2023

Class Activation Mapping and Uncertainty Estimation in Multi-Organ Segmentation

Md. Shibly Sadique
Old Dominion University, msadi002@odu.edu

Walia Farzana
Old Dominion University, wfarz001@odu.edu

Ahmed Temtam
Old Dominion University, atemt001@odu.edu

Khan Iftekharuddin
Old Dominion University, kiftekha@odu.edu

Khan Iftekharuddin (Ed.)

See next page for additional authors

Original Publication Citation

Sadique, M. S., Farzana, W., Temtam, A., & Iftekharuddin, K. (2023). Class activation mapping and uncertainty estimation in multi-organ segmentation. In K. Iftekharuddin & W. Chen (Eds.), *Medical Imaging 2023: Computer-Aided Diagnosis, Proceedings of SPIE 12465* (124650V). SPIE. https://doi.org/10.1117/12.2655508

Class activation mapping and uncertainty estimation in multi-organ segmentation

M. Sadique, W. Farzana, A. Temtam, K. Iftekharuddin

M. S. Sadique, W. Farzana, A. Temtam, K. M. Iftekharuddin, "Class activation mapping and uncertainty estimation in multi-organ segmentation," Proc. SPIE 12465, Medical Imaging 2023: Computer-Aided Diagnosis, 124650V (7 April 2023); doi: 10.1117/12.2655508

Event: SPIE Medical Imaging, 2023, San Diego, California, United States

Class Activation Mapping and Uncertainty Estimation in Multi-Organ Segmentation

M. S. Sadique, W. Farzana, A. Temtam, K. M. Iftekharuddin[*]
Vision Lab in Department of Electrical and Computer Engineering, Old Dominion University,
Norfolk, VA 23529

ABSTRACT

Deep learning (DL)-based medical imaging and image segmentation algorithms achieve impressive performance on many benchmarks. Yet the efficacy of deep learning methods for future clinical applications may become questionable due to the lack of ability to reason with uncertainty and interpret probable areas of failures in prediction decisions. Therefore, it is desired that such a deep learning model for segmentation classification is able to reliably predict its confidence measure and map back to the original imaging cases to interpret the prediction decisions. In this work, uncertainty estimation for multiorgan segmentation task is evaluated to interpret the predictive modeling in DL solutions. We use the state-of-the-art nnU-Net to perform segmentation of 15 abdominal organs (spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, pancreas, right adrenal gland, left adrenal gland, duodenum, bladder, prostate/uterus) using 200 patient cases for the Multimodality Abdominal Multi-Organ Segmentation Challenge 2022. Further, the softmax probabilities from different variants of nnU-Net are used to compute the knowledge uncertainty in the deep learning framework. Knowledge uncertainty from ensemble of DL models is utilized to quantify and visualize class activation map for two example segmented organs. The preliminary result of our model shows that class activation maps may be used to interpret the prediction decision made by the DL model used in this study.

Keywords: Uncertainty estimation, multi-organ segmentation and prediction, knowledge uncertainty, activation map

1. DESCRIPTION OF OBJECTIVE

One of the problems with modern deep neural networks is that they are poorly calibrated and tend to rely too heavily on predictions with inherent uncertainty [1]. There are different techniques to improve estimates of predictive uncertainty. A classical approach is called temperature scaling, where the model confidences are scaled using a post-hoc procedure on the retained validation set [2]. A popular approximate Bayesian approach is a dropout-based model, where the predictive uncertainty is computed based on multiple model outputs on a given image (with dropout enabled) [3]. Another sampling-based approach uses the agreement between a set of models as a measure of model uncertainty [4]. Interestingly, the use of ensembles has been shown to produce the best results in estimating uncertainty under distribution change [5,6]. The common configuration of ensembles is to use neural networks trained using different random initialization weights to induce diversity among the models [7]. This is because networks pretrained on the same dataset have been shown to stay in the same catchment in the loss landscape and thus reduce variation in the models [8]. In deep learning, dropout is designed as a regularization technique and can also be interpreted as an ensemble of multiple models [9]. The realization that dropout could be used to effectively quantify uncertainty [10] motivated a further exploration of ensembles in deep learning models for the same purpose. Deep ensembles have been shown to outperform Monte-Carlo (MC) dropout in quantifying uncertainty in a variety of datasets and tasks in regression and classification [11]. Additionally, deep ensembles have been shown to be state-of-the-art in out-of-distribution settings (e.g., perturbations of the data or the introduction of new classes unseen during training). They outperform MC dropout and other methods [12]. The reason why deep ensembles perform so well in out-of-distribution settings is that their weight values and loss trajectories are very different from one another, and, as a result, they lead to diverse predictions [13].

