Deep Learning in Computer Vision - Project 2

Till Ariel Aczél (s203216), Jan Piotr Latko (s193223), Jonas Søbro Christophersen (s153232), Technical University of Denmark, Deep Learning in Computer Vision (02514), 16/06/2021, Github Repository

Introduction

The purpose of this project is to implement at CNN using the U-Net architecture to segment lesions from lung CT images. Additionally, uncertainty of the model predictions are quantified using an ensemble method and an MC dropout model.

Data

The data consist of 12816 annotated 128-by-128 images of lesions from CT lung images. The data set is split into training/validation/test with the following split: 8843/1993/1980.

Annotations are binary with each pixels belonging to either a positive class (pixel covers a lesion) or a negative class (pixel does not cover lesion). For each image four annotations are recorded, labelled by four different expert radiologists. These annotations vary, and such the ground truth of the data is uncertain. These variances are illustrated in Figure 1 of a data-sample below.

Figure 1: .Sampledata illustrating differences in annotations across experts.

Network

The U-net architecture with skip-connect is used, using four downsampling and upsampling layers, with downsampling layers consisting of a max-pooling layer, double convolutional layers using batch-normalisation and relu-activation. Upsampling layers consist of bilinear upsampling layers paired with double convolutions, also utilizing max-pooling, batch-normalisation and relu-activation. During training, input-images were augmented using rotations and vertical and horizontal flips.

For uncertainty quantification two methods are used: An ensemble of U-nets; and a U-net utilizing drop-out in both training and validation combined with monte-carlo sampling.

In the ensemble, four U-nets are trained, one for each of the radiologist annotations. Combining these models in an ensemble allows us to quantify uncertainty of the predictions as the models will represent the bias in labelling for each of the four experts.

The U-net using dropout quantifies uncertainty through monte carlo sampling with drop-out during prediction. Generating several samples from the network using a different set of dropout neurons for each sample allows us to investigate the epistemic uncertainty of the model.

Results

Three different model types were trained and their ability to segment the the lesions were assessed using several metrics. The three models trained were:

One-Model: A model trained using data from a single experts annotations.
Ensemble: An ensemble of models, each trained on different expert annotated data.
Mean-Model: A model trained on the mean of the four experts annotations.

The model performances are shown in Table 1 below with the metrics for the best performing model (the Mean-Model) seen in Table 2. All metrics were calculated on each segmentation separately and then averaged over segmentations. The **Inter-Expert** measure is the average difference in annotation between the four experts.

	Sensitivity	Specificity	IoU
Inter-Expert	57.76%	99.77%	**47.43%**
One-Model	36.88%	99.85%	42.68%
One-Model (pos. weight=5)	**65.06%**	99.60%	32.36%
Ensemble	35.13%	99.86%	43.46%
Mean-Model	37.19%	**99.87%**	44.00%

Table 1: Model performances measured using Sensitivity, Specificity and intersection over Union.

Very high specificity scores come from the fact that most of the annotations are background and a very small change in specificity might still be large w.r.t. the size of the object. This results in models underpredicting the segmentations. To promote more sensitivity over specificity we upscale the positive term in the binary cross entropy loss by a factor of 5, which leads to lower IoU due to over-predicting.

Model predictions are visualised in Figure 2. Both MC Dropout and Ensemble produce segmentations that are more “blurry” on the sides due to averaging, which corresponds to lower model certainty in these regions.

