Practical CVOps: Datasets and Labels

Deepak Subramani
Assistant Professor
Dept. of Computational and Data Science
Indian Institute of Science Bengaluru

Creating Vision Datasets

- Collecting images can be done in multiple ways
  - Mounting a camera at the end of an almond processing line, traffic intersection etc
  - Extracting photographs from a digital catalog
  - Purchasing an archive of satellite imagery
  - Medical image devices
- Capturing photographs
  - Use the highest resolution that is necessary for the noise characteristics of your image and what your ML infra budget can handle
    - Higher-resolution outdoor images in low light have higher noise
    - Collecting high-res images needs a lot of time and bandwidth

$\frac{\Delta L}{\Delta \theta}$ → derivative of loss w.r.t. weight (parameter)

gradient

C1 C2 C3 D1 D2 → o/p

w_{c1}

b_{c1}

θ_{c1} θ_{c2} θ_{c3} θ_{D1} θ_{D2}

$$\Rightarrow \theta = \begin{bmatrix} \theta_{D2} \\ \theta_{D1} \\ \theta_{C10} \\ \vdots \\ \theta_{C2} \\ \theta_{C1} \end{bmatrix}$$

parameter of the model

Trainable parameters

Stochastic Gradient Descent

Optimizer

ADAM { Mini-batch SGD

Proof of Concept

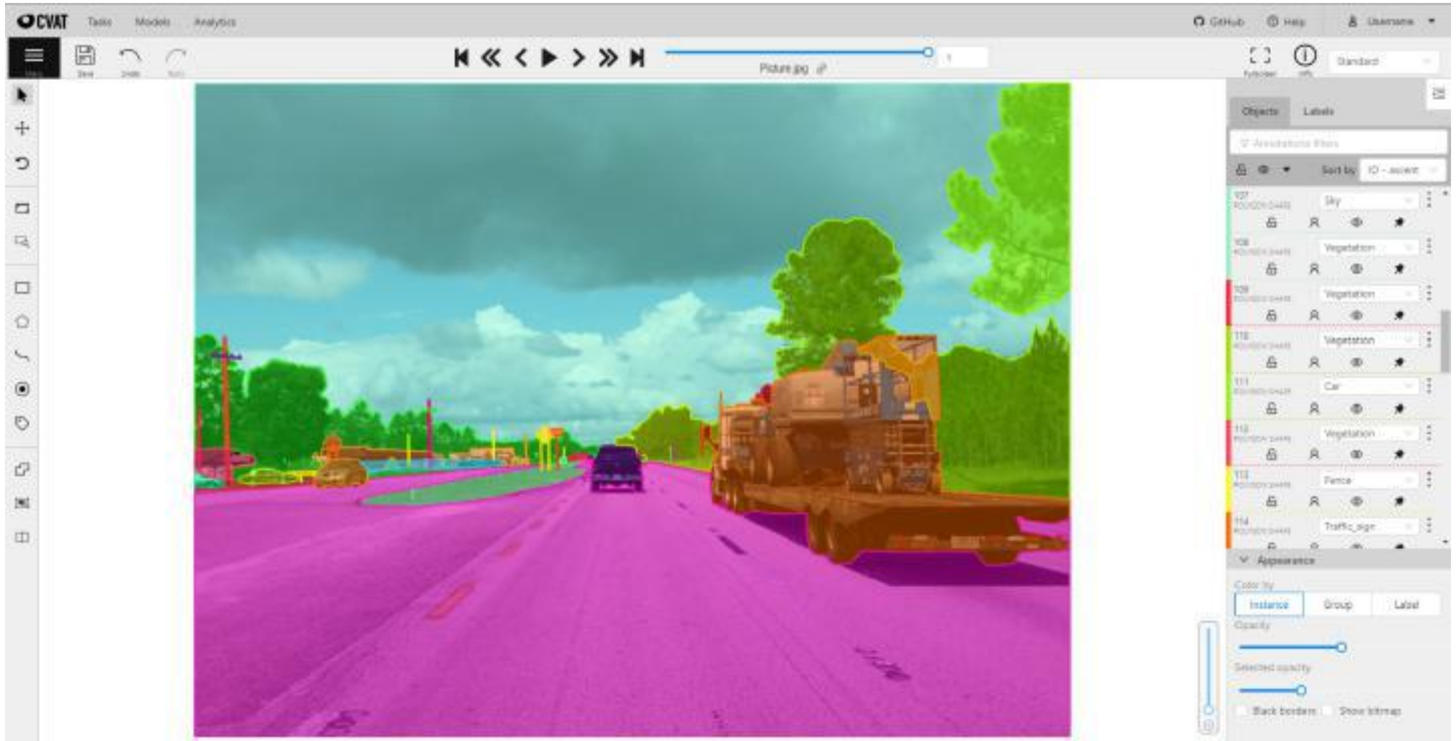
- In many situations, you may not have the data on hand, and collecting it for a PoC would take too long. What to do?
- Purchase similar data to understand the feasibility of a project before investing in routine data collection.
- When purchasing images,
  - acquire images that are similar in quality, resolution, etc. to the images that you will ultimately be able to use in the actual project.
- Simulate labeled images by modifying existing images
  - In advanced applications of crowd counting etc

Labeling Data

- For image classification there are two manual labeling approaches
  - Move the images to a folder whose name is the label
  - Create an excel file (spreadsheet) with first column having the path of the image and the other columns having the label(s)
- Object detection
  - Needs bounding box (usually counterclockwise starting from top-left)
- Segmentation
  - Needs labels of pixels

Labeling at scale

- OpenCV Computer Vision Annotation Tool
  - https://github.com/opencv/cvat



Noisy Student

- Manually label, say, 10,000 images.
- Use these images to train a small CV model. This is the teacher model.
- Use this model to predict the labels of, say, one million unlabeled images.
- Train a larger CV model, called the student model, on the combination of labelled and pseudo-labelled images.
- During the learning of the student model, employ dropout and random data augmentations so that this model generalizes better than the teacher.
- Iterate by putting the student model back as the teacher.
- Manually correct the pseudo-labels by choosing images where the models are not confident.

Labelling Service

- Crowdsourcing
- AI Labeling Services - https://cloud.google.com/vertex-ai/pricing#labeling

Computer Vision: Revision

Deepak Subramani
Assistant Professor
Dept. of Computational and Data Science
Indian Institute of Science Bengaluru

Lecture and Assignment Guide

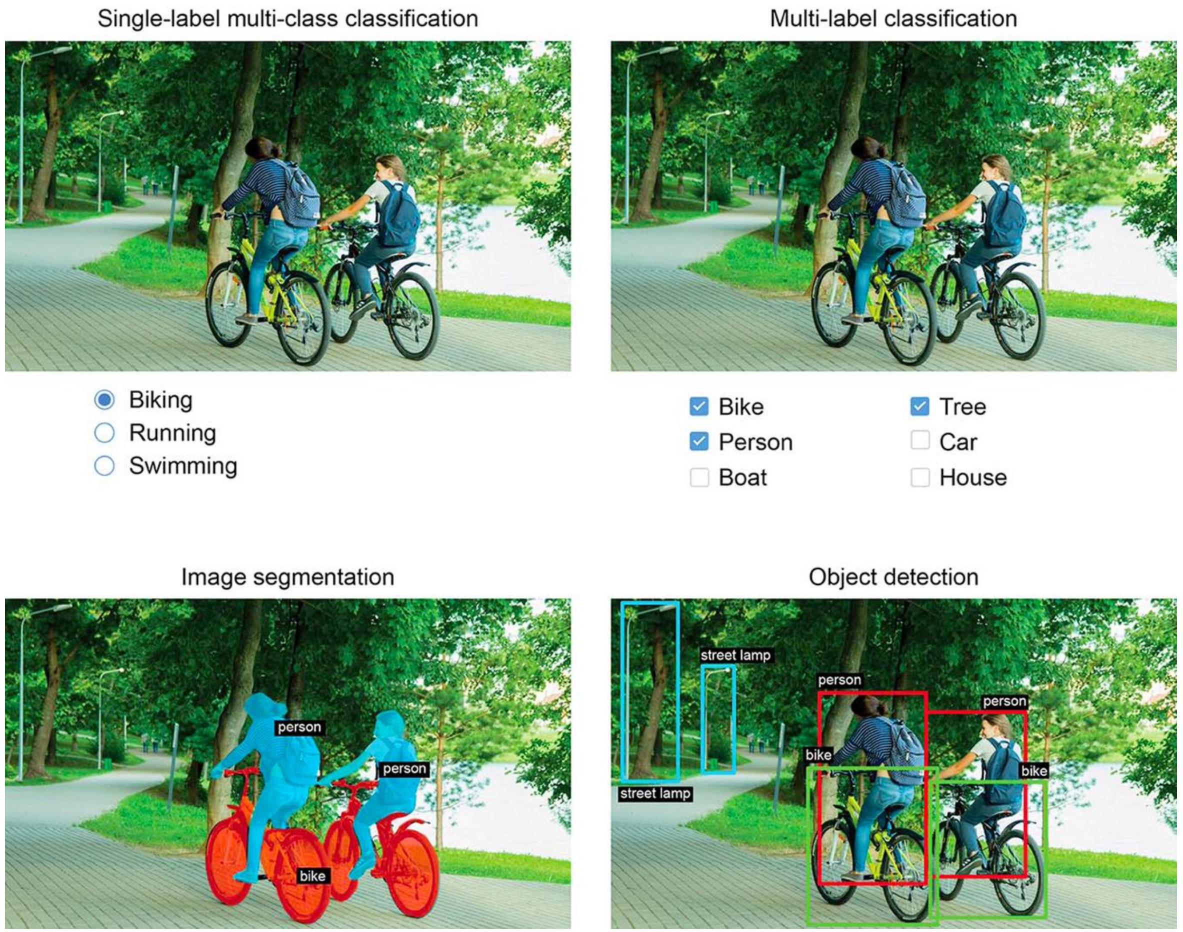
- This Slide Deck has Material for 6 hours of teaching divided into Parts 1-6
- We will go through
  - Week 01
    - Part 01 - Convolutional and Pooling Layers; AST 01
    - Part 02 - Transfer Learning and Modern CV Design Principle; AST 02
  - Week 02
    - Part 01 - Modern Convolutional Building Blocks for Image Classification; AST 03
    - Part 02 - Object Localization
    - Interpreting what convolutions learn (Advanced topic) – AST 03
  - Week 03
    - Part 01 - Object Detection (YOLO), Image Segmentation – Lec 05
    - Part 02 - Practical CVOps
    - AST04 – Object Detection with YOLO
  - Week 04
    - Revision
    - AST05 – Image Segmentation
- Additional Reading material to go in depth of math with references and code references are provided with the marking of “Additional Material” or “Additional Discussion” etc

