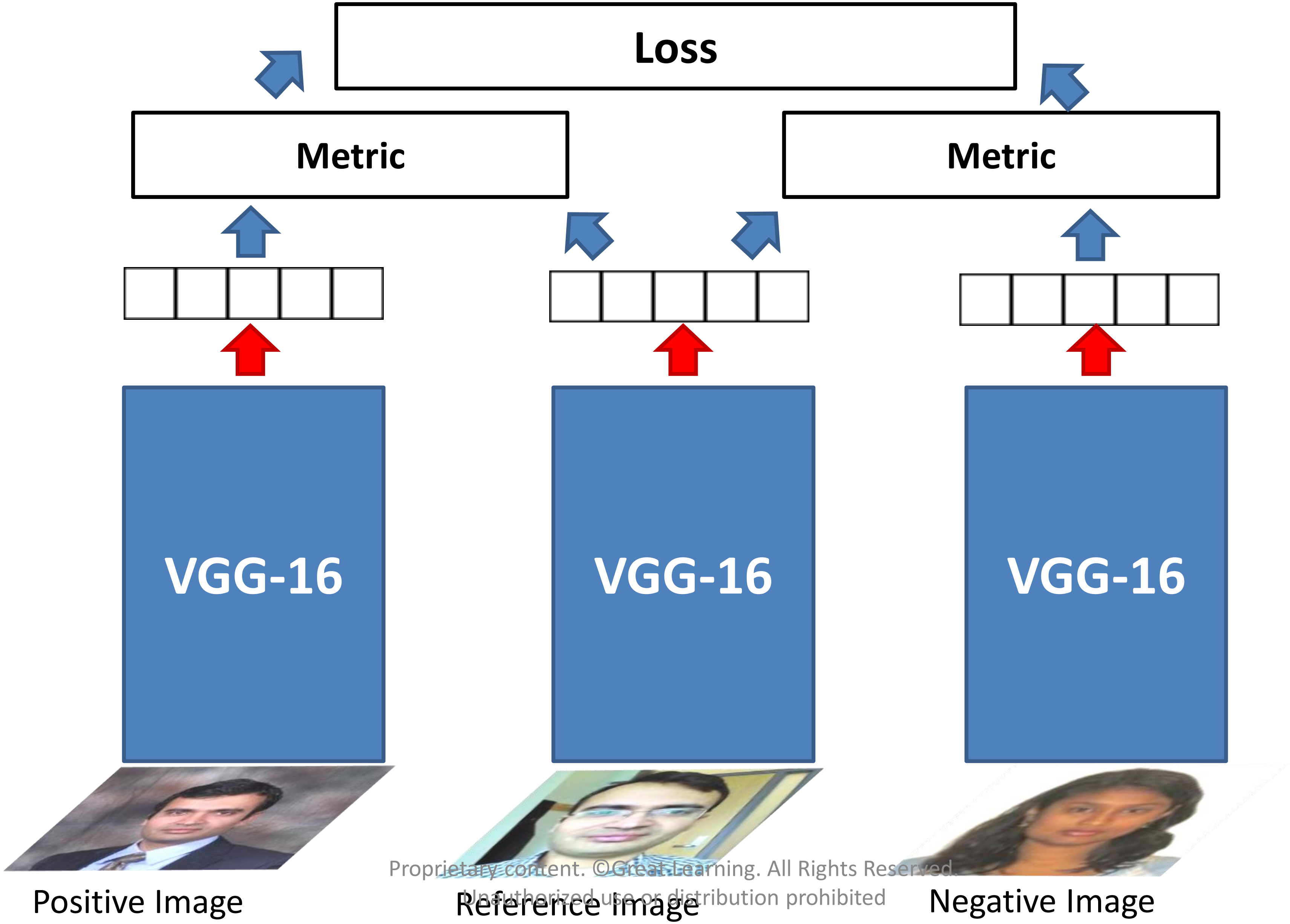
Loss gradient is propagated back



Loss gradient is propagated back



Pre-trained networks can be used



Some joint layers can also be added