Université de Paris

Deep learning and applications – Part 1

Sylvain Lobry

Analyse d’images - 29/09/23

Before we start…

https://app.wooclap.com/M2VMISS

Deep learning and applications

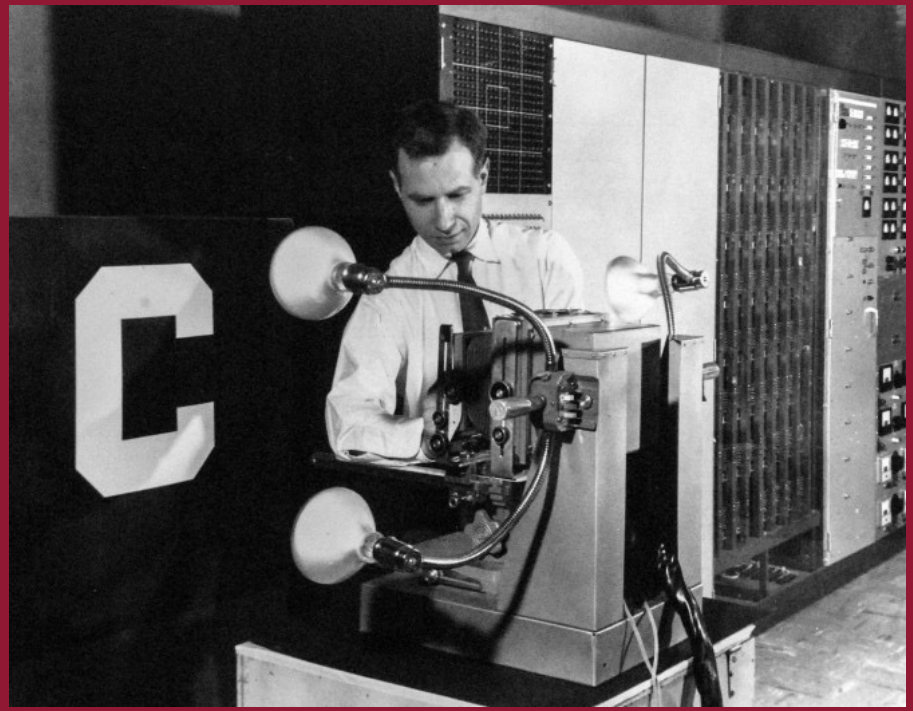
- Objective of Today:
  - Understanding better how does a deep learning model work and how to train one
  - See typical use cases and deep learning solutions

- Most current Machine learning research is deep learning based
- Driver of a transformation of our society: 1.75 Millions jobs in information and communication technology to be created in EU by 2030*

*estimation, from https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=58918

Deep learning, a new thing?

Source: https://www.youtube.com/watch?v=6fXNiJXUheI



Source: https://news.cornell.edu/stories/2019/09/professors-perceptron-paved-way-ai-60-years-too-soon

Article from the New York Times, July 1958:

« The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence »

Rise and fall of NN – Episode 1

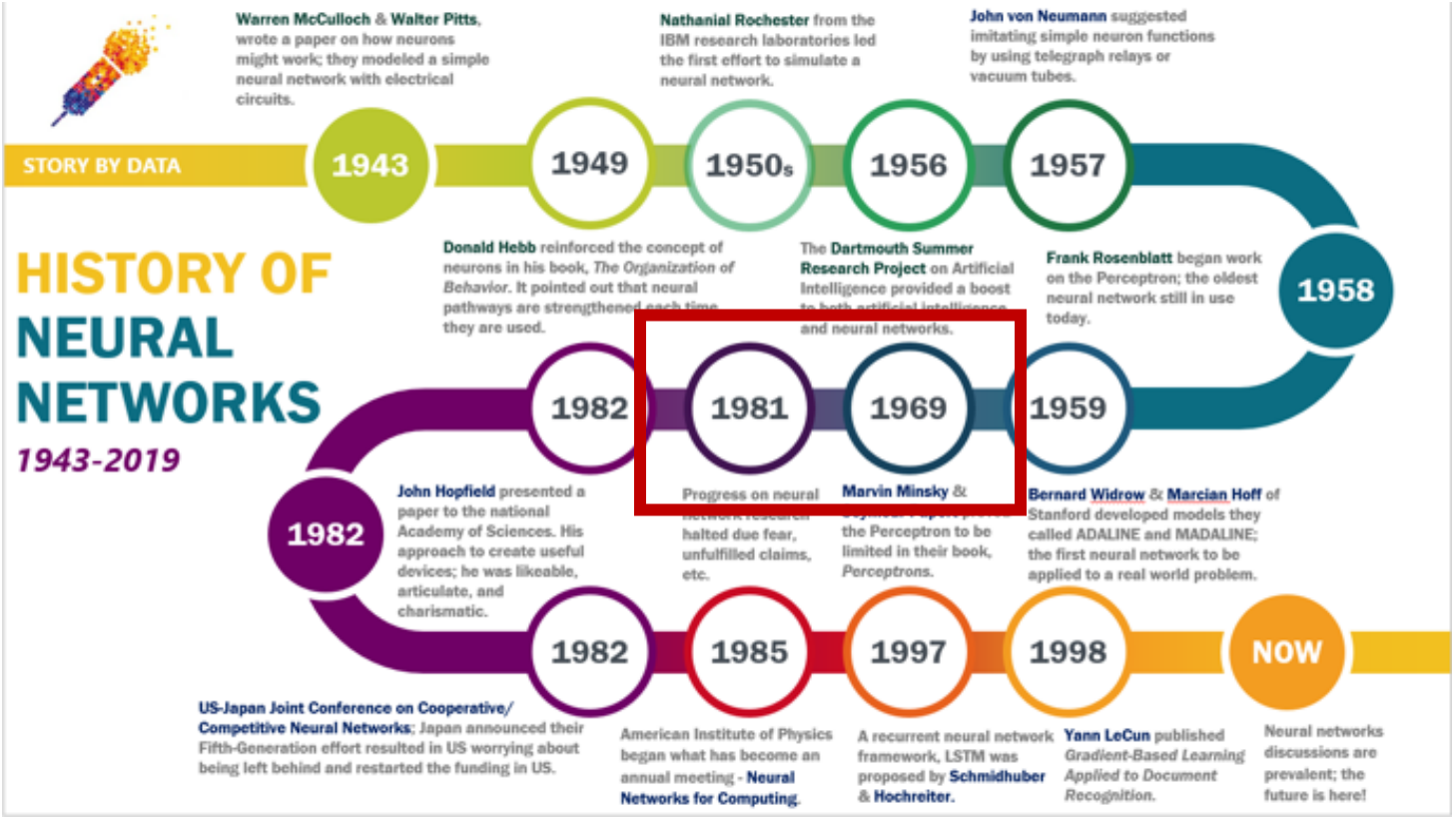
Article from the New York Times, July 1958:

« The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence »

The perceptron has shown itself worthy of study despite (and even because of!) its severe limitations. […] **There is no reason to suppose that any of these virtues carry over to the many-layered version**. Nevertheless, we consider it to be an important research problem to elucidate (or reject) our intuitive judgement that the **extension to multilayer systems is sterile**.

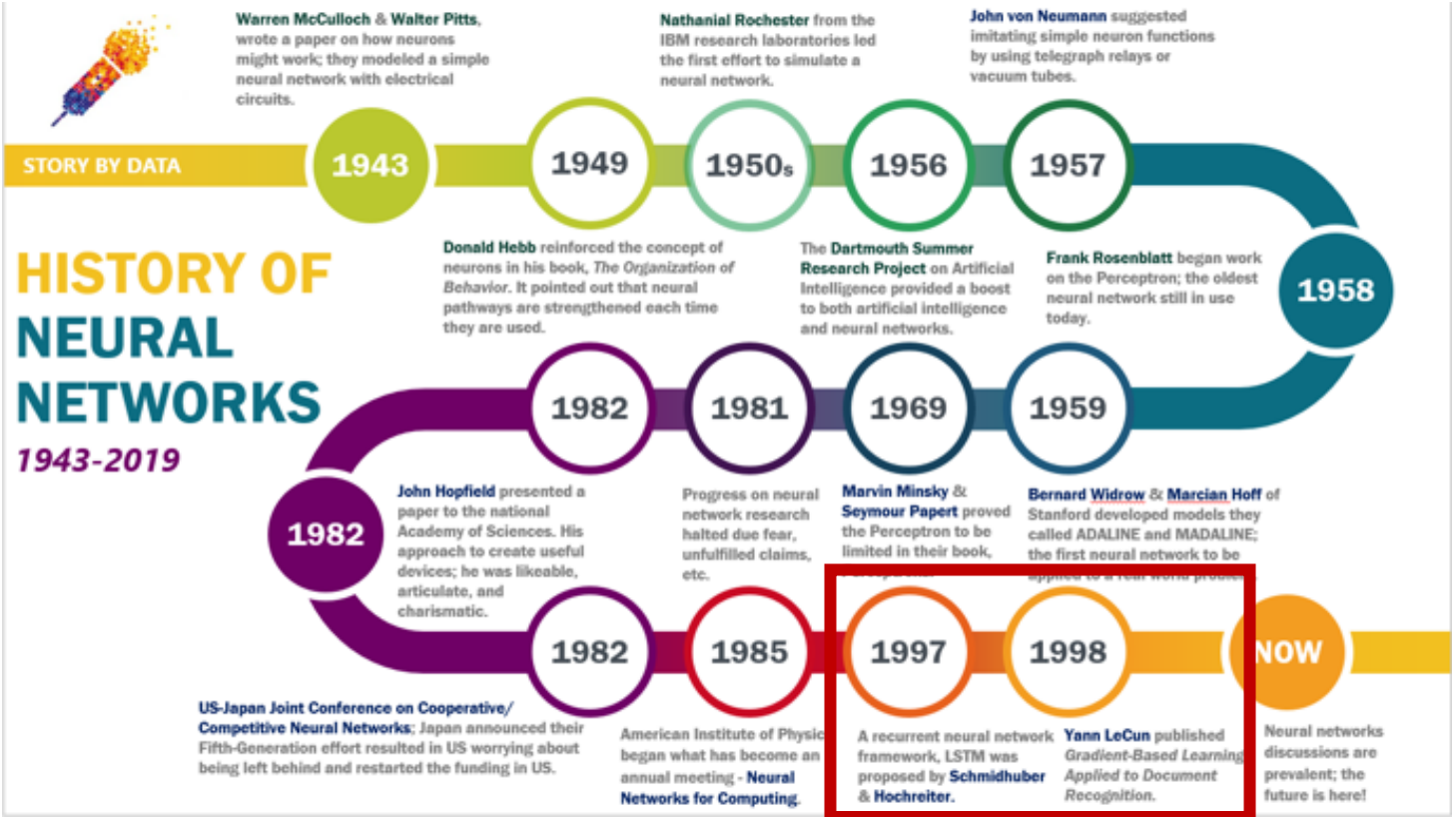
Minsky and Papert, 1969

Deep learning, a new thing?



