Machine Learning (XAI501) Term Project Final Report

*Created on* December 13, 2020

Bayesian Uncertainty Estimation for Ultrasound Medical Image Segmentation

Sunwoo Kim1, Kwanseok Oh2, Jinhyo Shin2, Sangjin Kim2

1Department of Computer Science and Engineering, 2Department of Artificial Intelligence

1. **Introduction**

Sonography is a diagnostic medical procedure to produce dynamic visual images of organs, tissues, blood flow, and nerves inside the body. Different structures in the body reflect ultrasound waves differently. With this characteristic, many doctors use it as a visual tool to find a specific tissue while performing the procedure or to detect nerves for anesthesia. However, Sonographers, who perform ultrasound examinations, require a lot of proficiency to distinguish different organs in the visual image of ultrasound. In addition, misdiagnosis may occur because the accuracy of the test is affected by the condition of the Sonographers. This misdiagnosis is very dangerous because it is able to cause fatal damage to a patient's tissues during surgery. In particular, the nerves are very difficult to recover once they are damaged, so more attention is required during the medical procedure. Therefore, accurate detection is necessary to reduce the side effects and failures during anesthesia or medical procedures.

We want to make a detector that segments nerves in an ultrasound image as a tool to assist the proficiency of Sonographers by using semantic segmentation technique. Semantic segmentation is the task of clustering parts of an image together which belong to the same object class. It is a form of pixel-level prediction in which each pixel in the image is classified according to category. However, limitations exist in semantic segmentation. This is because even if the network is not sure which class a pixel belongs to, the network must select the class. Moreover, even if the network has performed well, there may always be a question of whether the results are completely reliable, depending on whether the training data is reliable. Thus, the uncertainty is very important information[1]. Especially in medical images that are directly linked to the patient's health, to fully trust what the model finds can be more dangerous. Therefore, we made the network to play the same role as the Bayesian network, which can reflect uncertainty by using dropout when performing image segmentation[2]. Uncertainty discovered by the network will allow Sonograpers to pay more attention to that part during procedure.

1. **Method**

**2.1 Unet**

The backbone framework of our architecture is similar to standard U-Net[3] which is composed of an encoder and a decoder, and connected with skip connections. The encoder is a contracting path which consists of 5 convolutional blocks with 64, 128, 256, 512 and 1024 output channels and each block has two 3 × 3 convolutional layers, ReLU activation. After convolutional blocks, downsampling is performed by using 2 × 2 max-pooling operation with stride 2. The decoder is an expansive path which consists of 4 convolutional blocks with 512, 256, 128 and 64 output channels performing up-sampling every step after convolutional blocks. The difference between our architectures and the standard U-Net are follows:

1. We adjust the number of output channels of convolutional blocks (64, 128, 256, 512, 1024) to (32, 64, 128, 256, 512) across the network. Since the ultrasound image is a single channel with gray scale, the semantic segmentation task becomes a pixel-wise binary class clustering. Therefore, to prevent overfitting, we reduce the number of output channels to reduce the number of trainable parameters.
2. To perform uncertainty estimation, we use dropout after each convolutional block of decoder part except for the first and the second blocks[4]. A dropout rate was 0.2 in all the cases. And then, we conduct ablation study to find the optimal location of dropout rate and the optimal dropout rate.

**2.2 Loss function**

We use the Dice Loss based on Dice coefficient as an objective function to find the boundary of the nerve on the ultrasound image[5]. The Dice coefficient is widely used to measure the performance of segmentation. It calculates the similarity of overlapped two images: the ground truth and predicted output images[6, 7]. Since nerve segmentation in ultrasound images is a pixel-wise binary segmentation that separates each pixel from the nerve of the background, using Dice Loss makes it easy to determine where each pixel belongs through overlap of ground truth and predictive images.

