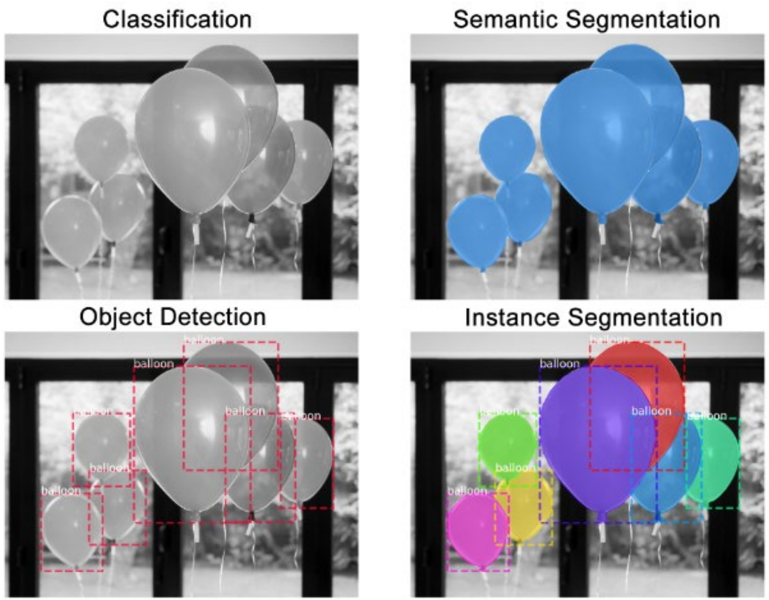
Machine Learning and Image Processing

Urban Information Lab

Computer Vision Tasks



Classification: What is in this image (There is a balloon in this image)

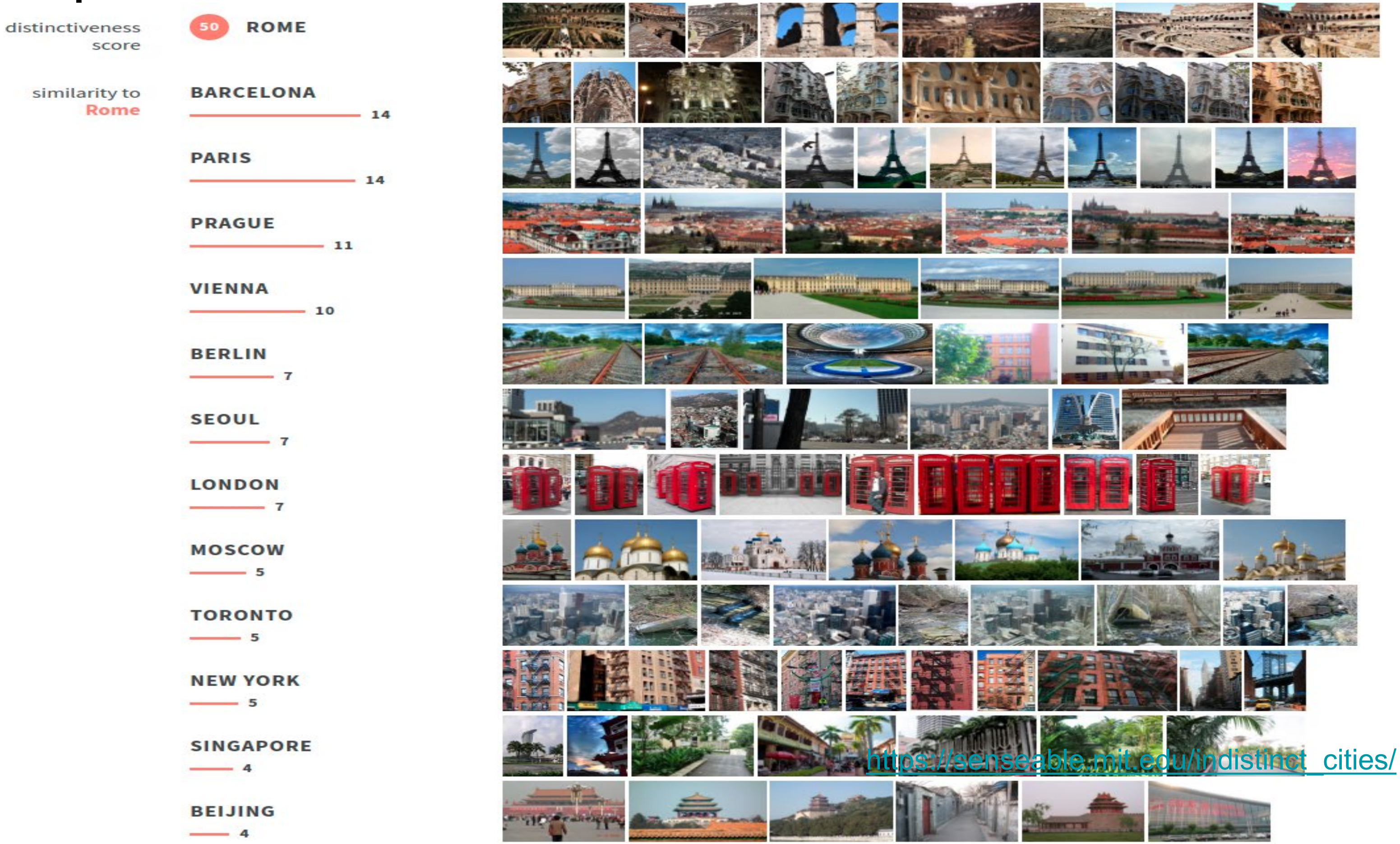
Semantic Segmentation: Masks of objects (These are all the ballon pixels)

Object Detection: Count and location of objects (These are 7 balloons in this image at these locations. We’re starting to account for objects that overlap.

Instance Segmentation: location and pixel for each object (There are 7 balloons at these locations, and there are the pixels that belong to each one)

https://medium.com/@tibastar/mask-r-cnn-d69aa596761f

Computer Vision Tasks - Classification



https://senseable.mit.edu/indistinct_cities/

Computer Vision Tasks - Classification

The model aims to predict from which city a given image comes. The percentage of images that the model correctly classifies from the city is its distinctiveness score. The most distinct images from each city are displayed below.

While each city has its own visual identity, they often share visual similarities with one another. We quantify the similarity between two given cities as the summed percentage of images that the model misclassified as the other city.

https://senseable.mit.edu/indistinct_cities/

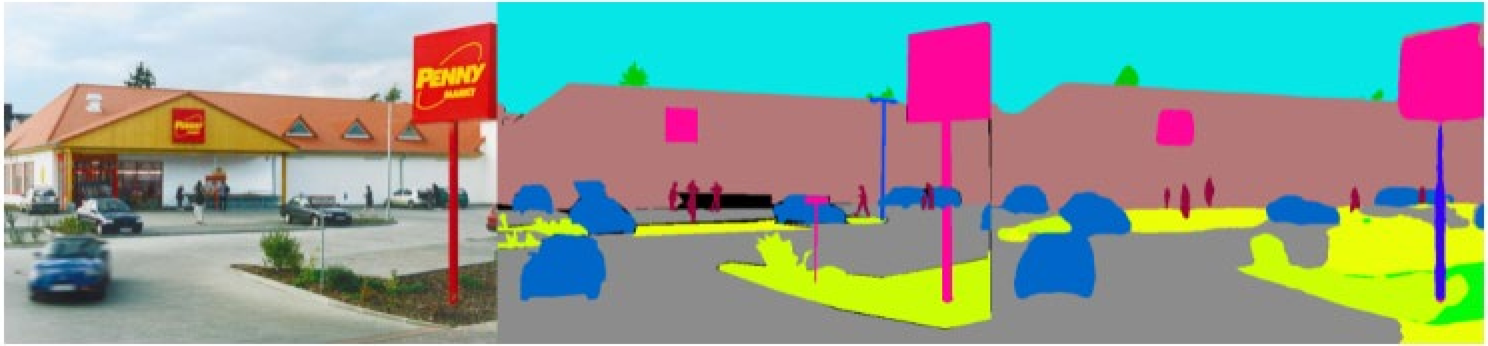
Computer Vision Tasks - Semantic Segmentation

Self-driving cars
Self-driving cars require image capturing sensors that could enable them to visualize the environment, make decisions and navigate accordingly. Semantic segmentation allows for an effective differentiation between various objects.

https://www.v7labs.com/blog/semantic-segmentation-guide#:~:text=Semantic%20Segmentation%20is%20a%20technique,task%20at%20a%20pixel%20level.

Computer Vision Tasks - Semantic Segmentation

Scene understanding
Scene understanding applications require the ability to model the appearance of various objects in the scene like buildings, trees, roads, billboards, pedestrians, etc. The model must learn and understand the spatial relationship between different objects.



https://www.v7labs.com/blog/semantic-segmentation-guide#:~:text=Semantic%20Segmentation%20is%20a%20technique,task%20at%20a%20pixel%20level.