Mini Projects

- Persons with face-masks
  - load the image dataset using ImageDataGenerator from the path directory
  - perform data augmentation on the fly and create batches of the dataset
  - build the convolutional neural networks for classification problem
  - visualize & interpret what CNN layers learn
  - use the transfer learning (pre-trained models) for classification problems
- Lungs Segmentation – Biomedical Image Analytics
  - understand, prepare, and visualize the ~~the~~ dataset containing image and corresponding masked image used for segmentation
  - implement DeepLabV3+ architecture
  - create a masked image (prediction)

Three Essential Tasks in Computer Vision

- Image Classification
  - Single Label
    - Binary
    - Multiclass
  - Multi Label
- Image Segmentation
  - Pixel wise identify the class
  - Example: Zoom background replacement
- Object Detection
  - Bounding box around objects
  - Self-driving cars, face detection in cameras



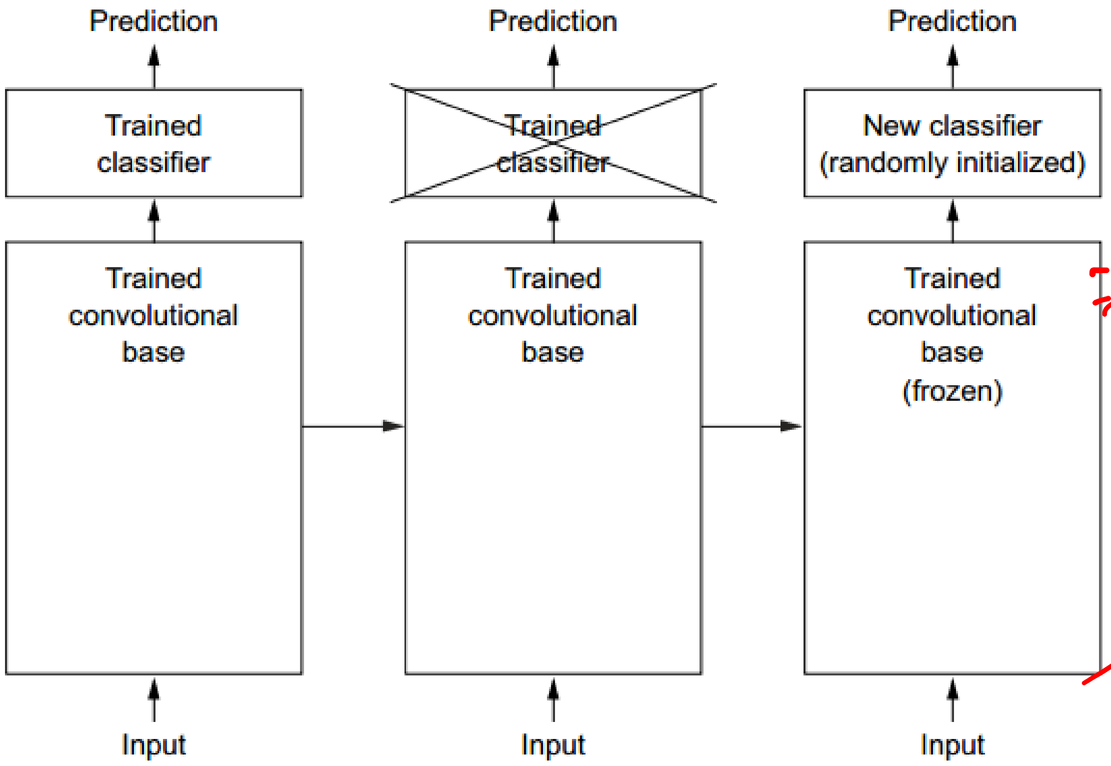
State of the Art as of May 2024

- https://paperswithcode.com/area/computer-vision
- Image Classification –
  - For speed: Efficient Net (CNN),
  - For accuracy: Vision Transformer
- Semantic Segmentation –
  - For speed: Unet, DeepLabv3
  - For accuracy: Deformable Convolution (InternImage), Vision Transformer (Segment Anything)
- Instance Segmentation - Mask-R-CNN, RetinaNet [Feature Pyramid]
- Object Detection –
  - For speed: YOLOv10
  - For accuracy: Deformable Convolution (InternImage)

Recommended Strategy

- Small Dataset (<1000 labelled images) – Use Transfer Learning
- Medium Dataset (Upto 5000-10000) – Use Fine Tuning
- Large Dataset (Beyond 10k) – Train from scratch
  - Rules of thumb!
- Edge Devices use MobileNet
- SoTA needed? – Use Efficient Net (or even ViT)
- Traditional firms who like time-tested methods
  - ResNet50, VGG19
- If training cost and inference time are not a concern, use all three and do an ensemble!