**2.3 Apply dropout to network**

Regularization techniques such as weight decay or dropout may not be sufficient to generalize networks used in various scenarios. These shortcomings can be addressed using uncertainty estimates. This is because uncertainty estimates can interpret the results and intuitively identify areas of ambiguity and uncertainty in predictions for the trained model. Therefore, we modified the standard U-Net framework to perform Bayesian inference through Monte Carlo (MC) sampling method using dropout at the training and test step. Each dropout layer is located at the end of the 3rd, 4th, and 5th convolution blocks corresponding to the low level, and the dropout rate *p* is set to 0.2.

**2.4 Metrics**

**2.4.1 Uncertainty Metrics**

There are two kinds of uncertainties which we encountered in the Bayesian network[8]. One is *epistemic uncertainty* and the other is *aleatoric uncertainty*. Epistemic uncertainty represents how much a model does not know because of insufficient data. Epidemic uncertainty can reduce by increasing data. Aesthetics uncertainty represents how many noises in data there are and is independent of the amount of data. This uncertainty can be reduced with increased sensor precision. The combination of these two uncertainties produces *predictive uncertainty* which refers to the model's confidence in its prediction.

In order to evaluate our model’s uncertainty, we adopted the following metrics. To measure pixel-wise uncertainty, [9] suggested uncertainty maps which can be used for semantic segmentation using two such metrics, *predictive entropy* \hat{\mathbb{H}}[y|\bold{x}, \mathcal{D}_{train}]
 which captures epistemic uncertainty and *mutual information* \hat{\mathbb{I}}[y, w|\bold{x}, \mathcal{D}_{train}] which captures predictive uncertainty. They are given \hat{\mathbb{H}}[y|\bold{x}, \mathcal{D}_{train}]

= - \sum_c \bigg( \frac{1}{T}\sum_t p(y=c|\bold{x}, \hat{w_t}) \bigg) 

log \bigg( \frac{1}{T} \sum_t p(y=c|\bold{x} , \hat{w_t}) \bigg) 

\hat{\mathbb{I}}[y, w|\bold{x}, \mathcal{D}_{train}]

=
\hat{\mathbb{H}}[y, |\bold{x}, \mathcal{D}_{train}]

+ \frac{1}{T} \sum_{c,t} p(y=c|\bold{x}, \hat{w_t}) 
log  p(y=c|\bold{x} , \hat{w_t})

where and are test input and output(prediction), is training data, is class which is binary in our task, is the number of Monte Carlo samples which is obtained during forward passes, and are the model parameters on the sample.

**2.4.2 Performance Evaluation Metrics**

We also followed performance evaluation metrics which is suggested in [9]. According to the hypothesis that if a model is confident about its predictions, it will be accurate on various parameters sampled and if a model is inaccurate on its any predictions, it would have been uncertain for the model, we can get two conditional probabilities. In our work, the detailed method of evaluation using a sliding window follows the originally proposed one.

1. P(accurate|certain) : The probability that the model is accurate on its output given that it is confident on the same.
2. P(uncertain|inaccurate) : The probability that the model is uncertain about its output given that it has made a mistake in its prediction (i.e., is inaccurate).
3. **Experiment**

**3.1 Dataset**

A set of total 11,144 ultrasound images of the neck was downloaded from the open database of the kaggle and used in our experiments[10]. This data set was divided into train and test sets, each of them comprising 5635 and 5509 samples. The provided train set comprised 47 different subjects with about 120 images for each subject. The nerve on the train images has been manually annotated by humans. Annotators were trained by experts and instructed to annotate images where they felt confident about the existence of the Brachial Plexus (BP) landmark.

**3.2 Hyperparameter setting**

The overall hyper parameter settings for model training are as follows. In the case of the learning rate, we tried all the experiments for 0.001, 0.0001, and 0.00001, but there was no significant difference in the performance of the selected model, and we could only confirm the difference in the training step. Also, we tried setting the dropout ratio to 0.5, but we confirmed that it does not train when the dropout is too high.

|  |  |
| --- | --- |
| Hyper-parameter setting for model training | |
| Learning rate | 0.001 |
| Learning decay (exponential) | 0.99 |
| Epoch | 150 |
| Batch size | 8 |
| Dropout rate | 0.2 |
| weight constraint | 0.0001 |

**3.3 Conducting base segmentation without dropout**

First we conduct the simple segmentation of our dataset to evaluate the base U-Net performance. We use our adjusted U-Net just without dropout. We conduct training and evaluate the model five times and use the f1 score metric to evaluate performance.

