

MO434 - Deep Learning

Applications in Image Analysis - Part II

Alexandre Xavier Falcão

Institute of Computing - UNICAMP

afalcao@ic.unicamp.br

- Fully Convolutional Neural Networks (FCNNs).
- Semantic segmentation with U-Net.
- Strategies for instance segmentation.

Fully Convolutional Neural Networks (FCNN)

- So far we have learned that convolutional layers (backbone, **encoder**) extract suitable image features while predictions (classification or regression) are done by dense layers.

Fully Convolutional Neural Networks (FCNN)

- So far we have learned that convolutional layers (backbone, **encoder**) extract suitable image features while predictions (classification or regression) are done by dense layers.
- The strategy is suitable for image classification and object detection. For segmentation, however, we have to classify pixels as belonging or not to each object of interest.

Fully Convolutional Neural Networks (FCNN)

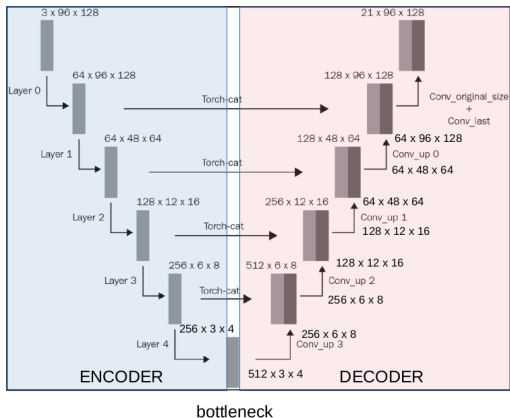
- So far we have learned that convolutional layers (backbone, **encoder**) extract suitable image features while predictions (classification or regression) are done by dense layers.
- The strategy is suitable for image classification and object detection. For segmentation, however, we have to classify pixels as belonging or not to each object of interest.
- You may wonder to train one classifier (dense layers) per pixel by using its values in the feature map, but well succeeded approaches substitute such predictors per pixel by a **decoder**.

Fully Convolutional Neural Networks (FCNN)

- So far we have learned that convolutional layers (backbone, **encoder**) extract suitable image features while predictions (classification or regression) are done by dense layers.
- The strategy is suitable for image classification and object detection. For segmentation, however, we have to classify pixels as belonging or not to each object of interest.
- You may wonder to train one classifier (dense layers) per pixel by using its values in the feature map, but well succeeded approaches substitute such predictors per pixel by a **decoder**.
- While the encoder reduces image size, the decoder must retrieve the original spatial dimension with **one output channel per class**.

Fully Convolutional Neural Networks (FCNN)

A decoder is also a sequence of convolutional layers, except that **up-sampling** is adopted to retrieve spatial dimension.



U-Net is one of the most well succeeded and popular FCNNs (<https://paperswithcode.com/method/u-net>).

Semantic segmentation with U-Net

At each step of a decoder, we have the following operations.

- Up-sampling as implemented by transposed convolution followed by activation.
- To retrieve precision at object borders, the output of up-sampling must be concatenated with the result of a corresponding layer at the encoder.
- The result of that concatenation is then processed by a convolutional layer to reduce the number of channels.



Transpose Convolution 2D

The transpose convolution between an image with (N_x, N_y) pixels and a kernel of size (K_x, K_y) , padding (P_x, P_y) and strides (S_x, S_y) , is an image with (O_x, O_y) pixels, where

$$O_* = (N_* - 1)S_* + K_* - 2P_*$$

Let's play with it in [▶ \(TRANPOSE CONVOLUTION\)](#) and understand its arithmetic in the next slide.