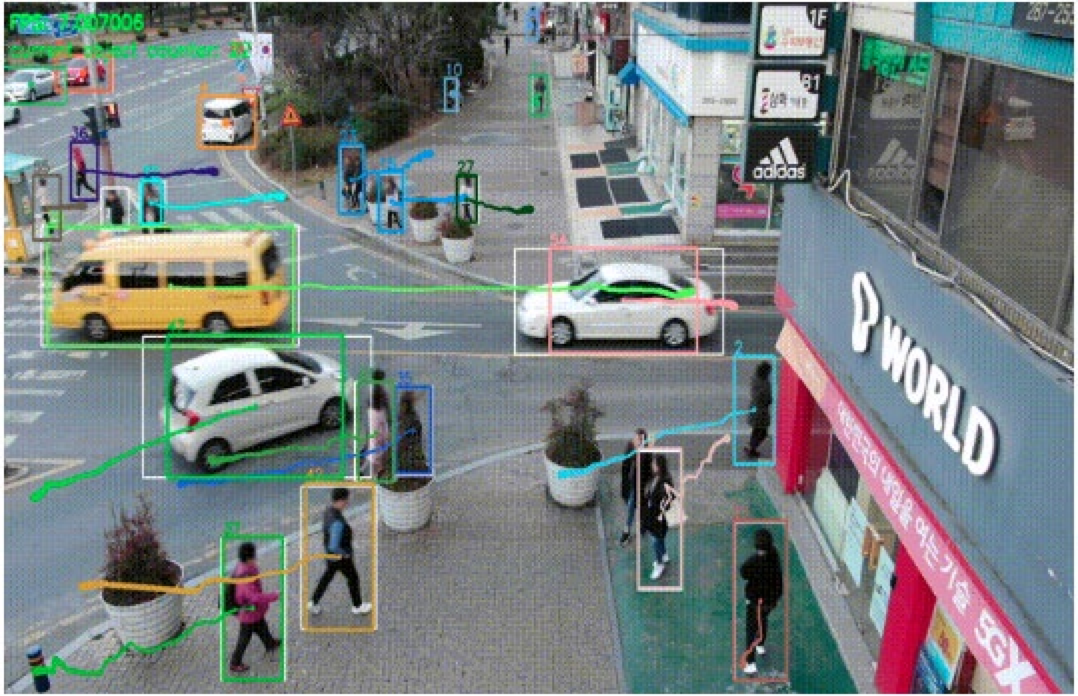
Computer Vision Tasks - Semantic Segmentation

Aerial image processing
Aerial image processing is similar to scene understanding, but it involves semantic segmentation of the aerial view of the landscape. This type of technology is very useful in times of crisis like a flood, where drones can spread to survey different areas to locate people and animals who need rescuers. Another area where aerial image processing can be used is the air delivery of goods.

https://www.v7labs.com/blog/semantic-segmentation-guide#:~:text=Semantic%20Segmentation%20is%20a%20technique,task%20at%20a%20pixel%20level.

Computer Vision Tasks - Object Detection

Video surveillance



https://laptrinhx.com/yolo-rcnn-object-detection-and-multi-object-tracking-872064671/

Computer Vision Tasks - Object Detection

Crowd counting

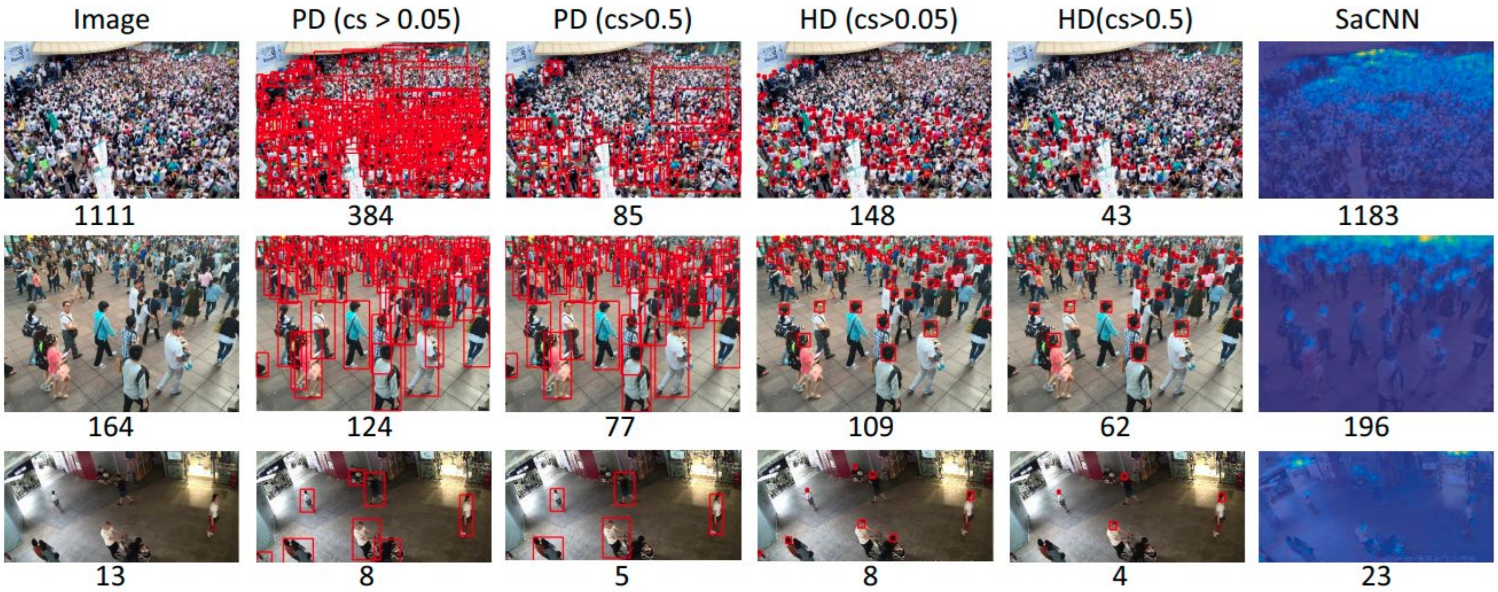
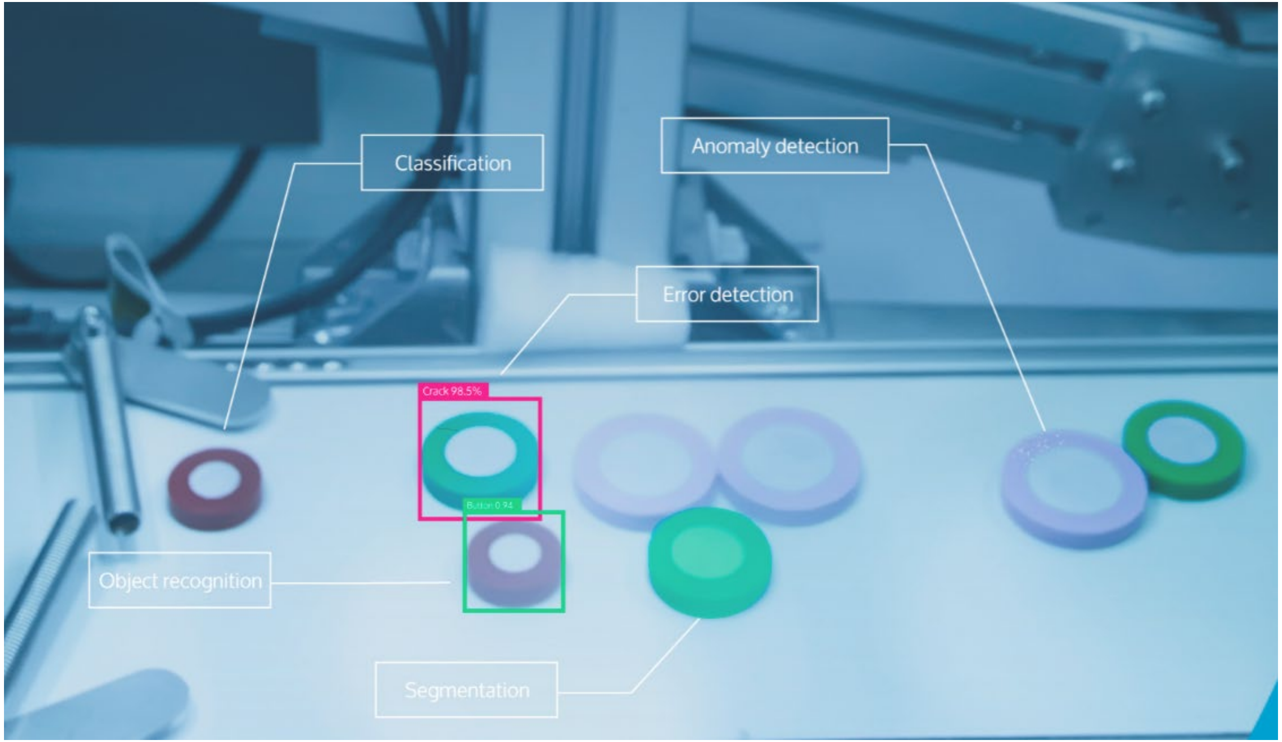


Figure 7. Comparison between YOLO9000 [22] and SaCNN. PD: pedestrian detector; HD: head detector; *cs* confidence score. The numbers below the real images (first column) are the ground truth. Numbers below other images are the estimated pedestrian counts.

