Deep Learning - MAI

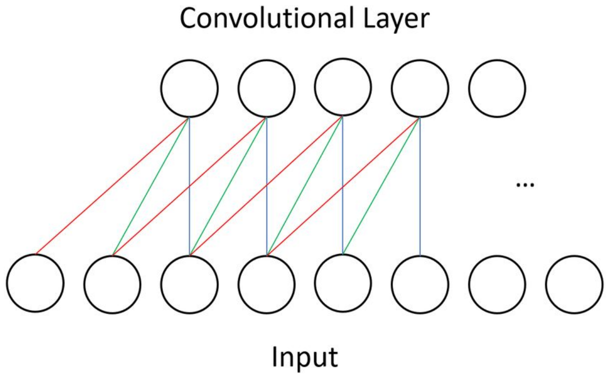
Convolutional neural networks

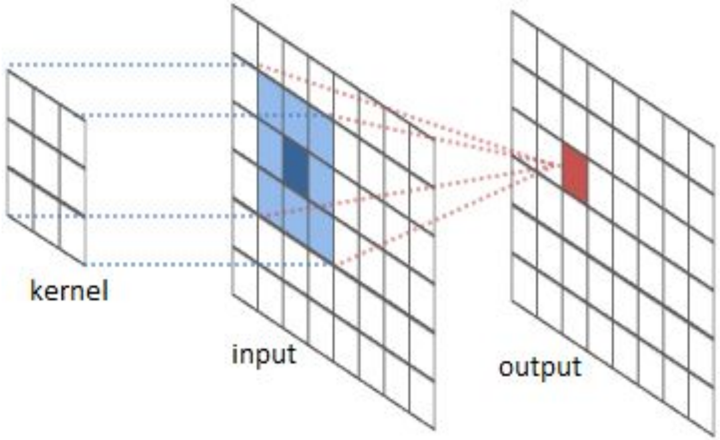
THEORY

Dario Garcia Gasulla
dario.garcia@bsc.es

Spatial Connectivity

- Some data has spatial correlations that can be exploited
  - 1D, 2D, 3D, ...
- Near-by data points are more relevant than far-away.
- Sparsify connectivity to reduce complexity and ease the learning





HIGH PERFORMANCE ARTIFICIAL INTELLIGENCE

Weight Sharing

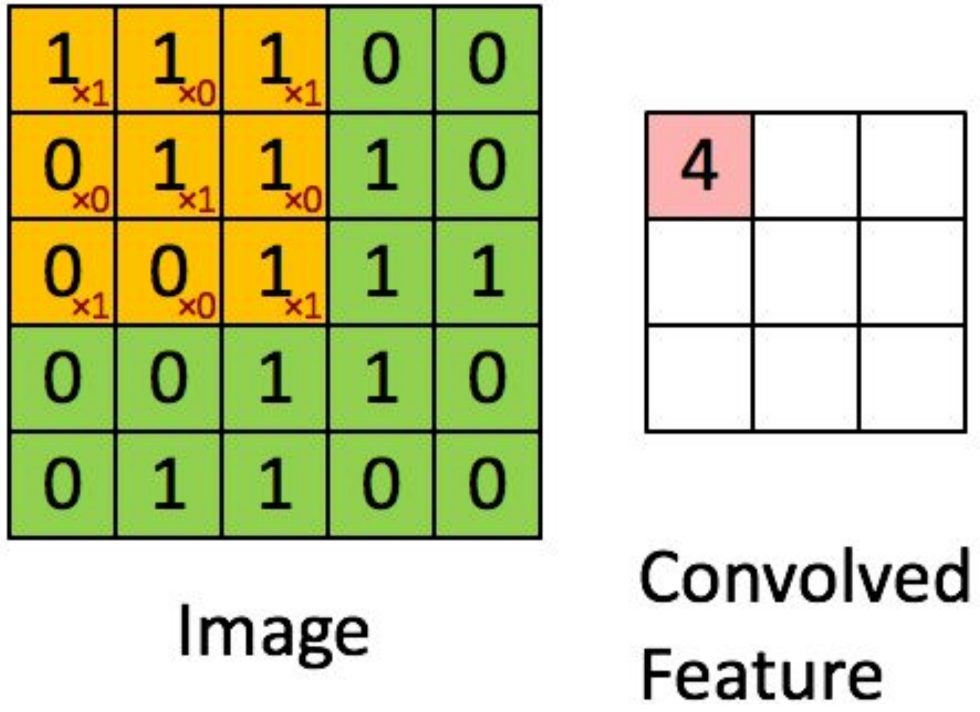
- Sparse connectivity is nice, but we want to apply filters everywhere.
- Each filter will get convolved all over the image: 2D activations matrix
- In static we have sets of neurons sharing weights
- In this context, what is a neuron?

Convolution in Action

Kernel size 3x3
(neuron input = 9)

1 0 1
0 1 0
1 0 1

Detect ‘X’



Filter convolution process

Activations (pre-func.)

Image Transformations

- Convolving filters transform the image
- Let the model learn the kernels it needs

Edge detection	$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$	
Sharpen	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	
Gaussian blur 3 × 3	$\frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$	

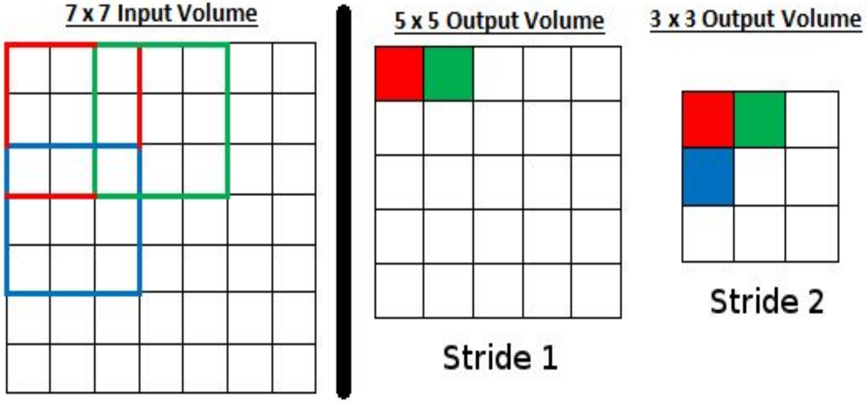
Convolution Details

Kernel size: Size of the receptive field of convolutional neurons

Stride: Steps size of convolution

Padding: Allows focus on border

- ❖ Most common fill: Zeros
- ❖ Valid (no padding): Internal only. May skip data. Reduces dimensionality
- ❖ Same: Keep dimensionality with stride 1

