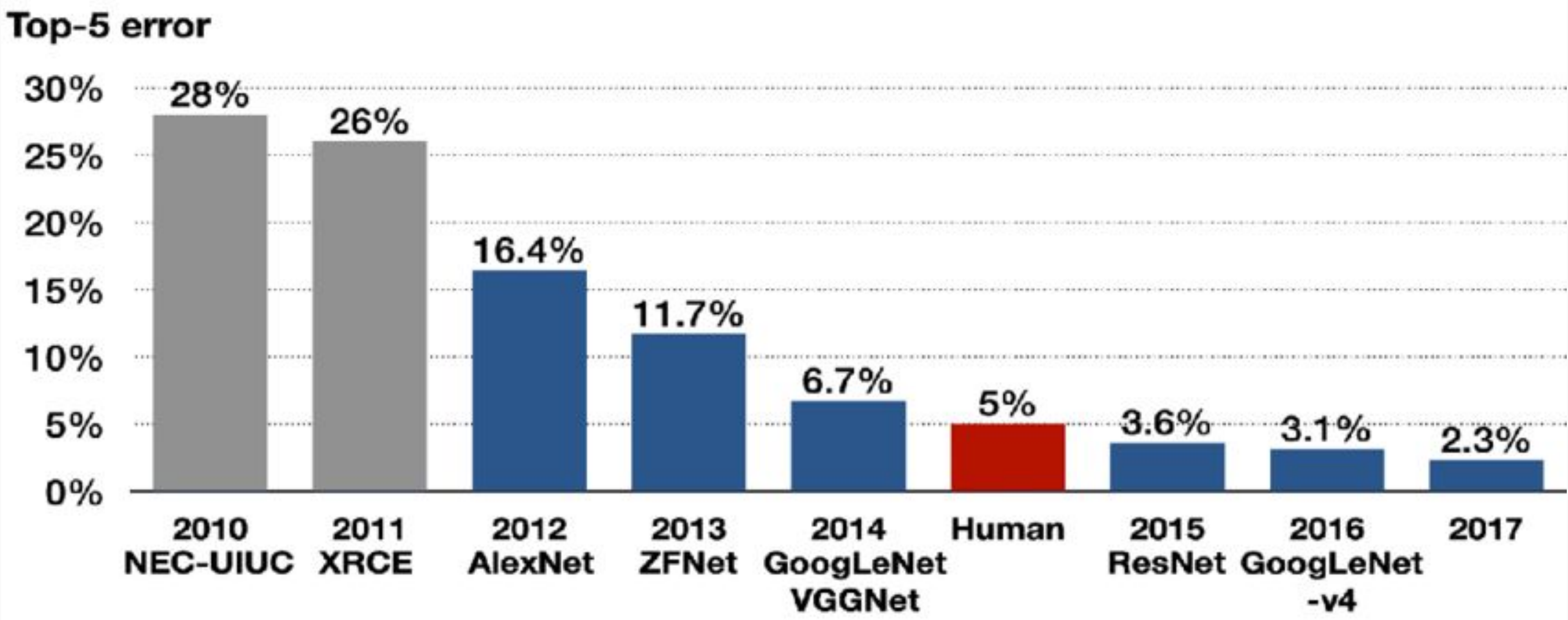
Neural Networks in Image Processing

Jakub Kuciński

Algorithms that won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

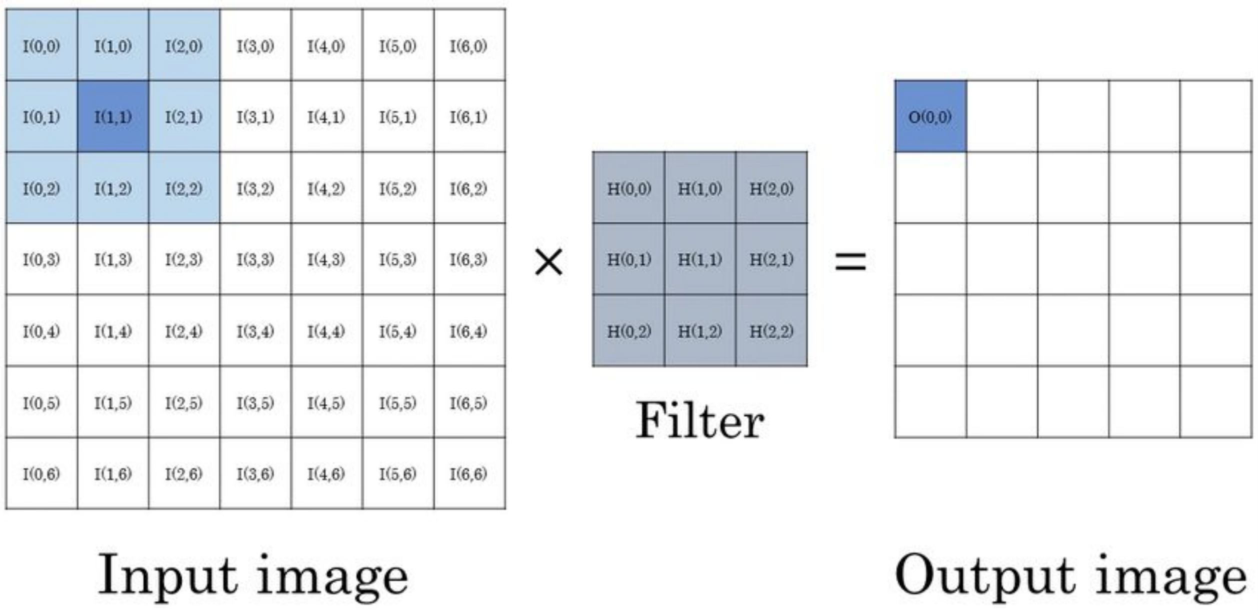


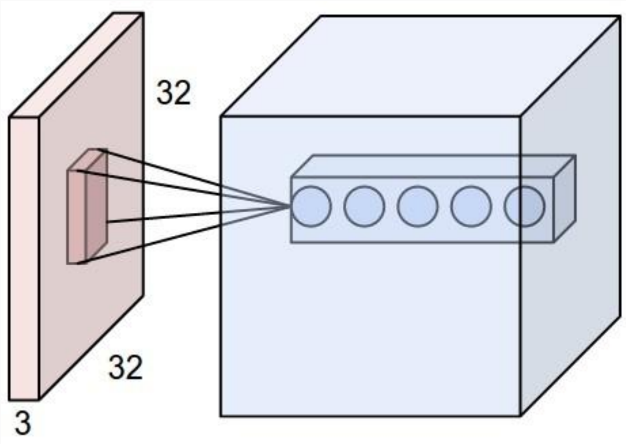
Kang, Dae-Young & Duong, Hieu & Park, Jung-Chul. (2020). Application of Deep Learning in Dentistry and Implantology. The Korean Academy of Oral and Maxillofacial Implantology. 24. 148-181. 10.32542/implantology.202015.

Convolutions

Why not fully connected layers?

(256 x 256 x 3)^2 = 38 654 705 664 parameters!!!