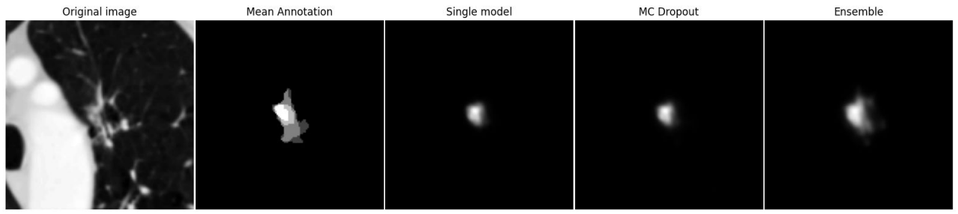

Figure 2: Model predictions using the three different methods

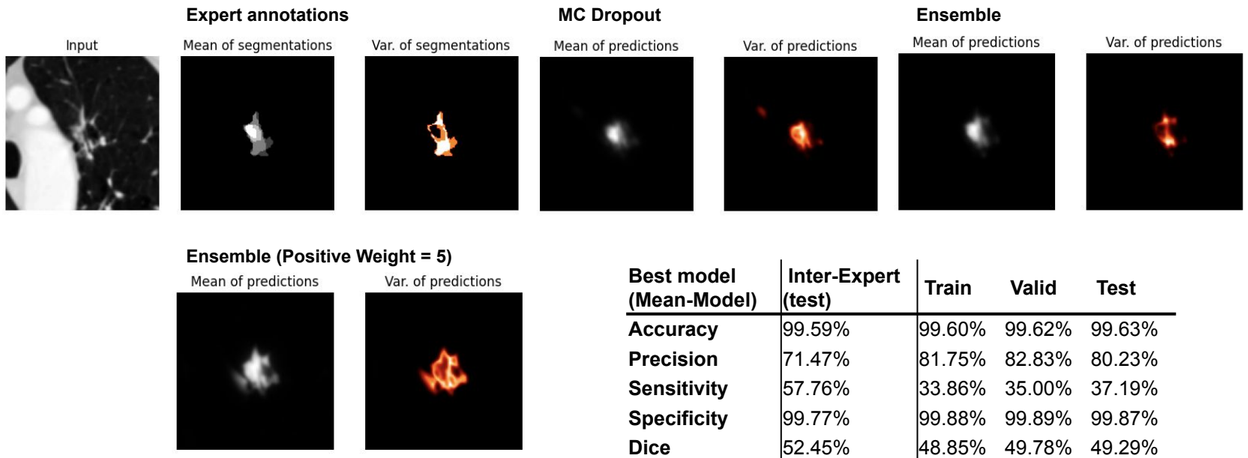

Figure 3: MC Dropout and Ensemble predictions and uncertainty visualisation.

In table 2 the performances of the best performing model is seen. From the scores it is seen that specificity is high, whereas the sensitivity is low. This indicates that the model is more inclined to predict pixels as background, compared to lesions. This could potentially be circumvented by weighting the positive class as was done for the One-Model.

Uncertainty Quantification (UQ)

Two different approaches are used to quantify uncertainty in predicted segmentations. The Ensemble Model and the MC Dropout

To assess the UQ of these models the Generalized Energy Distance (GED) is used to compare the uncertainty quantified by the ensemble model and dropout, with the uncertainty encoded by the four expert annotations. We obtain a QED of 0.547 and 0.437 for the Ensemble and MC Dropout models respectively, suggesting that the MC Dropout model achieved better UQ on this task.

Best model (Mean-Model)	Inter-Expert (test)	Train	Valid	Test
Accuracy	99.59%	99.60%	99.62%	99.63%
Precision	71.47%	81.75%	82.83%	80.23%
Sensitivity	57.76%	33.86%	35.00%	37.19%
Specificity	99.77%	99.88%	99.89%	99.87%
Dice	52.45%	48.85%	49.78%	49.29%
IoU	47.43%	43.05%	44.36%	44.00%

Table 2: Performance metrics for the best performing model (Mean-Model).

Additionally, uncertainty is visualized through mean and variance of the predictions (MC Dropout and Ensemble) and annotations in Figure 3. Both methods as well as annotators are more certain in the center of the object and less around it. The ensemble prediction looks a bit wider with the uncertainty visually more similar to the annotators.

Discussion & Further Work

As seen in the results, the model is not very proficient in capturing the finer contours of the lesions, but rather the overall size of the lesion. This is somewhat expected as especially the boundaries of the lesions are where the uncertainty is highest.

It should be noted that input images all have lesions centered, which likely is not the case in a real world use case. Altering the positions of the lesions using random crops could prove useful in combating this. However, for the purpose of this project random crops were left out of data augmentation.