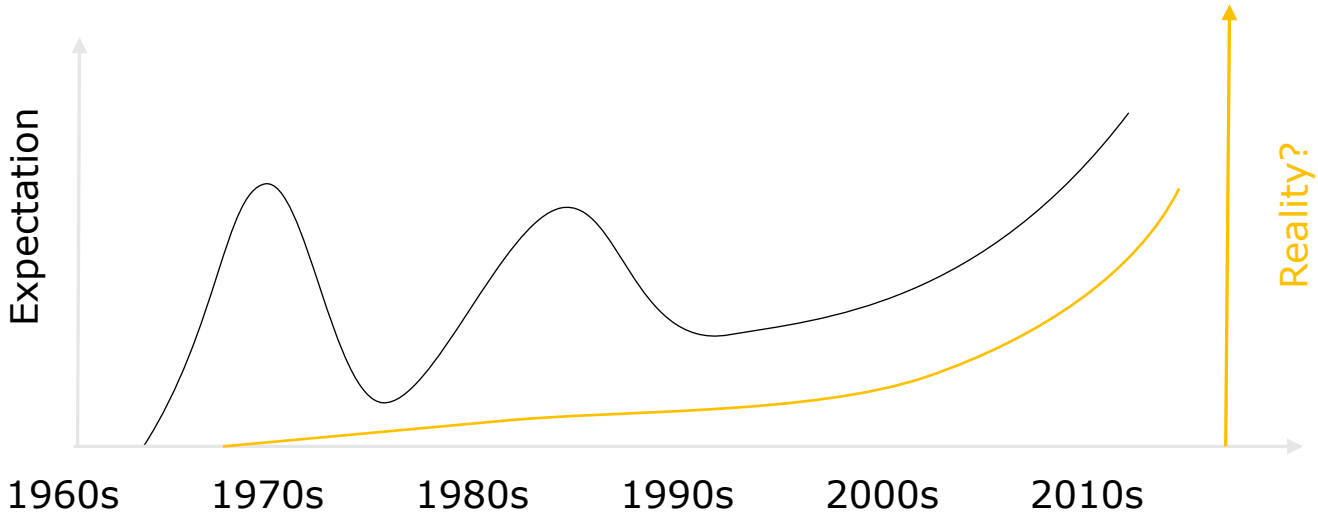
Source: https://www.youtube.com/watch?v=6fXNiJXUheI

Rise and fall of NN – Episode 2



Source: https://www.youtube.com/watch?v=6fXNiJXUheI

Rise and fall of NN

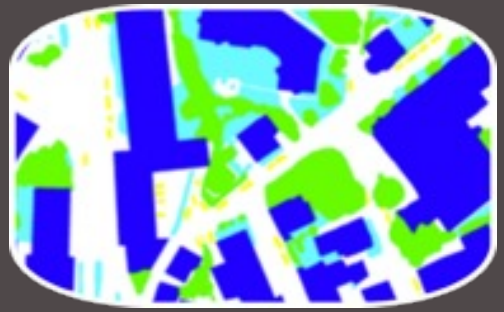


Learning from history

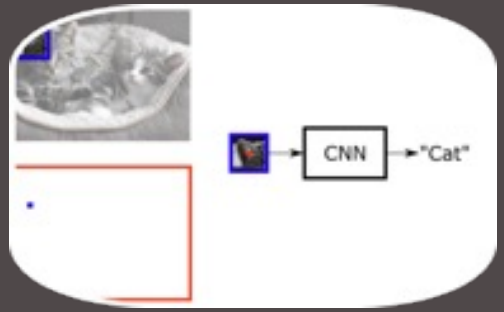
- Methods that we use for deep learning are old (Basis were invented ~70 years ago!)
- Two times through history: lot of expectations, lot of disappointment.
- Today: lot of expectations…

Menu of the day

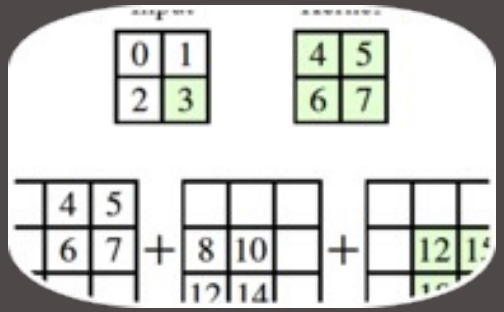
Intro to semantic segmentation



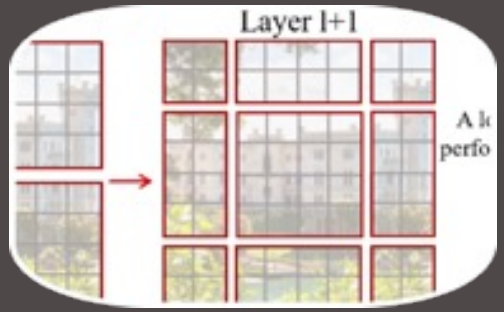
Remote sensing applications



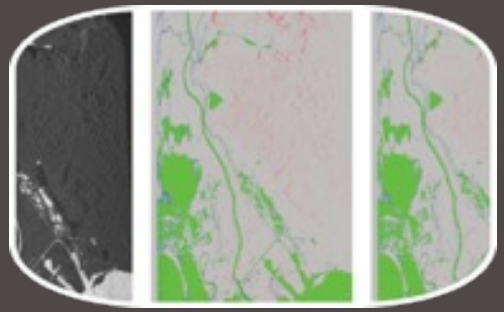
First models



Transposed convolution



Modern architectures



Besides deep learning

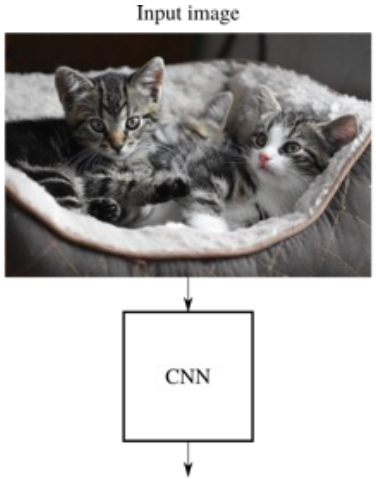
Semantic segmentation

What can you do with a CNN and an image?

Input image

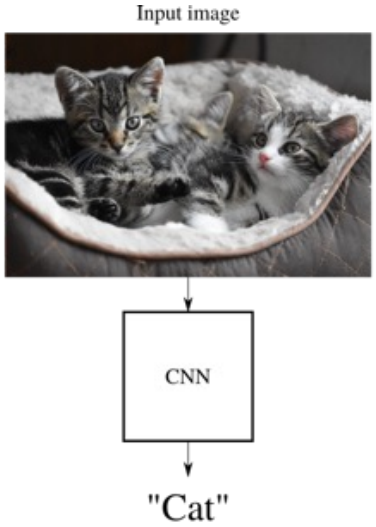
Semantic segmentation

What can you do with a CNN and an image?



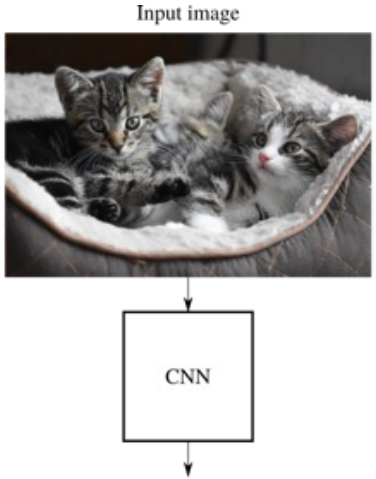
Semantic segmentation

What can you do with a CNN and an image?



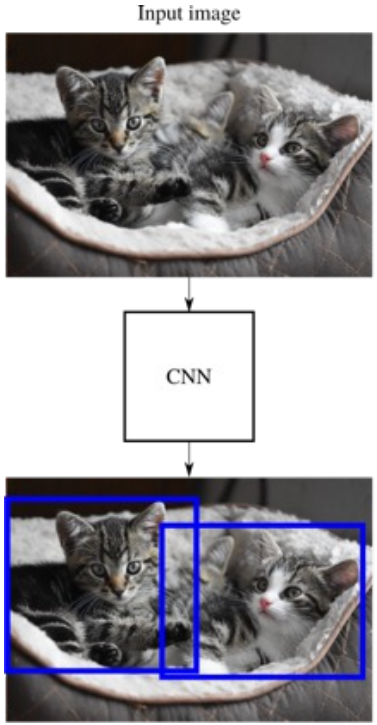
Semantic segmentation

What can you do with a CNN and an image?



Semantic segmentation

What can you do with a CNN and an image?

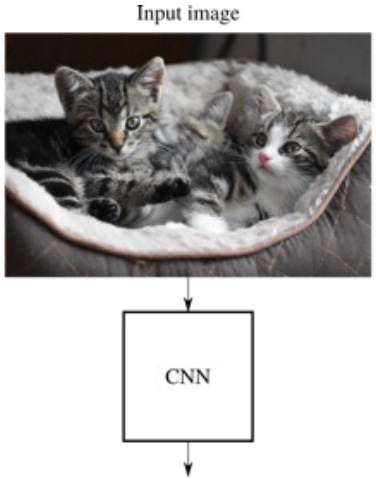


Classification

"Cat"

Semantic segmentation

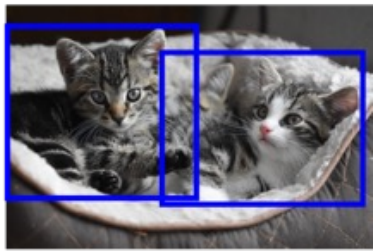
What can you do with a CNN and an image?