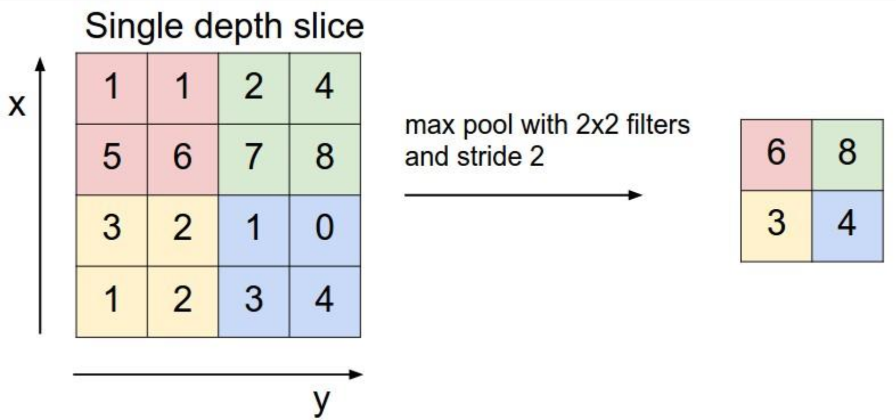


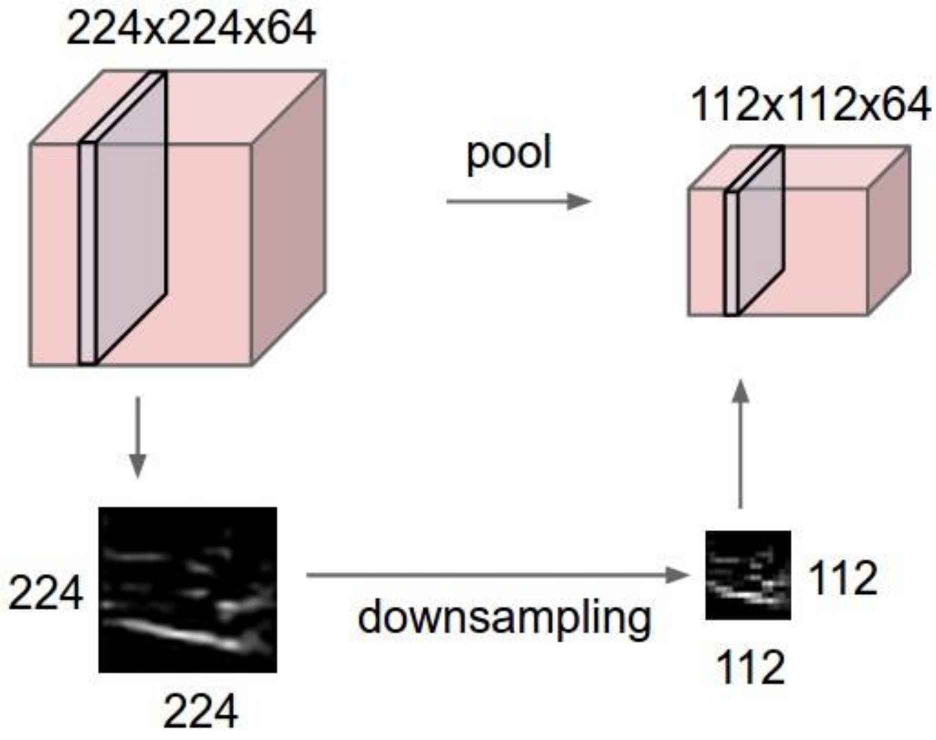
Baskin, Chaim & Liss, Natan & Mendelson, Avi & Zheltonozhskii, Evgenii. (2017). Streaming Architecture for Large-Scale Quantized Neural Networks on an FPGA-Based Dataflow Platform.

Pooling

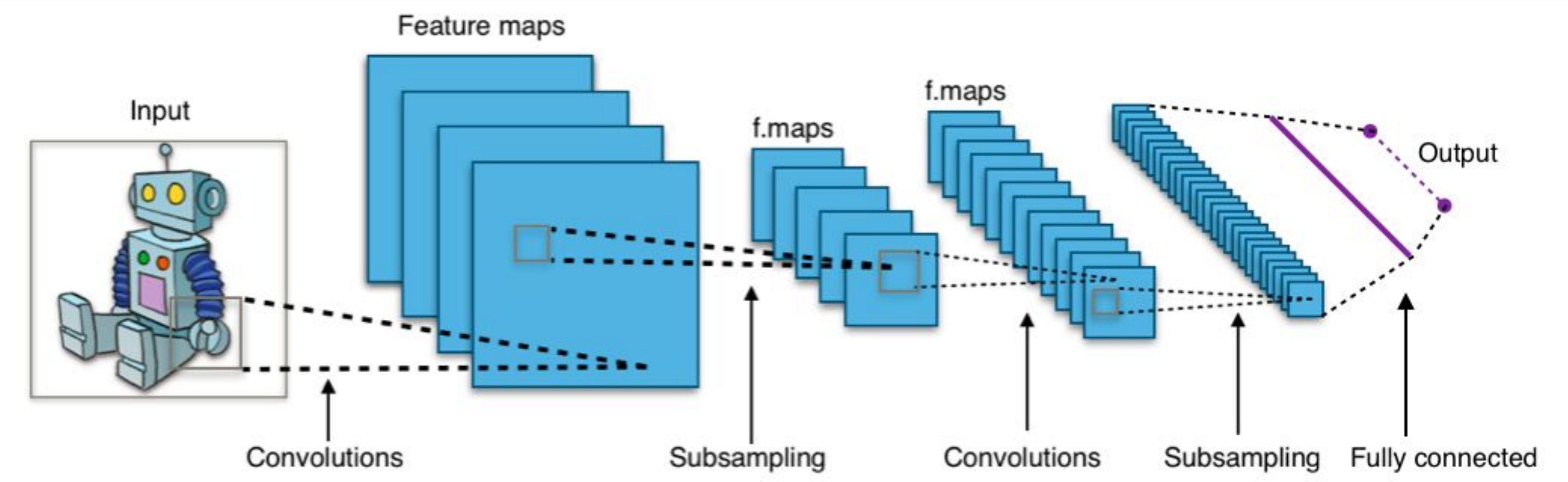
Role of pooling:

- decreases spatial size
- make network less sensitive to exact location of features
- enable convolutions to work on the bigger context of the image