https://hal.inria.fr/hal-01830946/file/wacv2018shi.pdf

Computer Vision Tasks - Object Detection

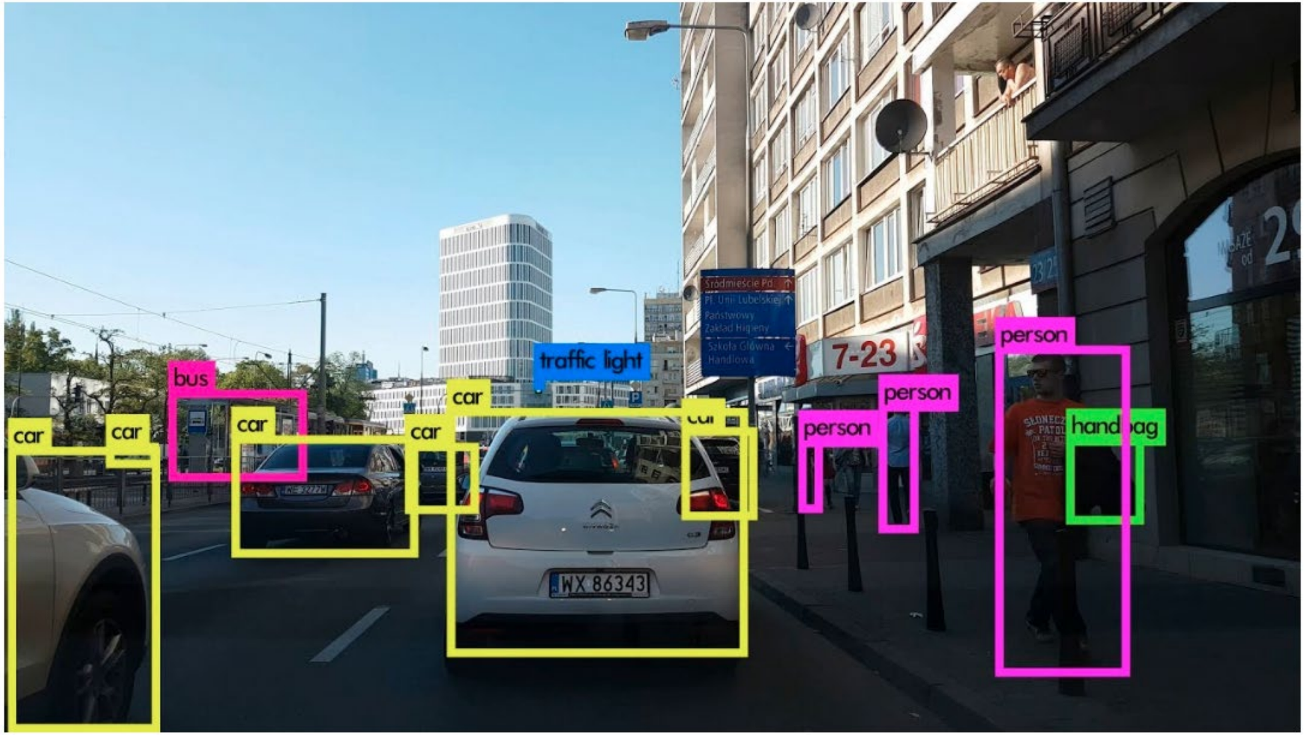
Anomaly detection



https://www.elunic.com/en/ai-see/

Computer Vision Tasks - Object Detection

Self-driving cars



https://neilnie.com/2018/11/18/implementing-yolo-v3-object-detection-on-the-autonomous-vehicle/

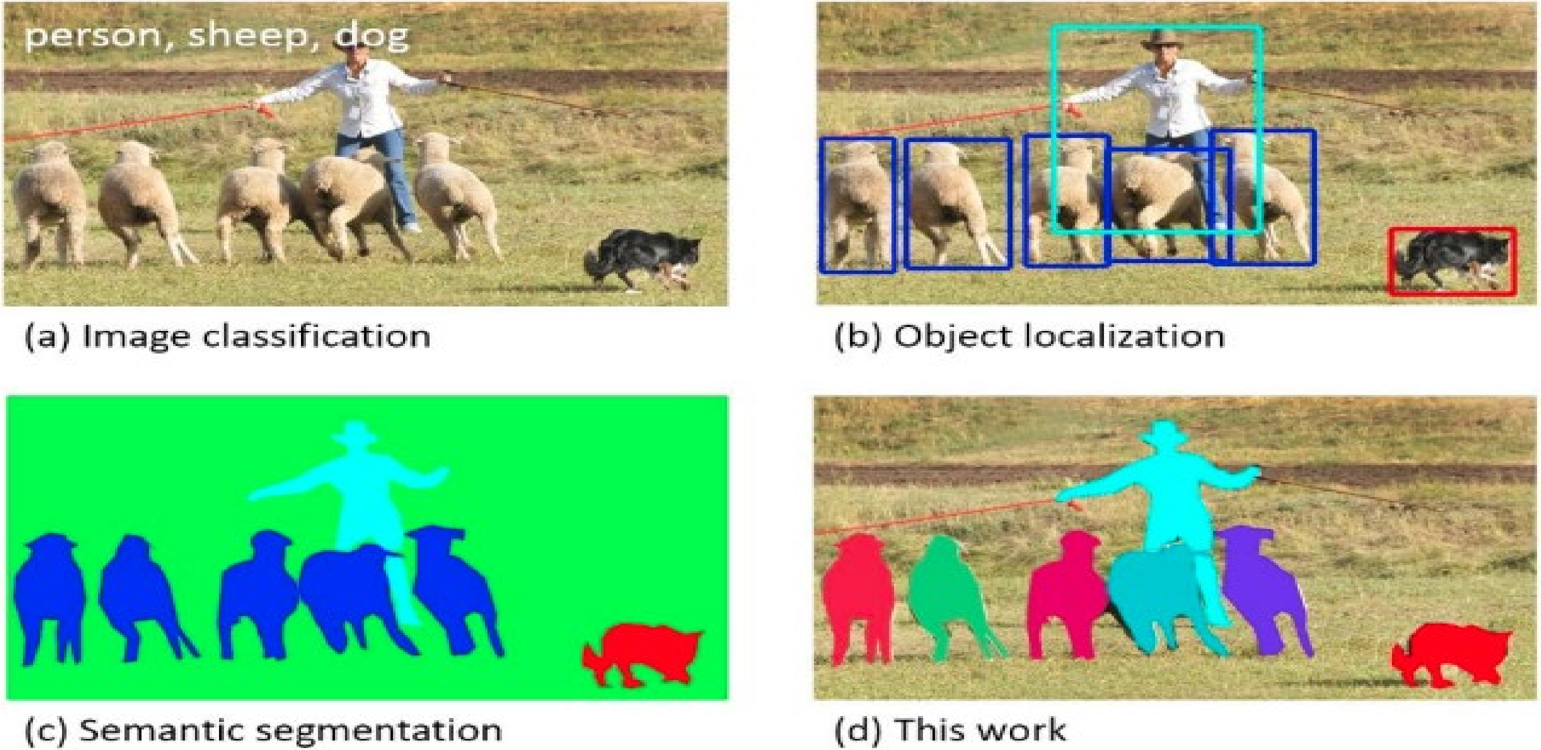
Computer Vision Tasks - Instance Segmentation

The basic difference between semantic segmentation and instance segmentation

Semantic segmentation associates every pixel of an image with a class label such as a person, flower, car and so on. It treated multiple objects of the same class as a single entity. In contrast, instance segmentation treats multiple objects of the same class as distinct individual instances.

https://analyticsindiamag.com/semantic-vs-instance-vs-panoptic-which-image-segmentation-technique-to-choose/#:~:text=Semantic%20segmentation%20associates%20every%20pixel,flower%2C%20car%20and%20so%20on.&text=In%20contrast%2C%20instance%20segmentation%20treats,class%20as%20distinct%20individual%20instances.

Computer Vision Tasks - Instance Segmentation



(a) Image classification

(b) Object localization

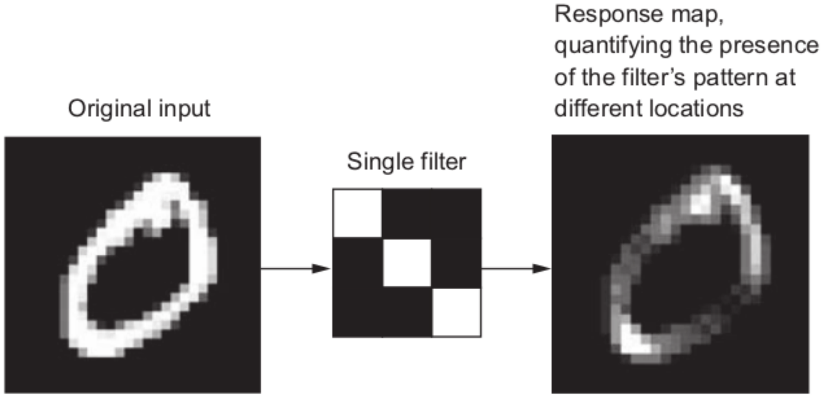
(c) Semantic segmentation

(d) This work

Convolutional Neural Network

Why CNN?

- Dense layers learn global patterns in their input feature space, but convolution layers learn local patterns.
- Characteristics of convnets:
  - The patterns they learn are translation invariant.
  - They can learn spatial hierarchies of patterns.