Classification

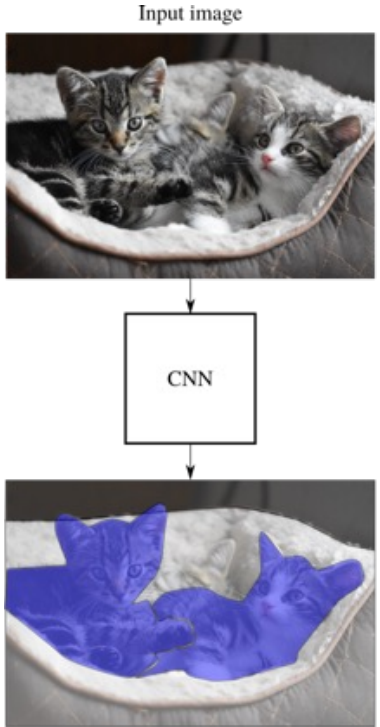
"Cat"

Object detection



Semantic segmentation

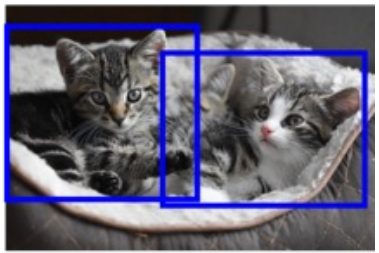
What can you do with a CNN and an image?



Classification

"Cat"

Object detection



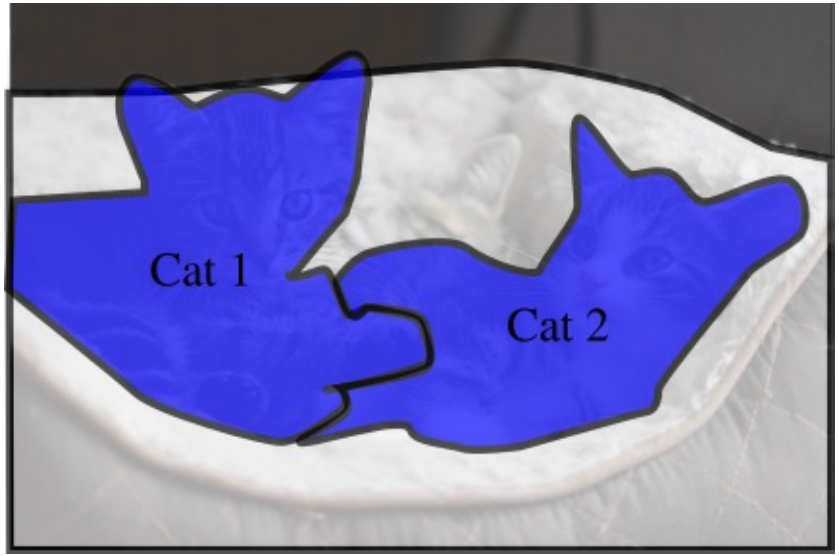
Semantic segmentation

Semantic segmentation

1 pixel = 1 label
+ Good precision
- More complex (high dimensional output)

Semantic segmentation

Side note: Instance segmentation

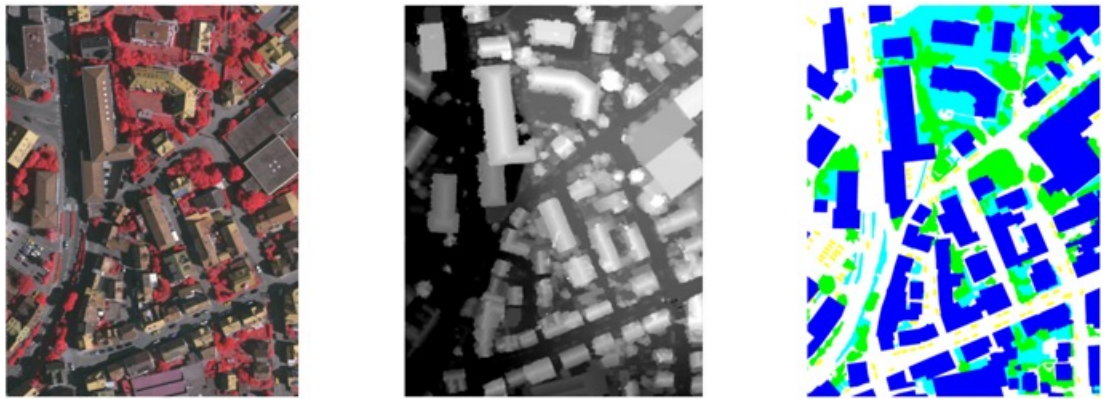


1 pixel = 1 label and 1 instance
We will NOT do that today

Semantic segmentation

Why do we need it?

- Maps



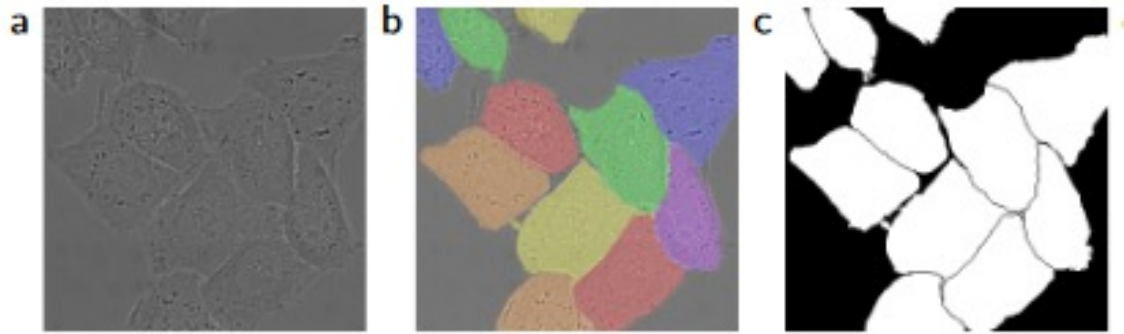
(a) (b) (c)

http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html

Semantic segmentation

Why do we need it?

- Maps
- Biology



Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.

Semantic segmentation

Why do we need it?

- Maps
- Biology
- Autonomous driving, …

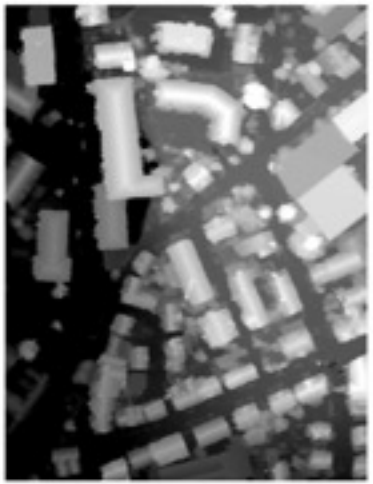
Zhu, Yi, et al. "Improving semantic segmentation via video propagation and label relaxation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

Semantic segmentation

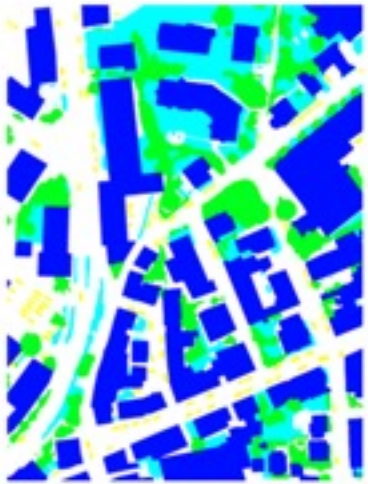
Remote Sensing datasets

- Semantic segmentation from very high resolution NIR/R/G + DSM

(a)



(b)



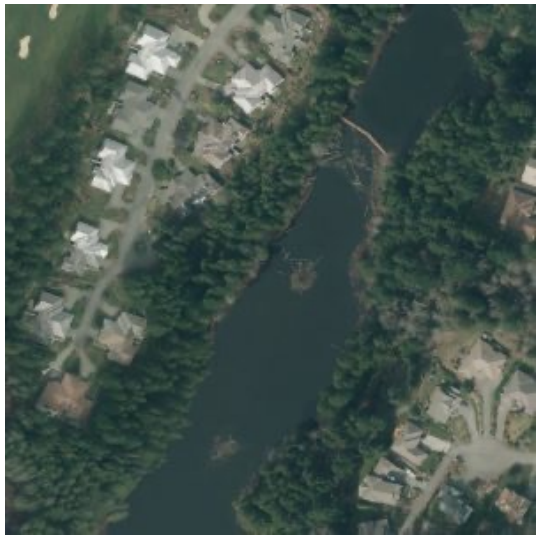
(c)

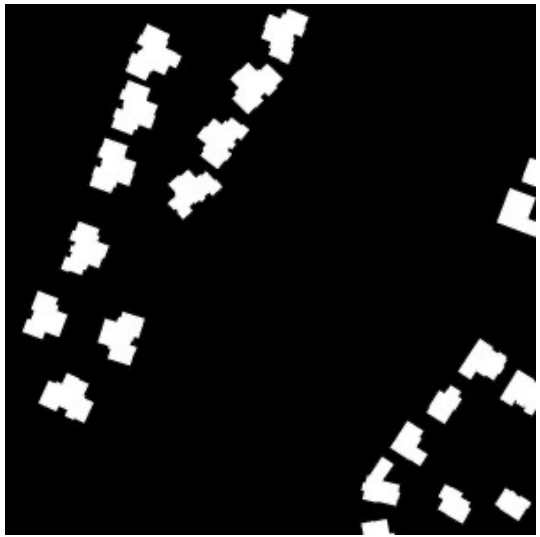
2D Semantic Labeling – Vaihingen - https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx

Semantic segmentation

Remote Sensing datasets

- Delineate buildings from very high resolution RGB imagery



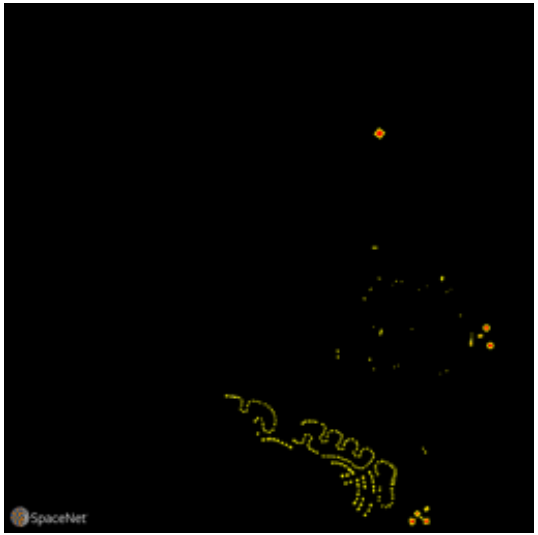


INRIA aerial image labeling dataset https://project.inria.fr/aerialimagelabeling/

Semantic segmentation

Remote Sensing datasets

- Semantic segmentation from high resolution RGB time series



SpaceNet 7- https://spacenet.ai/sn7-challenge/

Semantic segmentation

Remote Sensing datasets

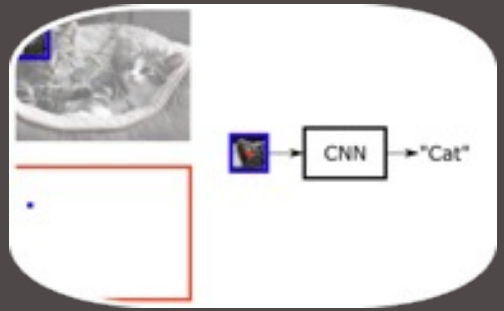
- Semantic segmentation of land cover from S1/S2

Schmitt, M., Hughes, L. H., Qiu, C., & Zhu, X. X. (2019). SEN12MS–A CURATED DATASET OF GEOREFERENCED MULTI-SPECTRAL SENTINEL-1/2 IMAGERY FOR DEEP LEARNING AND DATA FUSION. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *4*, 153-160.

