

Segmentation techniques

Önálló projekt III., 2021/22 1. félév

Bakos Bence

Témavezető: Lukács András





Eötvös Loránd Tudományegyetem

Budapest, 2021

Input images: $\{X_i\}_{i=1}^N \subset \mathbb{R}^{H \times W \times Ch}$.

- Classification
 - Annotation: $L_i \in \{1, 2, \dots, C\} = [C]$
 - C: number of classes
- Object detection
 - Annotation: $B_i \in \mathbb{R}^4 \times [C]$
 - 4 real numbers which define the bounding box + class label
- Semantic segmentation
 - Annotation: $M_i \in \{0, 1\}^{H \times W \times C}$
 - Pixel-level class labels
- Instance segmentation
 - Annotation: $I_i \in \{0, 1, \dots, k\}^{H \times W \times C}$
 - k is the maximum number of instances

Computer vision tasks

| Semantic Segmentation | Classification + Localization | Object Detection | Instance Segmentation |
|---|---|---|--|
|  |  |  |  |
| GRASS, CAT, TREE, SKY | CAT | DOG, DOG, CAT | DOG, DOG, CAT |
| No objects, just pixels | Single Object | Multiple Object | |

This image is CC0 public domain

Setup:

- Image dataset of N element: $\{X_i\}_{i=1}^N \subset \mathbb{R}^{H \times W \times C_h}$ and its annotations (masks): $\{M_i\}_{i=1}^N \subset \{0, 1\}^{H \times W \times C}$
- We want a model f s.t. $f(X_i) \sim M_i$.
- One-class case: $C = 1$ (multi-class case can be viewed as C times the one-class case)
- Suppose, that the output of our model is $f(X_i) = P_i \in \mathbb{R}^{H \times W}$, and $P_i^{h,w} \in [0, 1]$.
- Output: probability of the pixel belonging to the class (achieved by sigmoid layer at the end of the model)

- Cross-entropy loss function for segmentation:

$$\text{BCELoss}(P, M) = \frac{1}{H \cdot W} \sum_{i=(h,w)} \alpha M_i \log P_i + (1 - M_i) \log(1 - P_i)$$

With α we can address class imbalance.

- Focal-loss:

$$\begin{aligned} \text{BinaryFocalLoss}(P, M) = \frac{1}{H \cdot W} \sum_{i=(h,w)} \alpha (1 - P_i)^\gamma M_{h,w} \log P_i + \\ + P_i^\gamma (1 - M_i) \log(1 - P_i). \end{aligned}$$

Can magnify the signal of hardly classifiable pixels in the the loss.

- Classification metrics using the elements of the confusion matrix after thresholding the output of the model: Jaccard-score, Dice-score. This works as a metric for segmentation, but not as a loss-function
- Generalization of the elements of the confusion matrix:

$$TP = M \circ P, \quad FP = (J - M) \circ P, \quad FN = M \circ (J - P), \quad TN = (J - M) \circ (J - P),$$

$J \in \mathbb{R}^{H \times W}$ is the all-one matrix, a \circ is the Hadamard product.

- Tversky loss:

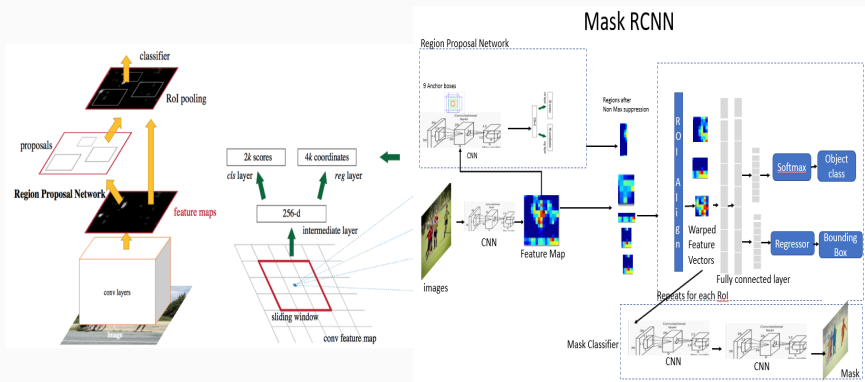
$$\text{TverskyLoss} = \frac{TP + \varepsilon}{TP + \alpha FN + \beta FP + \varepsilon}.$$

The special cases of the Tversky loss with the right α and β are i.e. Dice-loss and IoU-loss.