**3.4 Ablation study**

We conducted an ablation study on how the location of the dropout layer affects the performance of the model. As a result, there was no significant difference in performance depending on the location, and it was possible to confirm the presence or absence of training about the drop rate. From this result, we can infer that it is a network that learns smoothly even when the regularizer is not used. In other words, it can be assumed that the model's activation nodes for classification tasks are well highlighted.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Performance of our Baseline U-Net without dropout | | | | |
| Metric | Mean | | Std | |
| F1 | 0.8113 | | 0.000559 | |
| Ablation study for major component analysis | | | | |
|  | Loss | IoU | P(accurate|certain) | P(uncertain|inaccurate) |
| encoder | 0.4006 | 0.6046 | 0.9716 | 0.1225 |
| decoder | 0.3089 | 0.6241 | 0.9731 | 0.1104 |
| encoder+decoder | 0.3138 | 0.6015 | 0.9612 | 0.1475 |

Table 1. Performance of baseline U-Net and Ablation study to find optimal location of dropout

**3.5 Attempt to predict uncertainty map using soft mask**

To better estimate the uncertainty, we also conducted an experiment using a soft mask with a range of 0 to 1, not a binary mask. Min-max normalization was used, and the output of the decoder was sigmoid. As a result, it did not yield better results than when using a binary mask. The result images of segmentation images and uncertainty map are followed in Appendix.

**3.6 Evaluation and Analysis**

Our task is a binary segmentation and labels in each image are highly imbalanced. Most pixels are background and not our concern. The imbalance of labels interferes with accurate evaluation of our model. It is well known that accuracy is vulnerable to label imbalance. Since P(accurate|certain) is based on accuracy, we thought that this metric could not accurately evaluate our model. To overcome this problem, we have applied other well-known metrics and customized metrics.

The first one we can think of is IoU(Intersection on Union) which is one of the popular metrics in computer vision. It is calculated by ratio of intersection of prediction and Ground Truth and union of them. The denominator, which is union, constrains the area of interest in evaluation, excluding not meaningful background. Furthermore we want to evaluate the performance based on uncertainty and modified canonical IoU. We devised two custom metrics : certain IoU which uses only predicted pixels where the model is confident and uncertain IoU which uses prediction including pixels where the model is not confident. Whether confidence or not was divided on an arbitrary threshold.

Another way we used to evaluate uncertainty is using Cohen’s kappa coefficient which is a statistic and is used to measure inter-rater reliability. In this case, accuracy and certainty can be considered as raters. From a point of view that kappa coefficients considers agreement between accurate and certain and agreement between inaccurate-uncertain, it is similar to the P(accurate|certain) and P(uncertain|inaccurate) mentioned above. But, since the relative observed agreement is normalized by the probability of random agreement, kappa coefficient has a property robust to label imbalance.

Results are shown on table 1. We expected that certain IoU is higher than standard IoU because the predicted area would be small and the union would be smaller naturally. In the opposite we expected that uncertain IoU is lower than standard IoU by the same logic. However there were many cases contrary to our expectations. The deviation of IoU series is large. Inconsistent performance of predictions on test images might affect this result. Kappa coefficient is a value between 0 and 1. The closer it is to 1, the more significant it is. Though the uncertainty map shows that prediction of uncertainty is qualitatively significant, kappa coefficient is low. This result drove us to some discussions.

|  |  |  |
| --- | --- | --- |
| Certain IoU Mean. (std.) | Uncertain IoU Mean. (std.) | Kappa coefficient |
| 0.6450 (0.198043) | 0.6428 (0.193733) | 0.0365 |

Table 2. Performance evaluation on various metrics

1. **Discussion & Future works**

**4.1 Absence and necessity of metrics for uncertainty**

Though our various trials to evaluate the uncertainty properly, these Metrics were not enough to evaluate the uncertainty. When the model is confident, it is an intuitive evaluation criterion that the accuracy should be high. P(uncertain|inaccurate) and kappa coefficient evaluate the agreements between inaccurate and uncertain. However, a model’s lack of confidence does not mean that it must tend to be inaccurate. How to evaluate the uncertainty still remains unsolved.