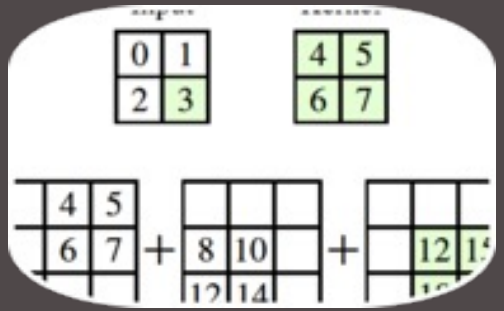
Menu of the day

Intro to semantic segmentation

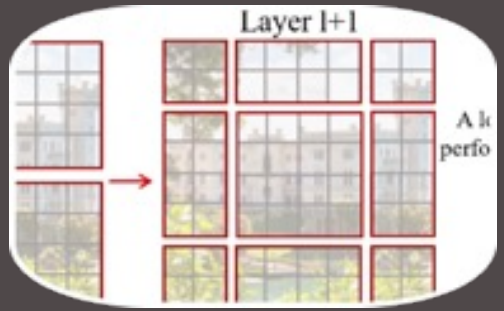
Remote sensing applications



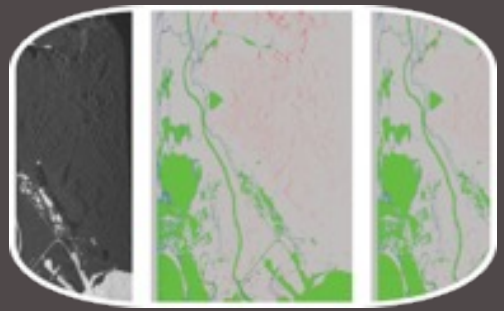
First models



Transposed convolution



Modern architectures



Besides deep learning

Semantic segmentation

Naïve semantic segmentation

- Idea 1: classify each pixel independently

Ciresan, Dan, et al. "Deep neural networks segment neuronal membranes in electron microscopy images." Advances in neural information processing systems, NIPS 2012

Semantic segmentation

Naïve semantic segmentation

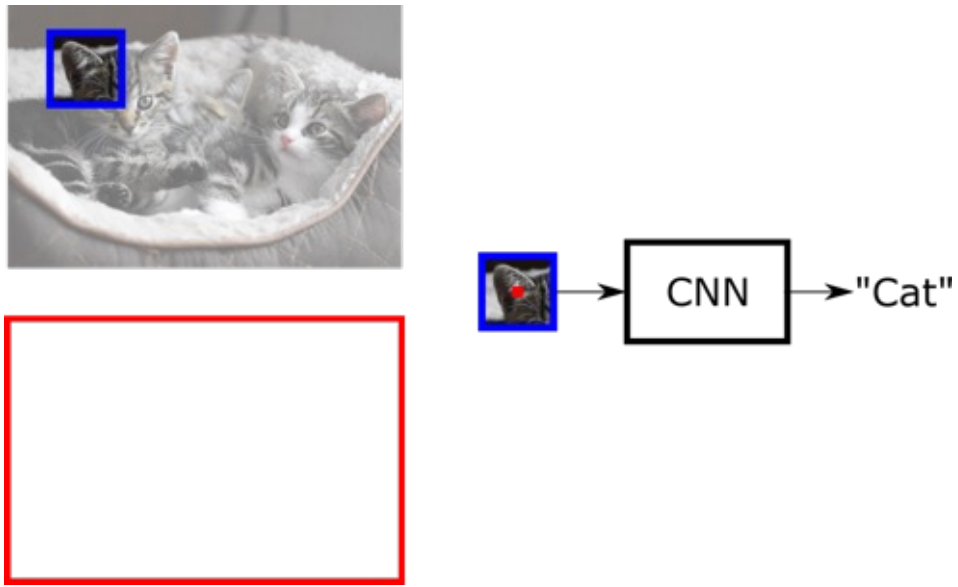
- Idea 1: classify each pixel independently

Ciresan, Dan, et al. "Deep neural networks segment neuronal membranes in electron microscopy images." Advances in neural information processing systems, NIPS 2012

Semantic segmentation

Naïve semantic segmentation

- Idea 1: classify each pixel independently



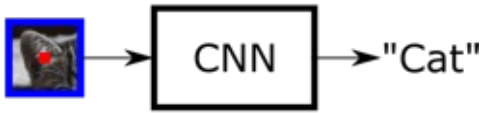
Ciresan, Dan, et al. "Deep neural networks segment neuronal membranes in electron microscopy images." Advances in neural information processing systems, NIPS 2012

Semantic segmentation

Naïve semantic segmentation

- Idea 1: classify each pixel independently (using a classical cross-entropy loss)

$$l(\mathbf{y_i}, \hat{y}_i) = -\log\left(\frac{exp(y_{i,\hat{y}_i})}{\sum_c exp(y_{i,c})}\right)$$, where $y_{i,c}$ is the score of class c at pixel i



Ciresan, Dan, et al. "Deep neural networks segment neuronal membranes in electron microscopy images." Advances in neural information processing systems, NIPS 2012

Semantic segmentation

Naïve semantic segmentation

- Idea 1: classify each pixel independently
- In 2012 -> Best model by a large margin (winner of ISBI 2012Electron Microscopy)
- Slow
- Trade-off between context and localization accuracy (because of max pooling)

Ciresan, Dan, et al. "Deep neural networks segment neuronal membranes in electron microscopy images." Advances in neural information processing systems, NIPS 2012

Semantic segmentation

How to evaluate?

- TP_c is the number of true positive for class c (e.g. the number of pixels correctly predicted as c),
- FP_c is the number of false positive for class c (e.g. the number of pixels incorrectly predicted as c),
- TN_c is the number of true negative for class c (e.g. the number of pixels correctly predicted as not c),
- FN_c is the number of false negative for class c (e.g. the number of pixels incorrectly predicted as not c).
- T is the total number of pixels.

- Per-class F1:

$$\mathrm{F1}_c = 2 \times \frac{\mathrm{Precision}_c \times \mathrm{Recall}_c}{\mathrm{Precision}_c + \mathrm{Recall}_c}, \text{ with } \mathrm{Precision}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c} \text{ and } \mathrm{Recall}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}$$

- Average F1:

$$\mathrm{AF1} = \frac{1}{C} \sum_c \mathrm{F1}_c$$

- Overall accuaracy

$$\mathrm{OA} = \frac{\sum_c \mathrm{TP}_c}{\mathrm{T}}$$

Semantic segmentation

How to evaluate?

- TP_c is the number of true positive for class c (e.g. the number of pixels correctly predicted as c),
- FP_c is the number of false positive for class c (e.g. the number of pixels incorrectly predicted as c),
- TN_c is the number of true negative for class c (e.g. the number of pixels correctly predicted as not c),
- FN_c is the number of false negative for class c (e.g. the number of pixels incorrectly predicted as not c).
- T is the total number of pixels.

- Intesection over Union (IoU)

$$\mathrm{IoU}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c+\mathrm{FP}_c+\mathrm{FN}_c}$$

- Mean IoU

$$\frac{1}{C}\sum_c \mathrm{IoU}_c$$

Semantic segmentation

How to evaluate?

- TP_c is the number of true positive for class c (e.g. the number of pixels correctly predicted as c),
- FP_c is the number of false positive for class c (e.g. the number of pixels incorrectly predicted as c),
- TN_c is the number of true negative for class c (e.g. the number of pixels correctly predicted as not c),
- FN_c is the number of false negative for class c (e.g. the number of pixels incorrectly predicted as not c).
- T is the total number of pixels.

  - Matthew's correlation coefficient (between – 1 and 1)

$$\mathrm{MCC} = \frac{\mathrm{TP}_c \times \mathrm{TN}_c - \mathrm{FP}_c \times \mathrm{FN}_c}{\sqrt{(\mathrm{TP}_c + \mathrm{FP}_c)(\mathrm{TP}_c + \mathrm{FN}_c)(\mathrm{TN}_c + \mathrm{FP}_c)(\mathrm{TN}_c + \mathrm{FN}_c)}}$$

Semantic segmentation

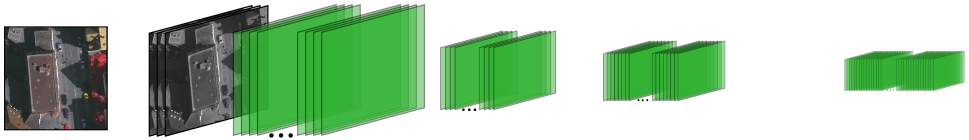
How to evaluate?