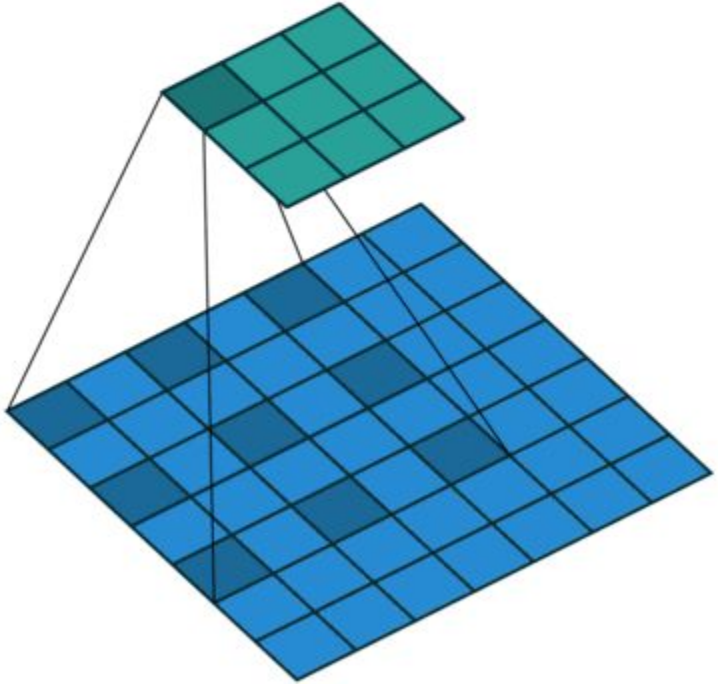


$$OutputSize = \frac{InputSize - KernelSize + 2 * Padding}{Stride} + 1$$

Dilated/Atrous Convolutions

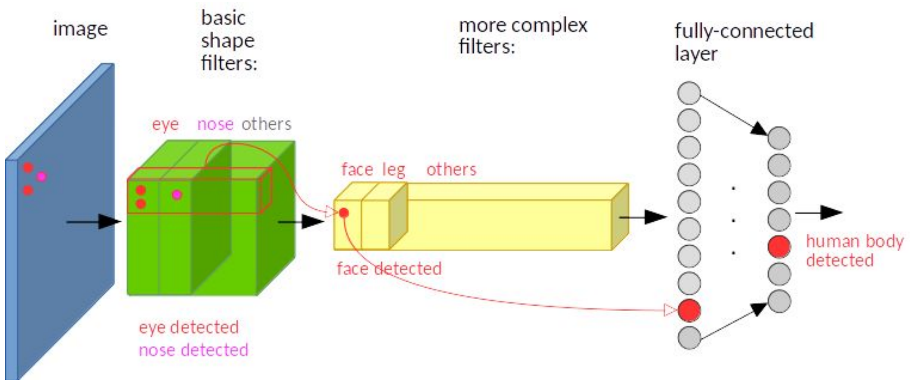
Sparsify the kernel

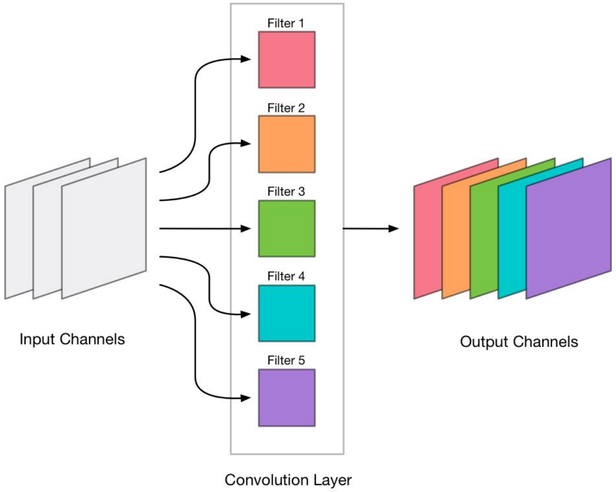
- Increases perceptive field without added complexity
- Loses details, gains context
- Another hyperparam :(
- Used for
  - Down/Upsampling (segmentation)
  - High Resolution inputs



Output Volumes

- Typically, conv filters are full depth (**N*N*input_depth**)
- Each conv filter (often 3D) convolved generates a 2D plane of data
- Depth provides all the *views* on a part of the input
- Output volume: New representation of input with different dimensions



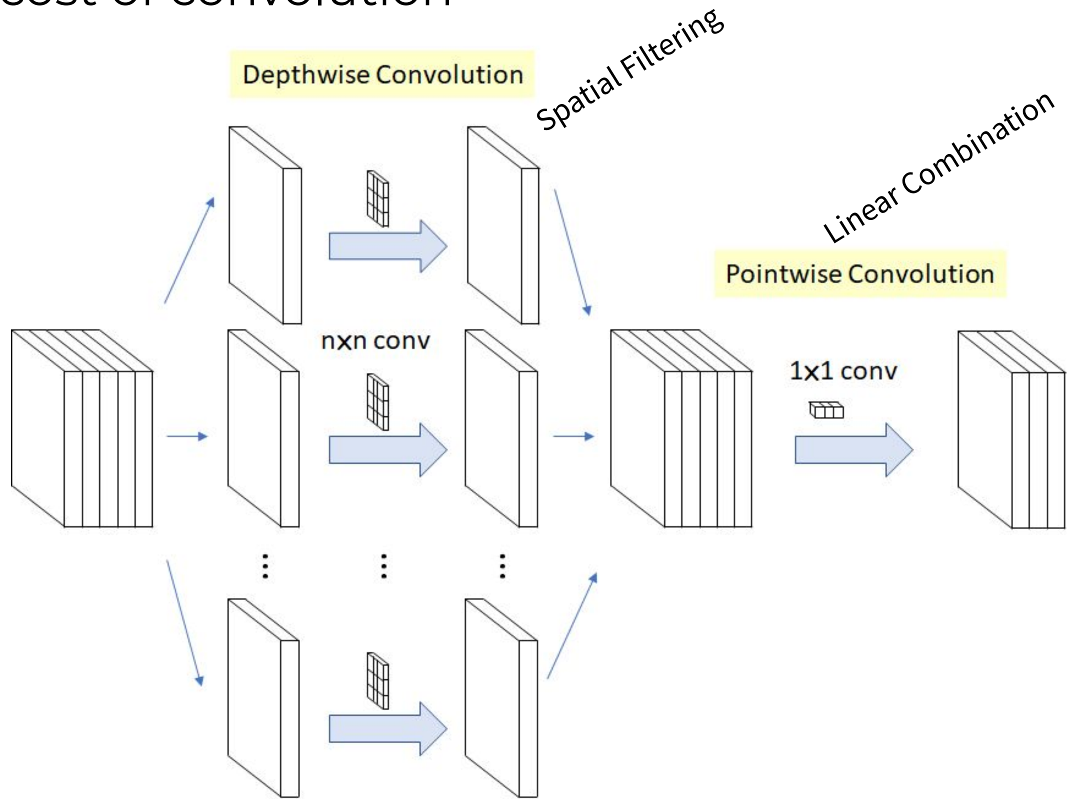


Depth-wise Separable Convolutions

Decreasing the complexity and cost of convolution

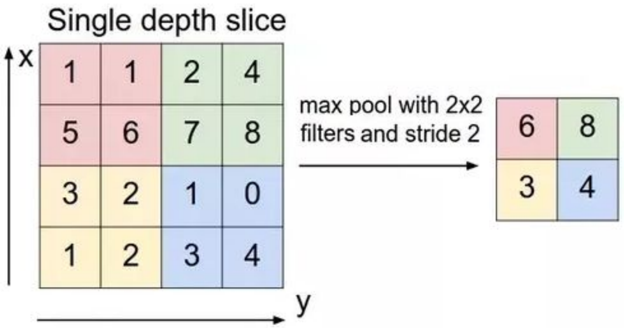
1. Depth-wise convolutions
   - Filters: **N*N*1**
2. Point-wise convolution
   - Filters: **1*1*input_depth**