Transfer Learning



Trainable params

Tricks of the Trade

Data Augmentation
- Keras – ImageDataGenerator

Batch Normalization

Fine tuning – unfreeze layer by layer

Deepak Subramani, deepakns@iisc.ac.in

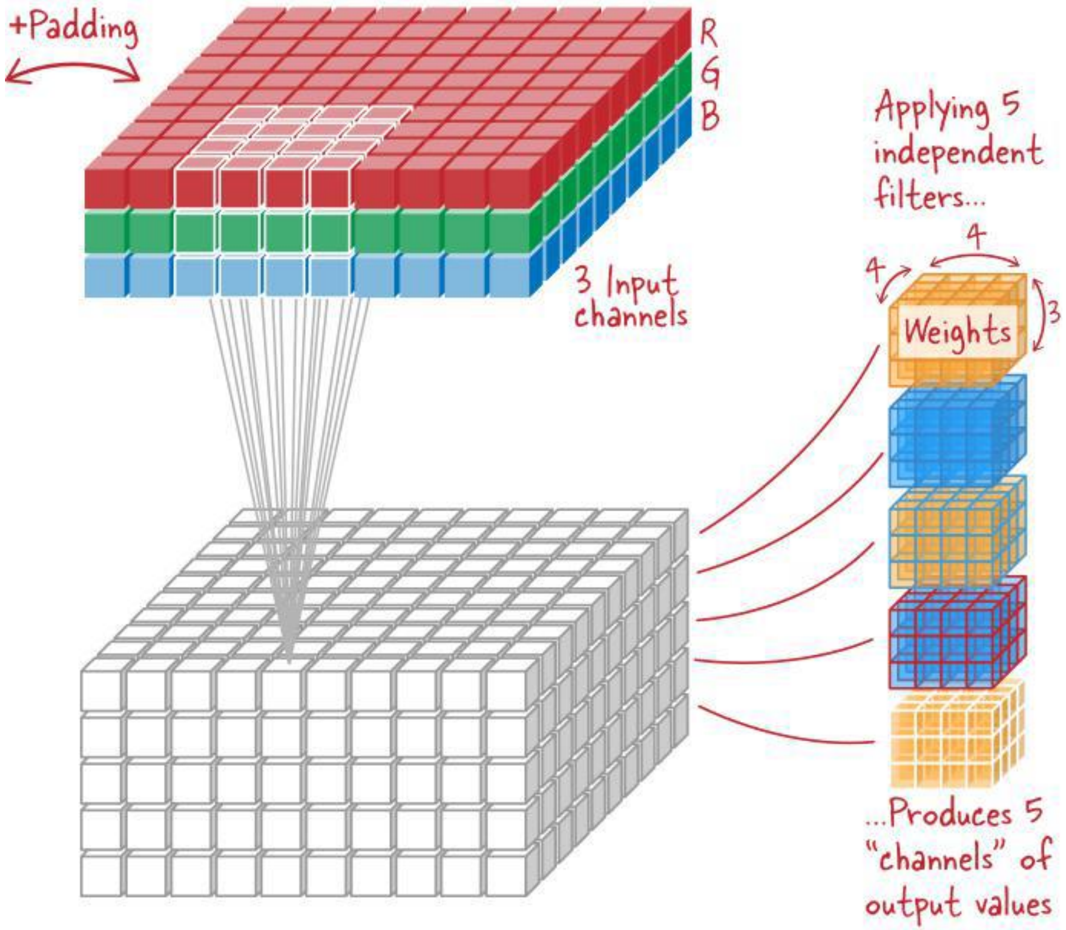
Labeling Data

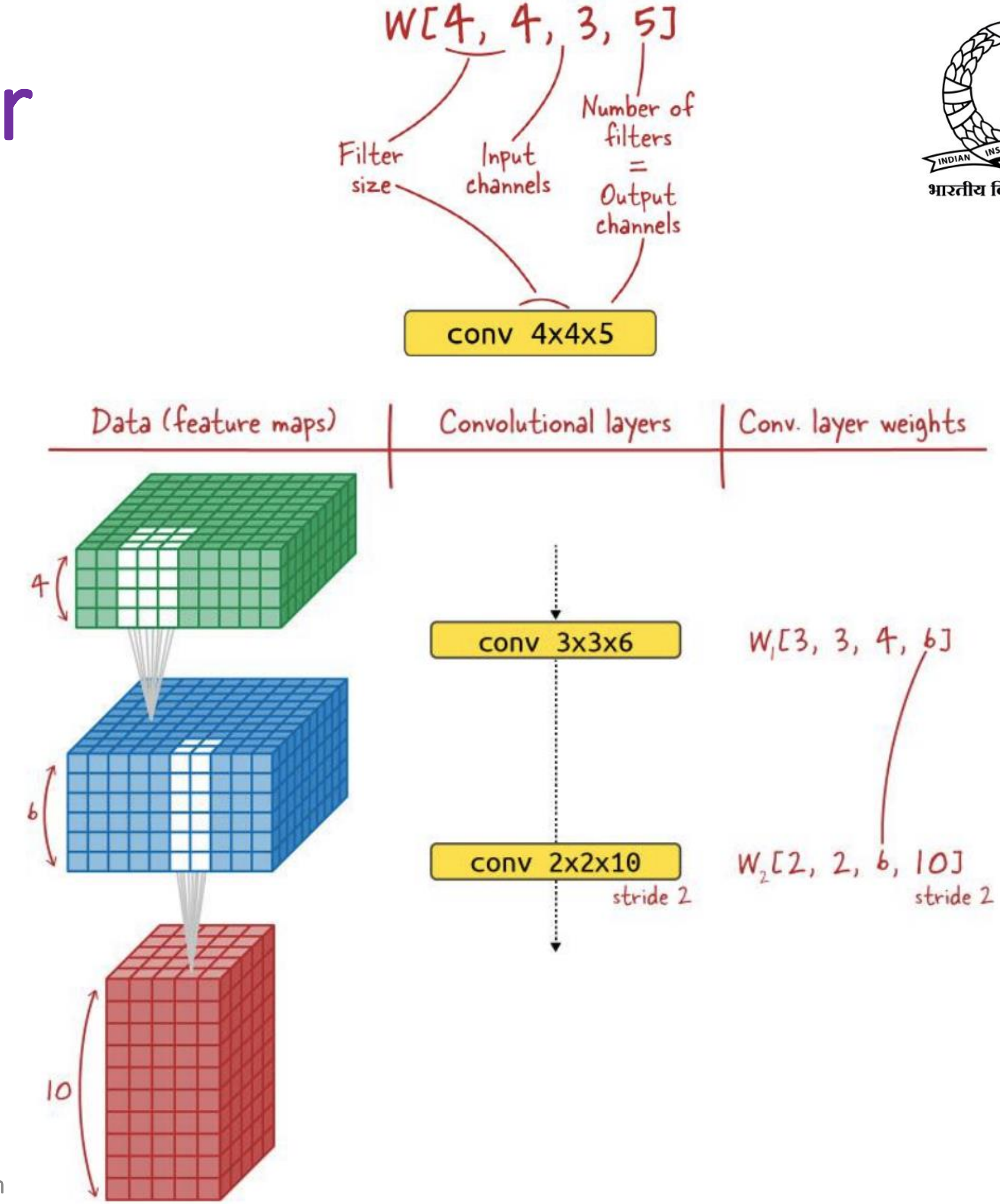
- For image classification there are two manual labeling approaches
  - Move the images to a folder whose name is the label
  - Create an excel file (spreadsheet) with first column having the path of the image and the other columns having the label(s)
- Object detection
  - Needs bounding box (usually counterclockwise starting from top-left)
- Segmentation
  - Needs labels of pixels
- OpenCV Computer Vision Annotation Tool
  - https://github.com/opencv/cvat
- AI Labeling Service from Cloud Providers

Theory Concepts

- Convolution
- Pooling
- Residual Connection
- Depthwise Separable Convolution
- Inverted Residual Bottleneck (Efficient Net)
- Transpose Convolution
- Atrous Convolution
- Batch Normalization
- Fully Convolutional Network
- Evaluation Metrics (IoU, mAP)