This work examines ensemble-based uncertainty-estimation for deep learning models. The contributions are as follows. First, we consider generating ensembles using different nnU-Net variants (2D, 3D full resolution, 3D low resolution and 3D cascade) for multi-organ segmentation. Second, we compute the total uncertainty (TU) and knowledge uncertainty (KU) from the softmax probabilities to predict the organs with highest accuracy. Third, to understand the attributes of

*K. M. Iftekharuddin: E-mail: kiftekha@odu.edu

Medical Imaging 2023: Computer-Aided Diagnosis, edited by Khan M. Iftekharuddin,
Weijie Chen, Proc. of SPIE Vol. 12465, 124650V · © 2023 SPIE
1605-7422 · doi: 10.1117/12.2655508

using ensembles-based uncertainty estimation in our models, we conduct extensive analysis associated with the highest and the lowest knowledge uncertainty values for different cases. Finally, we apply class activation maps technique to examine predictive uncertainty in two example organ classes and discriminative image regions used by our DL model to identify a specific class in the image.

2. METHODOLOGY

2.1 Datasets

For this study, we use the datasets provided by Multi-Modality Abdominal Multi-Organ Segmentation (AMOS) Challenge 2022 .The AMOS Challenge 2022 Task I provides a total of 200 CT scans with voxel-level annotations of 15 abdominal organs including the spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, pancreas, right adrenal gland, left adrenal gland, duodenum, bladder, prostate/uterus for training cases and 100 CT scans for the validation phase are presented [14].

2.2 Semantic Segmentation

Semantic segmentation can be viewed as a pixel-wise classification problem where the goal is to assign to each pixel a predicted category $c \in \{1, ..., C\}$. As it is now common in the visual recognition area, semantic segmentation models are mostly based on Convolutional Neural Networks (CNNs), for example, UNet [15]. Many different DL architectures have been developed for medical image segmentation. In our work, we use state-of-the-art nnU-Net which is an open-source tools [16,17,18,19]. The abdominal CT images are used for training each configuration of the nnU-Net with five-fold cross-validation. A more detailed overview of the performance scores for all organs and different DL configurations are provided in Table 1. We use 2D, 3D_lowres, and 3D_fullres models and find the best configurations obtained from the cross validation on the training cases as an ensemble (ensemble 1: 2D and 3D_lowres, ensemble 2: 2D and 3D_fullres, and ensemble 3: 3D_lowres and 3D_fullres) to predict the validation cohort. We generate the segmented mask from different variants of nnU-Net configurations.