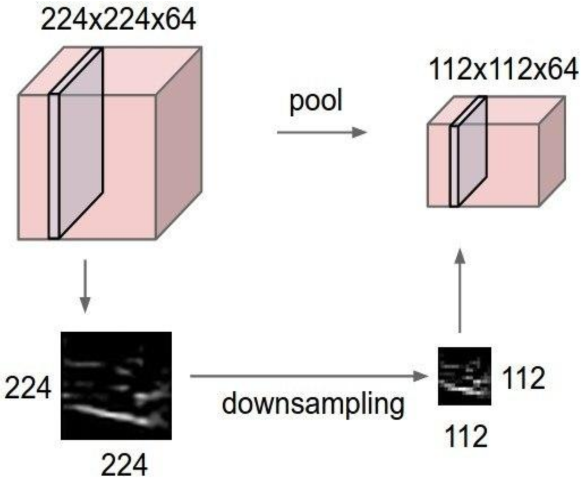
Params: **N*N+N**



To Pool Or Not To Pool

- Operation: **Max** or Avg
- Dimensionality reduction (along *x* and *y* only)
- Rarely applied full depth
- Parameter free layer
- Hyperparams: Size & Stride
- Loss in spatial precision / Robust to invariance



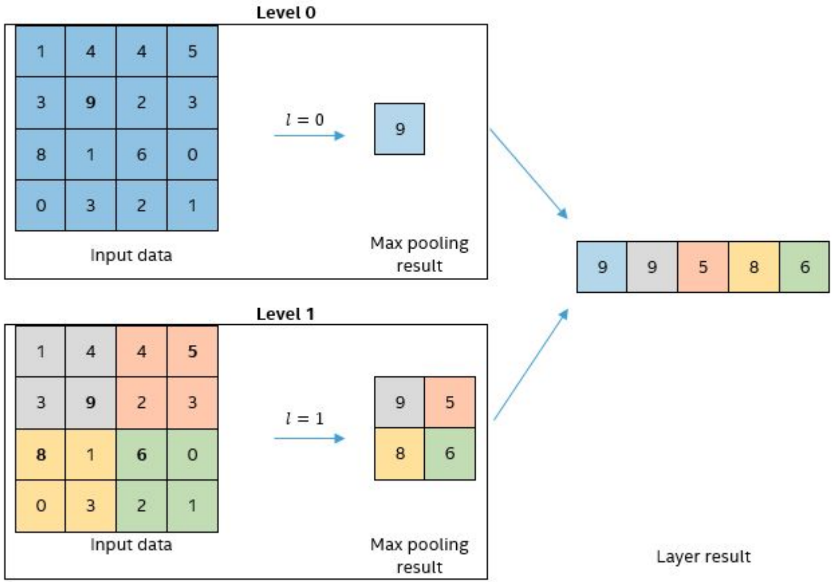


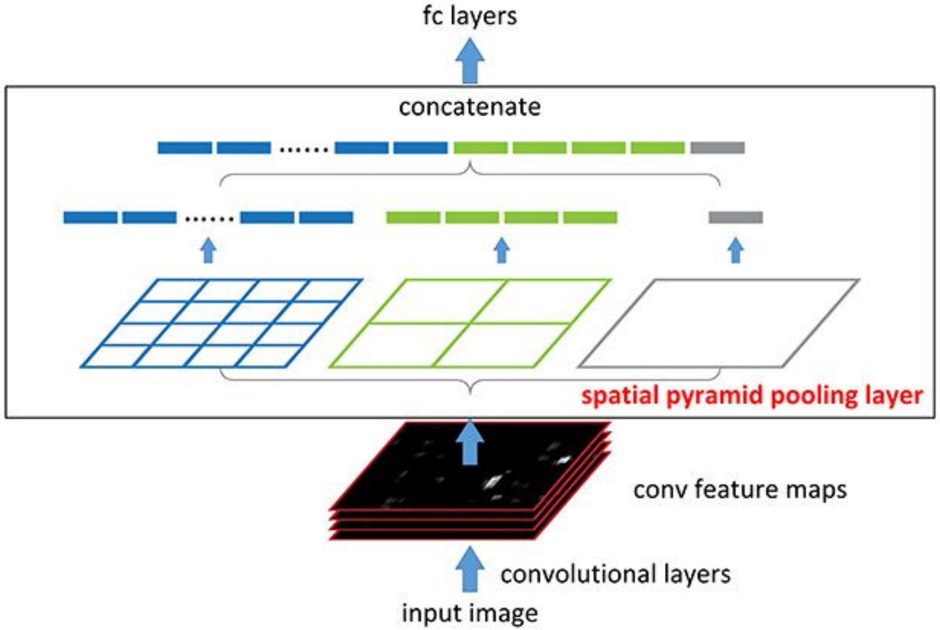
Other means to reduce complexity

- Depth-wise separable convs, bigger conv. strides

Spatial Pyramid Pooling (SPP)

- Multi-scale Pool (by powers of 2)
- Often used between conv and fc





More alternatives: Atrous spatial PP, Global average pooling, Pyramid pooling module, Adaptive PP

Practical Tips XI

Convolutional

- Small/big filters (**3x3**, 5x5, 7x7)
  - Cheap/Expensive
  - Local/General
  - Bigger/Smaller outputs (stride)
- Kernel Size = input size: fc
- Kernel size = 1x1: Alter depth)

Pooling

- 2x2, stride 1 is the least invasive

Hyperparameters incomplete list #4

- ❏ Kernel size (conv & pool)
- ❏ Stride (conv & pool)
- ❏ Padding (conv & pool)
- ❏ Num. filters
- ❏ Dilatation rate

HPAI
HIGH PERFORMANCE ARTIFICIAL INTELLIGENCE

The Challenge

ImageNet Large-Scale Visual Recognition Challenge 2012 (ILSVRC'12)

- ❖ Image Classification: 1,000 classes
- ❖ Training: 1.2M
- ❖ Living things + Human-made objects
  - ■ 120 breeds of dogs