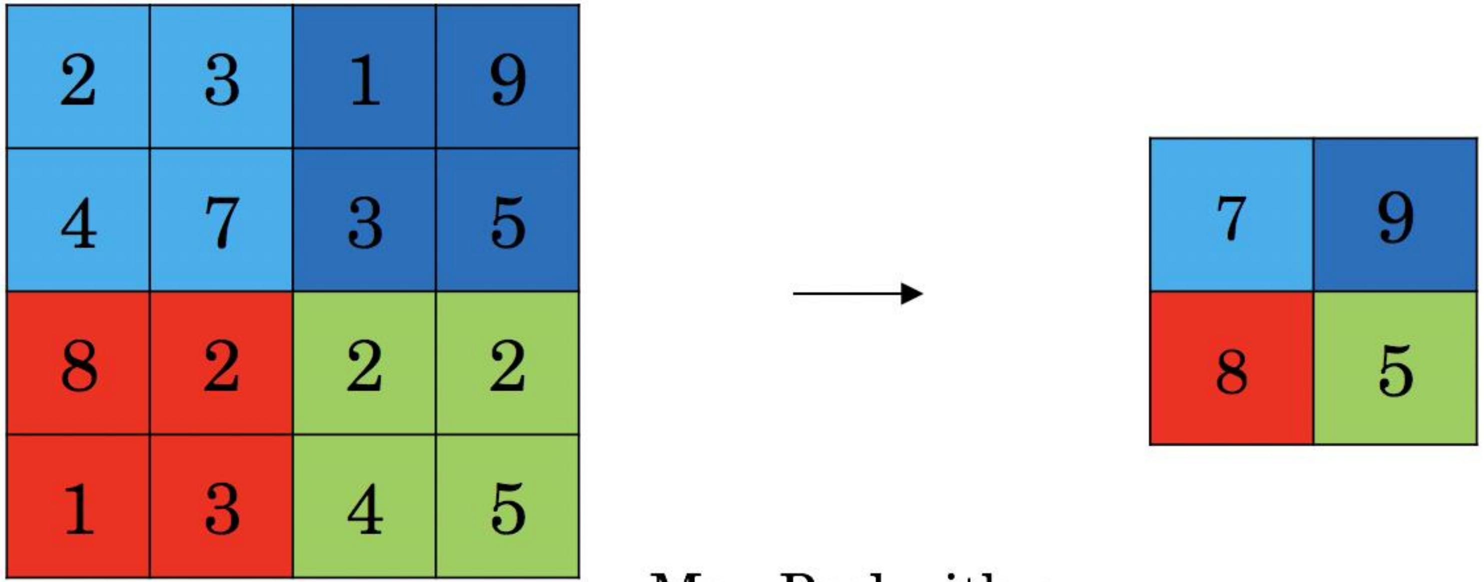
Convolutional Layer





Pooling Layer

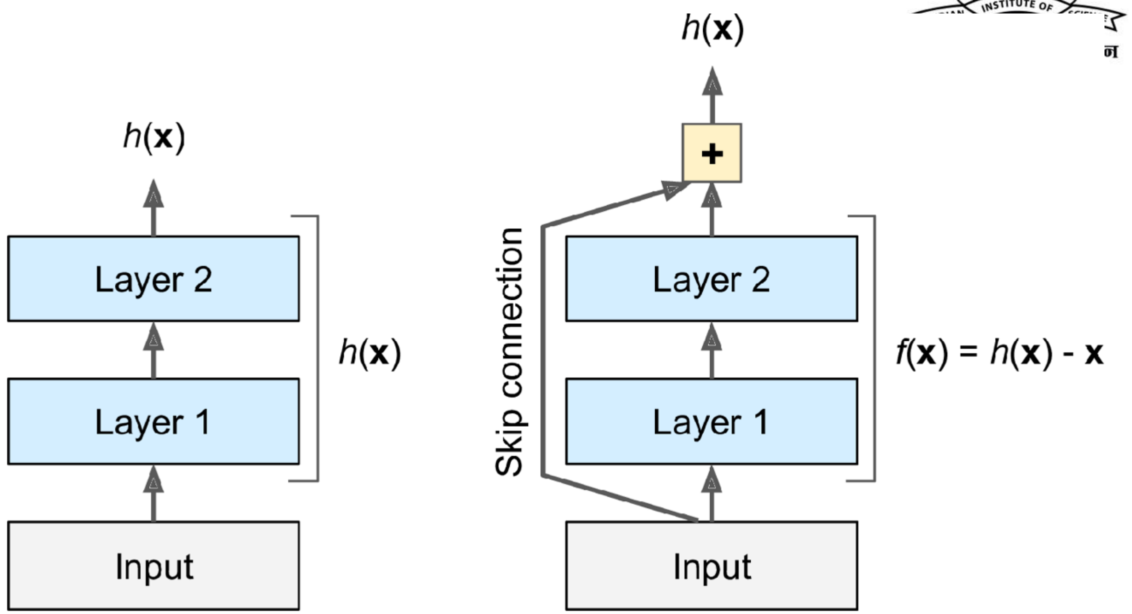
Max Pool

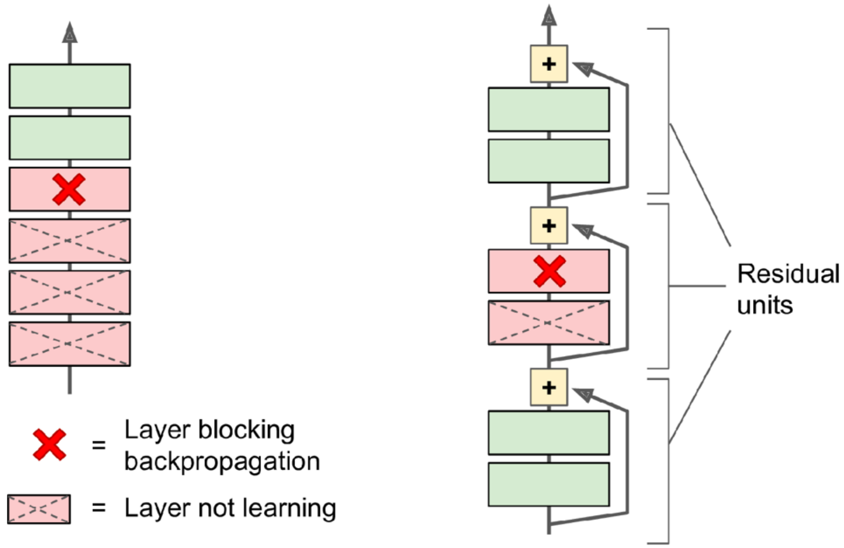


Max-Pool with a 2 by 2 filter and stride 2.

Residual Block

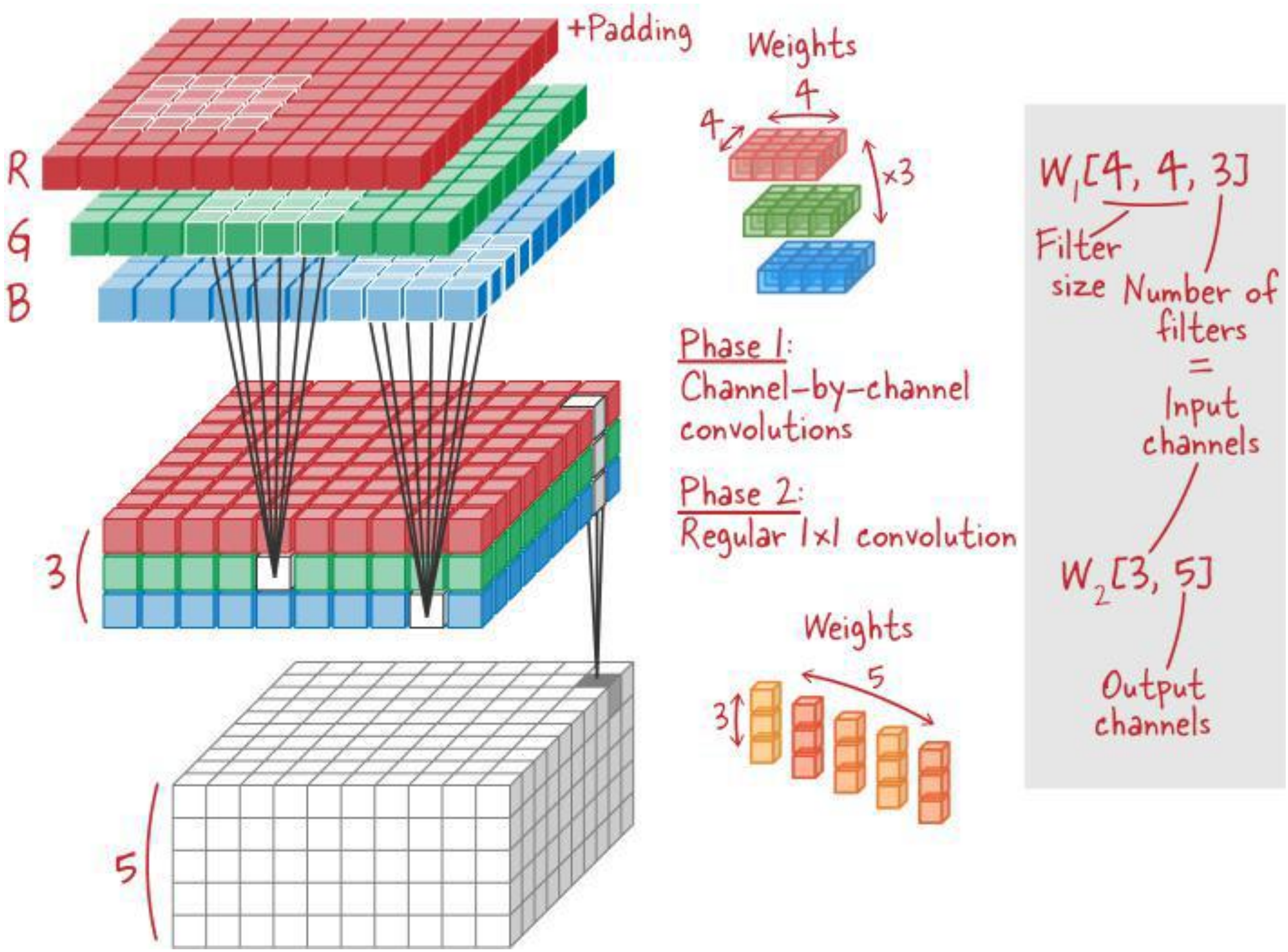
- 2015 winner is ResNet that used a residual block
- Networks were being deeper and residual (or skip connections) enabled training such deeper networks
- Usually networks are trained to learn a function $h(x)$
- By adding a skip connection, we are forcing the network to learn $f(x) = h(x) - x$
- When stacking several Residual Units, the signal can make its way to all the parts of network even if some layers experience a vanishing gradient