**4.2 Attempt to improve base segmentation model**

Assuming that training images have reliable annotations of nerve sections, we thought if the performance of simple semantic segmentation improves, the areas with uncertainty will be reduced. At the same time, we thought that the performance of assessing uncertainty would also be improved. Therefore, we will attempt to improve the base segmentation model as our future work by using mean and standard deviation(std) values of our training data. It would be able to produce predicted images which are dependent on the latent space of training set through the training process. Furthermore, since the std of predicted image depends on the std of training data, it would be able to help to assess the uncertainty of the predicted images.

1. **Conclusion**

We conducted nerve segmentation in ultrasound images by adjusting standard U-Net. To check the uncertainty area of segmentation results, we applied dropout to U-Net. It makes that general deep learning network perform bayesian network algorithms. Moreover, we conduct ablation studies to find the optimal location of dropout. And then, by using various evaluation metrics and soft mask labels, we visualize the uncertainty to image. Referring to Appendix, it can be seen that image segmentation was relatively well performed if the IoU score between the predicted image and the Ground Truth image was good. In addition, even if the IoU score was not good, the uncertainty where the area that the actual nerve does not exist was high. Therefore, these results may contribute to the Sonographer to identify areas of uncertainty in the nerves when detecting the nerve.

1. **References**

[1] [Alex Kendall](https://arxiv.org/search/cs?searchtype=author&query=Kendall%2C+A), [Yarin Gal](https://arxiv.org/search/cs?searchtype=author&query=Gal%2C+Y), What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?, arXiv:1703.04977, NIPS 2017

[2] [Yarin Gal](https://arxiv.org/search/stat?searchtype=author&query=Gal%2C+Y), [Zoubin Ghahramani](https://arxiv.org/search/stat?searchtype=author&query=Ghahramani%2C+Z), Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, arXiv:1506.02142, ICML 2016

[3] [Olaf Ronneberger](https://arxiv.org/search/cs?searchtype=author&query=Ronneberger%2C+O), [Philipp Fischer](https://arxiv.org/search/cs?searchtype=author&query=Fischer%2C+P), [Thomas Brox](https://arxiv.org/search/cs?searchtype=author&query=Brox%2C+T), U-Net: Convolutional Networks for Biomedical Image Segmentation, arXiv:1505.04597, MICCAI 2015

[4] [Alex Kendall](https://arxiv.org/search/cs?searchtype=author&query=Kendall%2C+A), [Vijay Badrinarayanan](https://arxiv.org/search/cs?searchtype=author&query=Badrinarayanan%2C+V), [Roberto Cipolla](https://arxiv.org/search/cs?searchtype=author&query=Cipolla%2C+R), Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding, arXiv:1511.02680, British Machine Vision Conference 2017

[5] [Fausto Milletari](https://arxiv.org/search/cs?searchtype=author&query=Milletari%2C+F), [Nassir Navab](https://arxiv.org/search/cs?searchtype=author&query=Navab%2C+N), [Seyed-Ahmad Ahmadi](https://arxiv.org/search/cs?searchtype=author&query=Ahmadi%2C+S), V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation, arXiv:1606.04797, Fourth International Conference on 3D Vision(3DV) 2016

[6] [Carole H Sudre](https://arxiv.org/search/cs?searchtype=author&query=Sudre%2C+C+H), [Wenqi Li](https://arxiv.org/search/cs?searchtype=author&query=Li%2C+W), [Tom Vercauteren](https://arxiv.org/search/cs?searchtype=author&query=Vercauteren%2C+T), [Sébastien Ourselin](https://arxiv.org/search/cs?searchtype=author&query=Ourselin%2C+S), [M. Jorge Cardoso](https://arxiv.org/search/cs?searchtype=author&query=Cardoso%2C+M+J), Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations, arXiv:1707.03237, 2017

[7] [Shruti Jadon](https://arxiv.org/search/eess?searchtype=author&query=Jadon%2C+S), A survey of loss functions for semantic segmentation, arXiv:2006.14822, SPIE Medical Imaging 2020

[8] GAL, Yarin. Uncertainty in deep learning. University of Cambridge, 2016, 1.3.