HIGH PERFORMANCE ARTIFICIAL INTELLIGENCE

Data Augmentation for CNNs

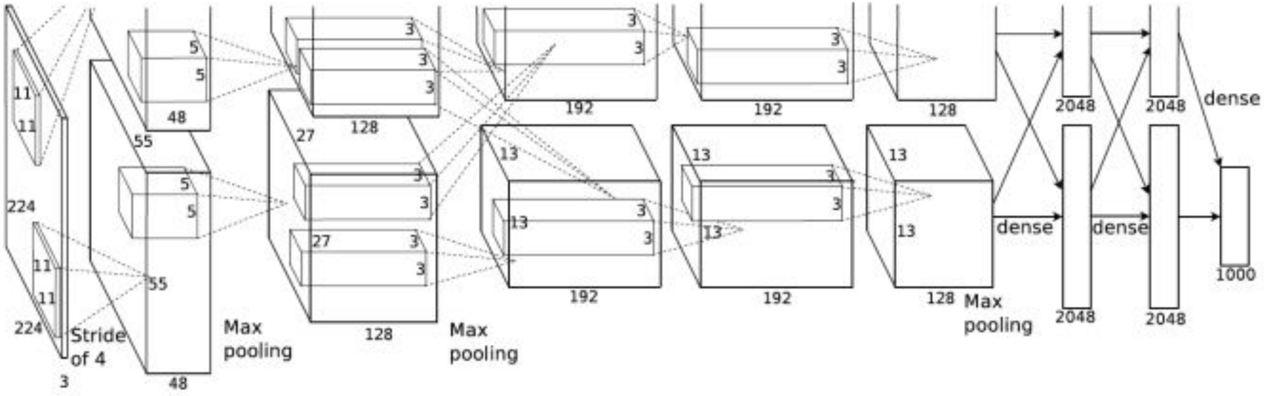
Apply what is safe for each case

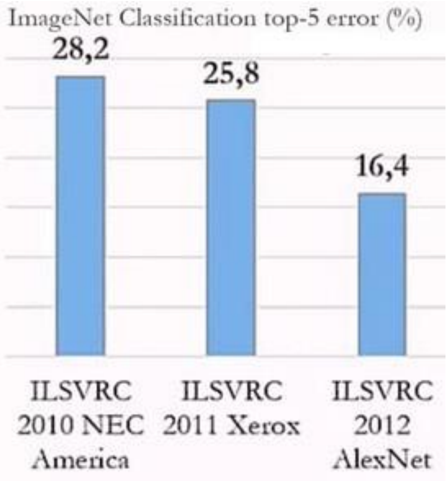
- Horizontal/Vertical flips
- Random Crops
- Color Jitter
- Rotation
- …

CNNs Big Bang

AlexNet (2012)

- ❖ Breakthrough in ILSVRC
- ❖ 5 convs+pools, ReLU, 2 dense, and dropout
- ❖ 62M parameters



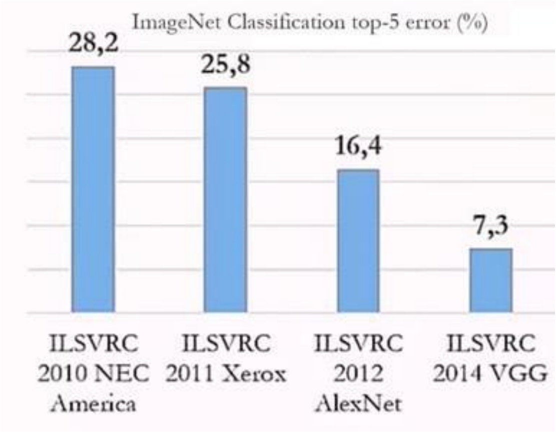


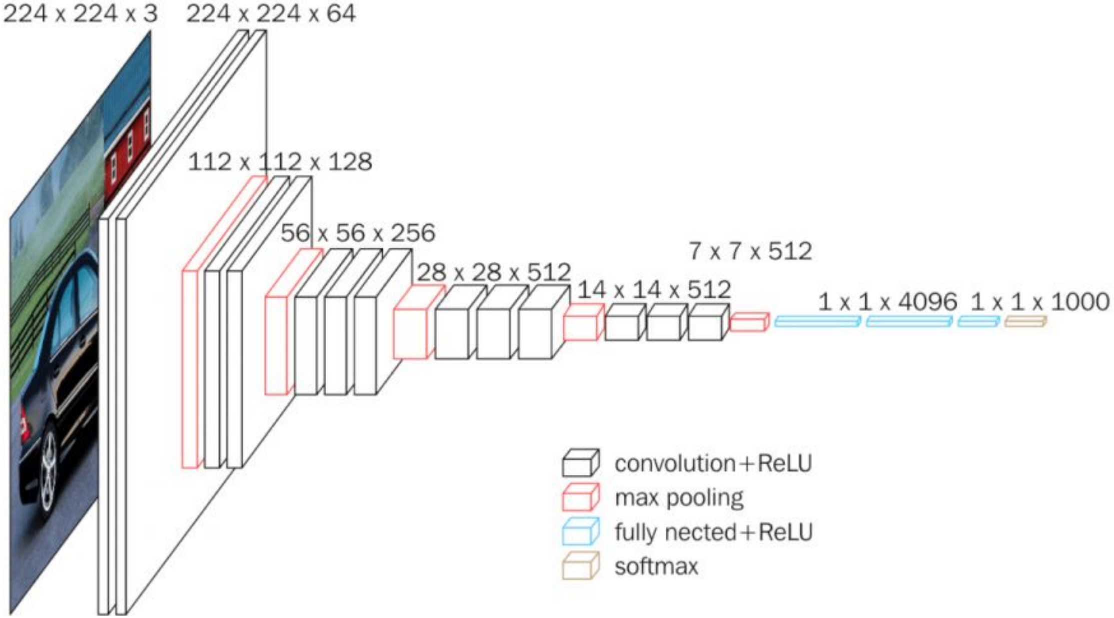
On the shoulders of giants

A new Standard for CNNs

VGG 11/13/16/19 (2014)

- Prototype of **(conv-pool)*+dense*** architecture
- 133-144M parameters
- 3x3 convs only