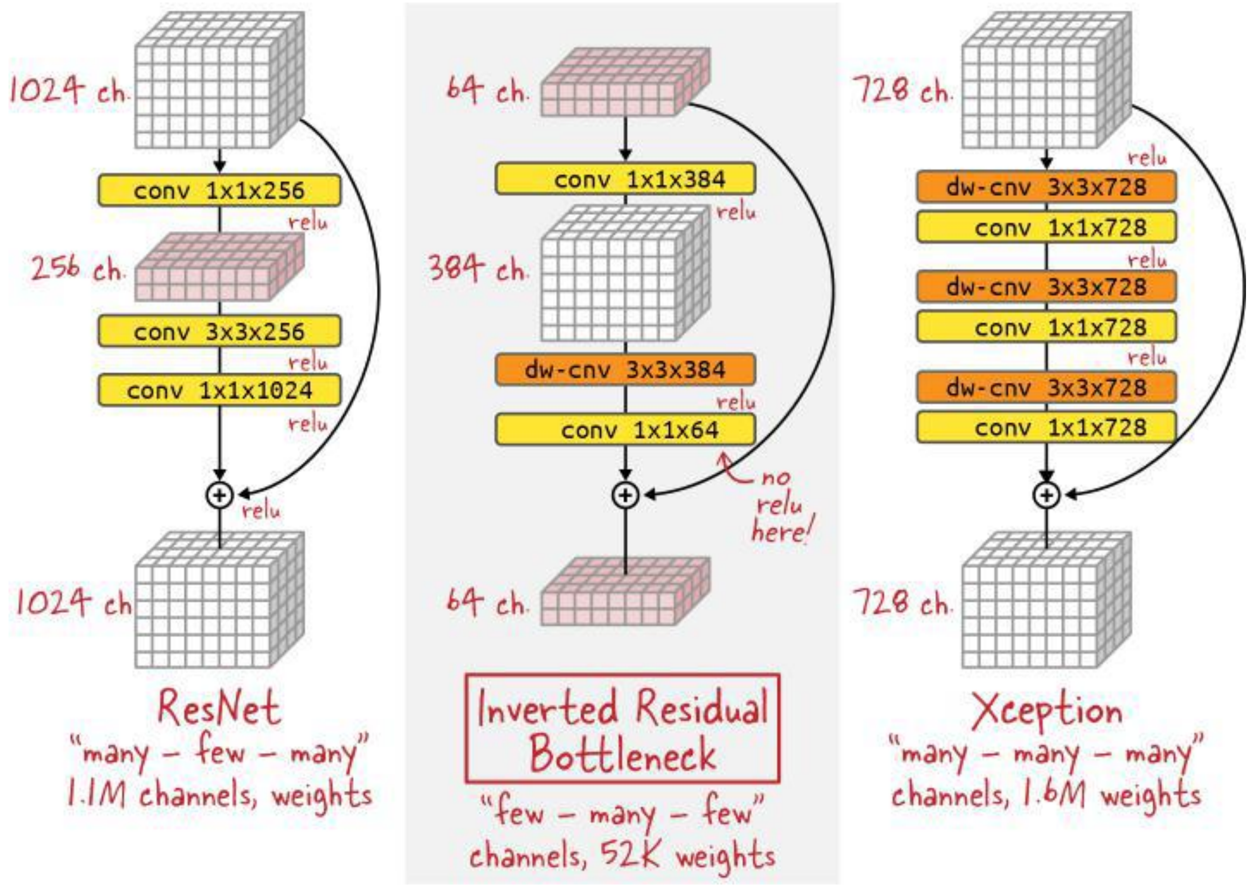
Depthwise Separable Convolutions

- Channel-by-channel convolutions followed by 1x1 Conv



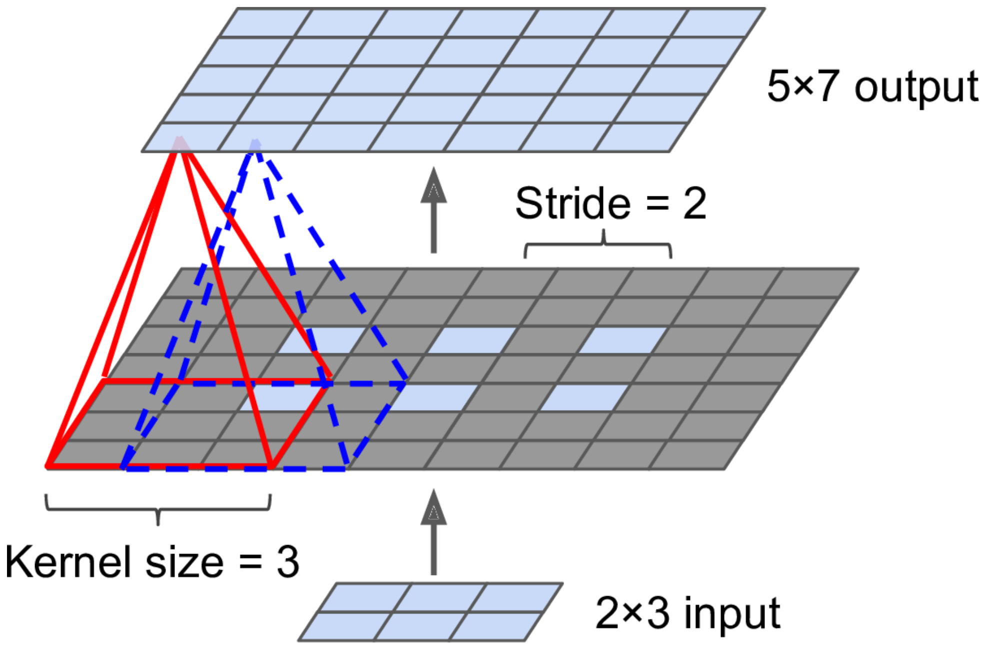
Inverted Residual Bottleneck

- Goal: Same expressivity as ResNet, Xception but with a dramatically reduced weight count and inference time
- Designed to be used on mobile phone where resources are scarce
- Argument: Information flow between residual blocks is low-dim in nature and can be represented by limited number of channels
- Important: Last 1x1 doesn't have any nonlinear activation as ReLU would destroy too much information



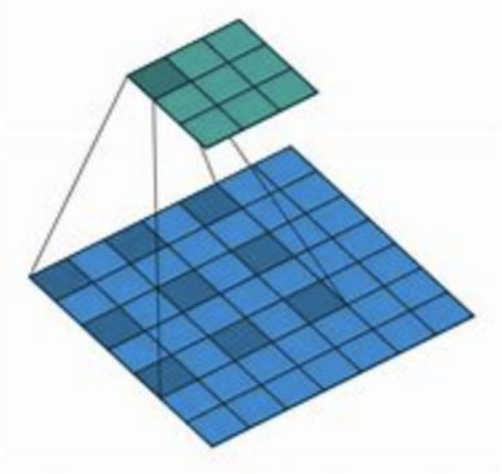
Transposed Convolution

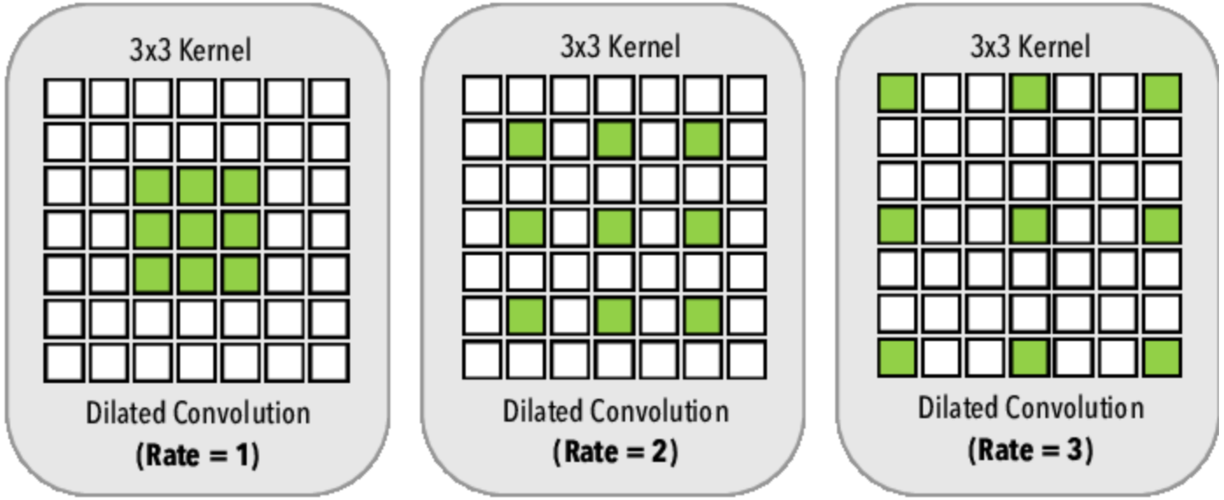
- Think of stretching an image by adding empty rows and columns
- Then on the stretched image do a regular convolution
- Initialize these kernels to do a linear interpolation
- But as the weights are learnable, it does better!