Segmentation models

Mask R-CNN:

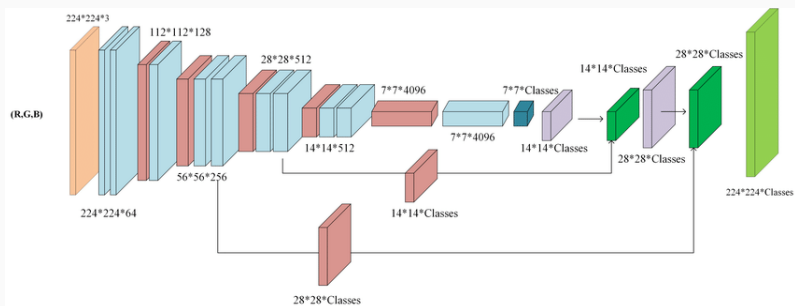
- Based on R-CNN, Fast R-CNN and Faster R-CNN models
- Region proposal block (increasing significance of conv. layers)
- Mask R-CNN: Faster R-CNN + RoI Align + Mask classifier



FCN

- Standard classification models (VGG, ResNet):
 - Some convolutional blocks + dense layers
 - In the i -th block the image has shape $H_i \times W_i \times Ch_i$, Ch_i is mon. increasing, H_i and W_i mon. decreasing
- For segmentation: Keep H_i and W_i fix and drop dense layers
- FCN:
 - encoder with decreasing spatial dimensions
 - decoder with increasing spatial dimensions (transpose convolutions/upsampling)
- To keep spatial information we add the outputs of some blocks in the encoder to the corresponding phase of the decoder

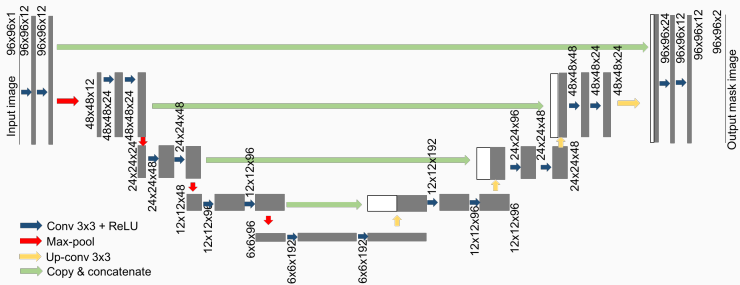
FCN-8 model



Segmentation models

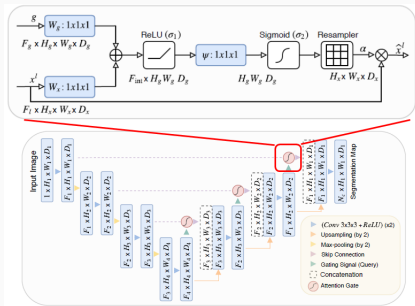
U-Net

- Encoder-decoder structure of convolutional blocks
- Losing precise spatial information during the encoder, but gaining higher level representation about the content
- Skip-connections:
 - Maintain spatial information in decoder from the encoder branch.
 - Concatenation (in channel dimension) instead of addition
- Especially popular in medical image segmentation

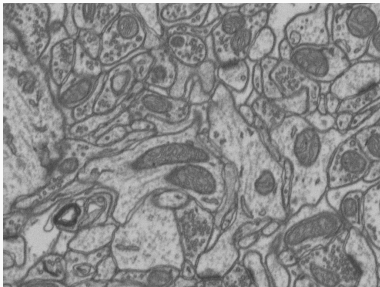


Attention U-Net

- Attention: selectively concentrating on some things, while ignoring others
- Attention gates in U-Net:
 - Attention mechanism in the decoder branch
 - 2 inputs: features from lower levels of the model and data through skip connection
 - Creates a weighting of the pixels of the image coming from the skip connection



- Tested U-Net and Attention U-Net with BCE and Focal loss
- Electron microscopy images from the hippocampus
- More than 90% accuracy with each method (easy dataset)



Other U-Net variants

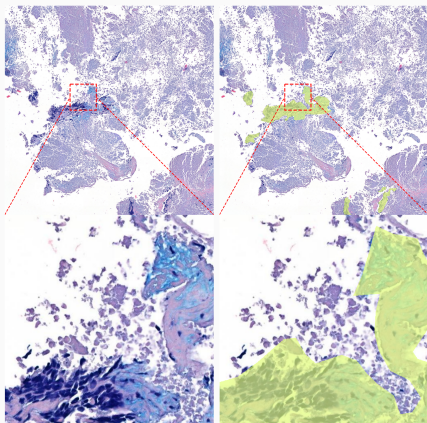
- Changing convolutional blocks: Residual U-Net, R2-UNet
- Combining modifications: Attention Res-UNet

Transformers

- First in NLP, great success with multihead self-attention blocks
- Vision Transformer (ViT): successful adaptation for image classification
- New results on transformer architectures for segmentation: DeTr, SeTr, Trans-UNet

Application

- H&E stained histopathology images of lung tissue
- 4096x4096 images divided into 512x512 patches
- Four different cancer type (ADC, SqCC, SCLC, SKMU), but only enough data for lung squamous cell carcinoma (SqCC) so far
- First experiments: Standard U-Net and weighted BCE for finding the optimal balancing method



Further plans

- Since changing the α factor in the loss function didn't improve the performance enough, we implement other class balancing methods like under- and oversampling
- Continuing the theoretical and experimental research in form of a Master's thesis
- Planning to try more advanced models (especially transformer based methods) and loss functions
- Possibly find out new techniques for datasets like this, and test them on benchmark datasets as well

Thank you for the attention!