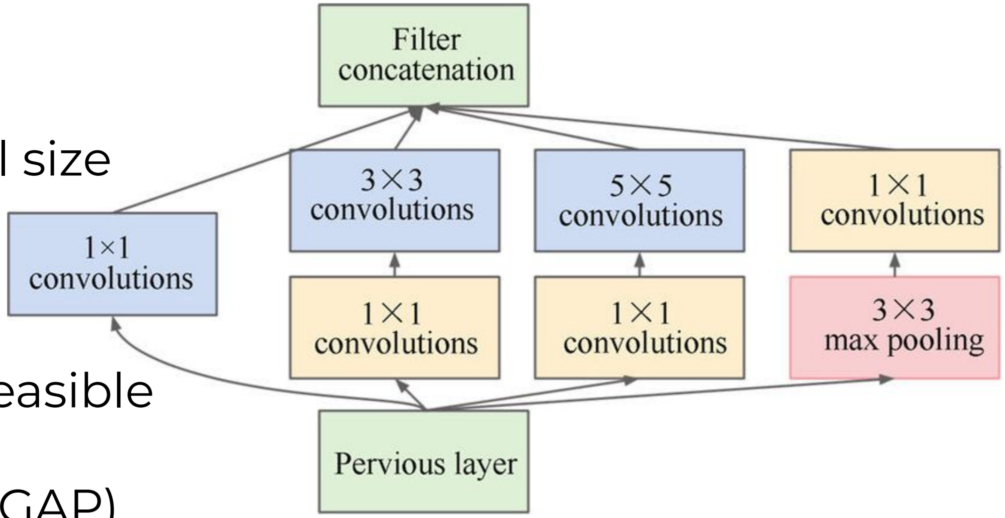




The Inception Family

GoogLeNet (2014)

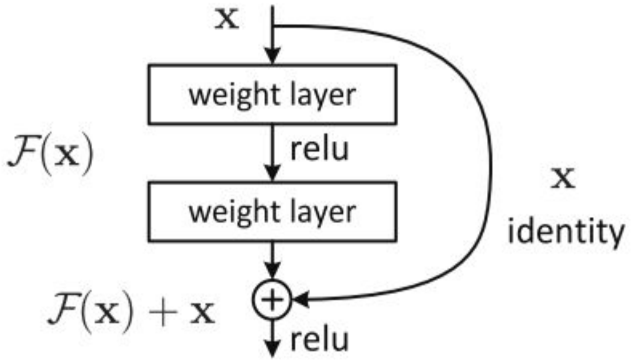
- ❖ The Inception block
- ❖ Let the model decide the kernel size
- ❖ Better scale adaptation
- ❖ Bottleneck 1x1 conv to make it feasible
- ❖ No FC: Global Average Pooling (GAP)

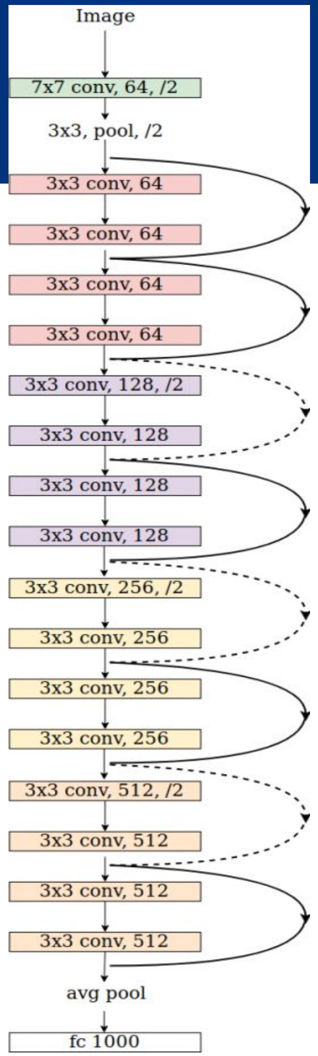


The Skipped Connection

ResNet (2015)

- ❖ Residual blocks / Skip connections
- ❖ Deeper should never be worse
  - ▪ Learning the identity is hard
  - ▪ Learning to cancel out is easy
- ❖ Shallow ensemble of nets
- ❖ Train up to 1K layers (do not!)
- ❖ ILSVRC'12 human level



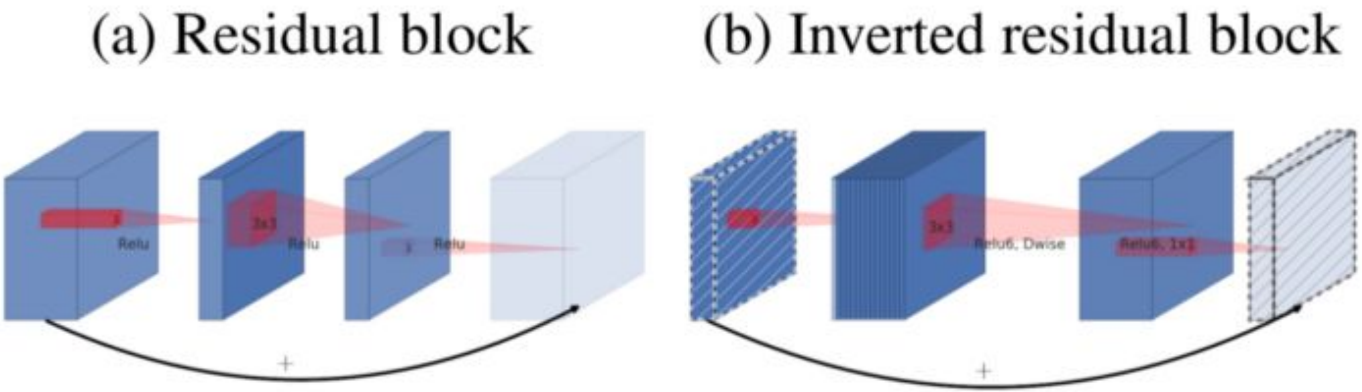


HIGH PERFORMANCE ARTIFICIAL INTELLIGENCE

[13]