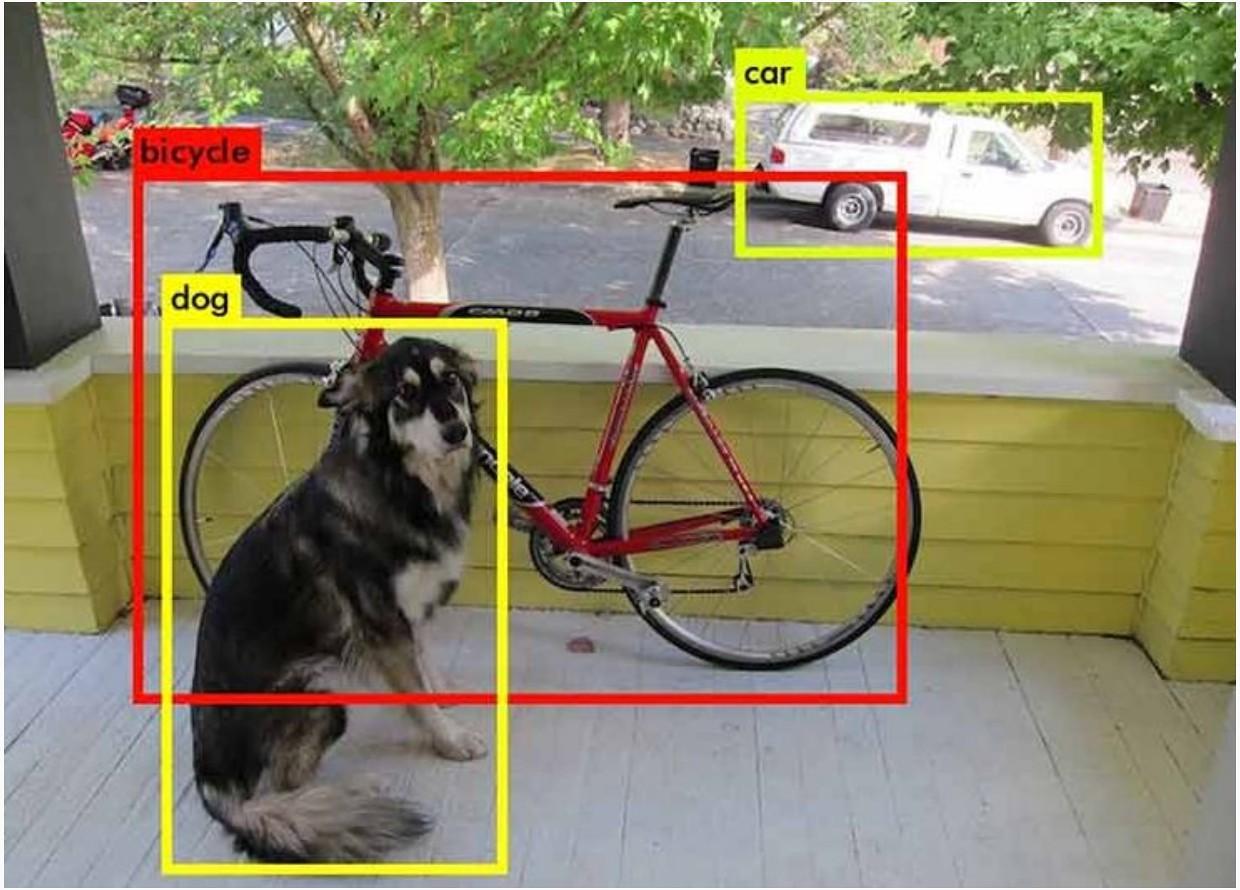




Typical CNN architecture for classification

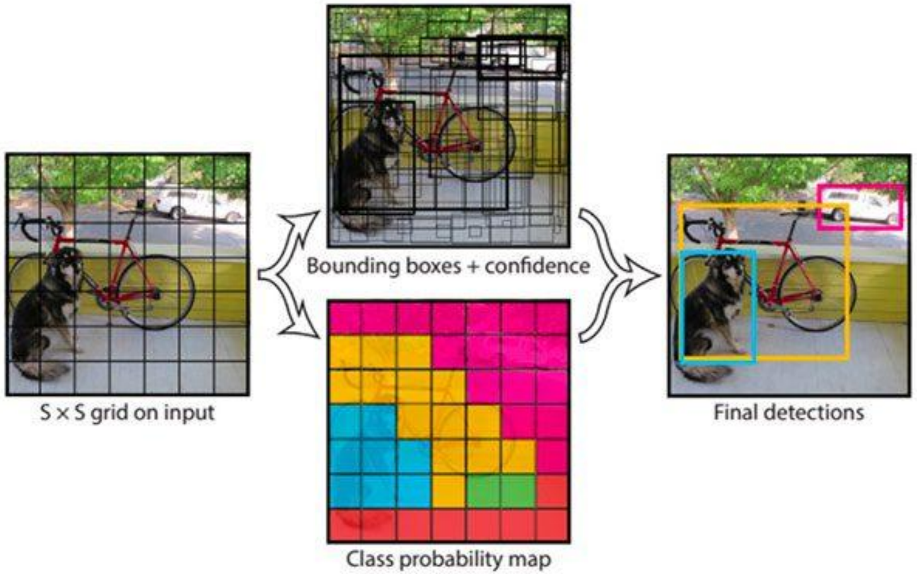


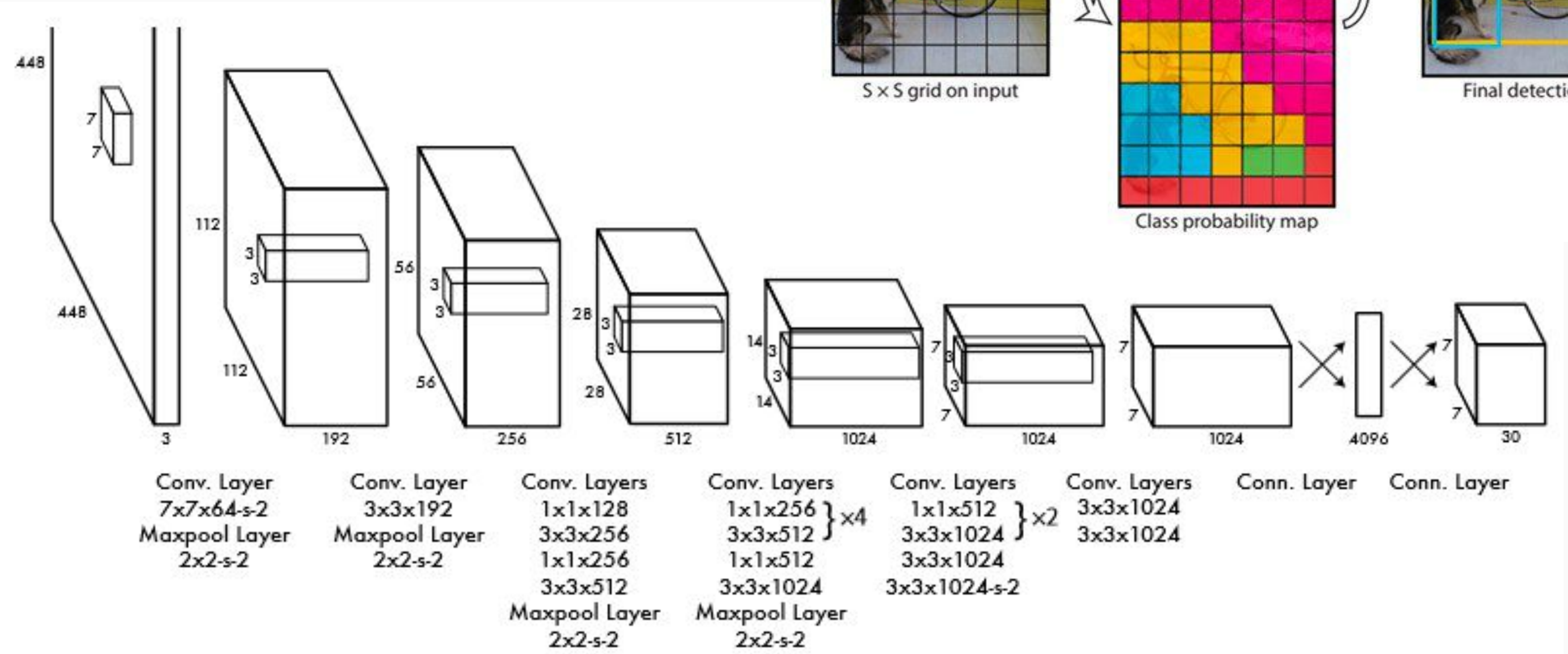
Object detection



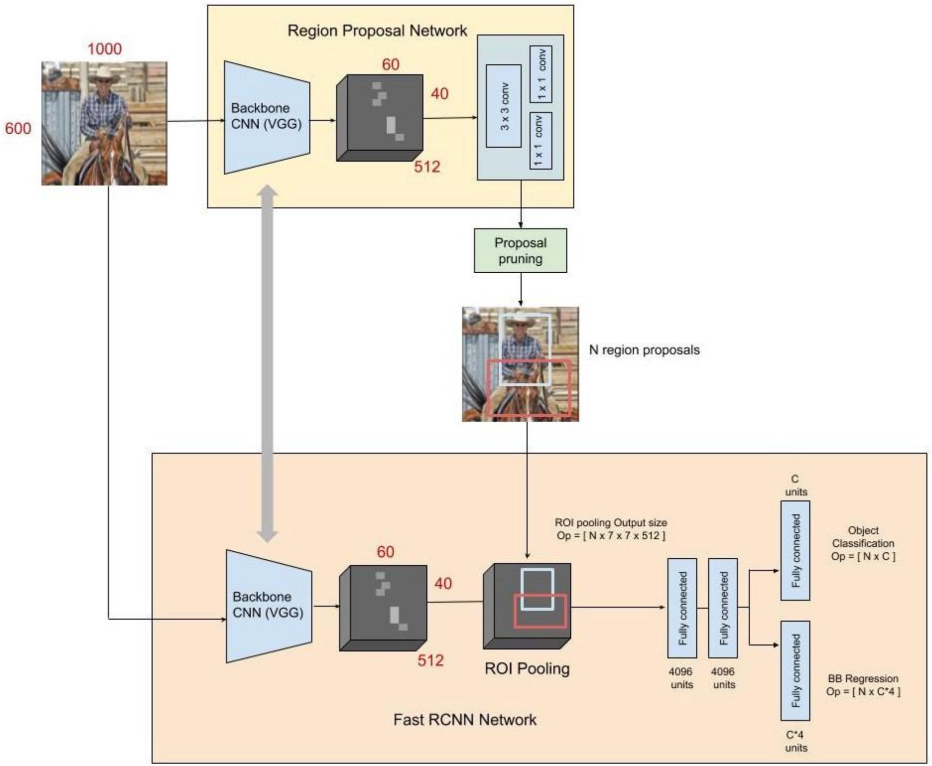
One stage detectors - YOLO

Joseph Redmon and Santosh Divvala and Ross Girshick and Ali Farhadi, [You Only Look Once: Unified, Real-Time Object Detection]()