Chollet, F. (2017). *Deep learning with python*. Manning Publications.

Convolutional Neural Network

Why CNN?

- Dense layers learn global patterns in their input feature space, but convolution layers learn local patterns.
- Characteristics of convnets:
  - The patterns they learn are translation invariant.
  - They can learn spatial hierarchies of patterns.

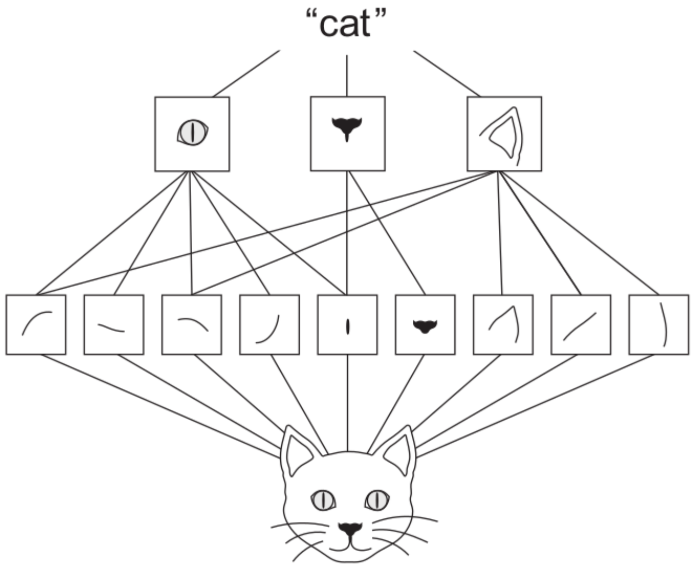


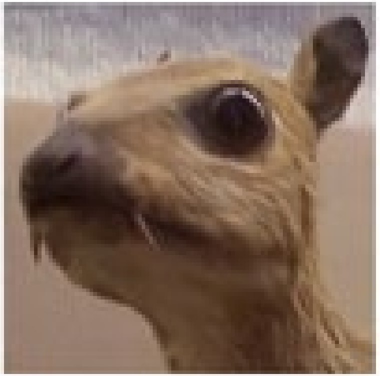
Figure 5.2 The visual world forms a spatial hierarchy of visual modules: hyperlocal edges combine into local objects such as eyes or ears, which combine into high-level concepts such as "cat."

Chollet, F. (2017). *Deep learning with python*. Manning Public

Convolutional Neural Network

Feature extraction

Input image



Convolution Kernel

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

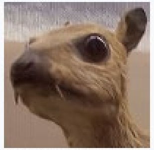
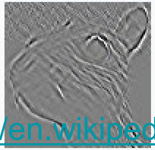
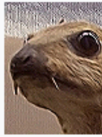
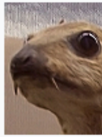
Feature map

https://timdettmers.com/2015/03/26/convolution-deep-learning/

Convolutional Neural Network

Filter/ Kernel

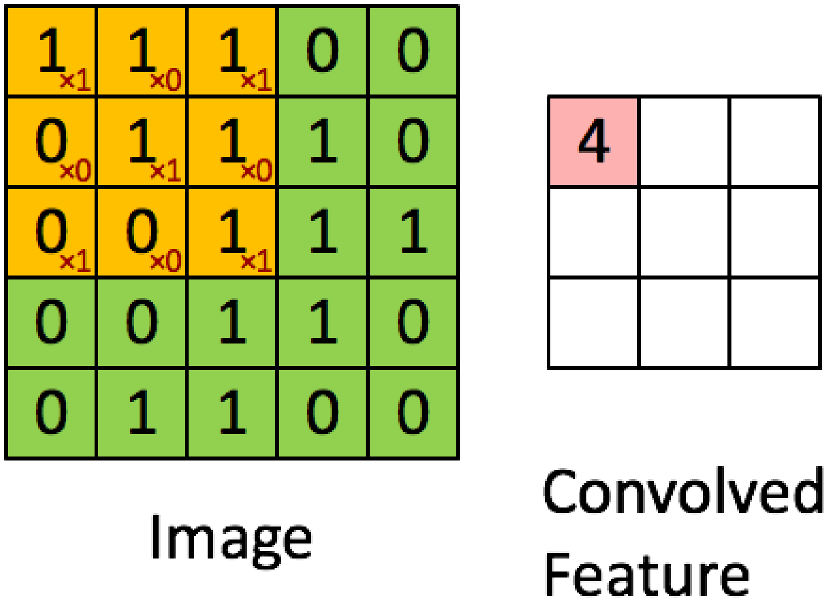
Depending on the element values, a kernel can extract different kinds of features and can cause a wide range of effects.

Operation	Kernel ω	Image result g(x,y)
Identity	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	
Ridge detection	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	
	$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$	
Sharpen	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	
Box blur (normalized)	$\frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	
Gaussian blur 3 × 3 (approximation)	$\frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$	
Gaussian blur 5 × 5 (approximation)	$\frac{1}{256}\begin{bmatrix} 1 & 4 & 6 & 4 & 1 \\ 4 & 16 & 24 & 16 & 4 \\ 6 & 24 & 36 & 24 & 6 \\ 4 & 16 & 24 & 16 & 4 \\ 1 & 4 & 6 & 4 & 1 \end{bmatrix}$	
Unsharp masking 5 × 5 Based on Gaussian blur with amount as 1 and threshold as 0 (with no image mask)	$\frac{-1}{256}\begin{bmatrix} 1 & 4 & 6 & 4 & 1 \\ 4 & 16 & 24 & 16 & 4 \\ 6 & 24 & -476 & 24 & 6 \\ 4 & 16 & 24 & 16 & 4 \\ 1 & 4 & 6 & 4 & 1 \end{bmatrix}$	

https://en.wikipedia.org/wiki/Kernel_(image_processing)

Convolutional Neural Network

Filter/ Kernel



https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

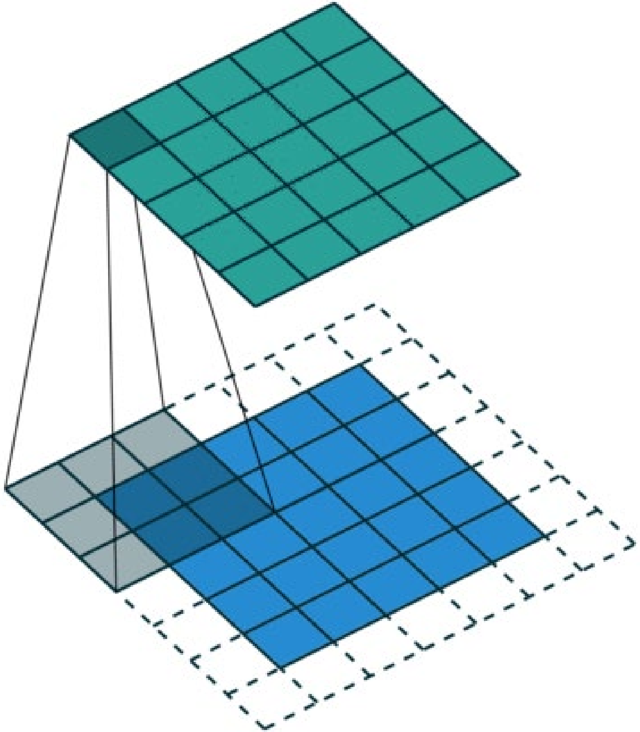
Convolutional Neural Network

Padding

Pad image to preserve the size of image

Preserve the edge information

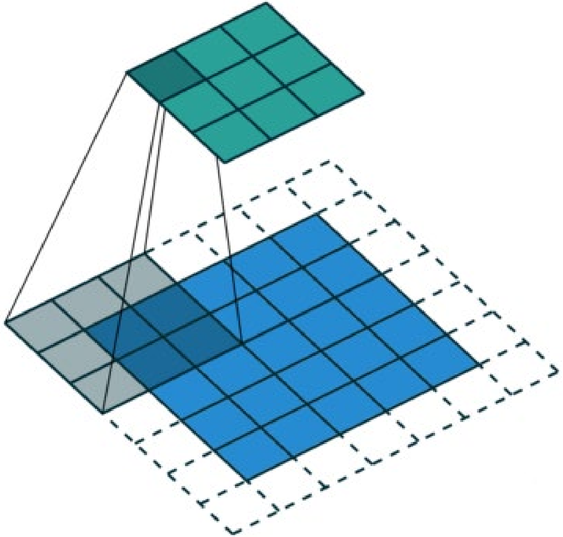
- Valid: nxn * fxf = n-f+1 * n-f+1 (no padding)
- Same: Pad so that output size is the same as the input size



https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Convolutional Neural Network