Atrous Convolution

- The convolution field of view is modified by considering a larger area with zeros added to the filter itself
- Number of learnable parameters is the same as regular convolution, but now the field of view has changed
- This is used in Deep Lab



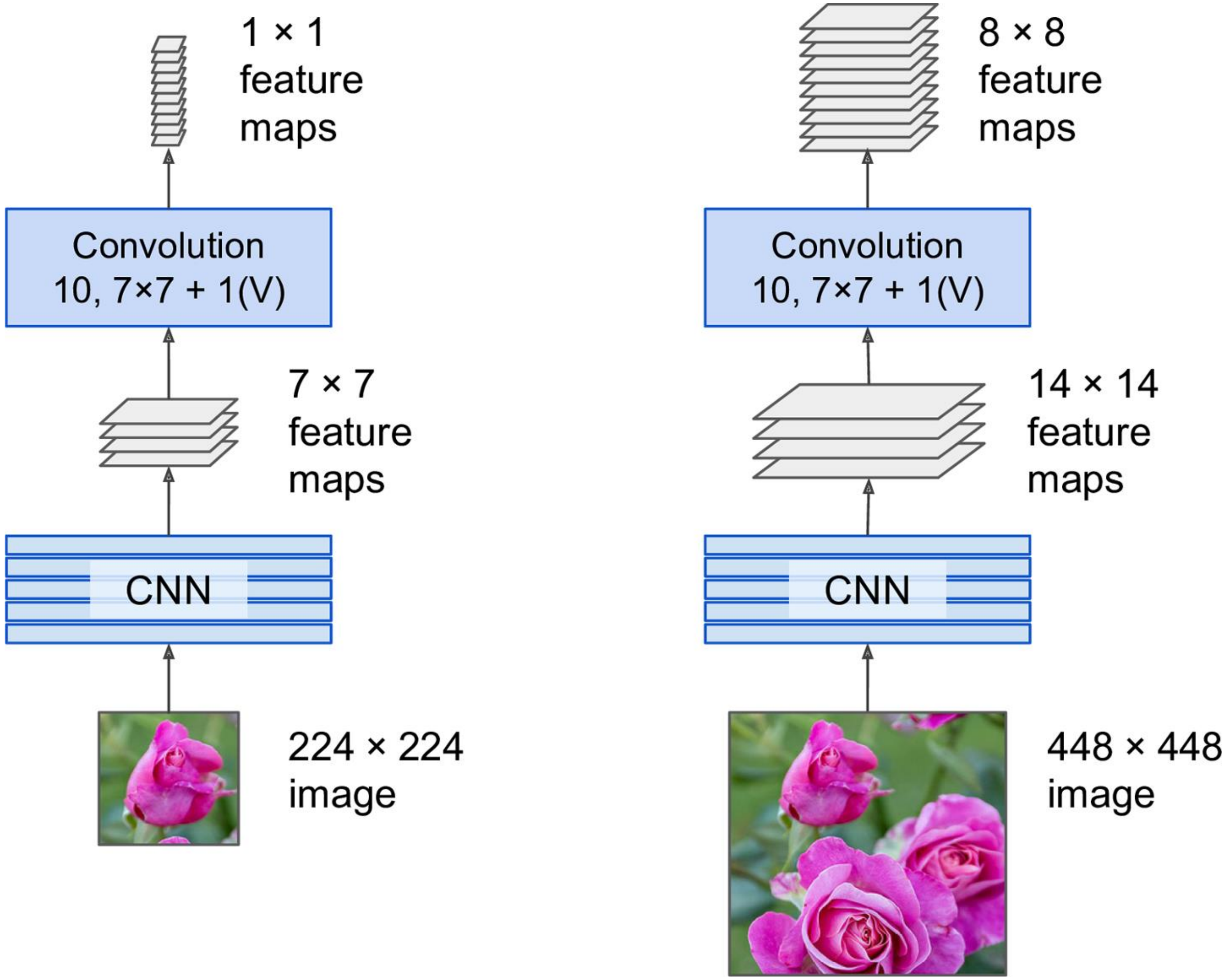


Batch Normalization

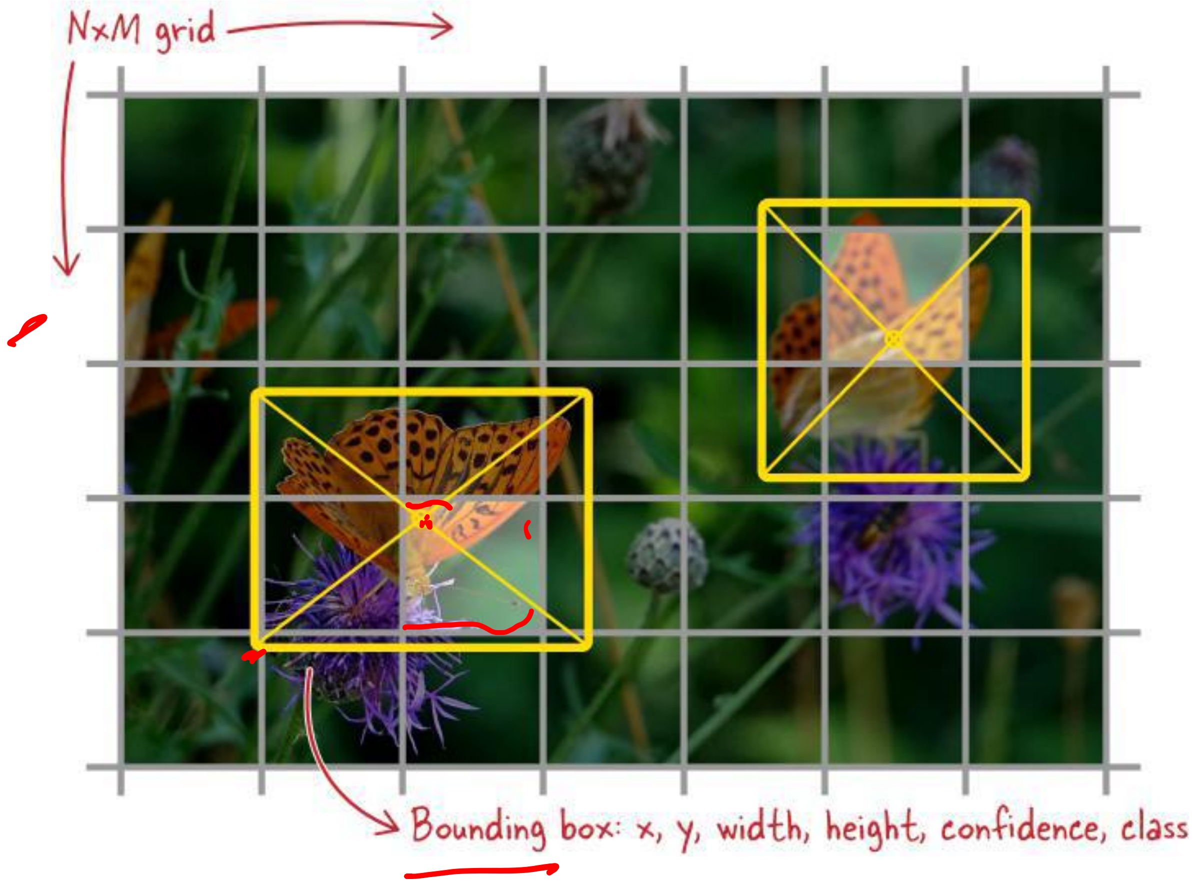
- He initialization + ELU can reduce vanishing/exploding gradient problem at the beginning, but problems can recur later during training
- Batch Normalization (Ioffe and Szegedy 2015) solves this problem
- Idea:
  - Zero center and normalize before or after activation function of every layer
  - Learn two parameter vectors (one set for every input) – output scaling and output shift – i.e., learn the optimal mean and scale of each of the layer's inputs!
- Question:
  - Need a batch to calculate the mean and std for scaling
  - Use the current mini batch to get the mean and std
- Note:
  - Add BN after input layer, then it is almost equivalent to applying StandardScaler, but only on the mini-batch and not the full train set