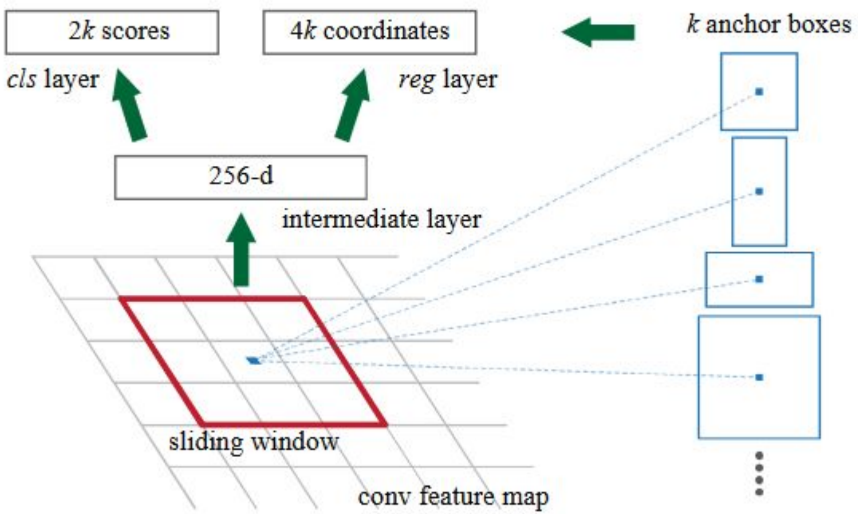




Two stage detectors - Faster R-CNN



Shilpa Ananth, [Faster R-CNN for object detection]



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

RoI Pool

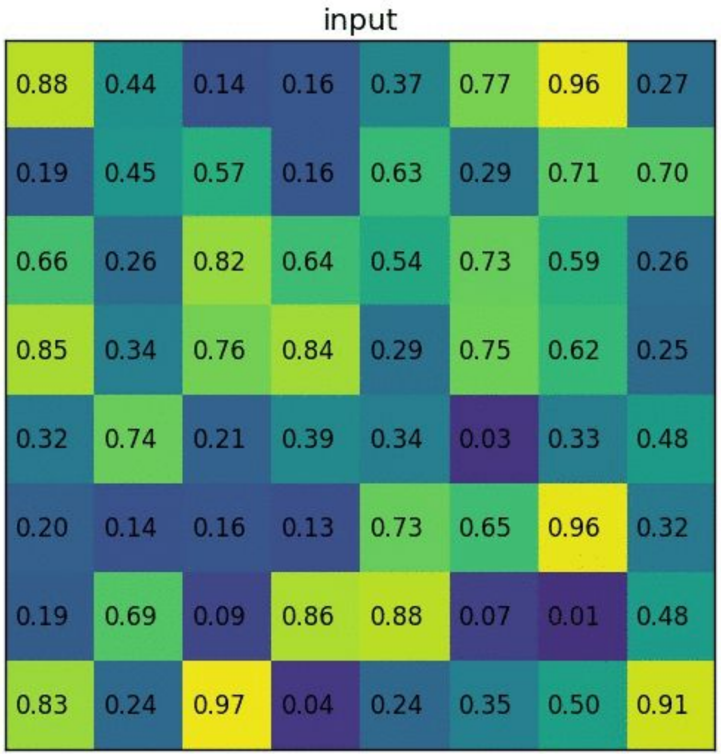
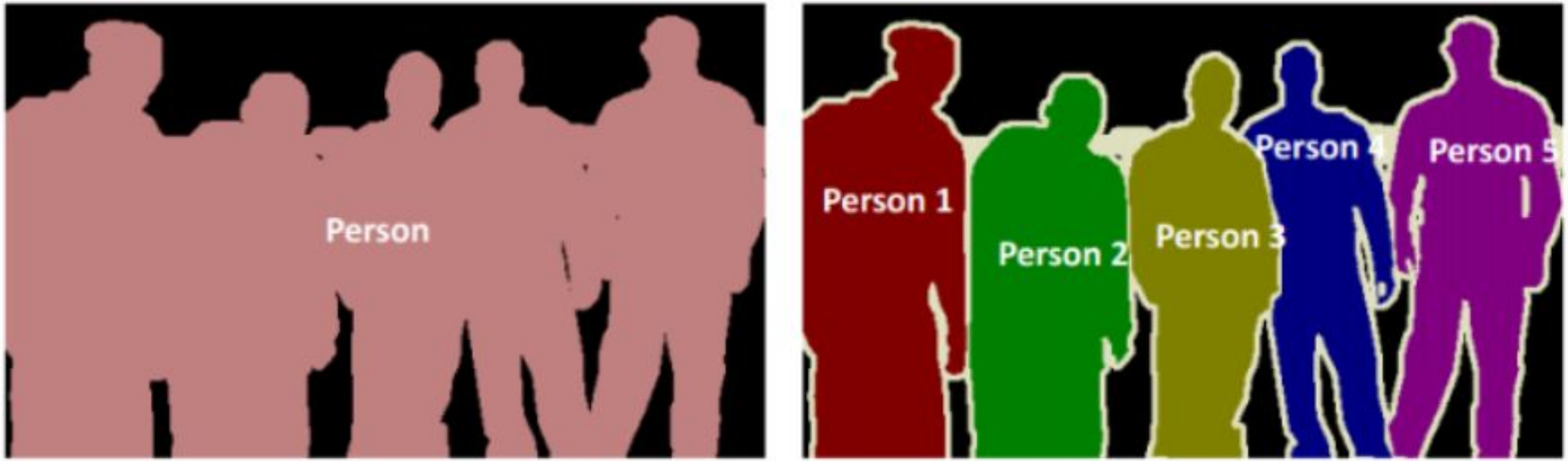

input
0.88 0.44 0.14 0.16 0.37 0.77 0.96 0.27
0.19 0.45 0.57 0.16 0.63 0.29 0.71 0.70
0.66 0.26 0.82 0.64 0.54 0.73 0.59 0.26
0.85 0.34 0.76 0.84 0.29 0.75 0.62 0.25
0.32 0.74 0.21 0.39 0.34 0.03 0.33 0.48
0.20 0.14 0.16 0.13 0.73 0.65 0.96 0.32
0.19 0.69 0.09 0.86 0.88 0.07 0.01 0.48
0.83 0.24 0.97 0.04 0.24 0.35 0.50 0.91