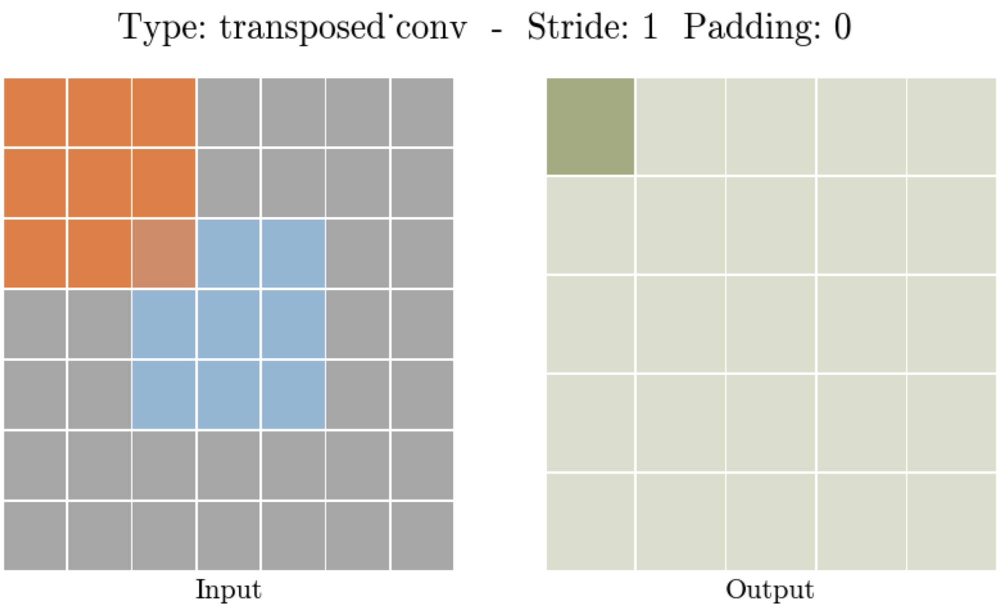
Strided convolution



Output feature size: (n + 2p - f)/s + 1 * (n + 2p - f)/s +1

https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

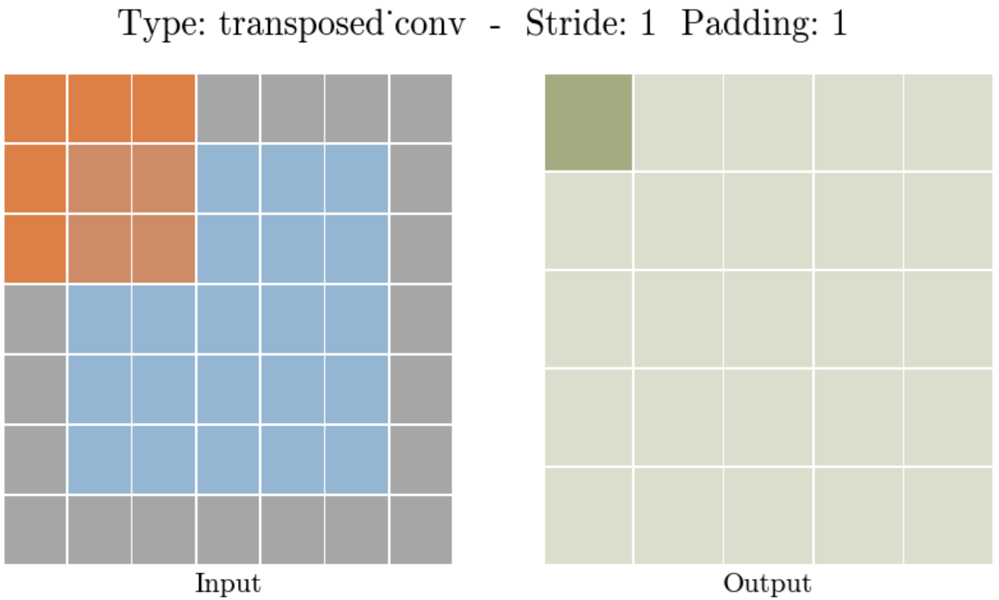
Convolutional Neural Network



$((n + 2p - f)/s + 1) * ((n + 2p - f)/s +1) = 5 * 5$

https://towardsdatascience.com/a-visualization-of-the-basic-elements-of-a-convolutional-neural-network-75fea30cd78d

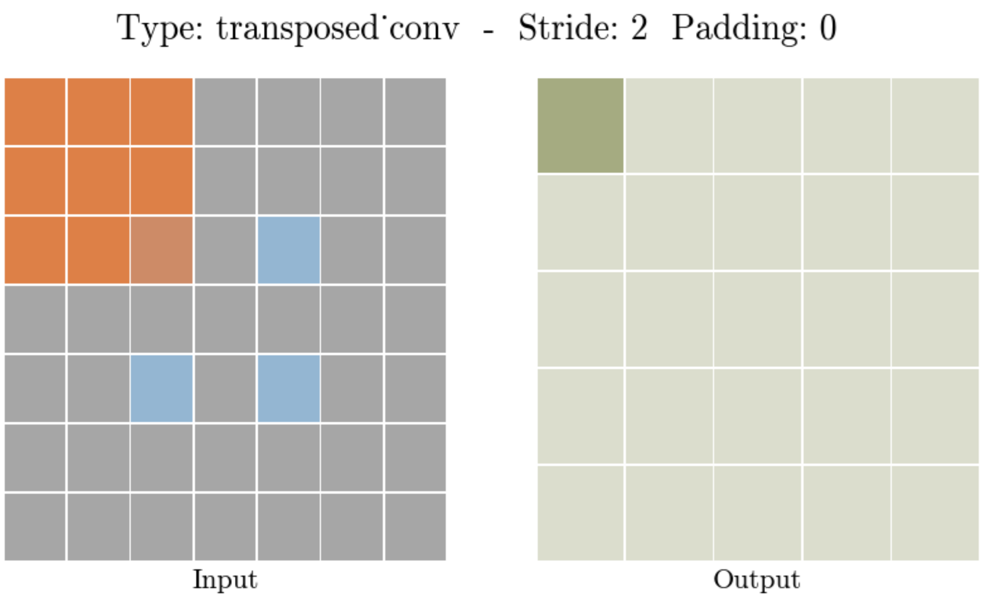
Convolutional Neural Network



((n + 2p - f)/s + 1) * ((n + 2p - f)/s +1) = 5 * 5

https://towardsdatascience.com/a-visualization-of-the-basic-elements-of-a-convolutional-neural-network-75fea30cd78d

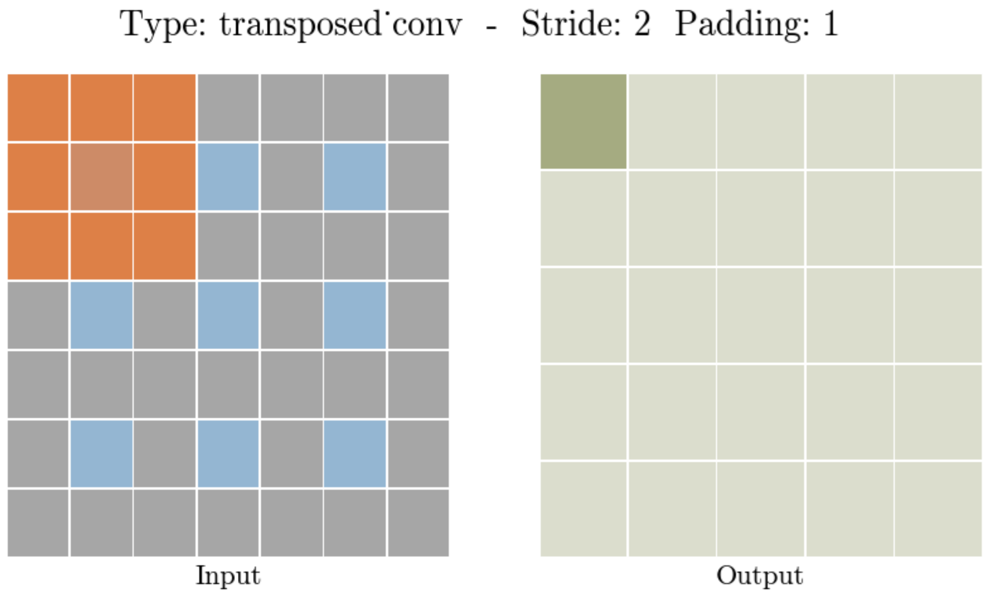
Convolutional Neural Network



((n + 2p - f)/s + 1) * ((n + 2p - f)/s +1) = 5 * 5

https://towardsdatascience.com/a-visualization-of-the-basic-elements-of-a-convolutional-neural-network-75fea30cd78d

Convolutional Neural Network



((n + 2p - f)/s + 1) * ((n + 2p - f)/s +1) = 5 * 5

https://towardsdatascience.com/a-visualization-of-the-basic-elements-of-a-convolutional-neural-network-75fea30cd78d