Inverted Residuals and Linear Bottlenecks

1. Upsample depth
2. Depth-wise conv
3. Point-wise conv



- ❖ Linear act at end
- ❖ Non-linear mid
- ❖ Residual link
- ❖ Efficient

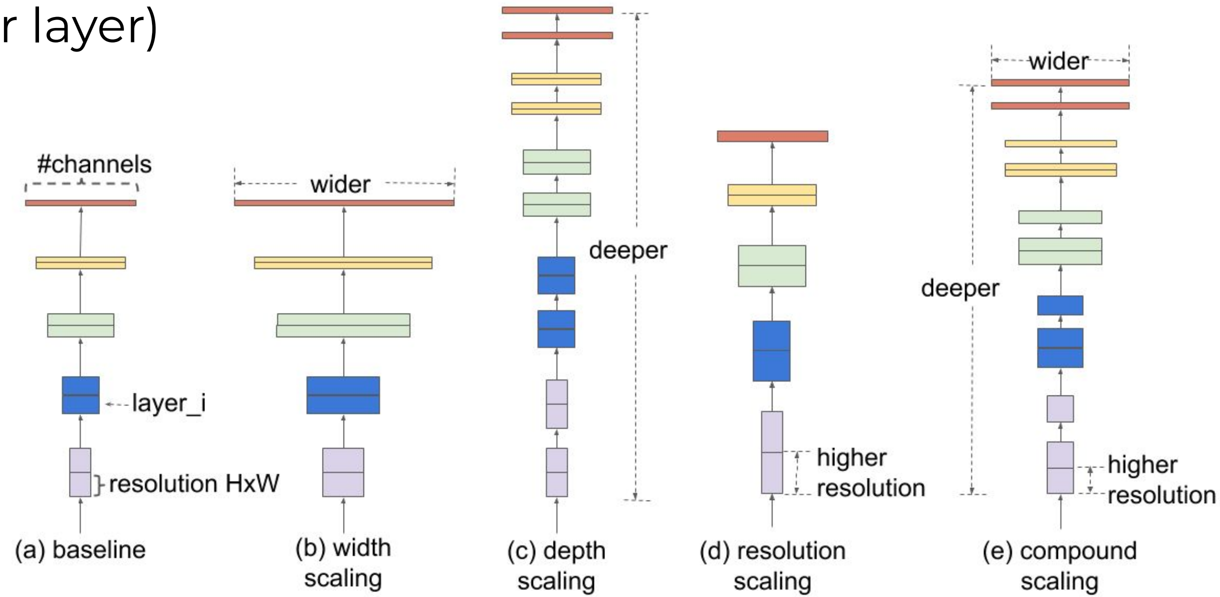
Sponsored by:
The manifold hypothesis

[38]

EfficientNet

Should I go deeper, wider or bigger?

- ❖ Find a balance between them (they are all related!)
  - ■ Width (neurons per layer)
  - ■ Depth (layers)
  - ■ Resolution (input)
- ❖ Choose a size
  - ■ EfficientNetB0-B7



HIGH PERFORMANCE ARTIFICIAL INTELLIGENCE

Noisy Student

A semi-supervised training paradigm

1. Train model A (teacher) with the labeled data
2. Use A to generate pseudo-labels for an unlabeled data set
3. Train model B (student) with both labeled and pseudo-labeled data

- Iterate, re-labeling the unlabeled data each time
- Highly regularized (noise!) student to guarantee improvement
- Each student has more capacity than the previous

Visualizing CNNs

The Basics

- NN are representation learning techniques
- CNNs build hierarchically complex features
  - From Gabor filters to dog faces
  - Induced by convolution
  - Tend to focus on the “non obvious for humans”
    - Backgrounds, textures
- The closer to the loss, more classifier (task) and less representation (data)

HIGH PERFORMANCE ARTIFICIAL INTELLIGENCE

[25,26]

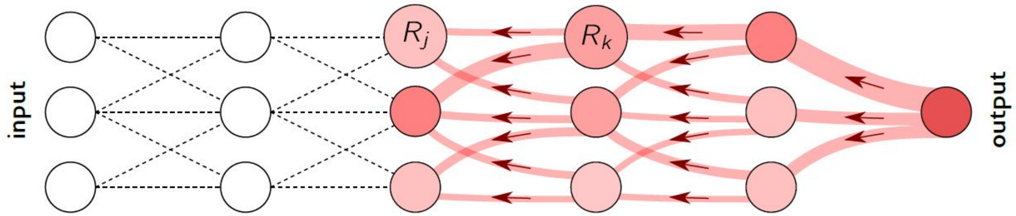
Ways of Looking at CNNs

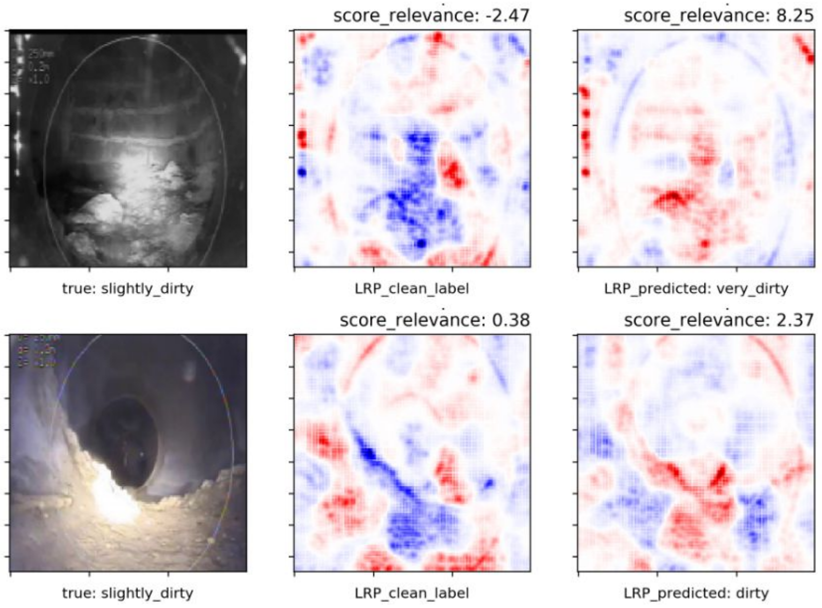
- *Attribution*: Where is the network looking?
  - Grounded. Instance based.
  - Explainability in practice.
- *Feature Visualization*: What is the network seeing?
  - Uncontextualized. Maximization based.
  - Diagnosys & Insight
- *Exemplification*: Which images cause a maximum activation?
  - Samples from a distribution

HIGH PERFORMANCE ARTIFICIAL INTELLIGENCE

Attribution

- Finding the importance of pixels
- Layerwise Relevance Propagation (LRP)
  - Backpropagate *an* output. Find the relevance of each neuron
    - Weighted by CNN parameters



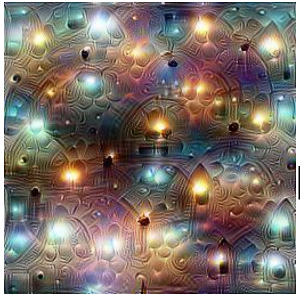


[29]

Feature Visualization

- Optimizing the input to maximize the output
    - A neuron
        - A channel
    - A layer (DeepDream)