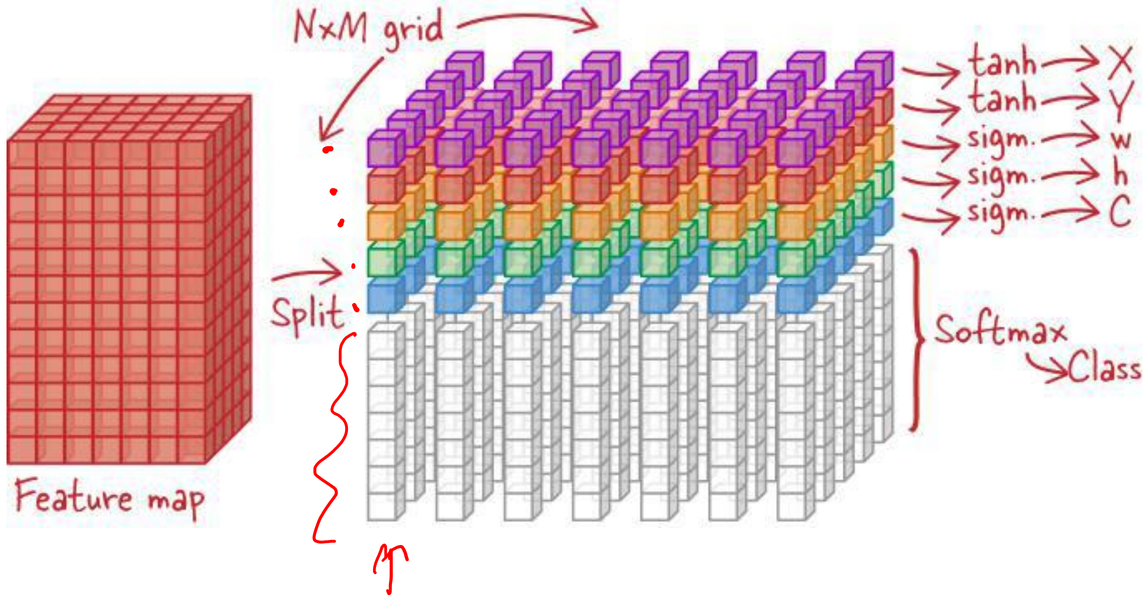
FCN Example (Cont)

- What happens if we feed 448x448 images to this FCN?
  - Last conv layer is 14x14, and it will produce a 8x8 map
  - What is this 8x8 map? – It is equivalent to sliding the original CNN across the image
- Now the network has to be run only once
- You Only Look Once (YOLO)