Convolutional Neural Network

Pooling

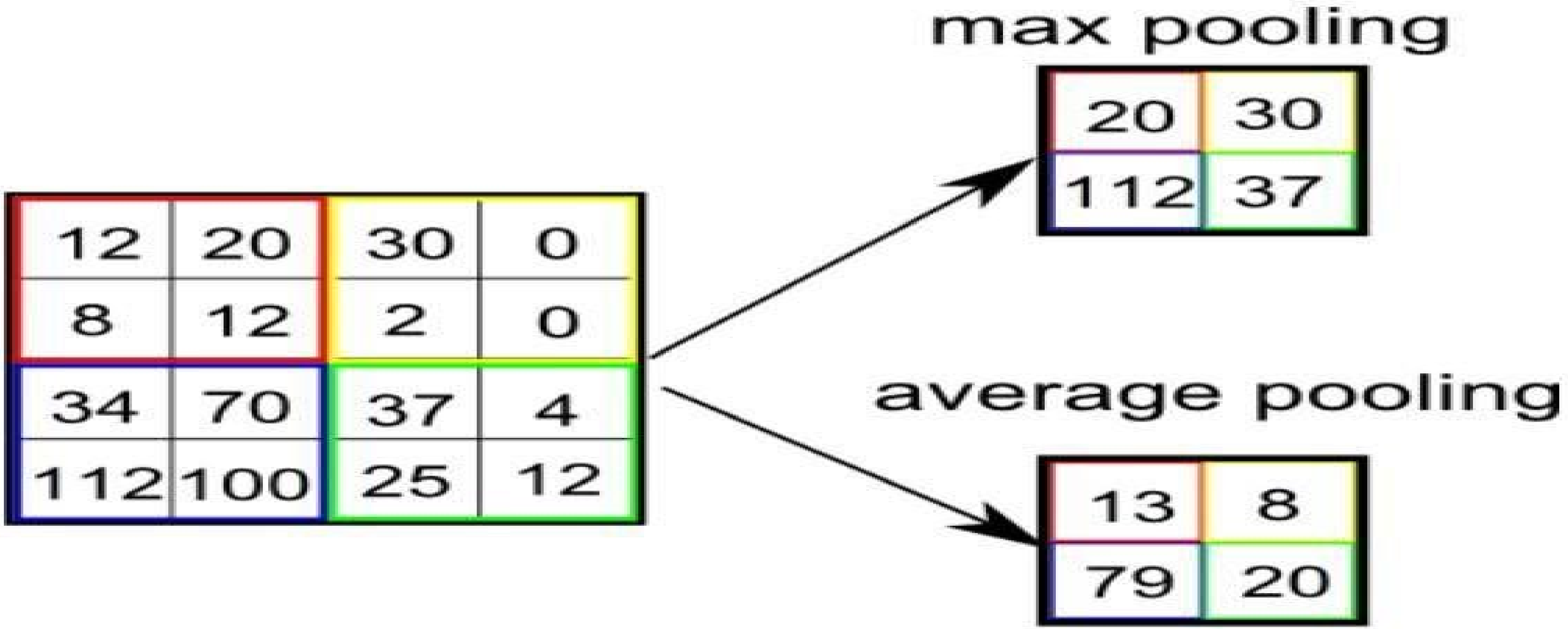
The pooling layer is used to reduce the spatial size of the convolved feature.

- Decrease the computational power required to process the data through dimensionality reduction.
- Extract dominant features which are rotational and positional invariant, thus maintaining the process of effective training of the model.

Two types of pooling

- Max pooling
- Average pooling

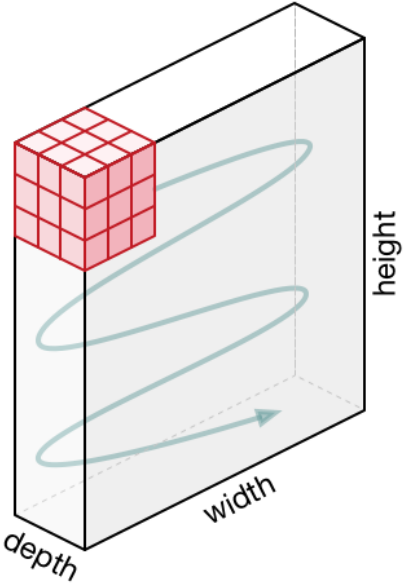
Convolutional Neural Network

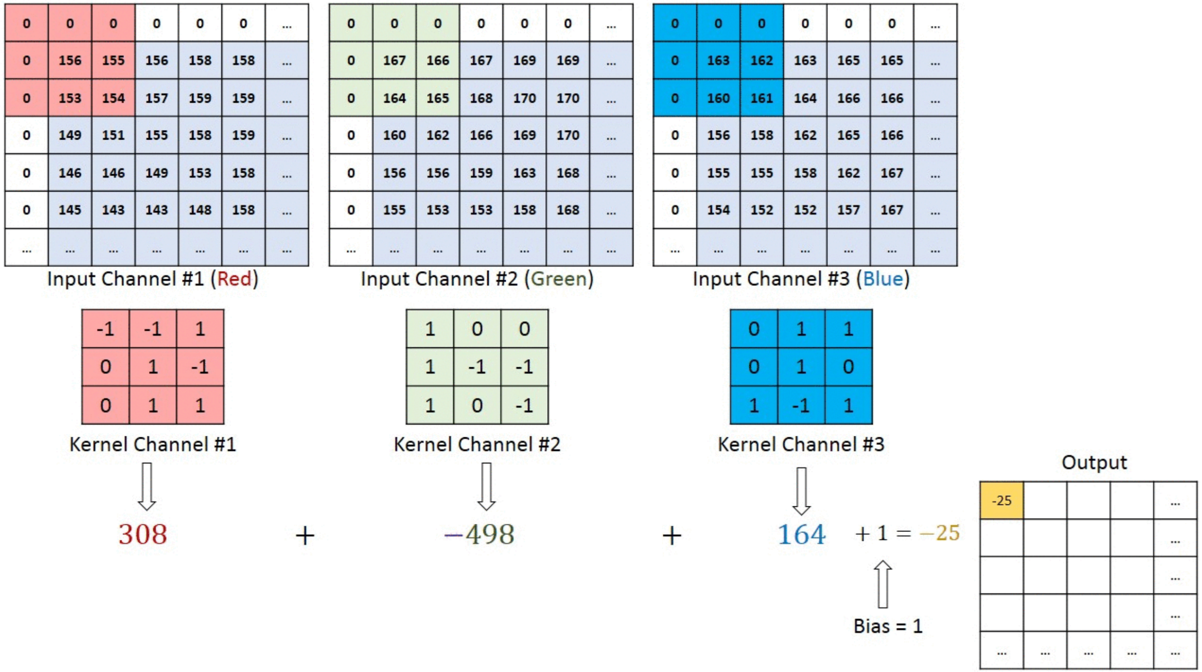


https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Convolutional Neural Network

3D feature extraction - process of one filter

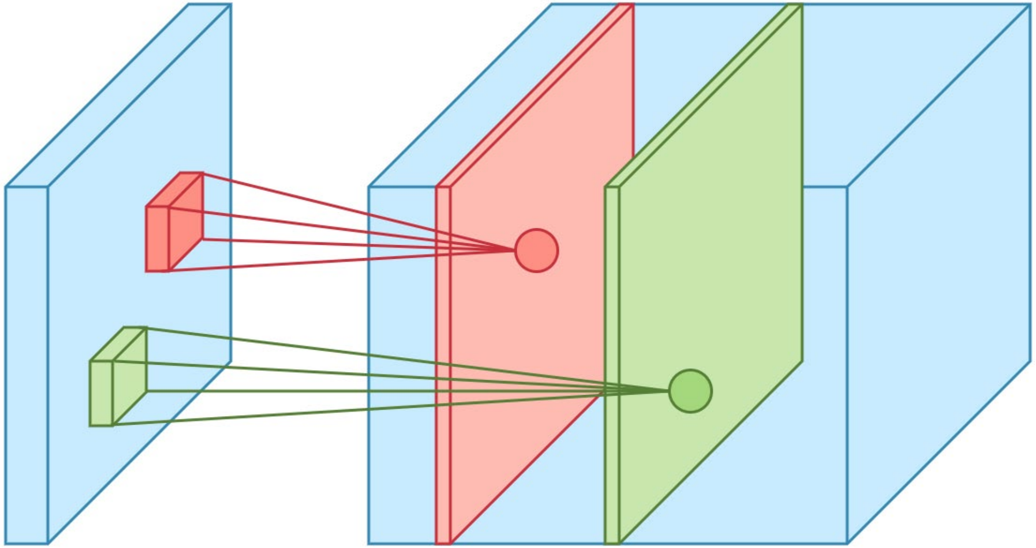




https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Convolutional Neural Network

3D feature extraction - process of multiple filters



https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2

Convolutional Neural Network

Hidden Layers

- **Layer function**: Basic transforming function such as convolutional or fully connected layer.
- **Pooling**: Used to change the spatial size of the feature map either increasing (up-sampling) or decreasing (most common) it. For example maxpooling, average pooling, and unpooling.
- **Normalization**: This subfunction normalizes the data to have zero mean and unit variance. This helps in coping up with problems such as vanishing gradient, internal covariate shift, etc. (more information). The two most common normalization techniques used are local response normalization and batch normalization.
- **Activation**: Applies non-linearity and bounds the output from getting too high or too low.

https://towardsdatascience.com/a-visualization-of-the-basic-elements-of-a-convolutional-neural-network-75fea30cd78d

Convolutional Neural Network

Activation Functions

Applies non-linearity and bounds the output from getting too high or too low

$$\textbf{Sigmoid}:\ f(x) = \frac{1}{1+e^{-x}};\quad f'(x) = f(x)(1-f(x)).$$

$$\textbf{tanh}:\ f(x) = \frac{e^x-e^{-x}}{e^x+e^{-x}};\quad f'(x) = 1-f(x)^2.$$

$$\boxed{\textbf{ReLU}:\ f(x) = \begin{cases} 0 & x < 0 \\ x & x \geq 0 \end{cases} \qquad f'(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}}$$