[9] MUKHOTI, Jishnu; GAL, Yarin. Evaluating bayesian deep learning methods for semantic segmentation. arXiv preprint arXiv:1811.12709, 2018.

[10] “Ultrasound Nerve Segmentation.” kaggle. last modified May 19, 2016, accessed Oct 22, 2020, <https://www.kaggle.com/c/ultrasound-nerve-segmentation/data>.

[11] Teye, Mattias, Hossein Azizpour, and Kevin Smith. "Bayesian uncertainty estimation for batch normalized deep networks." arXiv preprint arXiv:1802.06455 (2018).

[12] Harakeh, Ali, Michael Smart, and Steven L. Waslander. "Bayesod: A bayesian approach for uncertainty estimation in deep object detectors." 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020.

[13] Maddox, Wesley J., et al. "A simple baseline for bayesian uncertainty in deep learning." Advances in Neural Information Processing Systems. 2019.

**Appendix**

**Chapter A. Uncertainty map generated by Dropout**

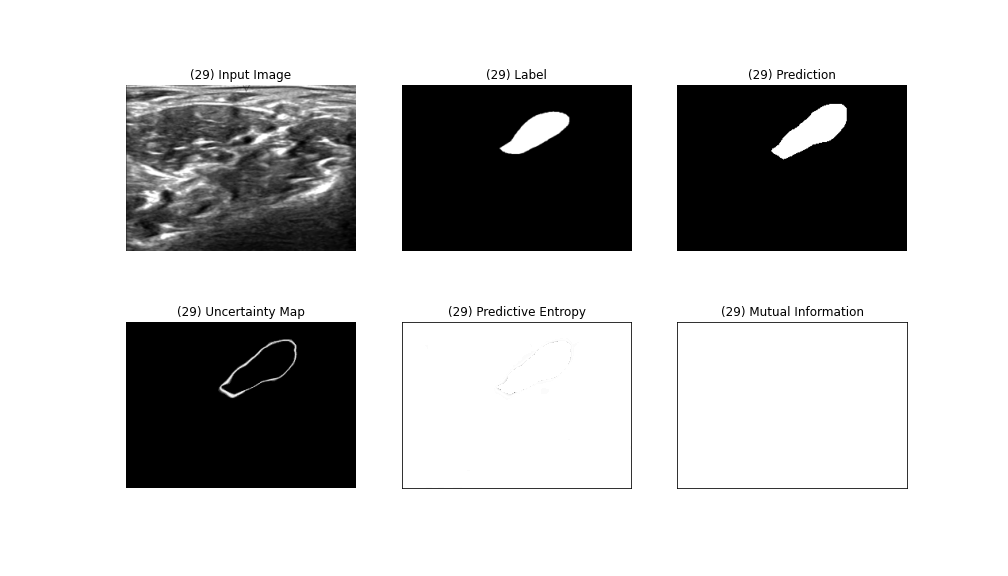
****

Figure 1. Prediction Image and Uncertainty map generated by dropout U-Net with high IoU score : 0.7834

