Exercise 8: Segmentation and Object Detection

Arnas Šniokaitis

February 2, 2026

1 Computational & comprehensive questions

- Explain the primary difference between semantic segmentation and instance segmentation in image processing.

  In semantic segmentation all objects of the same class are not differentiated. As in if we have an image with two stop signs. They will both be marked as a stop sign where as with instance segmentation they would be treated as separate objects/instances of the same class.

- Briefly describe how a Region-Based Convolutional Neural Network (R-CNN) for object detection works.

  An R-CNN creates multiple region proposals where an object may be, then each object is classified independently and after classifications are made the bounding boxes are adjusted to better encompass the objects in question.

- Explain the role of anchor boxes for object detection and instance segmentation.

  Anchor boxes are simmilar to bounding boxes. Anchor boxes are usually fixed in size and, simmilarly to convolutions, overlap with each other. Their goal is to act as a regional detection from a feature map to an output. After the likelyhood of an object being in the anchor boxes is calculated, those with a high prediction act as a sort of guide for other algorithms where to check for an object; hence its use in region proposal.

- Explain why single stage architectures are typically faster compared to two-stage approaches. Try to be as accurate as possible in explaining which computations are "saved".

  Two stage models typically are slower than single stage architectures because a lot of the calculations need to be redone.(for example if some proposals overlap). Single stage models save the region proposal calculation step and they do not have to redo calculations per proposals.

2 Working on an object detection task: Mitosis detection

- Discuss why approaching this task with an object detection is feasible and may pose benefits compared to a regression approach (that directly

regresses the number of mitotic figures per image), a segmentation, or an instance segmentation approach.

The main upside of having a segmentation model is that with it we not only gain the information (by counting the bounding boxes) but we would also have the added benefit of being able to see where mitosis occoured.

- Discuss whether you would suggest to use a single-stage or a two-stage approach for this task.

  I would use a two-stage architecture, because; even though it would be slower, it would allow use some more higher accuracy and, since I assume mitosis relatively speaking is a quick proccess and doesnt occour for a while during normal cell life, it would also help by having localization since our cells may not match nicely on a grid.

- For many architectures, it is required to define anchor boxes for possible detections. Describe how you would select the anchor boxes and whether you would use rather few or rather many different anchor box sizes/shapes for this task.

  Since all cells of the same type are roughly the same size, to me it seems redundant to have multiple anchor box sizes with the exception if we are training the model to recognize mitosis in multiple different type of cells which are different size.

- In the lecture, we have briefly talked about non-maximum suppression (see also Fig. 2), which combines multiple overlapping predictions into one prediction. Explain why non- maximum suppression is important for the downstream task and what risks there could be when using a very aggressive / too little non-maximum suppression.

  One important reason is that it prevents the case, that a single object is classified twice. If the suppresion is very aggressive we have the risk of not detecting the object at all, if for instance, a part of is covered and losing detail in the bounding boxes. If the suppression is too weak, then we again have the risk of classifying the same object twice.

3 Implementing biomedical segmentation with U-Net

- We have a binary segmentation task: What are the two classes in this exercise?

  By looking at the data as well as seeing some examples from the bagls.org website, the two classes are: 1) Is the gap between the vocal chords; 2) Is not the gap between the vocal chords (Is it the glottis or is it not the glottis).

- In the lecture, we discussed that segmentation can be understood as a pixel-wise segmentation task - we can therefore use (binary) cross entropy as a loss function. However, this can cause sub-par segmentation results, with the glottis not well segmented. Describe what could be a reason for this and how this is tackled in the code.

One reason may be due to the fact that the opening of the vocal chords is relatively small, as such there is a ‘bias‘ to ‘prefer‘ to not choose it. The way that it is mitigated via code is by having a weighted loss function

- In the data loader, you can see that we use two data augmentation transforms: horizontal flipping and cropping to predefined patches. In addition to increase variability in the training, we need the cropping also for efficient training - discuss briefly what benefit this approach has.