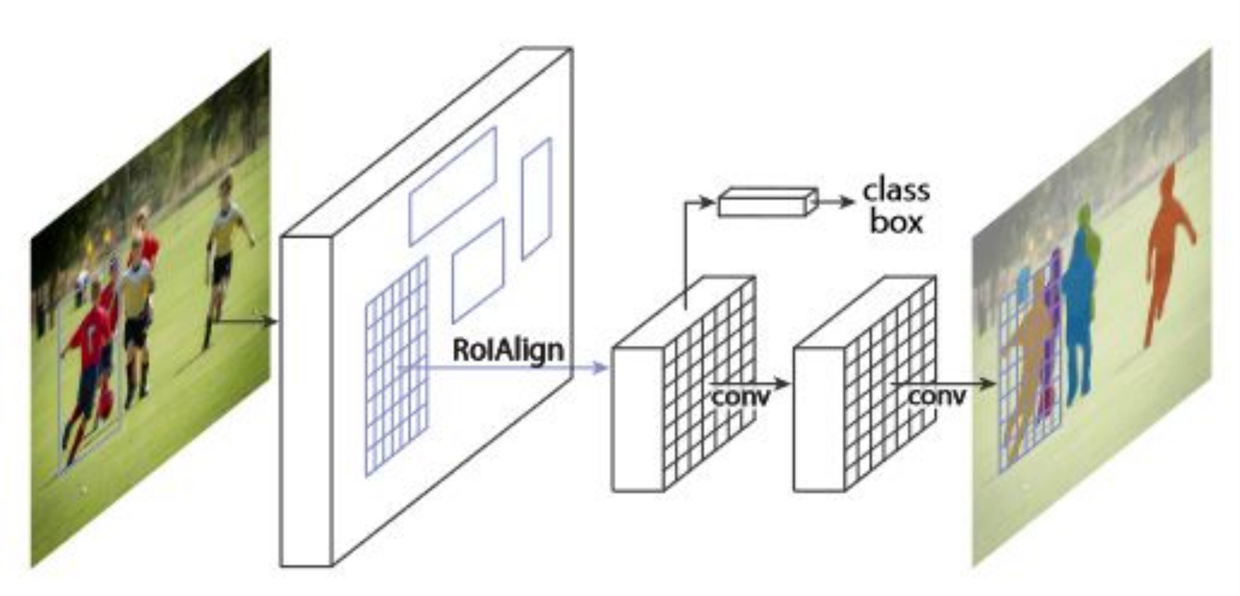
Image segmentation



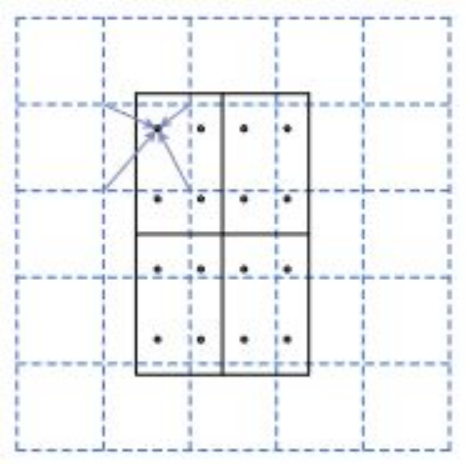
Semantic Segmentation

Instance Segmentation

Mask R-CNN

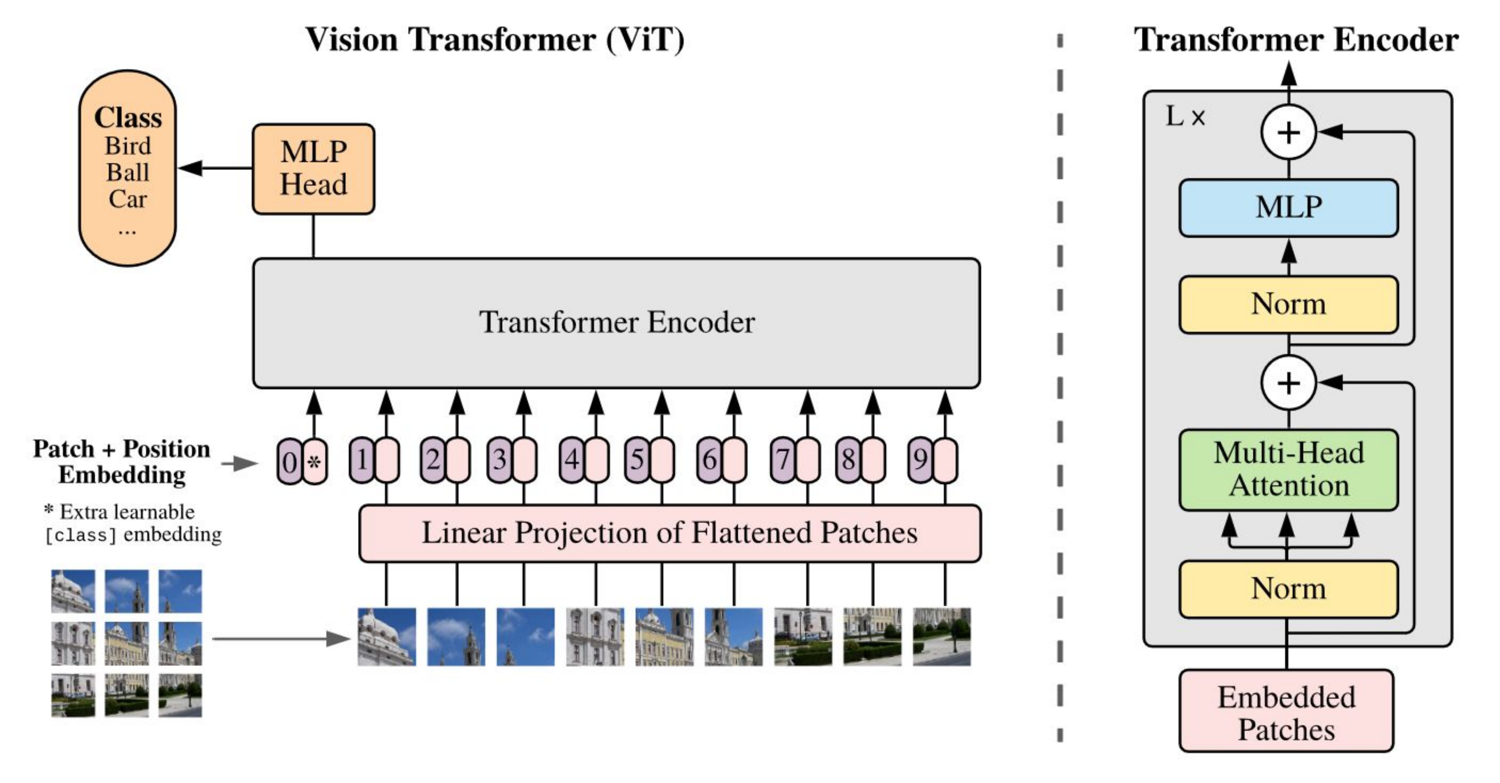


RoIAlign



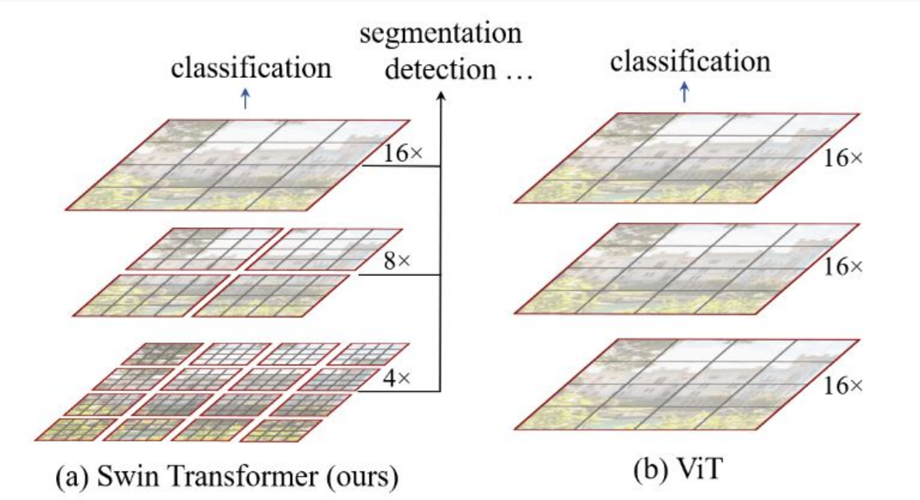
Kaiming He and Georgia Gkioxari and Piotr Dollár and Ross Girshick, [Mask R-CNN]

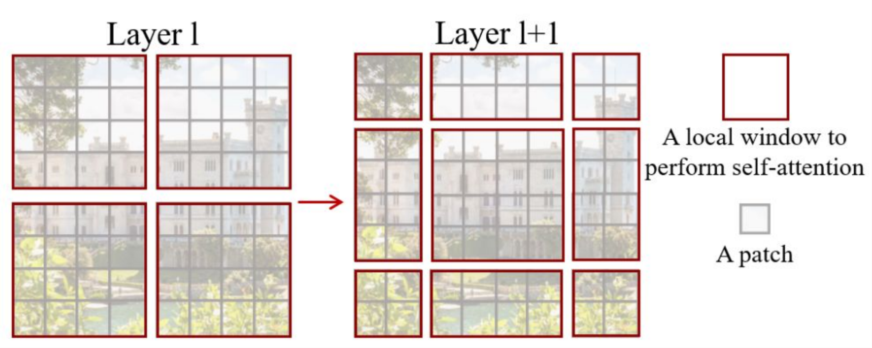
Vision Transformer



Alexey Dosovitskiy et al. (3 Jun 2021), [An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale]