Low level

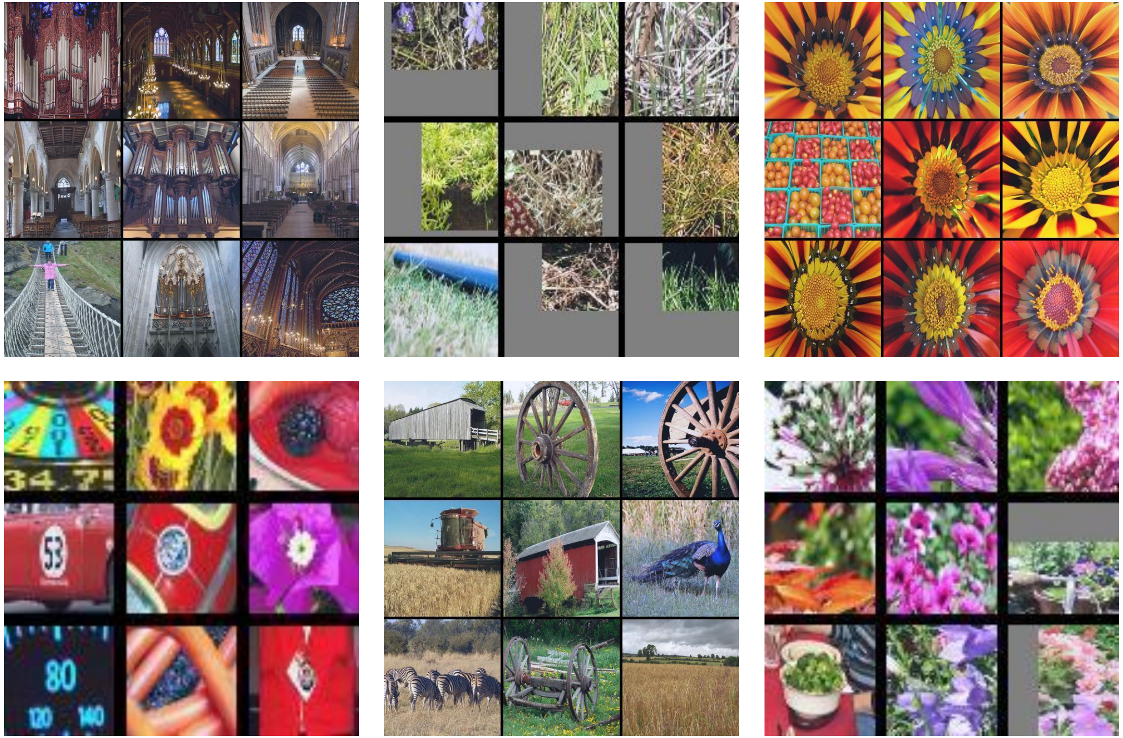


High level

HIGH PERFORMANCE ARTIFICIAL INTELLIGENCE

Exemplification

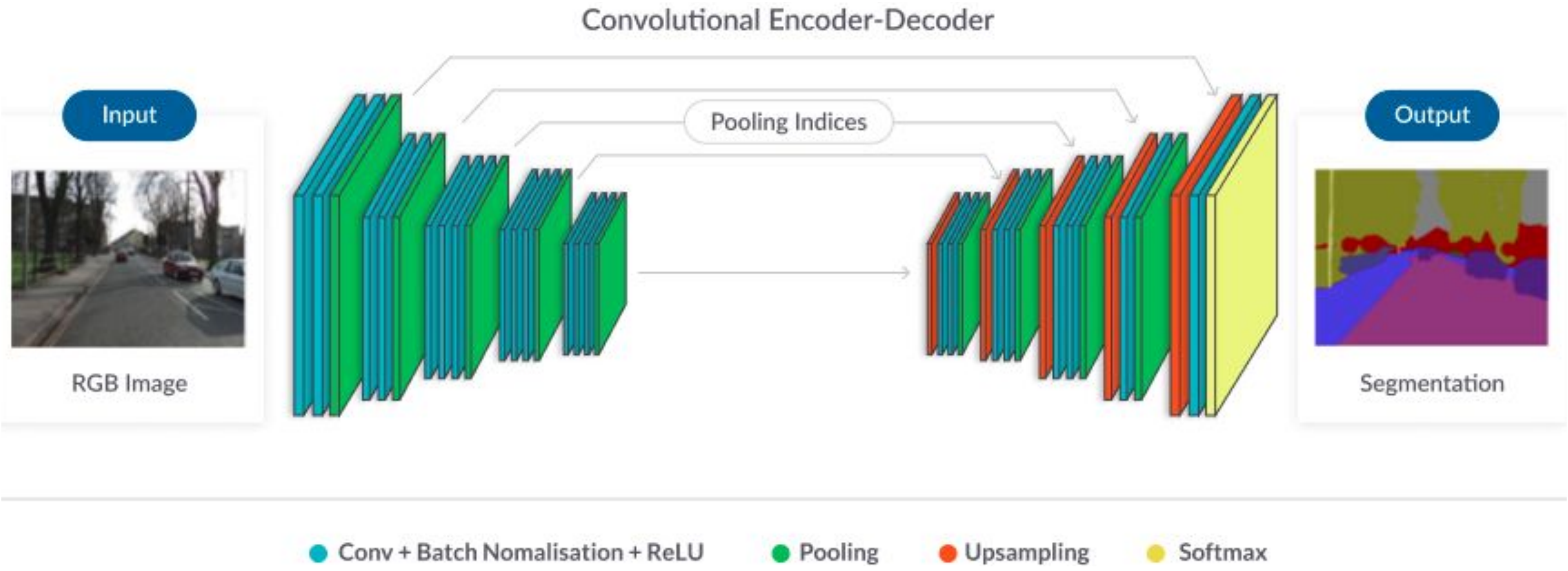
- Finding images within a dataset maximizing outputs
  - Subjective
  - Partial
  - Stochastic



Playing with CNNs

Encoder-Decoder CNNs

- Pixel-wise classification task (image reconstruction loss)
- Bottlenecking makes it cheaper



[14]

Automatic Image Colorization

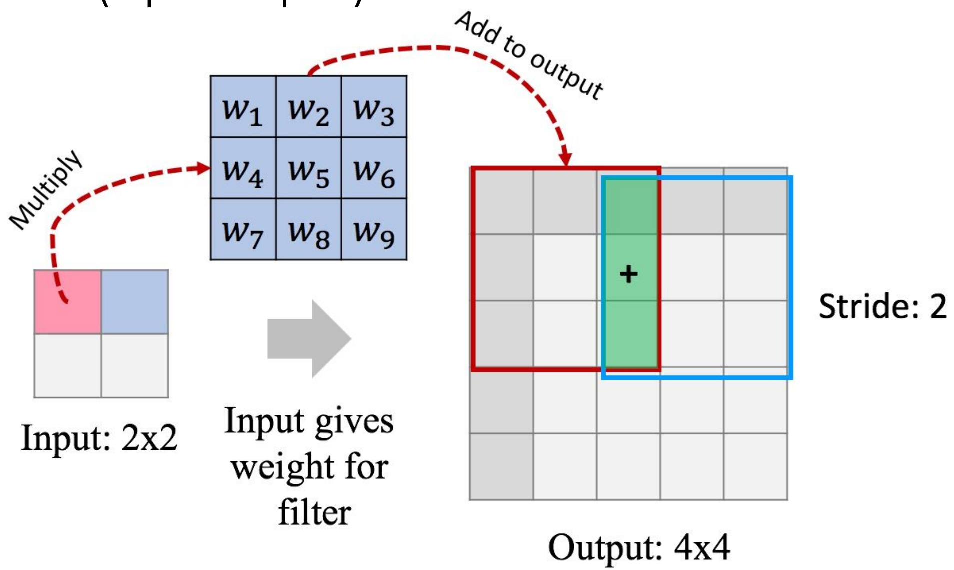
- Another pixel-wise classification application

