Multimedia Project

SoSe 23

Multimedia Computing Lab
Prof. Dr. Rainer Lienhart, Katja Ludwig, Julian Lorenz

Assignment Submission 5

Jan Claar

Exercise 5.1

When using convolutional layers, only the spatial height and width of the inputs are treated separately, while all other feature dimensions are combined into one output channel dimension. Consequently, the output must be reshaped to fit the dimensions of the label grid. Furthermore, since the model outputs a measure of likelyhood for each class (background, human) being contained in the corresponding bounding box, it has an additional dimension compared to the label grid, which only contains the class index of the actual class.

Exercise 5.3

Negative Mining is done to reduce the imbalance of positive to negative samples. Since most of the bounding boxes are negative samples, the model would mainly learn to predict negative samples with high confidence, which would reduce the loss, but has limited use for the actual task. To counteract this, the model is trained on a subset of the negative samples proportional to the number of positive samples, by multiplying the loss tensor with a mask that sets a random set of negative loss terms to zero. This should shift the focus of the training more on actually recognizing the positive samples.

To measure the model performance, a metric similar to the Intersection over Union is calculated. In this case, the number of correctly identified bounding boxes (intersection) is divided by the union of all positive targets and positive predictions. This is only done for the target class, as most of the bounding boxes are negative samples (background class), meaning a high metric for that class would not have any meaning.

The model is trained for 50 epochs with a batch size of 16, an initial learning rate of 0.01 a momentum of 0.9 and a weight decay of 0.0005. Negative mining is done with a ratio of $2:1$ negative to positive samples.

The model with negative mining enabled achieves a quasi IoU of 0.01275 on the validation set, which is very low. However the model without negative mining achieves a quasi IoU of 0.0000, meaning even after 50 epochs of training it is not able to identify a single bounding box correctly. Interestingly, the overall loss of the regular model is much lower ($\approx$0.008) than that of the model trained with negative mining ($\approx$0.39). Since the loss is calculated over all bounding boxes during validation, it is not surprising that the regular model performs better in this regard, since it has seen much more of the overwhelmingly negative samples.
The low overall score is likely due to suboptimal hyperparameters and no actual data augmentation, as well as the simple classifier module of the model, which consists only of an upsampling step and one convolutional layer. Nonetheless, negative mining proved to be effective in improving training results.