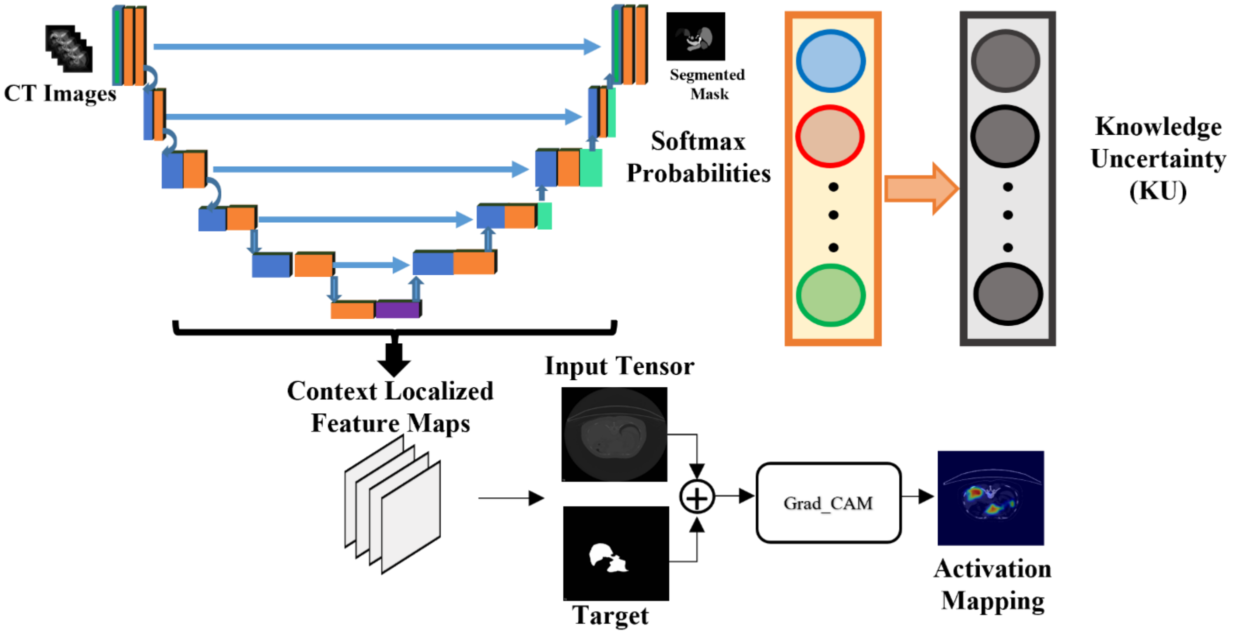

Figure 1. Overall pipeline for class activation map and uncertainty estimation for multi-organ segmentation

2.3 Uncertainty Estimation

Estimation of predictive uncertainty can be used to detect errors. Ideally, the model indicates a high level of uncertainty in situations where an error is likely to be made. That allows us to detect errors and take safer actions. Fundamentally, the choice of action may depend on why the model is uncertain. There are two main sources of uncertainty: data uncertainty (also known as random uncertainty) and knowledge uncertainty (also known as epistemic uncertainty).

Knowledge uncertainty arises when the model receives input from a region that is poorly covered by the training data or far from the training data. In these cases, the model knows very little about this region and is likely to make an error. Estimating knowledge uncertainty requires a set of models. If all models understand an input, they will give similar predictions (low knowledge uncertainty). However, if the models do not understand the input, they are likely to provide diverse predictions and strongly disagree with each other (high knowledge uncertainty).

Uncertainty represents the reliability of our inferences. Some statistics that proxy or approximate uncertainty include the softmax probability, predictive variance, and Shannon's entropy of the softmax vector. Consider samples of softmax probabilities from models can yield different predictions. Therefore, estimates of knowledge uncertainties can be obtained by analyzing the diversity of predictions. Consider an ensemble of softmax probabilistic models $\{P(y|x,\theta^{(m)})\}_{m=1}^{M}$ sampled from the model's predictions. Each model $P(y|x,\theta^{(m)})$ yields a different estimate of data uncertainty, represented by the entropy of its predictive distribution [20,21,22]. Uncertainty in predictions due to knowledge uncertainty is expressed as the level of spread, or disagreement of models in the ensemble.

2.4 Activation Map

Class activation maps may be used to interpret the prediction decision made by the DL methods. We apply Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize and interpret decisions from our DL-based models [23]. A class activation map for a particular category indicates the discriminative image regions used by the DL model to identify that category (e.g., Fig. 2).

3. EXPERIMENTAL RESULTS

We evaluate our model at two levels to estimate uncertainty: segmentation of multi-organ, and class activation mapping based on knowledge uncertainty. The 200 CT scans are used for training each configuration of our DL model followed

Table 1. 5-fold cross-validation on Training dataset

Model	Organs	2D	3D_fullres	3D_lowres	3D_cascade_fullres
Training Dice 5-fold CV	Backround	0.9528	0.9595	0.9588	0.9588
	Spleen	0.945	0.9536	0.9557	0.9523
	Right Kidney	0.9364	0.9479	0.9477	0.9417
	Left Kidney	0.7898	0.8288	0.825	0.8028
	Gallbladder	0.8379	0.8434	0.8394	0.8071
	Esophagus	0.9692	0.9714	0.9741	0.9718
	Liver	0.8682	0.9027	0.9028	0.8893
	Stomach	0.9509	0.9542	0.9491	0.9419
	Aorta	0.8807	0.9102	0.9043	0.8938
	Inferior Vena	0.8192	0.8589	0.8448	0.8426
	Pancreas	0.7525	0.7861	0.763	0.7404
	Right Adrenal	0.7579	0.8009	0.7714	0.7469
	Left Adrenal	0.7567	0.8228	0.8053	0.78
	Duodenum	0.8692	0.8911	0.8926	0.8709
	Bladder	0.8077	0.8434	0.8386	0.8336
	Mean Dice Score	0.8596	0.885	0.8782	0.8649