Swin Transformer





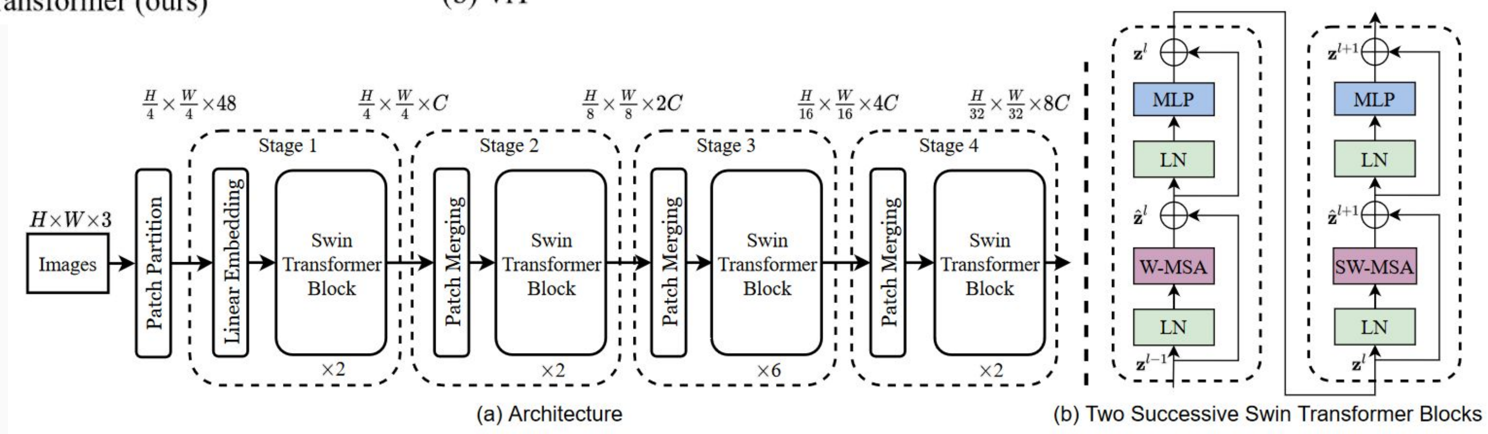


Image Classification on ImageNet

Rank	Model	Top 1 Accuracy	Top 5 Accuracy	Number of params	Extra Training Data	Paper	Code	Result	Year	Tags
1	CoAtNet-7	90.88%		2440M	✓	CoAtNet: Marrying Convolution and Attention for All Data Sizes			2021	Conv+Transformer JFT-3B
2	ViT-G/14	90.45%		1843M	✓	Scaling Vision Transformers			2021	Transformer JFT-3B
3	CoAtNet-6	90.45%		1470M	✓	CoAtNet: Marrying Convolution and Attention for All Data Sizes			2021	Conv+Transformer JFT-3B
4	ViT-MoE-15B (Every-2)	90.35%		14700M	✓	Scaling Vision with Sparse Mixture of Experts			2021	Transformer JFT-3B
5	Meta Pseudo Labels (EfficientNet-L2)	90.2%	98.8%	480M	✓	Meta Pseudo Labels			2021	EfficientNet JFT-300M
6	SwinV2-G	90.17%			✓	Swin Transformer V2: Scaling Up Capacity and Resolution			2021	Transformer

https://paperswithcode.com/sota/image-classification-on-imagenet

Object Detection on COCO test-dev

Rank	Model	box AP	AP50	AP75	APS	APM	APL	AP	Extra Training Data	Paper	Code	Result	Year	Tags
1	**Florence-CoSwin-H**	62.4							×	Florence: A New Foundation Model for Computer Vision			2021	Swin-Transformer
2	**GLIP** (Swin-L, multi-scale)	61.5	79.5	67.7	45.3	64.9	75.0		×	Grounded Language-Image Pre-training			2021	multiscale, Vision Language, Dynamic Head, BERT-Base
3	**Soft Teacher + Swin-L** (HTC++, multi-scale)	61.3							×	End-to-End Semi-Supervised Object Detection with Soft Teacher			2021	multiscale, Swin-Transformer
4	**DyHead** (Swin-L, multi scale, self-training)	60.6	78.5	66.6		64.0	74.2		×	Dynamic Head: Unifying Object Detection Heads with Attentions			2021	multiscale, Swin-Transformer
5	**Dual-Swin-L** (HTC, multi-scale)	60.1							×	CBNetV2: A Composite Backbone Network Architecture for Object Detection			2021	multiscale, Swin-Transformer

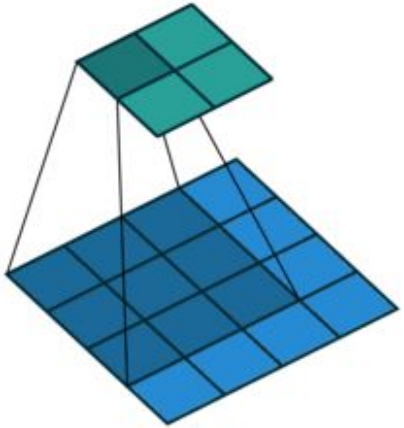
https://paperswithcode.com/sota/object-detection-on-coco

Recent trends and findings

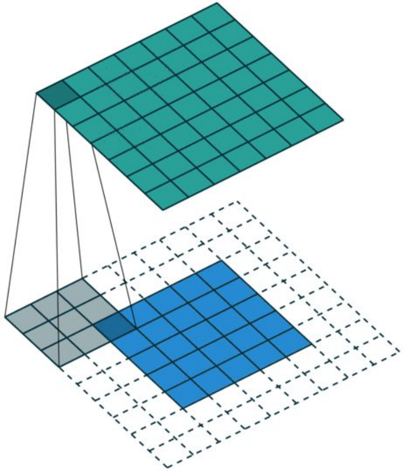
- ViTs needs very long training (probably due to big sparsity of attention layers and thus difficulty to learn first layers - because of fading gradient) and strong data augmentations to achieve their potential.
- More stable architectures, training schedules and proper data augmentations for ViTs are findings of the last year.
- Training CNNs to reasonably good performance is easier than ViTs, but training ViTs to close State-of-the-Art performance is easier than CNNs (probably due to inductive bias of CNNs in sense of local connectivity).
- Convnets if designed and trained carefully and precisely can achieve similar performance as ViTs.
  A ConvNet for the 2020s
- Data augmentation is crucial for proper training any architecture for specific task.
  Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation
- New strong data augmentations for mixing data samples: CutMix, MixUp, FMix. Especially useful when training CNNs on limited data.
  Cutmix-vs-Mixup-vs-GridMask-vs-CutOut
  FMix: Enhancing Mixed Sample Data Augmentation
- Creating special architectures and training models on different tasks or even different domains (e.g. images and text) are becoming more and more popular as it can allow to increase general performance and improve accuracy in few or zero-shot learning due to access to additional data and advance generalization.
  Grounded Language-Image Pre-training, Florance, CLIP, Multimodal Neurons in NN, DALL·E, Perceiver IO