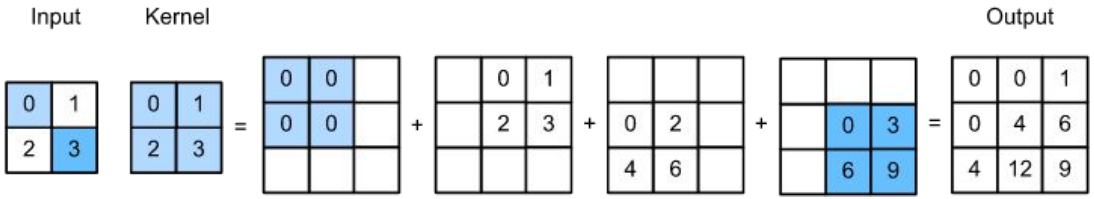
(a) Colorado National Park, 1941 (b) Textile Mill, June 1937 (c) Berry Field, June 1909 (d) Hamilton, 1936

[15,16]

Transposed Convolution ~~Deconvolution~~

- Reverse effect of regular convolution (upsample)
- Learnt interpolation
- Applications
  - Segmentation
  - GANs
  - Super-Resolution
  - Conv. Autoencoders





Faster Segmentation

- ~~Pixel-wise classification~~ Object detection (bounding box)
  - Can be done with a “regular” CNN
- R-CNN: Propose crops (SVM). Extract features (CNN). Classify crops (SVM)
- Fast R-CNN: Extract features. Propose crops. Classify/Bounding Box (CNN)
- Faster R-CNN: Propose crops through a specific sub-net (RPN)
- YOLO v? (no regions, faster, less accurate)
  - Divide into grid. Predict class and bounding box for each cell.

Better Segmentation

- Mask R-CNN
  - Faster R-CNN for object detection
  - FCN for instance segmentation (pixel classification)
- Xception
  - Depth-wise separable Convs (inverted order & w/o non-linearity)
  - Skip connections
  - Atrous SPP

HIGH PERFORMANCE ARTIFICIAL INTELLIGENCE

[17,18]

Style Transfer

- What do the correlation of activations intra-layer tell us?
  - What if we force it on another image?

- Gram matrix represents the *style*
  - Channel-wise (cXc)
  - Several mid layers

- Activations represents the *content*
  - One mid layer

- Optimize the **input** to minimize 2 losses
- Use a pre-trained net frozen
- Improved and extended

[19,20,21,22]

Image Generation

- StyleGAN2 + pix2pixHD
  - Pixel-wise generative models
- Flow-edge Guided Video Completion
- https://colab.research.google.com/drive/1KznIbRyNdiNBrrVbD7uolccdf9rngVUE?usp=sharing

HIGH PERFORMANCE ARTIFICIAL INTELLIGENCE

[34,35,36,37]

Handwritten Generation

- https://github.com/sjvasquez/handwriting-synthesis
- https://arxiv.org/abs/1308.0850
- https://www.calligrapher.ai/

[34,35,36,37]

References

[1] http://vordenker.de/ggphilosophy/mcculloch_a-logical-calculus.pdf
[2]http://www-public.tem-tsp.eu/~gibson/Teaching/Teaching-ReadingMaterial/Rosenblatt58.pdf
[3] http://www.dtic.mil/dtic/tr/fulltext/u2/236965.pdf
[4] https://en.wikipedia.org/wiki/Perceptrons_(book)
[5]http://www.andreykurenkov.com/writing/a-brief-history-of-neural-nets-and-deep-learning/
[6] https://en.wikipedia.org/wiki/Perceptrons_(book)
[7] Werbos et al. "Beyond regression:" new tools for prediction and analysis in the behavioral sciences." Ph. D. dissertation, Harvard University (1974).
[8] Rummelhart et al. "Learning Internal Representations by Error Propagation". MIT Press (1986).

HIGH PERFORMANCE ARTIFICIAL INTELLIGENCE

References

[9]https://towardsdatascience.com/effect-of-gradient-descent-optimizers-on-neural-net-training-d44678d27060
[10] https://arxiv.org/abs/1711.05101
[11] https://bbabenko.github.io/weight-decay/
[12] https://towardsdatascience.com/weight-decay-l2-regularization-90a9e17713cd
[13] Veit, Andreas, Michael J. Wilber, and Serge Belongie. "Residual networks behave like ensembles of relatively shallow networks." Advances in neural information processing systems. 2016.
[14] https://thegradient.pub/semantic-segmentation/
[15] https://arxiv.org/pdf/1603.08511
[16] https://pdfs.semanticscholar.org/5c6a/0a8d993edf86846ac7c6be335fba244a59f8.pdf

HIGH PERFORMANCE ARTIFICIAL INTELLIGENCE

References

[17] https://arxiv.org/pdf/1606.00915.pdf
[18] https://arxiv.org/pdf/1610.02357.pdf
[19] https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Gatys_Image_Style_Transfer_CVPR_2016_paper.pdf
[20] https://arxiv.org/abs/1603.08155
[21] https://arxiv.org/abs/1603.03417
[22] https://ai.googleblog.com/2016/10/supercharging-style-transfer.html
[23] https://arxiv.org/pdf/1903.07291.pdf
[24] http://nvidia-research-mingyuliu.com/gaugan
[25] Geirhos, Robert, et al. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." arXiv preprint arXiv:1811.12231 (2018).

References

[26] Beery, Sara, Grant Van Horn, and Pietro Perona. "Recognition in terra incognita." Proceedings of the European Conference on Computer Vision (ECCV). 2018.
[27] https://distill.pub/2017/feature-visualization/
[28] https://distill.pub/2018/building-blocks/
[29] Montavon, Grégoire, et al. "Layer-wise relevance propagation: an overview." Explainable AI: interpreting, explaining and visualizing deep learning. Springer, Cham, 2019. 193-209.
[30] https://medium.com/machine-intelligence-report/how-do-neural-networks-work-57d1ab5337ce
[31] Hebb, D.O. (1949), The organization of behavior, New York: Wiley

References

[32] Dauphin, Yann N., et al. "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization." Advances in neural information processing systems. 2014.
[33] Ruder, Sebastian. "An overview of gradient descent optimization algorithms." arXiv preprint arXiv:1609.04747 (2016).
[34] Viazovetskyi, Yuri, Vladimir Ivashkin, and Evgeny Kashin. 'StyleGAN2 Distillation for Feed-Forward Image Manipulation'. 7 March 2020. http://arxiv.org/abs/2003.03581.
[35] https://medium.com/@jonathan_hui/gan-stylegan-stylegan2-479bdf256299
[36] https://www.justinpinkney.com/making-toonify/
[37] http://chengao.vision/FGVC/files/FGVC.pdf
[38] Sandler, Mark, et al. "Mobilenetv2: Inverted residuals and linear bottlenecks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

References

[39] Howard, Andrew G., et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." arXiv preprint arXiv:1704.04861 (2017).

Dario Garcia-Gasulla (BSC)

dario.garcia@bsc.es

HIGH PERFORMANCE ARTIFICIAL INTELLIGENCE