Computer Vision Exercise 10

1.

a. In the beginning of the publication the authors argue about the relevance and novelty of their work. Summarise their main arguments and the evidence they present to support their arguments.

The paper emphasizes a deep neural network method called SSD (Single Shot MultiBox Detector) which is used for object detection. They argue it is relevant, as it is easy to train and cost-effective, works faster while providing an unified framework for both training and inference. The authors compare it to Faster R-CNN, and YOLO, which is generally more computationally expensive, and not sufficient for real time detection. By reducing the need for region proposal and pixel resampling stages, SSD achieves better performance.

Furthermore, SSD operates in a single pass, using default boxes in different scales and aspect ratios to better match the objects detected, accumulating previous predictions from multiple feature maps with different resolutions into the network.

The authors use performance and accuracy metrics in different scales to prove the effectiveness in the paper. Specifically, they show experiments across PASCAL, VOC, COCO and ILSVRC datasets.

b. SSD consists of two networks: a "truncated base network" and a network of added"convolutional feature layers to the end of the truncated base network." What is the purpose of the base network? Which base network do the authors of the SSD publication use, and what dataset was used to train the base network?

The purpose of the base network is to be the foundational structure—the VGG16— for feature extraction, which supplies critical low-level and mid-level visual features that assist in identifying objects within the image, therefore predicting offsets to default boxes of scales/aspect ratios and their associated confidence. The VGG6 is pretrained on the ILVRC dataset, which, when truncated, can be adapted to SSD's detection structure.

c. SSD has it's own loss function defined in chapter 2.2 in the original publication. What are the two attributes this loss function observes? How are these defined (short explanation without any formulas is sufficient) and how do they help the network to minimise the object detection error?

The paper mentions two main attributes: localization and confidence loss.

Localization loss observes the error in predicting the bounding box coordinates for detected objects. Comparing this with correct bounding boxes helps the model adjust and improve its accuracy of the placement overtime. This leads to better localization. It is also a Smooth L1 loss.

Confidence loss observes the error of predicted class probabilities for every detected object. Overtime, the model is encouraged to assign higher confidence to correct object classes, and vice versa to incorrect object classes. This leads to a better ability to do classifications. It is also a softmax loss.