- Look at the confusion matrix
- Look at the results

Semantic segmentation

Hypercolumn model

- Observation: classifying each pixel independently might not be optimal…
- Idea 2: Upsample the features to the original resolution and classify based on them

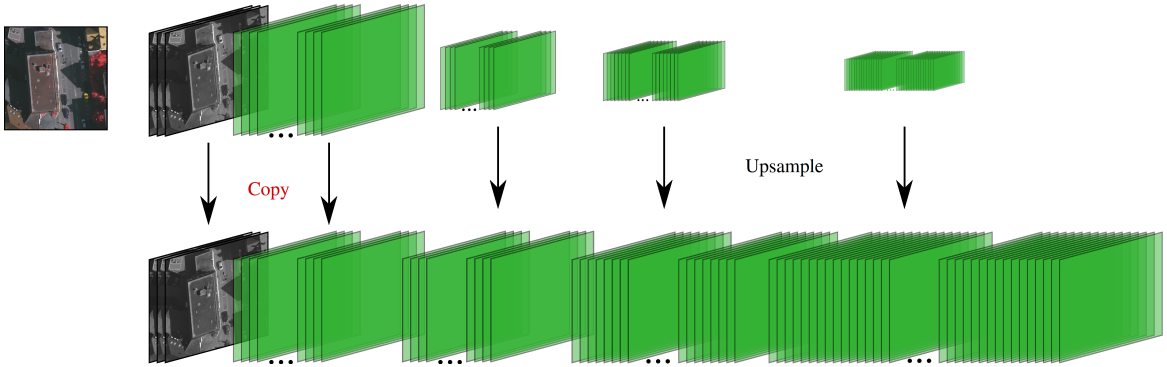


Hariharan B, Arbeláez P, Girshick R, Malik J. Hypercolumns for object segmentation and fine-grained localization. In Proceedings of the IEEE conference on computer vision and pattern recognition 2015 (pp. 447-456).

Semantic segmentation

Hypercolumn model

- Observation: classifying each pixel independently might not be optimal…
- Idea 2: Upsample the features to the original resolution and classify based on them

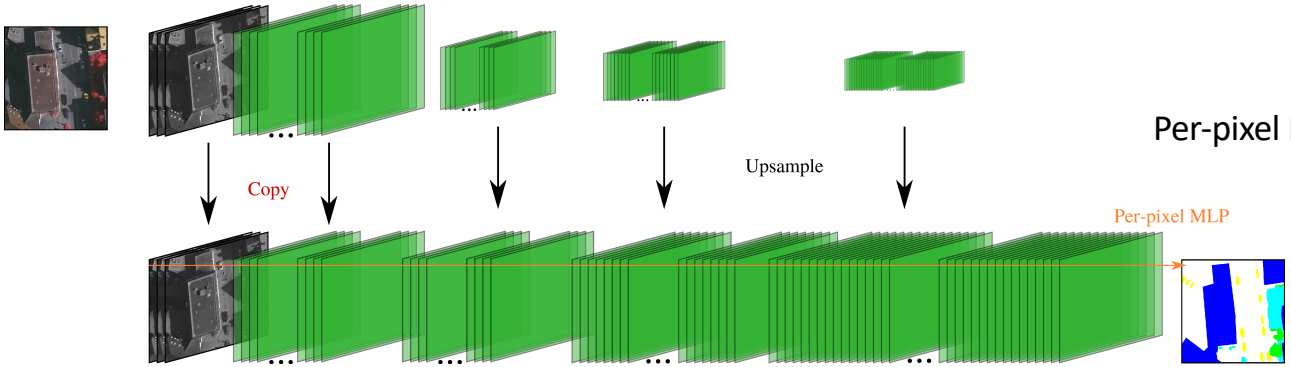


Hariharan B, Arbeláez P, Girshick R, Malik J. Hypercolumns for object segmentation and fine-grained localization. In Proceedings of the IEEE conference on computer vision and pattern recognition 2015 (pp. 447-456).

Semantic segmentation

Hypercolumn model

- Observation: classifying each pixel independently might not be optimal…
- Idea 2: Upsample the features to the original resolution and classify based on them



Per-pixel MLP = 1x1 Conv2D

Hariharan B, Arbeláez P, Girshick R, Malik J. Hypercolumns for object segmentation and fine-grained localization. InProceedings of the IEEE conference on computer vision and pattern recognition 2015 (pp. 447-456).

Semantic segmentation

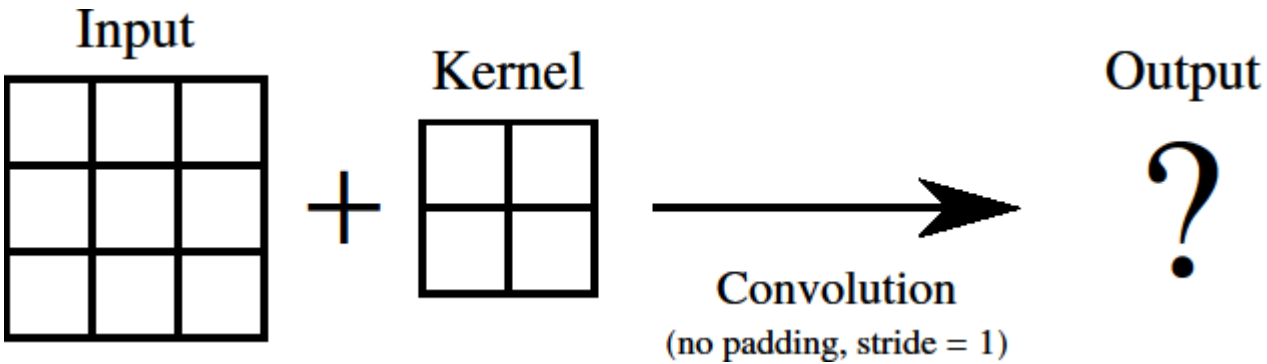
Hypercolumn model

- Observation: classifying each pixel independently might not be optimal…
- Idea 2: Upsample the features to the original resolution and classify based on them
- Better localization, take into account low and high-level features.
- BUT: interpolation is a fixed operation
- Can we learn the interpolation?

Semantic segmentation

A new tool: transposed convolution

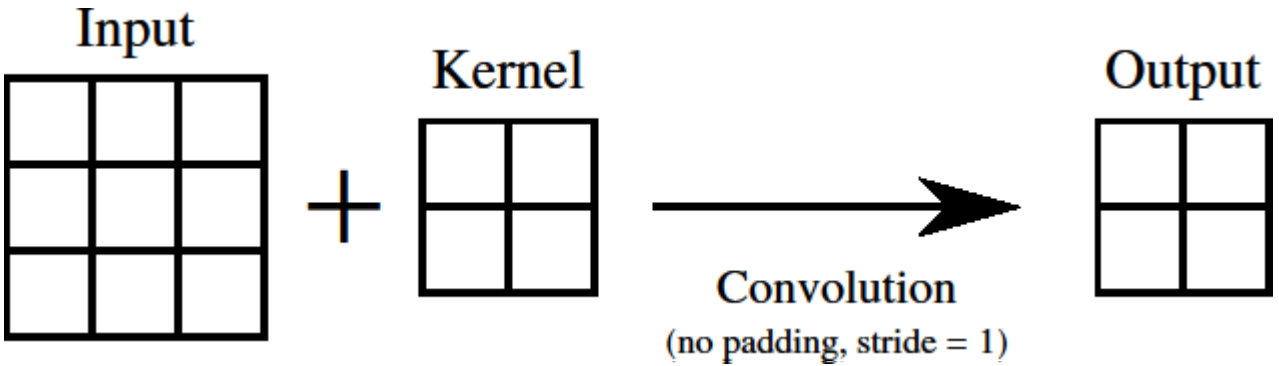
- Transposed convolution = convolution with reverted spatial transformation
- WARNING: it is NOT a deconvolution



Semantic segmentation

A new tool: transposed convolution

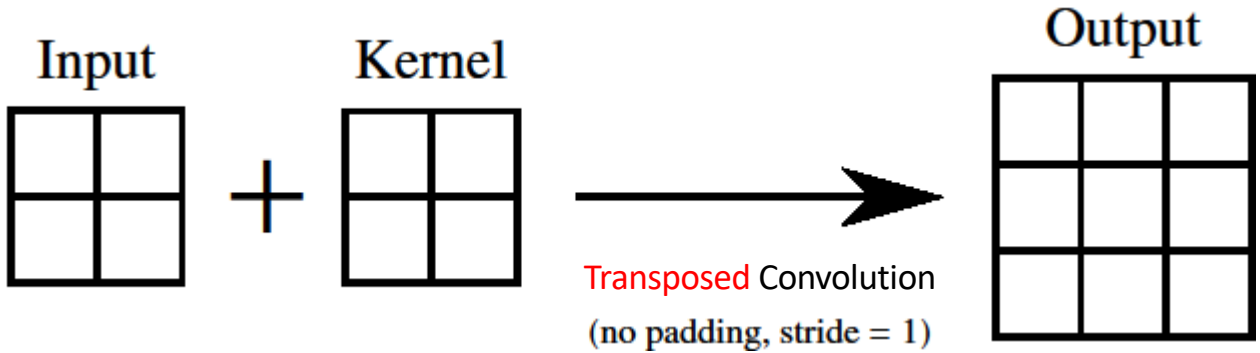
- Transposed convolution = convolution with reverted spatial transformation
- WARNING: it is NOT a deconvolution



Semantic segmentation

A new tool: transposed convolution

- Transposed convolution = convolution with reverted spatial transformation
- WARNING: it is NOT a deconvolution



Semantic segmentation

A new tool: transposed convolution

- Transposed convolution = convolution with reverted spatial transformation
- WARNING: it is NOT a deconvolution

Input

0	1
2	3

Kernel

4	5
6	7

Semantic segmentation

A new tool: transposed convolution

- Transposed convolution = convolution with reverted spatial transformation
- WARNING: it is NOT a deconvolution

Input

0	1
2	3

Kernel

4	5
6	7

0		

Semantic segmentation

A new tool: transposed convolution

- Transposed convolution = convolution with reverted spatial transformation
- WARNING: it is NOT a deconvolution

Input

0	1
2	3

Kernel

4	5
6	7

0	0	
0	0	

Semantic segmentation

A new tool: transposed convolution

- Transposed convolution = convolution with reverted spatial transformation
- WARNING: it is NOT a deconvolution

Input

0	1
2	3

Kernel

4	5
6	7

0	0	
0	0	

+

	4	5
	6	7

Semantic segmentation

A new tool: transposed convolution

- Transposed convolution = convolution with reverted spatial transformation
- WARNING: it is NOT a deconvolution