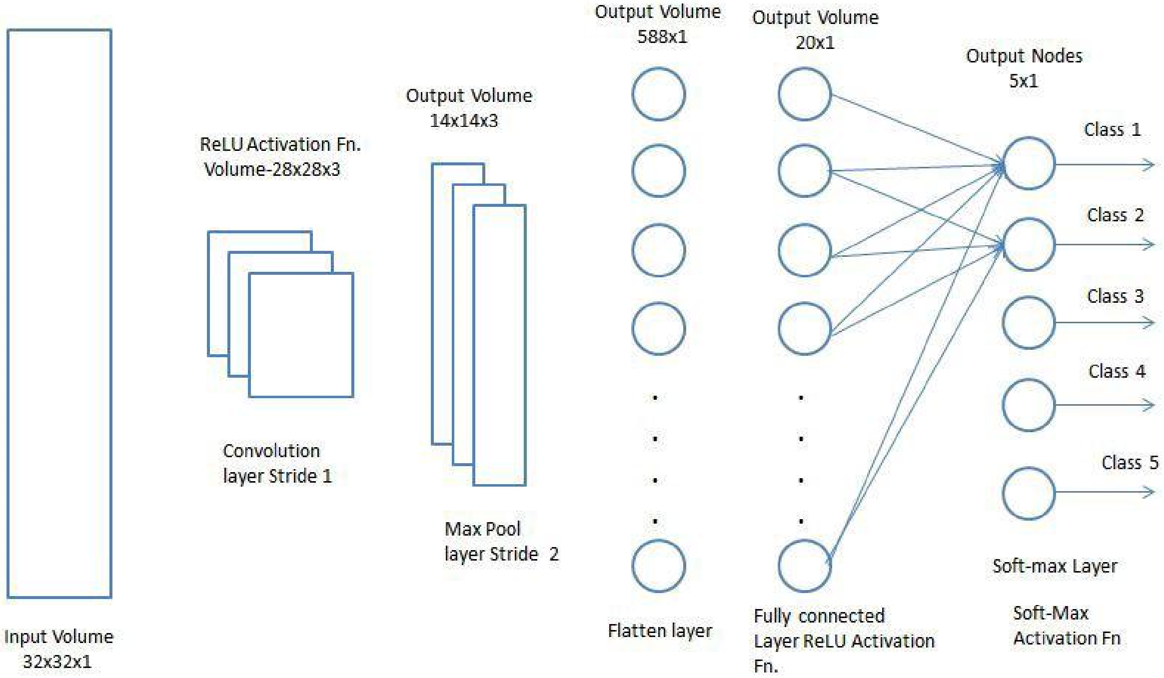
$$\textbf{Leaky ReLU}:\ f(x) = \begin{cases} 0.01x & x < 0 \\ x & x \geq 0 \end{cases} \qquad f'(x) = \begin{cases} 0,01 & x < 0 \\ 1 & x \geq 0 \end{cases}$$

$$\textbf{Softmax}:\ f(x_j) = \frac{e^{x_j}}{\sum_{k=1}^{d} e^{x_l}}$$

Convolutional Neural Network

Fully connected layers

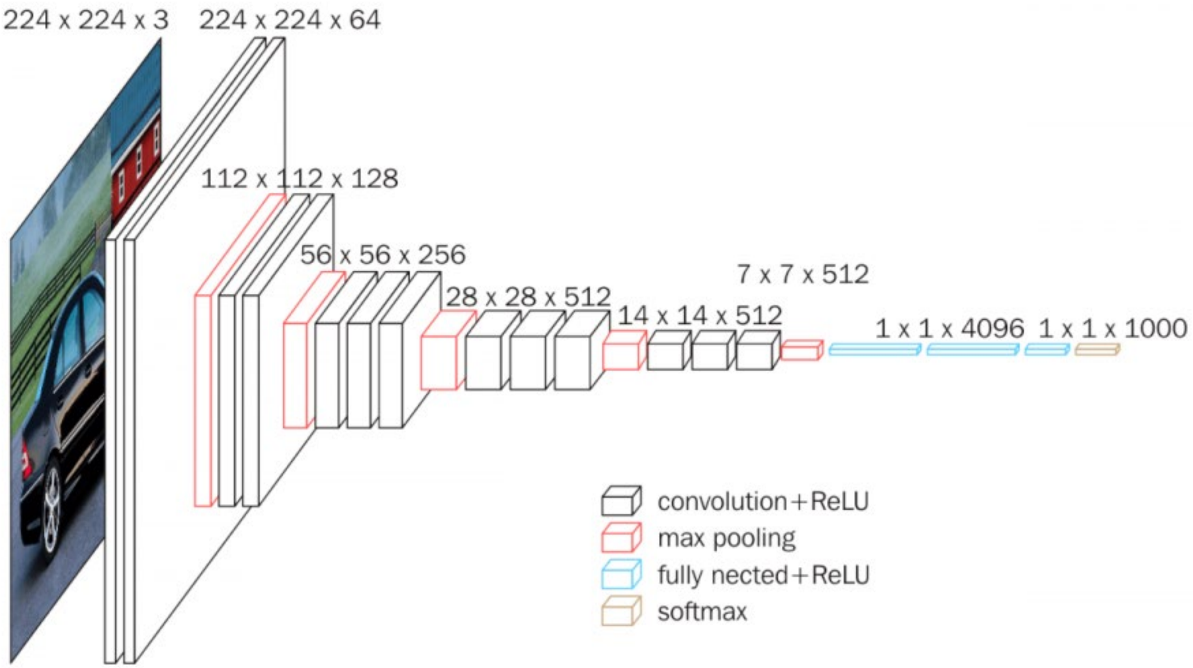
Adding a fully-connected layer is a cheap way of learning non-linear combinations of the high-level features as represented by the output of the convolutional layer.



https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Convolutional Neural Network

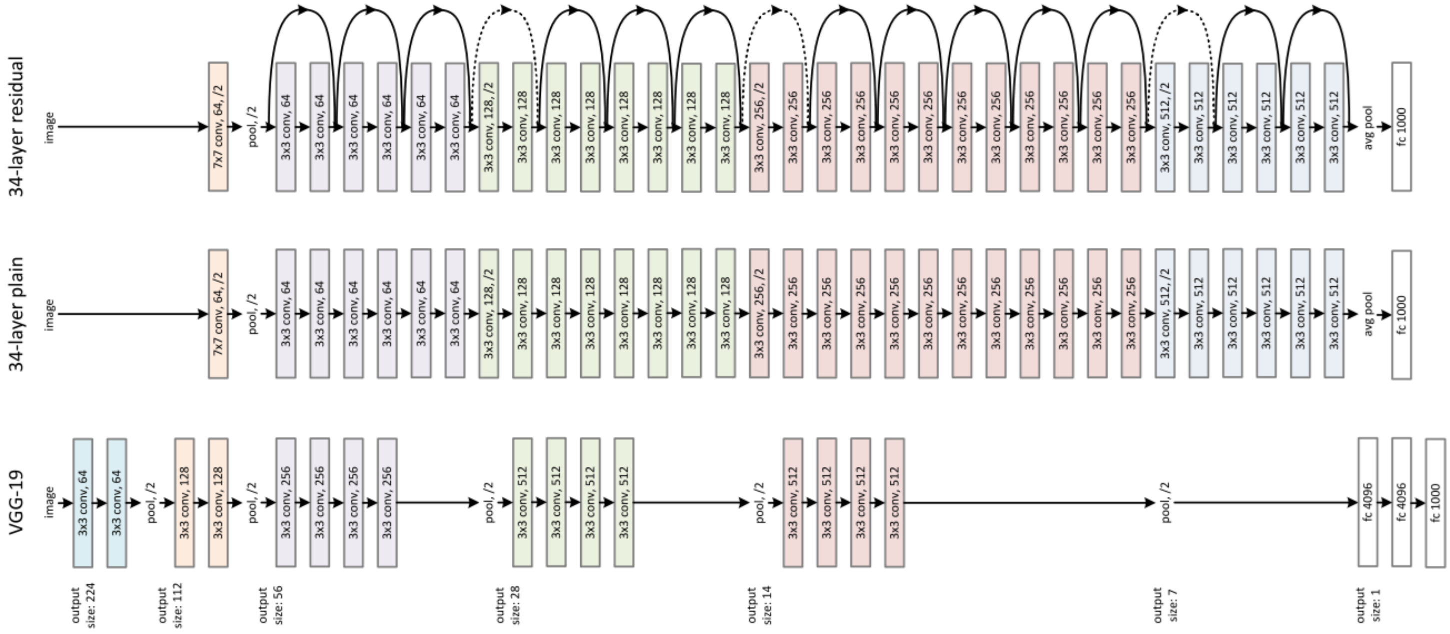
VGG16



https://medium.com/@mygreatlearning/what-is-vgg16-introduction-to-vgg16-f2d63849f615