by five-fold cross-validation. We generate the segmented mask from different variants of nnU-Net configurations. Table 1 summarizes the results for 5-fold cross validation with mean Dice Similarity Co-efficient (DSC) for different configurations of nnU-Net (2D, 3D_fullres, 3D_lowres, 3D_cascade_fullres). We further compute the total uncertainty (TU) and knowledge uncertainty (KU) from the softmax probabilities of different class categories to understand the attributes of using ensembles-based uncertainty estimation in our models. We conduct extensive analysis associated with the highest and the lowest KU values for different patient datasets.

Table 2. Total Uncertainty and Knowledge Uncertainty for each fold (mean)

Fold Number	Total Uncertainty (TU) %	Knowledge Uncertainty (KU) %
1	21.6308	0.2305
2	20.7892	0.0063
3	19.8342	0.0078
4	19.6178	0.0097
5	21.8820	0.0056

Table 2 shows the mean value of total uncertainty and knowledge uncertainty for test cases in each fold for 5-fold cross validation for different DL (2D, 3D_fullres, 3D_lowres, 3D_cascade_fullres) models. To estimate total uncertainty, we calculate entropy of expected value of softmax probabilities from different models in the ensemble model and knowledge uncertainty is evaluated from mutual information which is the difference between total uncertainty (entropy of expected) and the expected entropy of each model in the ensemble. The range of total uncertainty and knowledge uncertainty is 0-100 and the low value indicates lower uncertainty. However, as knowledge uncertainty provides insights into the disagreement in predictive probabilities of ensemble model for same test cases, we focus on knowledge uncertainty score.

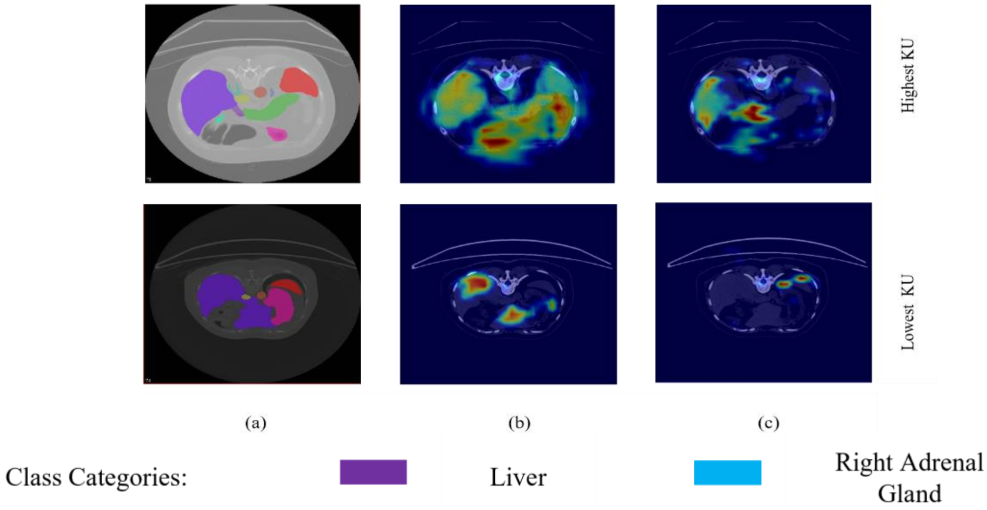

Figure 2. Visualization of Activation Mappings for the two examples associated with Lowest and Highest Knowledge Uncertainty. (a) Image overlaid with predicted mask, Activation mapping for (b) Liver and (c) Right Adrenal Gland

We select two example cases with the highest and the lowest knowledge uncertainty scores, respectively. Consequently, for two example cases, we consider the predicted softmax probabilities for 16 abdominal organs and focus on the organs which have the highest and the lowest predictive probabilities. This yield liver (class 6) with the highest and right adrenal gland (class 11) with the lowest predictive probabilities. Figure 2 shows activation maps of predictions for fold 1 for the two example cases. The first row represents the case with high knowledge uncertainty while the second row depicts the case with low knowledge uncertainty. The second and third column represents the two organs (liver and right adrenal gland) with highest and lowest predictive probabilities for the two example organ cases. For the activation mappings we extract the context localization layers from our 2D DL model. Then we preprocess the input tensor and extract the corresponding class categories (Liver and Right Adrenal Gland) for visualizing the activation maps with GRAD-CAM. Note for the case with high knowledge uncertainty, the focus of the model presented by yellow and red pixels in the figure is not concentrated on the respective organs (liver and right adrenal gland). Hence, the model is more uncertain to make predictive decision. However, for the case with low knowledge uncertainty, the focus of the model (presented by yellow and red pixels) is localized to the two respective organs. these results show that activation map may be used to interpret the uncertainty of prediction decision made by the proposed DL model.