Input	
0	1
2	3

Kernel	
4	5
6	7

0	0	
0	0	

+

	4	5
	6	7

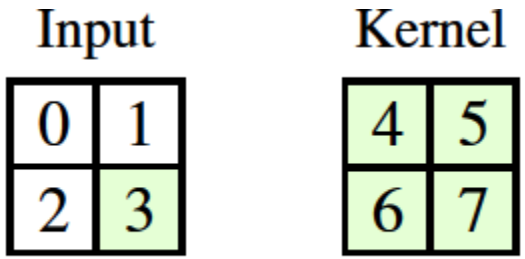
+

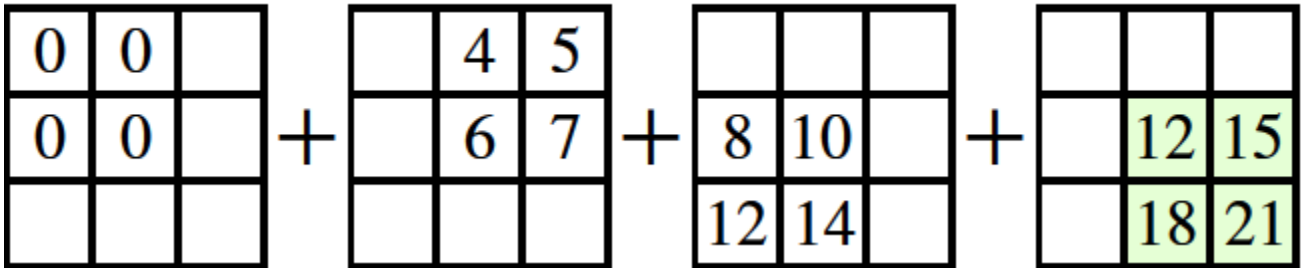
8	10	
12	14	

Semantic segmentation

A new tool: transposed convolution

- Transposed convolution = convolution with reverted spatial transformation
- WARNING: it is NOT a deconvolution





Semantic segmentation

A new tool: transposed convolution

- Transposed convolution = convolution with reverted spatial transformation
- WARNING: it is NOT a deconvolution

Input	
0	1
2	3

Kernel	
4	5
6	7

0	0			4	5												
0	0		+	6	7	+	8	10		+		12	15	=			
							12	14				18	21				

Semantic segmentation

A new tool: transposed convolution

- Transposed convolution = convolution with reverted spatial transformation
- WARNING: it is NOT a deconvolution

Input	
0	1
2	3

Kernel	
4	5
6	7

0	0		+		4	5	+				+				=	0	4	5
0	0				6	7		8	10				12	15		8	28	22
								12	14				18	21		12	32	21

Semantic segmentation

Half a break

- Let’s consider the following problem:

Input

0	1	2
1	2	3
2	3	4

Kernel

0	1
2	3

- What will be the spatial size of the transposed convolution (stride 1, padding 0)?
- What will be the result of the transposed convolution?

Semantic segmentation

Half a break

- Let’s check on PyTorch:

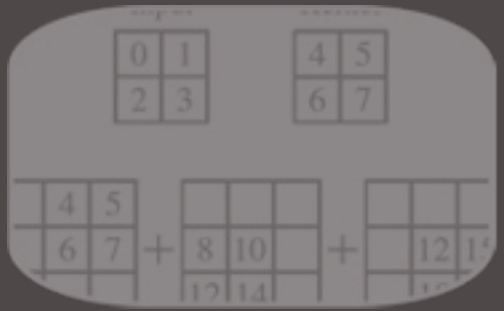
```
>>> m = torch.nn.ConvTranspose2d(1, 1, 2)
>>> m.weight = torch.nn.Parameter(torch.tensor([[[[0, 1], [2, 3]]]], dtype=torch.float32))
>>> input = torch.reshape(torch.tensor([[0, 1, 2], [1, 2, 3], [2, 3, 4]], dtype=torch.float32), (1, 1, 3, 3))
>>> m(input)
tensor([[[[-0.3173, -0.3173,  0.6827,  1.6827],
          [-0.3173,  2.6827,  8.6827,  8.6827],
          [ 1.6827,  8.6827, 14.6827, 12.6827],
          [ 3.6827, 11.6827, 16.6827, 11.6827]]]],
       grad_fn=<SlowConvTranspose2DBackward>)
```

Menu of the day

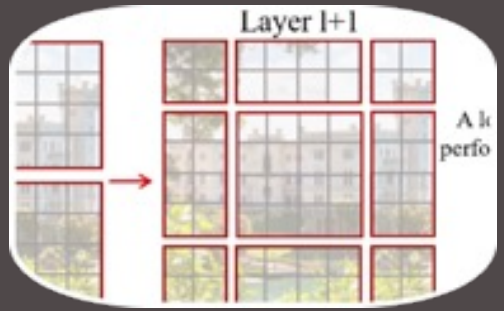
Intro to semantic segmentation

Remote sensing applications

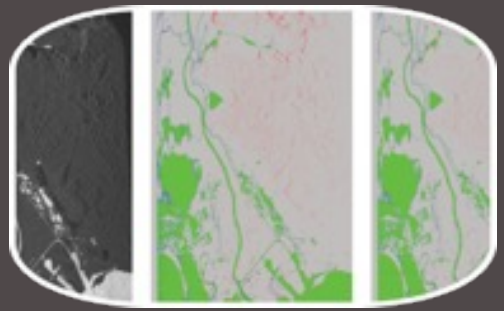
First models



Transposed convolution



Modern architectures

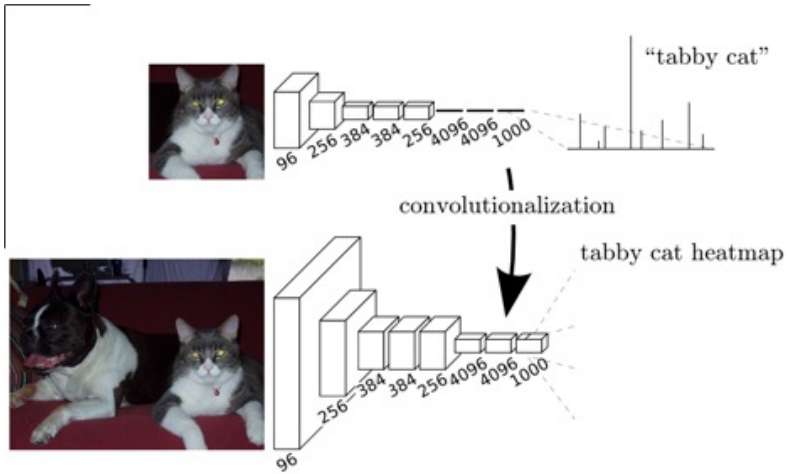


Besides deep learning

Semantic segmentation

Fully Convolutional Network (FCN)

- Classical classification architecture: fixed size input -> vector of class scores
- This can be seen as a convolution itself to obtain a heatmap of classes.
- Recover the initial spatial size with a transposed convolution

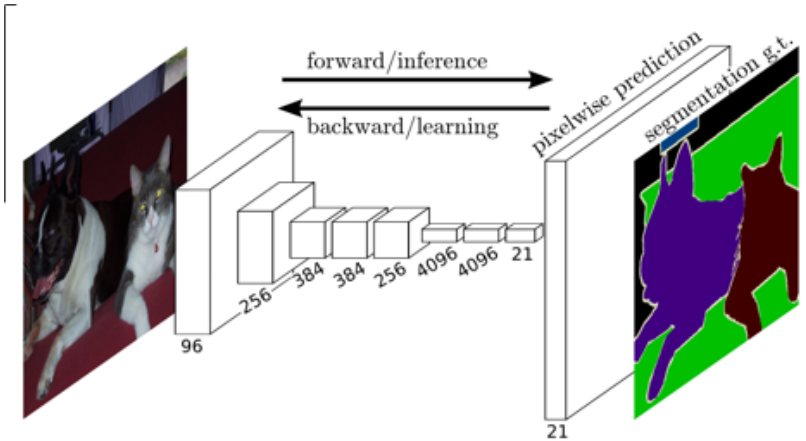


Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

Semantic segmentation

Fully Convolutional Network (FCN)

- Classical classification architecture: fixed size input -> vector of class scores
- This can be seen as a convolution itself to obtain a heatmap of classes.
- Recover the initial spatial size with a transposed convolution

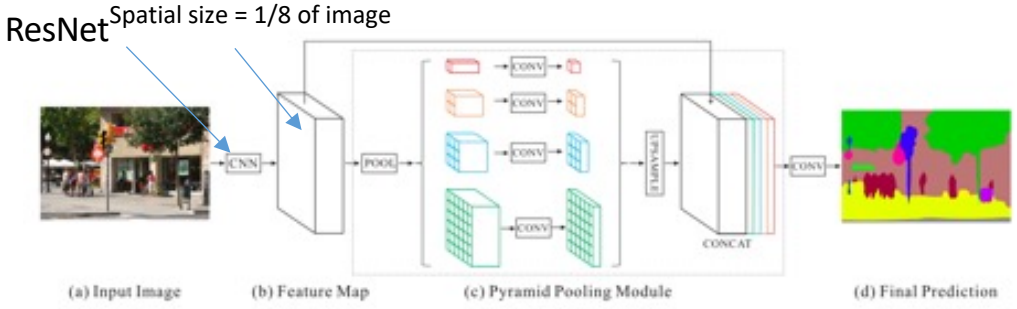


Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

Semantic segmentation

Pyramid Scene Parsing Network (PSP)