Additional slides

Different types of convolutions

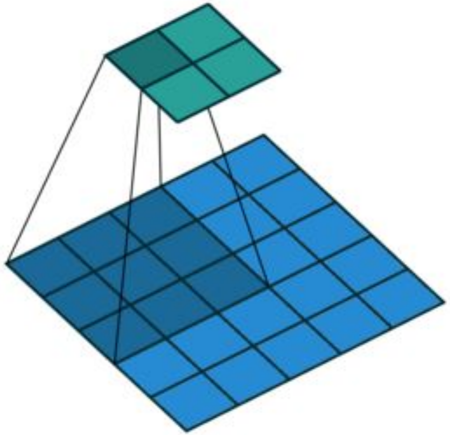


No padding,
no stride

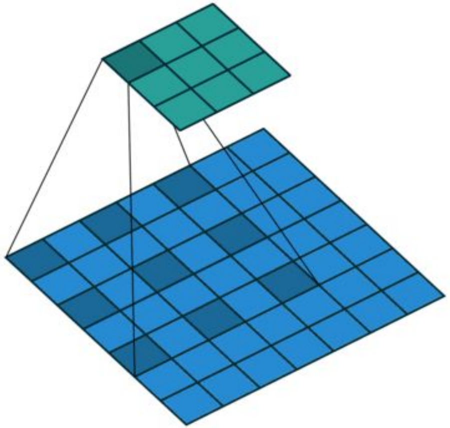


Full padding,
no stride

No padding,
strides

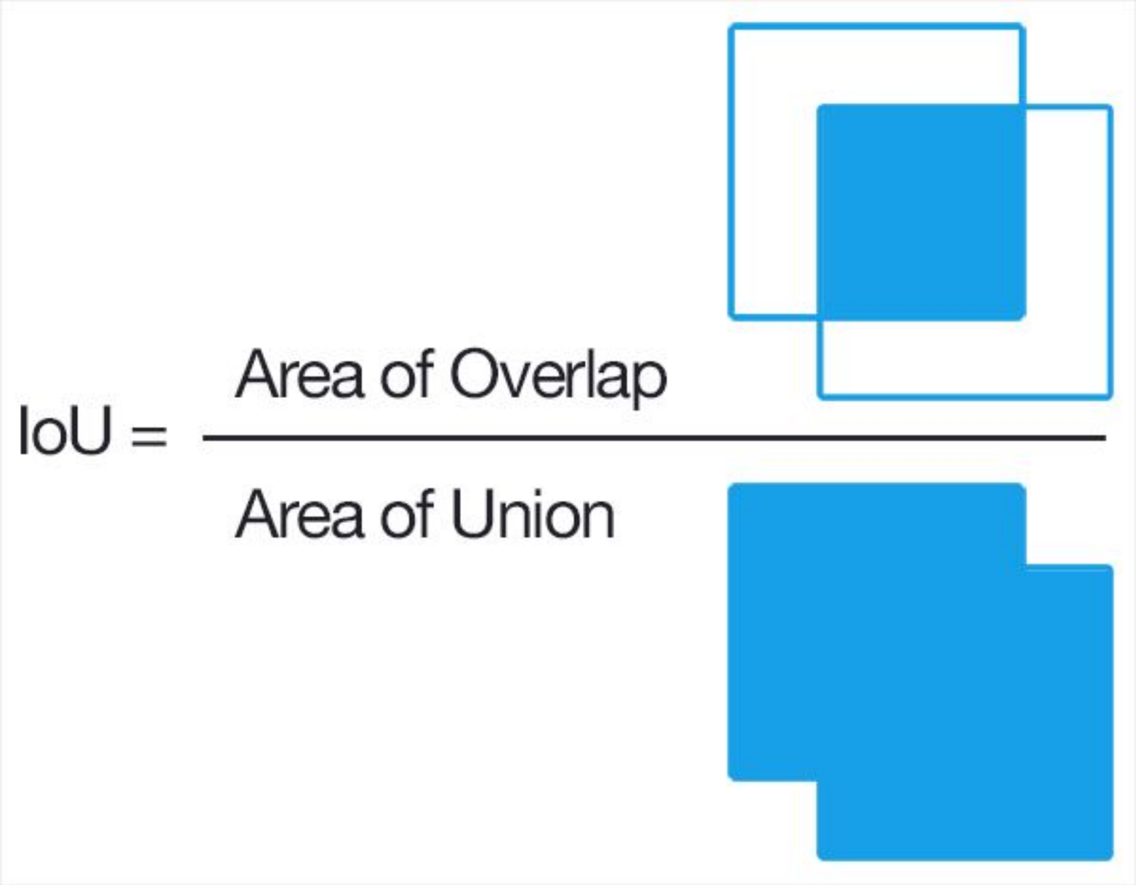


Dilation



Vincent Dumoulin, Francesco Visin - [A guide to convolution arithmetic for deep learning]

Intersection over Union (IoU) - training



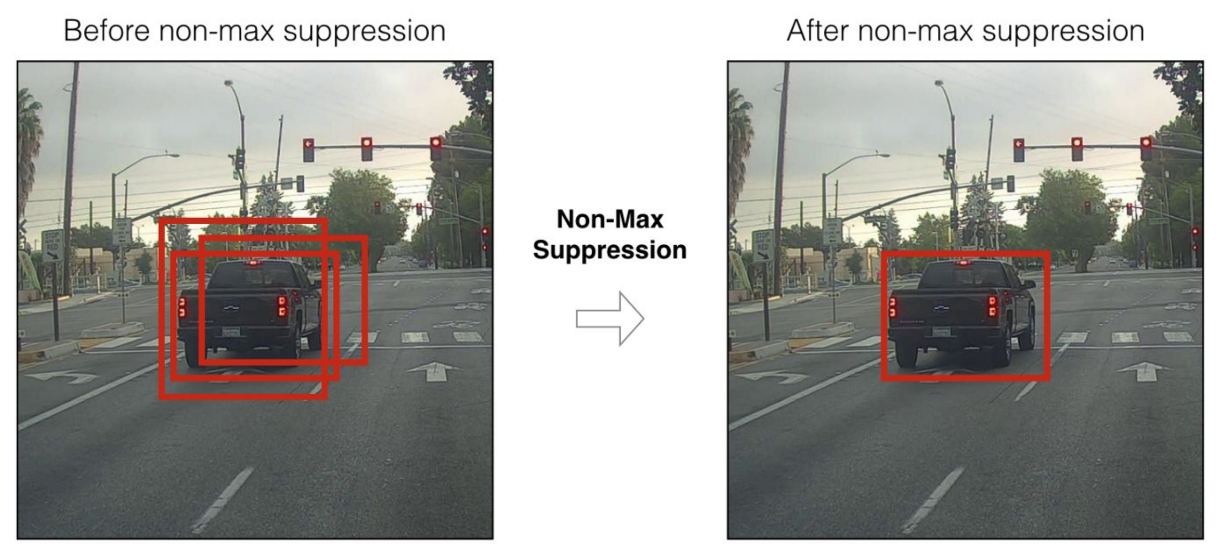
In YOLO training:

- If the center of an object falls into a grid cell, that grid cell is responsible for detecting that object.
- We bind ground truth box with the predicted box of greatest IoU and train only cells with such selected boxes.

In Faster R-CNN RPN training:

- We train only on positive and negative predicted boxes.
- Positive sample if it has the highest or greater than 0.7 IoU with any ground truth box.
- Negative sample if IoU with all ground truth boxes is less than 0.3.

Non-maximum Suppression (NMS) - testing



At inference we assume that we don’t have ground truth boxes and our model returns much more boxes than there are objects at the image. We need to somehow select the right box and discard excessive ones. One can use NMS for that purpose. It can be described as follows:

- Select box with highest confidence from proposals, add it to results and remove from proposals.
- Remove all proposal boxes that have IoU greater than the threshold with the selected box.
- Repeat until there are no more proposals.

NMS can be applied class-wise. It is also used in the proposal selection from RPN network in Faster-RCNN. There are some improved variants of NMS e.g. Soft-NMS.

Average Precision (AP) and mean Average Precision (mAP)

Rank	Correct?	Precision	Recall
1	True	1.0 ↑	0.2 ↑
2	True	1.0 –	0.4 ↑
3	False	0.67 ↓	0.4 –
4	False	0.5 ↓	0.4 –
5	False	0.4 ↓	0.4 –
6	True	0.5 ↑	0.6 ↑
7	True	0.57 ↑	0.8 ↑

Box prediction is correct if IoU > threshold

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

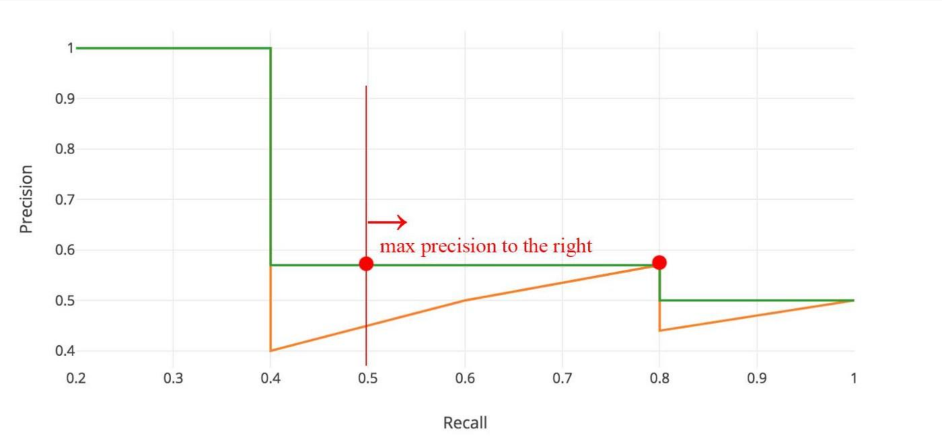
$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