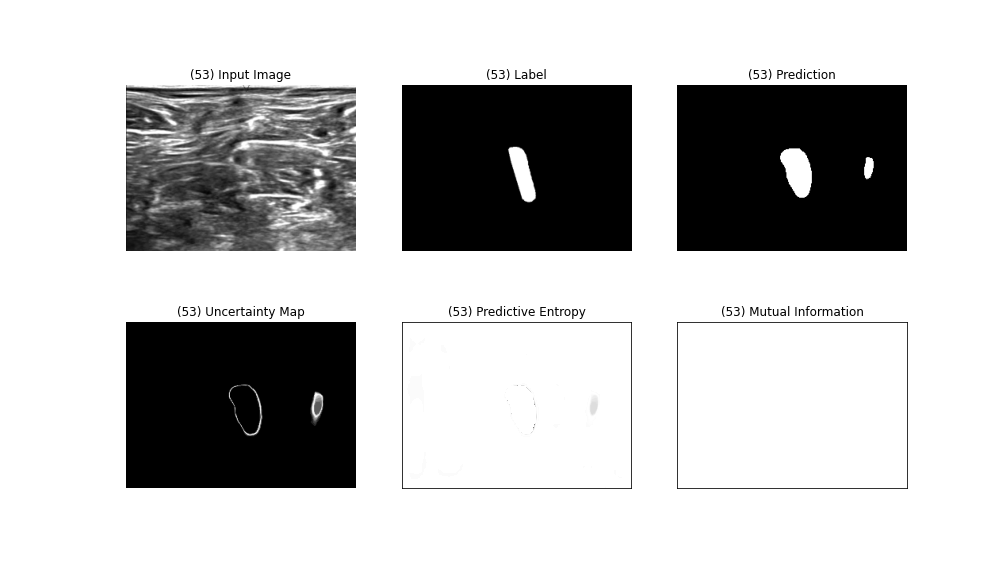


Figure 2. Prediction Image and Uncertainty map generated by dropout U-Net with low IoU score :0.6201 but well performed segment uncertainty area

**Chapter B. Uncertainty map generated by soft mask**

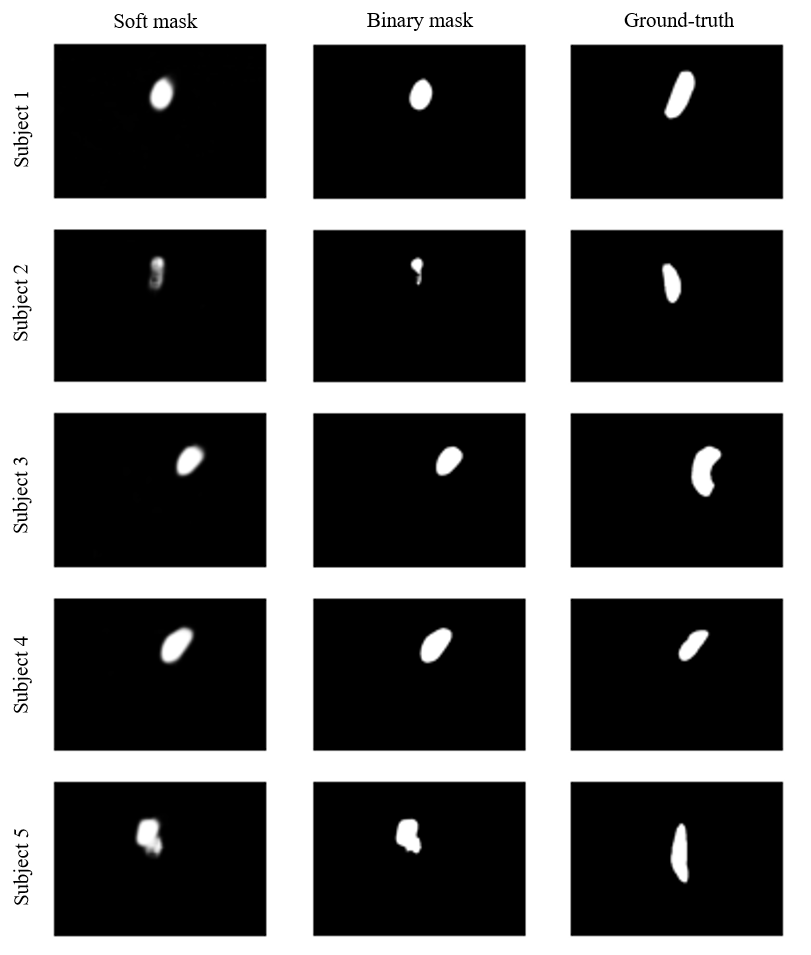


Figure 3. Comparison of soft mask and binary mask with ground-truth



Figure 4. Overlaying contour obtained from binary mask on soft mask