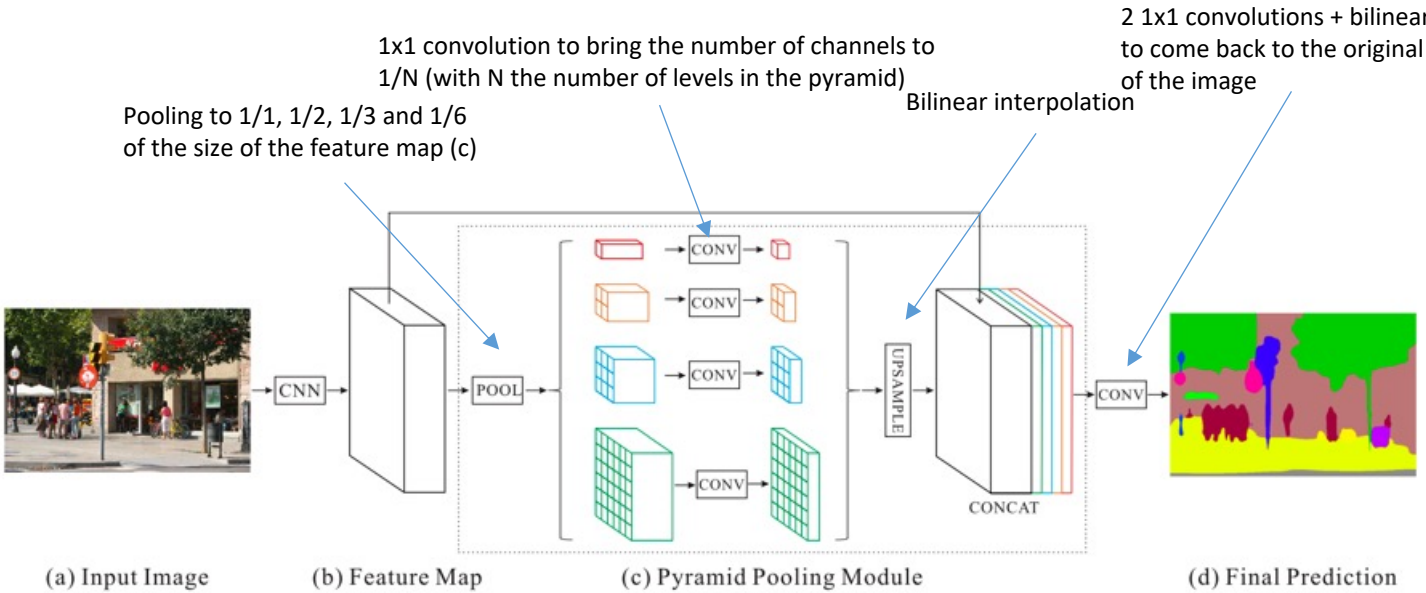
- Built for scene parsing (≃ fine-grained, hierarchical semantic segmentation + image-level scene descriptor)
- Observation: FCN lacks context for complex scenes.
- Objective: Extracting information at different scales



Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2881-2890).

Semantic segmentation

Pyramid Scene Parsing Network (PSP)



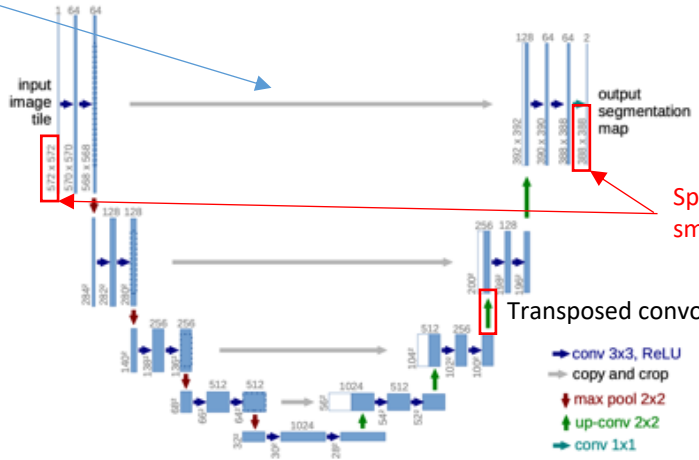
Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2881-2890).

Semantic segmentation

U-Net

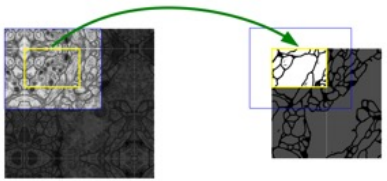
- Main idea: several layers in the expansive part of the network -> expansive part is the mirror of the contracting part

Valid part of the contracting path
is concatenated to the expansive path



Spatial size of the output is much smaller than the input's !!!

Transposed convolution



Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.

Semantic segmentation

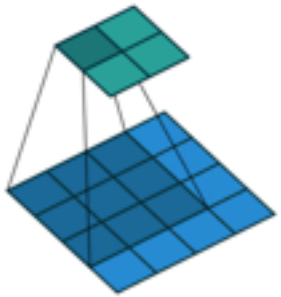
DeepLabv3

- Conv/pool:
  - + increase the receptive field
  - - loss of the spatial resolution
- Transposed convolutions to recover spatial resolution

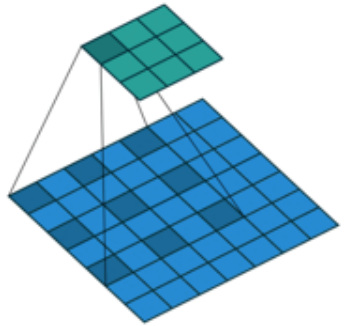
Semantic segmentation

DeepLabv3

- ~~Transposed convolutions to recover spatial resolution~~
- Do not lose the spatial resolution, use à trous convolution to increase the receptive field



Classical convolution



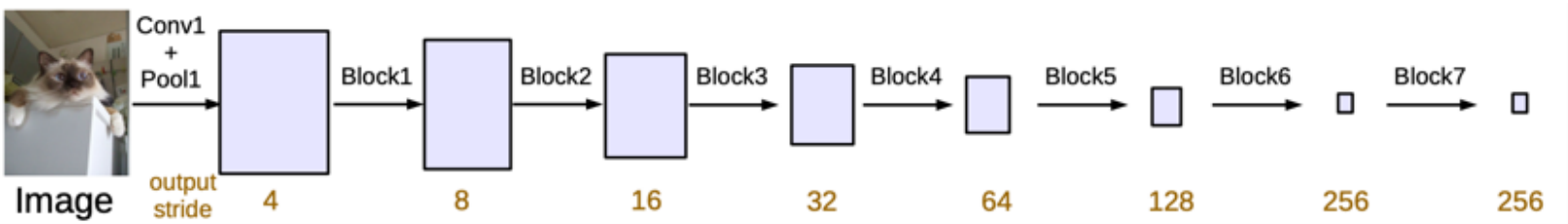
A trous convolution
(rate = 2)

Blue = input, cyan = output. Source https://github.com/vdumoulin/conv_arithmetic

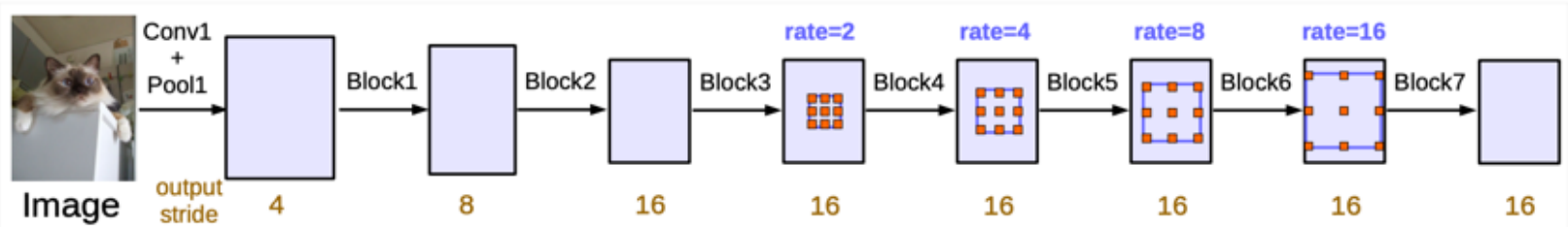
Semantic segmentation

DeepLabv3

- Use à trous convolution to increase the receptive field



Classical CNN, increase of receptive field and loss of spatial resolution

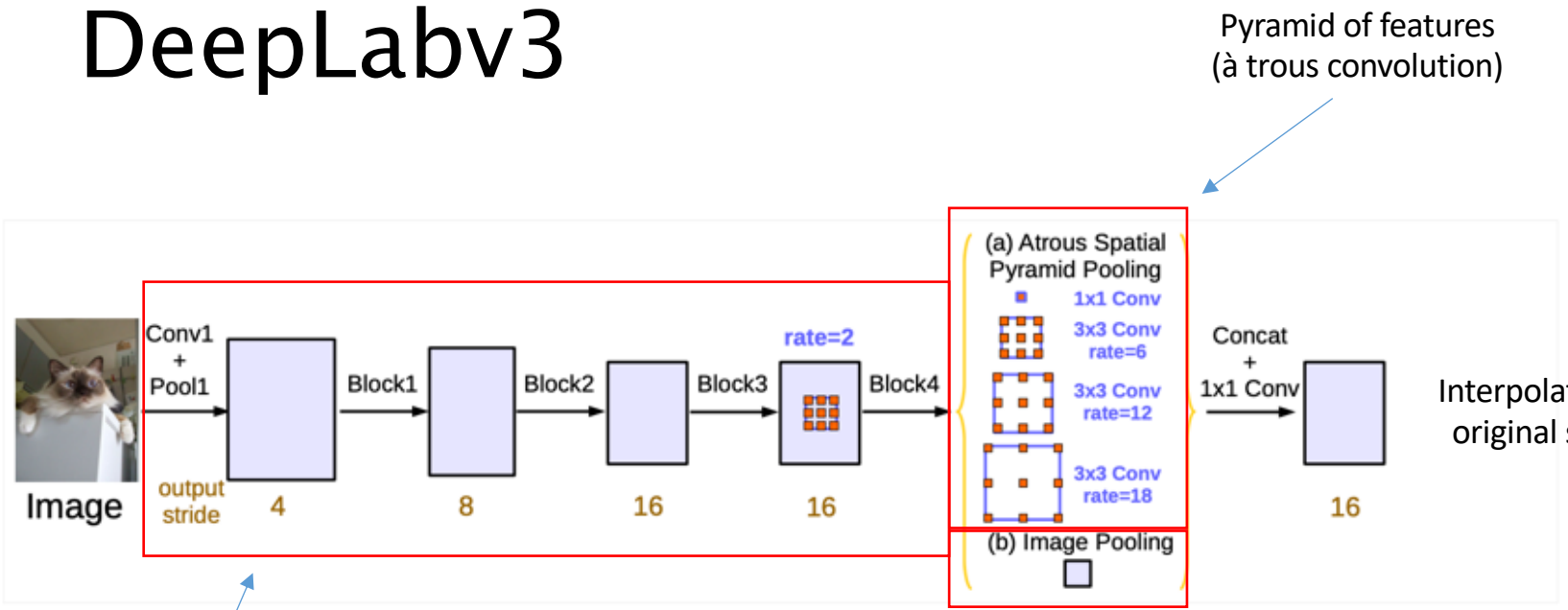


CNN with à trous convolution, increase of receptive field **without** loss of spatial resolution

Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Semantic segmentation

DeepLabv3

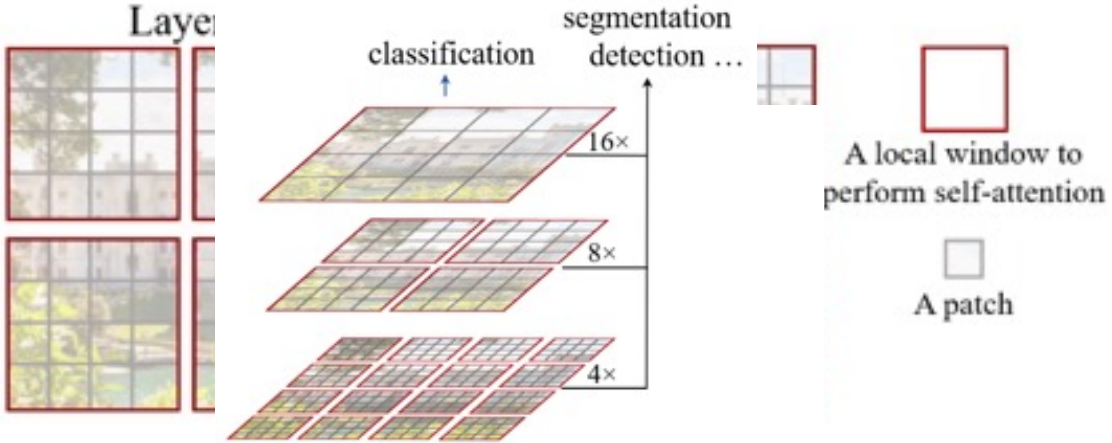


Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Semantic segmentation

Swin-Unet

- Swin transformers proposed as an architecture to use Transformers in vision
- Main idea: compute the self attention in shifted windows in-between layers
- Can be used as a general-purpose backbone… which people did

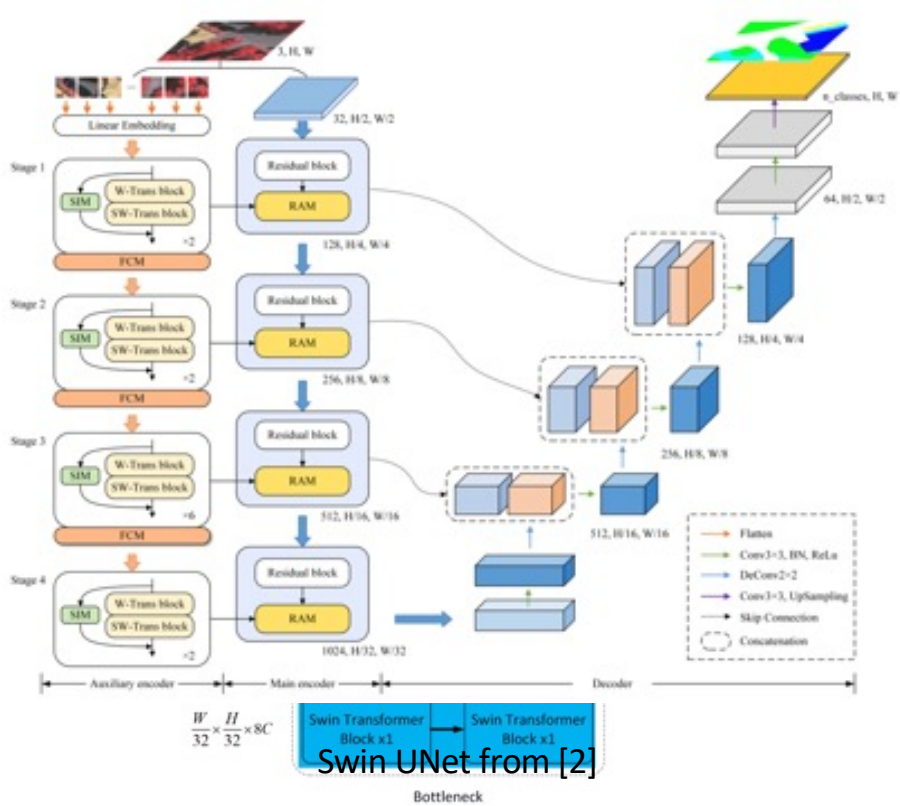


Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10012-10022).

Semantic segmentation

Swin-Unet

- In medical imagery [1] with a Swin expanding path
- In remote sensing [2] with a convolutional expanding path



Swin UNet from [2]

Swin UNet from [1]

[1] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2021). Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*.
[2] He, X., Zhou, Y., Zhao, J., Zhang, D., Yao, R., & Xue, Y. (2022). Swin Transformer Embedding UNet for Remote Sensing Image Semantic Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1-15.

Besides deep learning

Other methods

- Deep learning might not be the ideal tool if:
  - You do not have a lot of data
  - You do not have a lot of labeled data
  - You have prior knowledge on the classes

Besides deep learning

K-Means

- Iterative algorithm to partition the image in K clusters:
  1. Pick K cluster centers (e.g. randomly) and assign all of the pixels to the closest cluster
  2. Re-compute the cluster centers
  3. Repeat until convergence

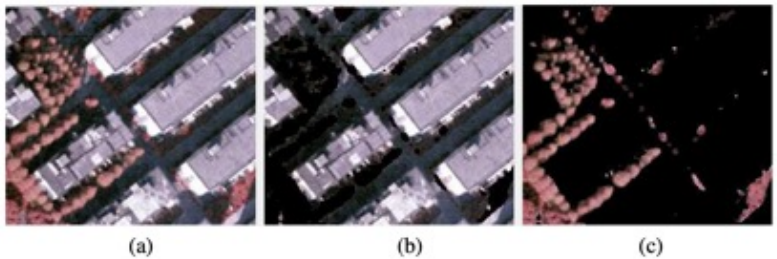


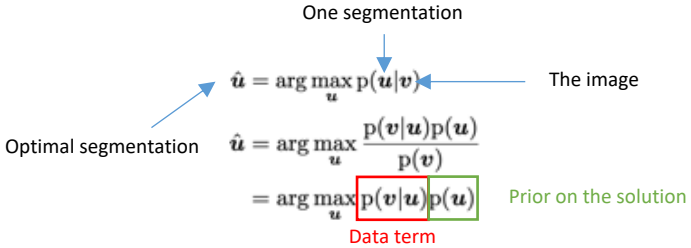
Fig. 1. Results of sequential K-Means, where K is two. (a) is the original image. (b) is one of its clusters and (c) is another cluster.

Lv, Zhenhua, et al. "Parallel k-means clustering of remote sensing images based on mapreduce." *International Conference on Web Information Systems and Mining*. Springer, Berlin, Heidelberg, 2010.

Besides deep learning

Markov Random Fields

- Maximum A Posteriori framework:



- MRF: prior for a given pixel only depends on the state of its neighbors
- Can be optimized using Iterative Conditional Modes, Simulated Annealing, or graph cuts

Besides deep learning

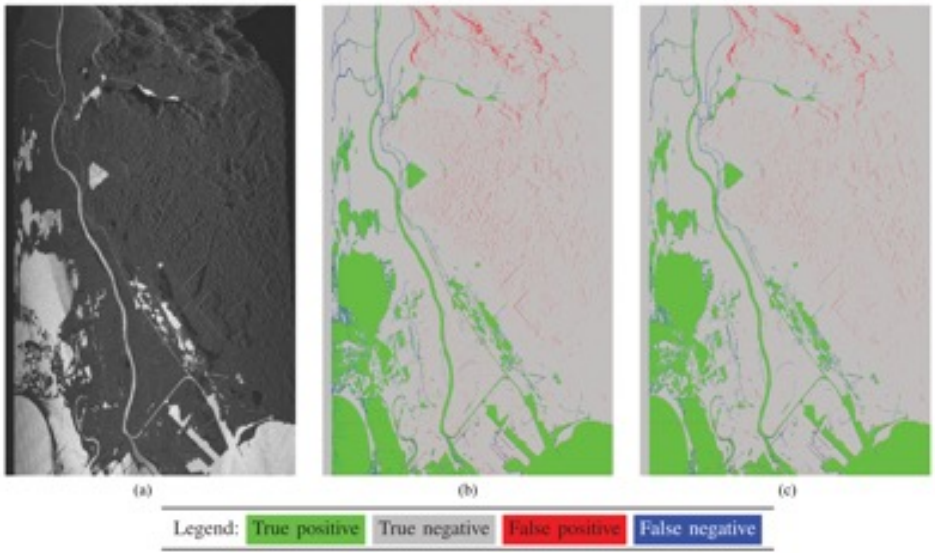
Markov Random Fields

- Application to SAR

The image should be smooth

$$\arg\max_{\boldsymbol{u}} \mathrm{p}(\boldsymbol{v}|\boldsymbol{u})\mathrm{p}(\boldsymbol{u})$$

Data term: known statistics of the signal

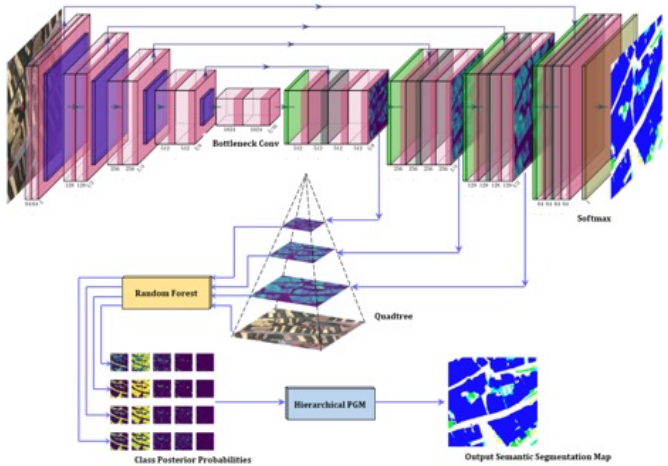


Lobry, S., Denis, L., Williams, B., Fjørtoft, R., & Tupin, F. (2019). Water detection in SWOT HR images based on multiple Markov random fields. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(11), 4315-4326.

Besides deep learning

Markov Random Fields

- Bonus: MRF can be used to improve your results from a deep learning model



Pastorino, Martina, et al. "Semantic Segmentation of Remote-Sensing Images Through Fully Convolutional Neural Networks and Hierarchical Probabilistic Graphical Models." *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022): 1-16.

Semantic segmentation

Wrap-up

- First thing to do: look at the data!
- Transposed convolution used in most modern semantic segmentation architectures
- Ensemble models are efficient

- Questions?

Université de Paris

Let’s code!

Notebook on moodle

Notes:

- You do not need a colab pro account
- Try to understand the code, not just press play!!