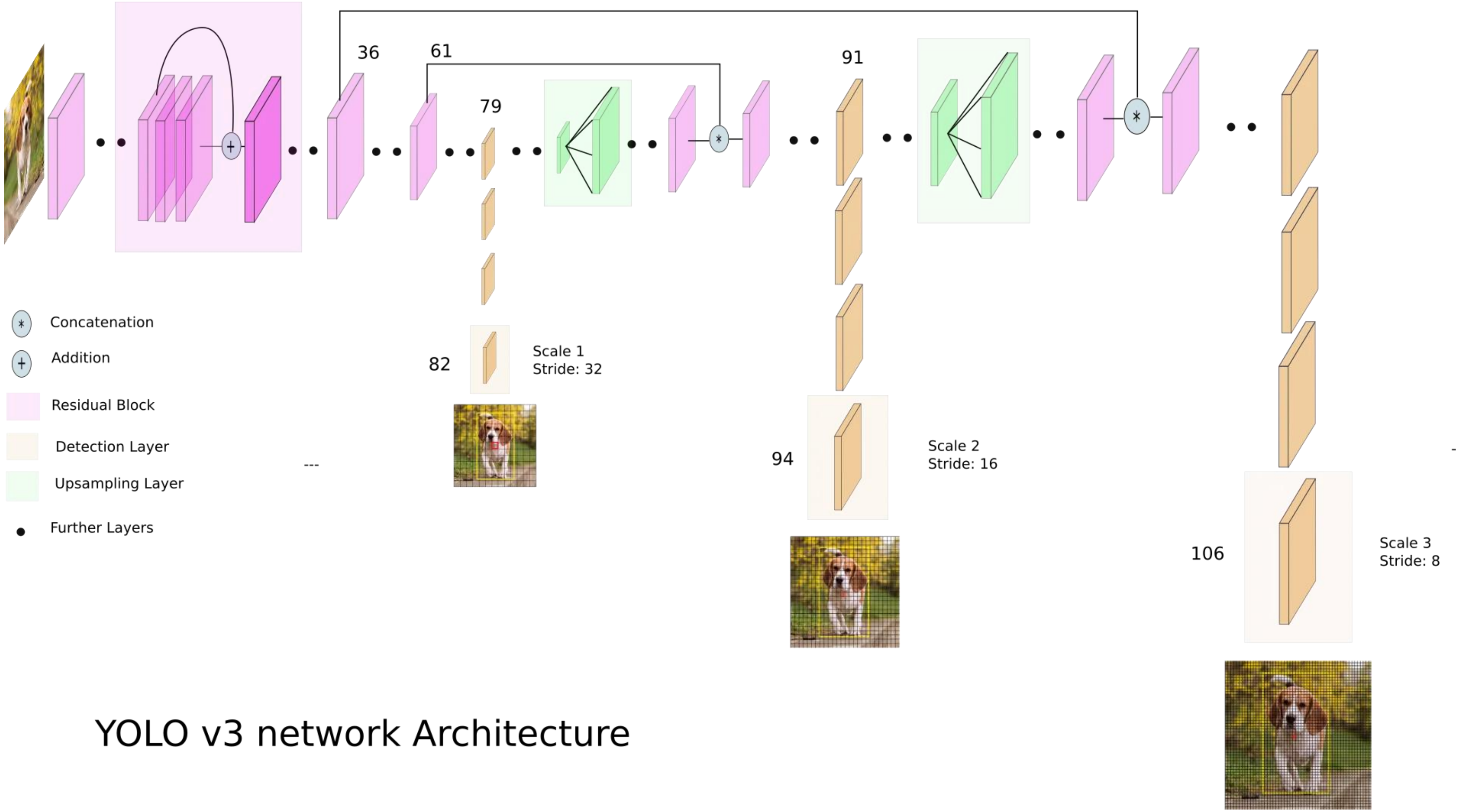


YOLO Visually



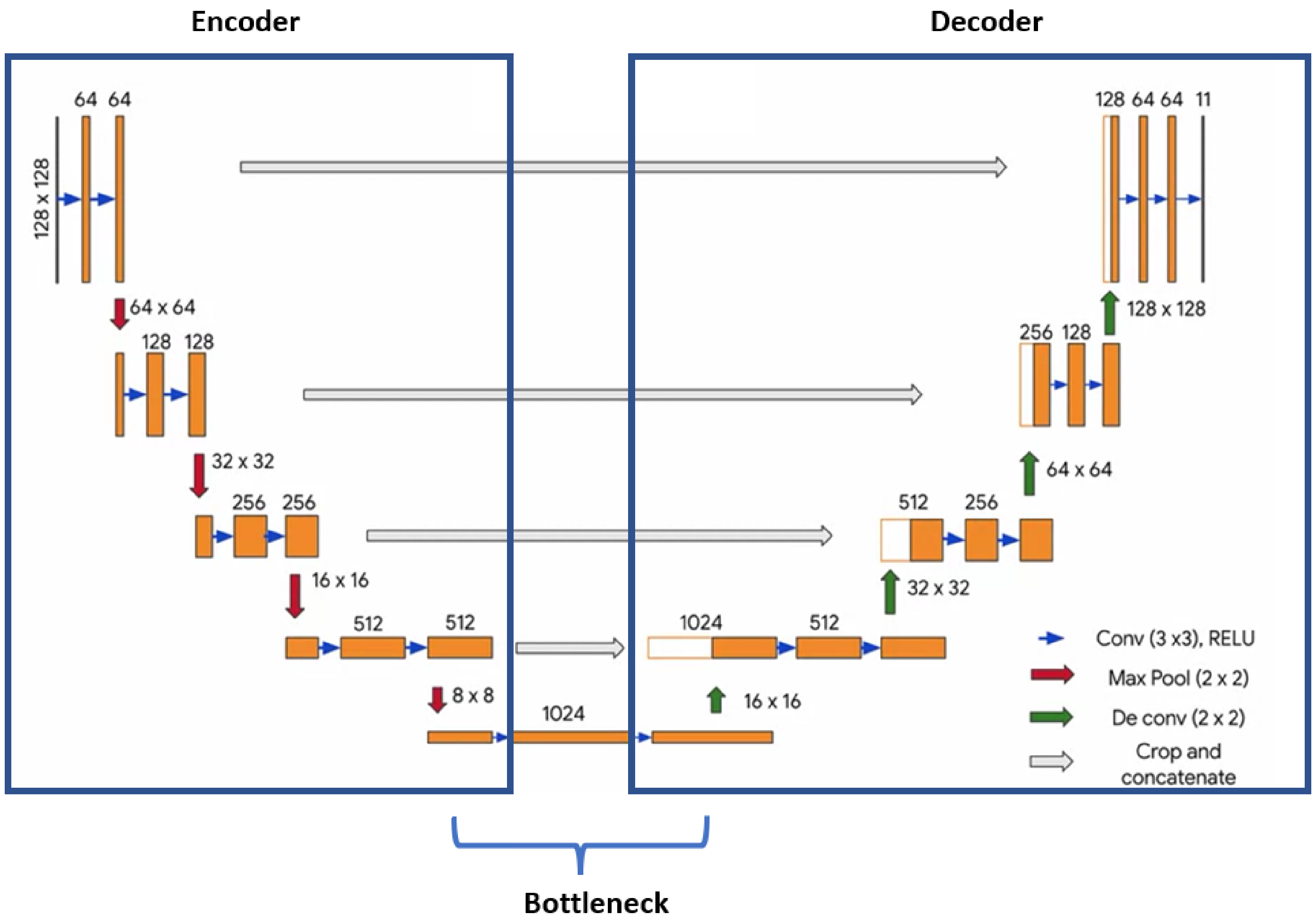


YOLO v3



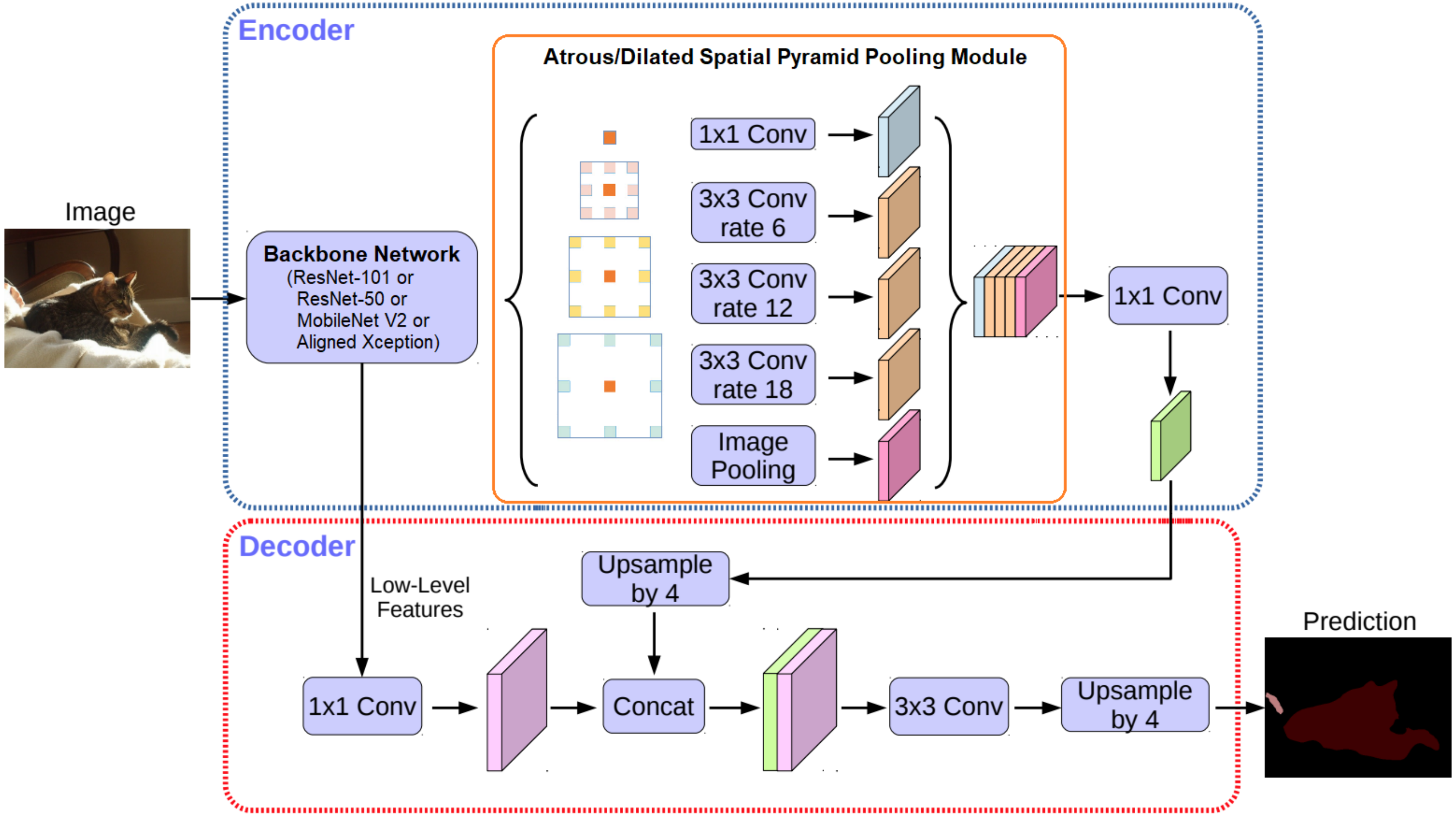
YOLO v3 network Architecture

U-Net



Encoder
Decoder
64 64
128 x 128
64 x 64
128 128
32 x 32
256 256
16 x 16
512 512
8 x 8
1024
1024 512
16 x 16
32 x 32
512 256
64 x 64
256 128
128 x 128
128 64 64 11
Conv (3 x3), RELU
Max Pool (2 x 2)
De conv (2 x 2)
Crop and concatenate
Bottleneck

DeepLabv3



Encoder
Atrous/Dilated Spatial Pyramid Pooling Module
Image
Backbone Network
(ResNet-101 or
ResNet-50 or
MobileNet V2 or
Aligned Xception)
1x1 Conv
3x3 Conv
rate 6
3x3 Conv
rate 12
3x3 Conv
rate 18
Image
Pooling
1x1 Conv
Decoder
Low-Level
Features
Upsample
by 4
1x1 Conv
Concat
3x3 Conv
Upsample
by 4
Prediction

Evaluation Metric

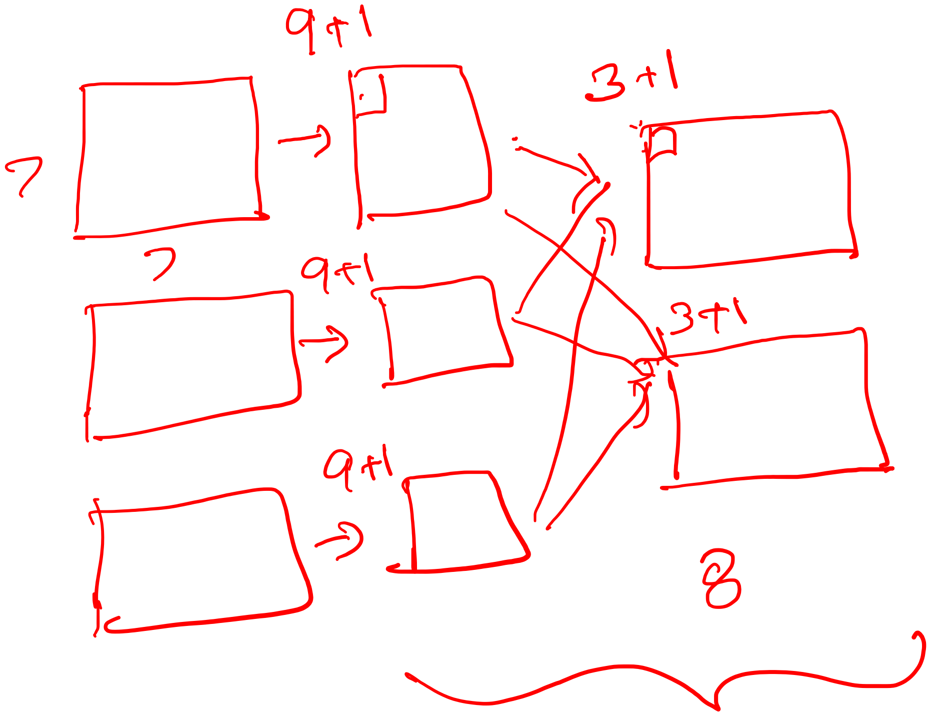
- IoU – Intersection over union
- mAP - Mean Average Precision

Other Tasks in CV

- Pose Estimation
- Object Tracking
- Action Recognition
- Motion Estimation
- Monocular Depth
- Content-aware Image Editing
- Scene Reconstruction (NeRF Neural Radiance Fields)
  - novel views of complex scenes

DW Separable

7x7x3   2 filters   3x3



30

38

DL 100

$50 \times 5 \times 5 \rightarrow 100$

5×5

$50 \times 5 \times 5$

1   5   5 $\rightarrow$ $\frac{25 \times 50}{}$

2

⋮

50

0
0
0
⋮
0
100

$100 \times 1 \times 1$

$\rightarrow 5 \times 5 \times 50 \times 100$

$\rightarrow 100$

GT		P	
0	$\rightarrow$	0	48%
		1	52%
		2	