TP = True positive

TN = True negative

FP = False positive

FN = False negative

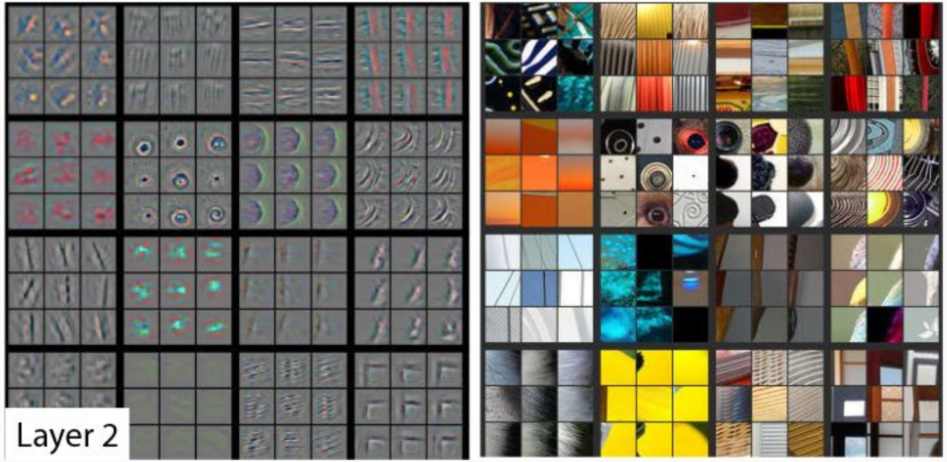


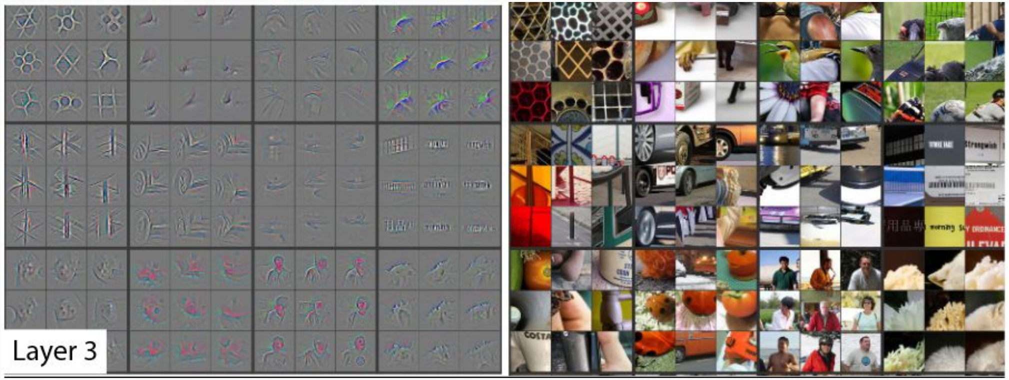
Orange curve represents the data in order of increasing recall. Green is orange curve after projection. AP is the integral under green curve.

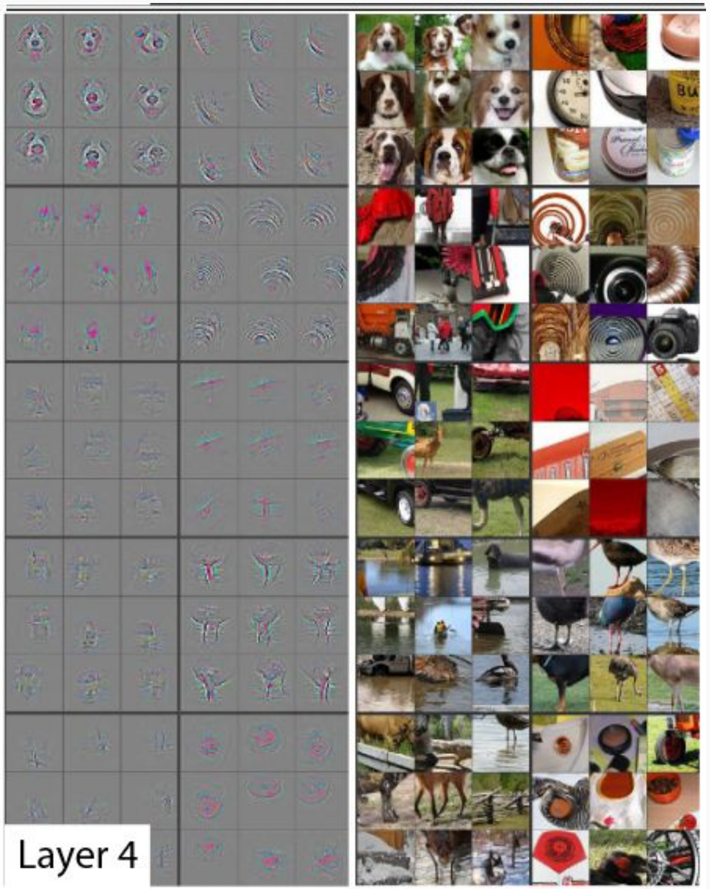
mAP is the mean of AP for different values of IoU threshold. For COCO benchmark it is of 0.5 to 0.95 with a step of 0.05.

https://jonathan-hui.medium.com/map-mean-average-precision-for-object-detection-45c121a31173

CNNs visualization

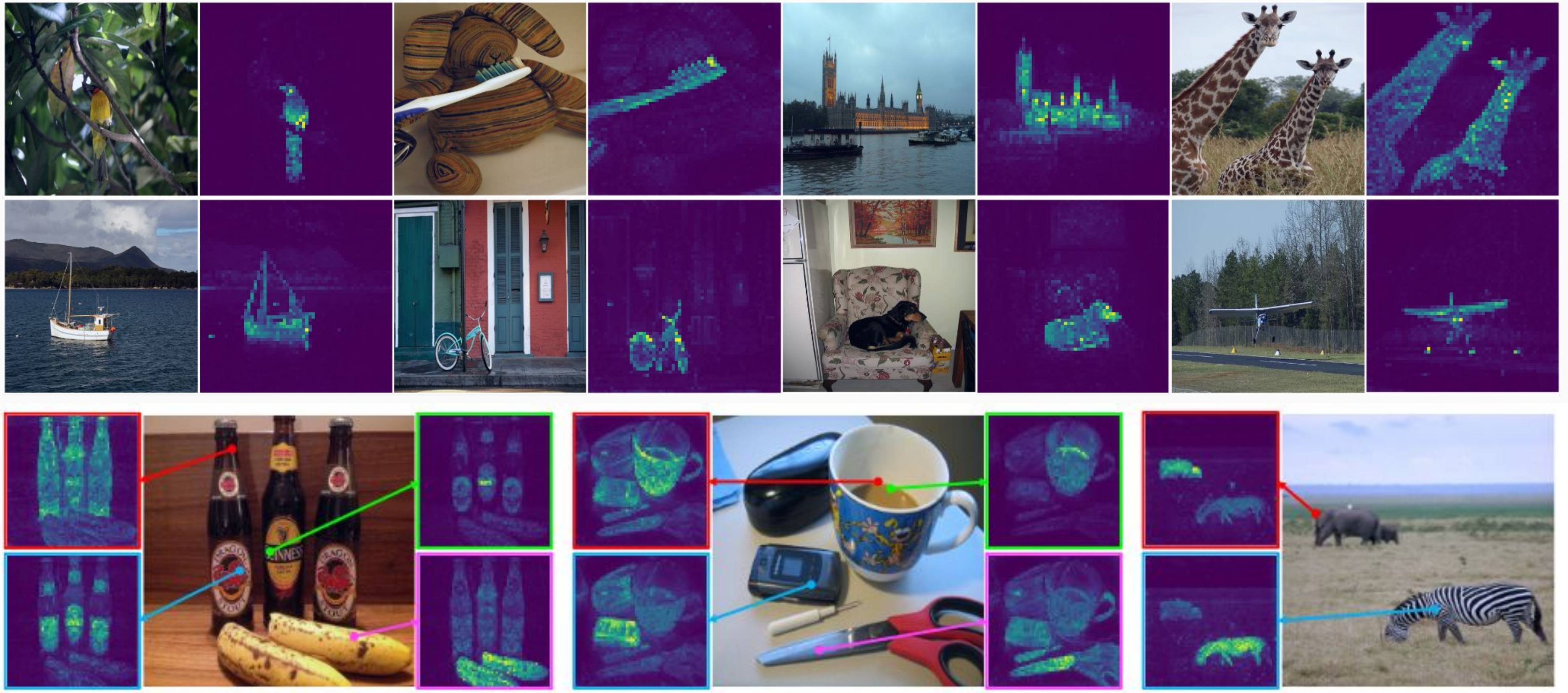






Matthew D. Zeiler and Rob Fergus, Visualizing and Understanding Convolutional Networks

ViTs visualization



Mathilde Caron et al., Emerging Properties in Self-Supervised Vision Transformers