4. DISCUSSION AND CONCLUSION

This work represents that knowledge uncertainty from an ensemble of models for quantifying and managing uncertainty in our DL multiorgan segmentation framework. To understand the attributes of using ensembles-based uncertainty estimation in our models, we conduct analysis associated with the highest and the lowest knowledge uncertainty values for different class categories. In addition, we apply class activation maps technique to examine predictive uncertainty in two example segmentation organ classes obtained by our DL model. These results suggest the feasibility of explaining model uncertainty in prediction decision of DL segmentation models.

ACKNOWLEDGEMENTS

This work is partially supported by NIBIB/NIH grant# R01 EB020683.

REFERENCES

[1] Guo, C., Pleiss, G., Sun, Y. and Weinberger, K.Q., 2017, July. On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321-1330). PMLR.

[2] Platt, J., 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, *10*(3), pp.61-74.

[3] Gal, Y. and Ghahramani, Z., 2016, June. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning* (pp. 1050-1059). PMLR.

[4] Lakshminarayanan, B., Pritzel, A. and Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, *30*.

[5] Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B. and Snoek, J., 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, *32*.

[6] Gustafsson, F.K., Danelljan, M. and Schon, T.B., 2020. Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 318-319).

[7] Hansen, L.K. and Salamon, P., 1990. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, *12*(10), pp.993-1001.

[8] Neyshabur, B., Sedghi, H. and Zhang, C., 2020. What is being transferred in transfer learning? *Advances in neural information processing systems*, *33*, pp.512-523.

[9] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), pp.1929-1958.

[10] Gal, Y. and Ghahramani, Z., 2016, June. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning* (pp. 1050-1059). PMLR.

[11] Lakshminarayanan, B., Pritzel, A. and Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, *30*.

[12] Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B. and Snoek, J., 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Advances in neural information processing systems, 32.

[13] Fort, S., Hu, H. and Lakshminarayanan, B., 2019. Deep ensembles: A loss landscape perspective. arXiv preprint arXiv:1912.02757.

[14] Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X. and Luo, P., 2022. AMOS: A Large-Scale Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation. arXiv preprint arXiv:2206.08023.

[15] Ronneberger, O., Fischer, P. and Brox, T., 2015, October. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

[16] Sadique, M.S., Rahman, M.M., Temtam, A.G., Farzana, W.,and Iftekharuddin, K.M., 2023, Brain Tumor Segmentation using Neural Ordinary Differential Equations with UNet-Context Encoding Network, forthcoming.

[17] Rahman, M.M., Sadique, M.S., Temtam, A.G., Farzana, W., Vidyaratne, L. and Iftekharuddin, K.M., 2022, July. Brain Tumor Segmentation Using UNet-Context Encoding Network. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I (pp. 463-472). Cham: Springer International Publishing.

[18] Isensee, F., Jäger, P.F., Kohl, S.A., Petersen, J. and Maier-Hein, K.H., 2019. Automated design of deep learning methods for biomedical image segmentation. arXiv preprint arXiv:1904.08128.

[19] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021 Feb;18(2):203-211. doi: 10.1038/s41592-020-01008-z. Epub 2020 Dec 7. PMID: 33288961.

[20] Malinin, A. and Gales, M., 2019. Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. Advances in Neural Information Processing Systems, 32.

[21] Malinin, A., Prokhorenkova, L. and Ustimenko, A., 2020. Uncertainty in gradient boosting via ensembles. arXiv preprint arXiv:2006.10562.

[22] Farzana, W., Shboul, Z.A., Temtam, A. and Iftekharuddin, K.M., 2022, April. Uncertainty estimation in classification of MGMT using radiogenomics for glioblastoma patients. In Medical Imaging 2022: Computer-Aided Diagnosis (Vol. 12033, pp. 365-371). SPIE.

[23] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer *vision* (pp. 618-626).