  Firstly all images are of the same size which has an advantage that our model becomes invariant of the input image size. It also creates the benefit, that the images don't take up that much space which allows computations to be completed faster ant take up less memory space.

- Report your results (curves in Tensorboard, illustrative images) and briefly discuss main insights.

  NOTE: due to the time it takes to run the training, I have decreaced the training to only two epochs. Also due to the fact, if I tried to increase the layer count Linux would terminate the proccess due to free memory deficit I will analize the decrease of the parameters.

4 Autoencoders

These are the graphs for changing the bottleneck for the AE.

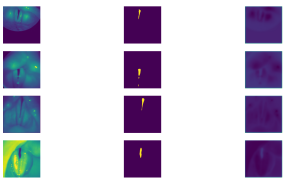

Figure 1: Detection with 2 layers and 16 filters

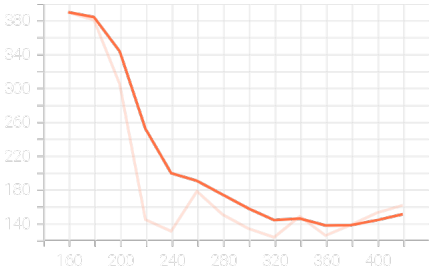

Figure 2: Error with 2 layers and 16 filters

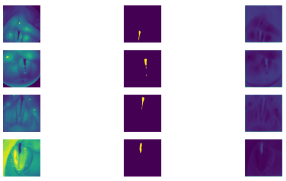

Figure 3: Detection with 2 layers and 32 filters

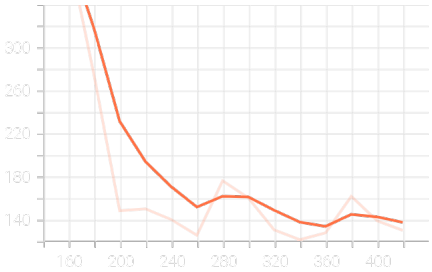

Figure 4: Error with 2 layers and 32 filters

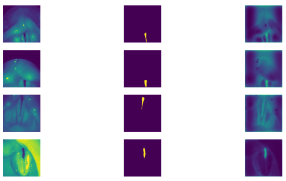

Figure 5: Detection with 5 layers and 16 filters

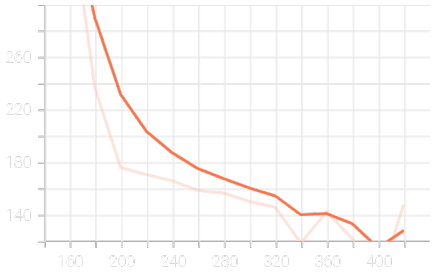

Figure 6: Error with 5 layers and 16 filters

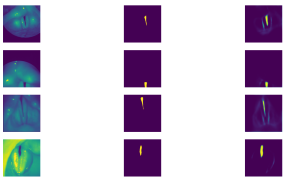

Figure 7: Detection with 5 layers and 32 filters

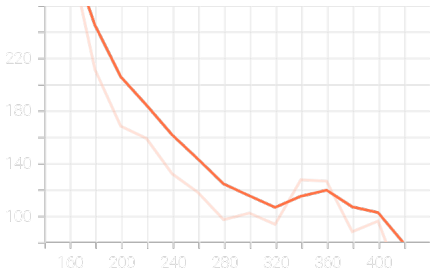

Figure 8: Error with 5 layers and 32 filters

The model with two layers and 16 filters is too weak to produce results. Not only do the rezults of the segmentation yeald no results, the error appears to be increasing. A simmilar effect is happening with 32 layers, although the images have a higher contrast, only the last image makes an accurate guess.

The architecture with 5 layers shows drastically better results. With 16 filters the last image seems to have, although weakly, predicted the data abd in all of the images the glottis is the brightest part or tied. The model with 5 layers and 32 filters produces the best results so far and makes accurate predictions except with the first and third images, where the true vocal folds seem to be missinterpreted as the glottis. However, this data alone is not enough as it is possible that the model with 5 layers and 16 filters could have achieved better results if it were allowed to train further.