Transpose Convolution 2D

Input 2x2

x1	x2
x3	x4

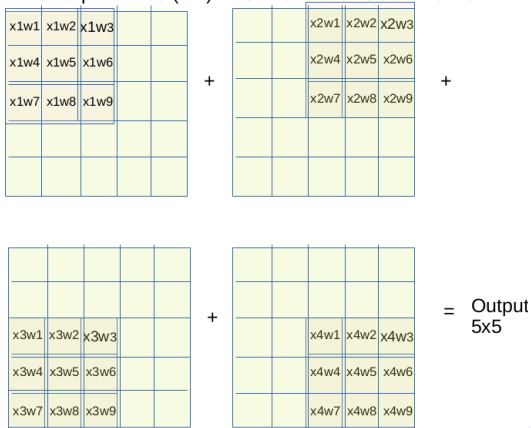
Kernel 3x3

w1	w2	w3
w4	w5	w6
w7	w8	w9

Padding = (0,0)

Strides = (2,2)

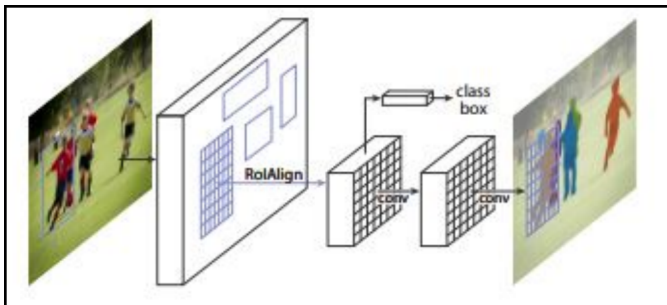
The output will be $(2-1) \times 2 + 3 - 2 \times 0$ in each side $\Rightarrow 5 \times 5$



Now, let's play with U-Net [▶ \(SEMANTIC SEGMENTATION\)](#)

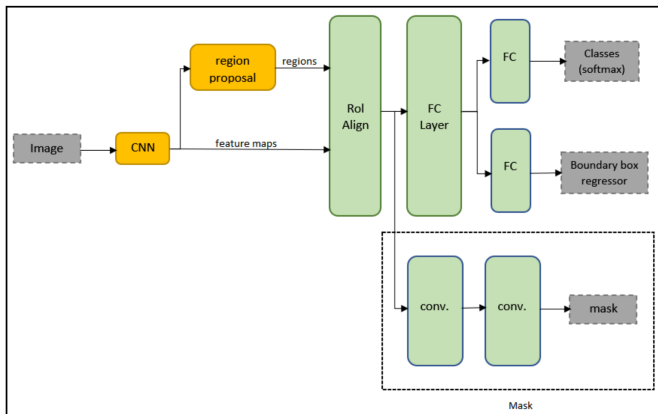
Strategies for instance segmentation

Instance segmentation combines object detection with object delineation in each detected region.



Mask R-CNN is an example that leverages Faster R-CNN and FCNN for instance segmentation (<https://paperswithcode.com/method/mask-r-cnn>).

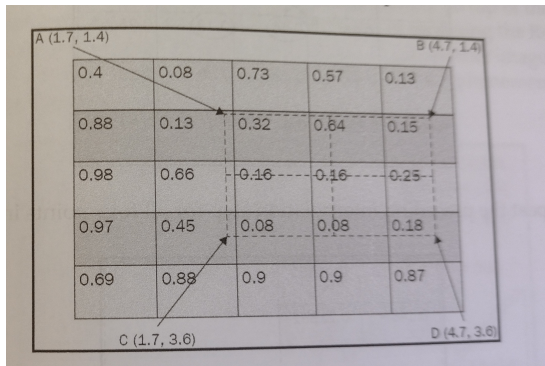
Strategies for instance segmentation



Mask R-CNN substitutes ROI pooling by **ROI align** in Faster R-CNN and adds a **decoder** (head) to estimate the mask of the detected object.