Convolutional Neural Network

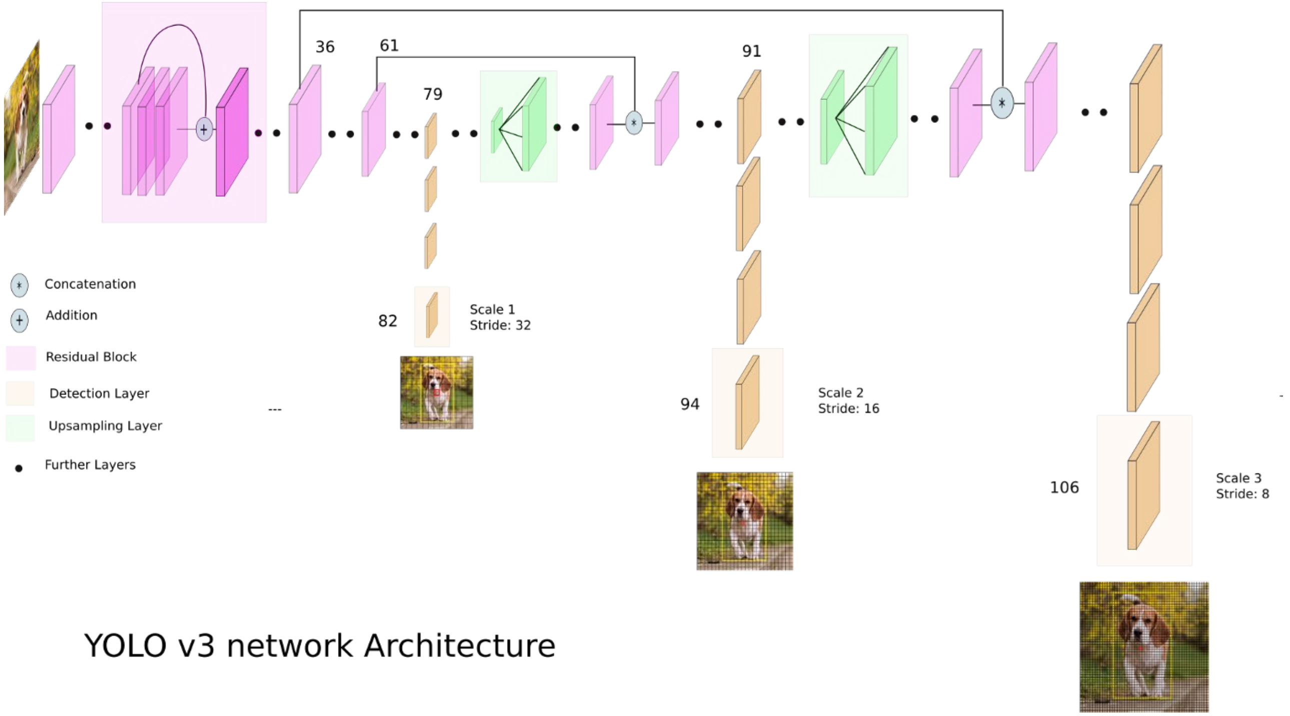
ResNet



https://medium.com/@pierre_guillou/understand-how-works-resnet-without-talking-about-residual-64698f157e0c

Convolutional Neural Network

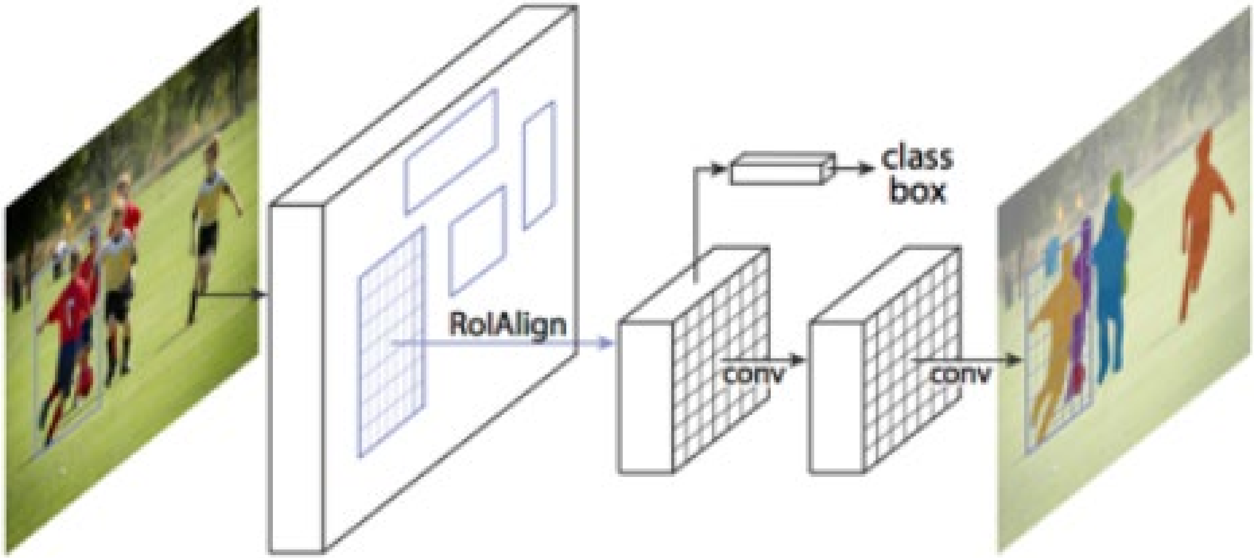
YOLOv3



https://towardsdatascience.com/yolo-v3-object-detection-53fb7d3bfe6b

Convolutional Neural Network

Mask-RCNN



https://aditi-mittal.medium.com/instance-segmentation-using-mask-r-cnn-7f77bdd46abd

Datasets - COCO

What is COCO?

COCO is a large-scale object detection, segmentation, and captioning dataset. COCO has several features:

- Object segmentation
- Recognition in context
- Superpixel stuff segmentation
- 330K images (>200K labeled)
- 1.5 million object instances
- 80 object categories
- 91 stuff categories
- 5 captions per image
- 250,000 people with keypoints

Collaborators

Tsung-Yi Lin Google Brain
Genevieve Patterson MSR, Trash TV
Matteo R. Ronchi Caltech
Yin Cui Google
Michael Maire TTI-Chicago
Serge Belongie Cornell Tech
Lubomir Bourdev WaveOne, Inc.
Ross Girshick FAIR
James Hays Georgia Tech
Pietro Perona Caltech
Deva Ramanan CMU
Larry Zitnick FAIR
Piotr Dollár FAIR

Sponsors

CVDF
Microsoft
facebook
Mighty Ai

Research Paper

Download the paper that describes the Microsoft COCO dataset.


Microsoft COCO: Common Objects in Context

Download paper here

Dataset examples

https://cocodataset.org/#home

Datasets - ImageNet

IMAGENET

14,197,122 images, 21841 synsets indexed

Home Download Challenges About

Not logged in. Login | Signup

About ImageNet

News and updates

- March 11 2021: ImageNet website update and a new paper on privacy preservation.
- October 10, 2019: The ILSVRC 2012 classification and localization test set has been updated. The Kaggle challenge and our download page both now contain the updated data.
- September 21, 2019: ImageNet 10th Birthday Party
- September 17, 2019: Research update on filtering and balancing the ImageNet person subtree
- June 2019: ImageNet wins the PAMI Longuet-Higgins Prize (Retrospective Most Impactful Paper from CVPR 2009)
- July 2017: ImageNet: Where have we been? Where are we going? at a CVPR 2017 workshop.
- July 2017: ImageNet is covered by Quartz.
- June 2, 2015: Follow-up update regarding status of the server
- May 19, 2015: Annoucement regarding the submission server

Overview

Welcome to the ImageNet project! ImageNet is an ongoing research effort to provide researchers around the world with image data for training large-scale object recognition models.

What is ImageNet?

ImageNet is an image dataset organized according to the WordNet hierarchy. Each meaningful concept in WordNet, possibly described by multiple words or word phrases, is called a "synonym set" or "synset". There are more than 100,000 synsets in WordNet; the majority of them are nouns (80,000+). In ImageNet, we aim to provide on average 1000 images to illustrate each synset. Images of each concept are quality-controlled and human-annotated. In its completion, we hope ImageNet will offer tens of millions of cleanly labeled and sorted images for most of the concepts in the WordNet hierarchy.

Why ImageNet?

The ImageNet project was inspired by two important needs in computer vision research. The first was the need to establish a clear North Star problem in computer vision. While the field enjoyed an abundance of important tasks to work on, from stereo vision to image retrieval, from 3D reconstruction to image segmentation, object categorization was recognized to be one of the most fundamental capabilities of both human and machine vision. Hence there was a growing demand for a high quality object categorization benchmark with clearly established evaluation metrics. Second, there was a critical need for more data to enable more generalizable machine learning methods. Ever since the birth of the digital era and the availability of web-scale data exchanges, researchers in these fields have been working hard to design more and more sophisticated algorithms to index, retrieve, organize and annotate multimedia data. But good research requires good resources. To tackle this problem at scale (think of your growing personal collection of digital images, or videos, or a commercial web search engine's database), it was critical to provide researchers with a large-scale image database for both training and testing. The convergence of these two intellectual reasons motivated us to build ImageNet.

https://www.image-net.org/about.php

Thank You!