Strategies for instance segmentation

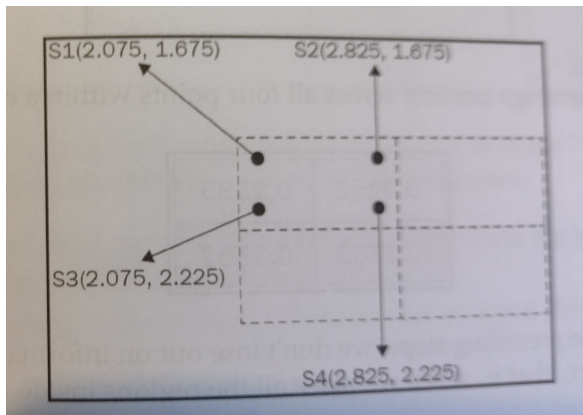
Different from ROI pooling, ROI align uses a same number of representation points per region.



Suppose the 3x3 dashed region above has to be transformed into a 2x2 region. It is first equally divided into 2x2 subregions.

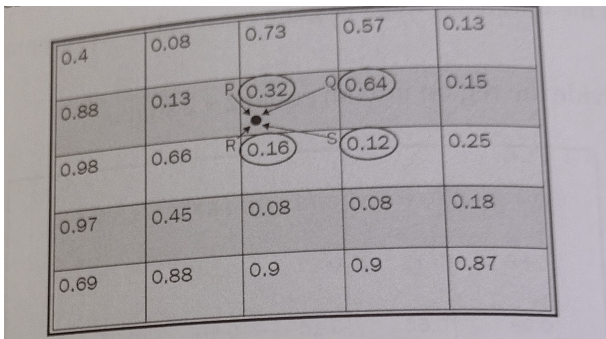
Strategies for instance segmentation

For each subregion (cell), four equally spaced points are defined and the distances between each point and its four nearest pixels are used as weights to estimate the feature value at that point.



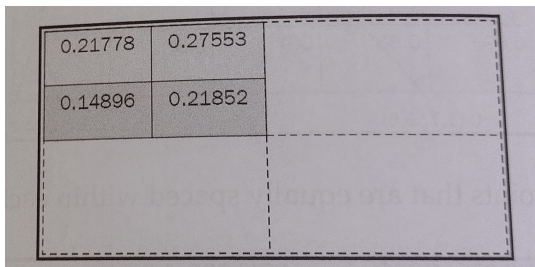
Strategies for instance segmentation

For each subregion (cell), four equally spaced points are defined and the distances between each point and its four nearest pixels are used as weights to estimate the feature value at that point.



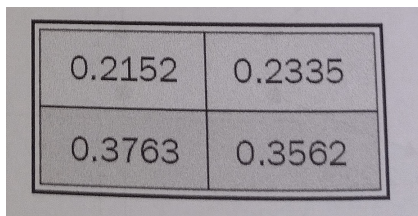
Strategies for instance segmentation

For each subregion (cell), four equally spaced points are defined and the distances between each point and its four nearest pixels are used as weights to estimate the feature value at that point.



Strategies for instance segmentation

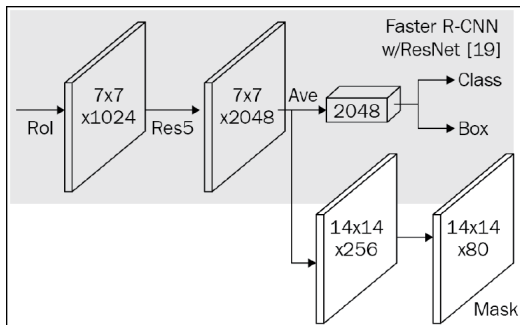
After estimating values to the four points in each of the 2×2 cells, the 2×2 region is obtained by average pooling among the points of each cell.



0.2152	0.2335
0.3763	0.3562

Strategies for instance segmentation

Using ResNet50 as backbone, the decoder may convert 7×7 regions with 2048 channels into 14×14 masks of 80 channels when you have 80 classes.



Conclusion

- At this point, you should be able to design CNNs and FCNNs for image classification, object detection, semantic and instance segmentation.
- Solutions for several image analysis applications are also available at <https://paperswithcode.com/>.
- Next, we will understand the fundamentals for text analysis using deep learning.