Propagating Uncertainty in Multi-Stage Bayesian Convolutional Neural Networks with Application to Pulmonary Nodule Detection

Onur Ozdemir, Benjamin Woodward, Andrew A. Berlin
Draper
{oozdemir,bwoodward,aberlin}@draper.com

Abstract

Motivated by the problem of computer-aided detection (CAD) of pulmonary nodules, we introduce methods to propagate and fuse uncertainty information in a multi-stage Bayesian convolutional neural network (CNN) architecture. The question we seek to answer is "can we take advantage of the model uncertainty provided by one deep learning model to improve the performance of the subsequent deep learning models and ultimately of the overall performance in a multi-stage Bayesian deep learning architecture?". Our experiments show that propagating uncertainty through the pipeline enables us to improve the overall performance in terms of both final prediction accuracy and model confidence.

1 Introduction & Related Work

We introduce techniques for fusing model confidence information with pixel-level image information to enable uncertainty estimates to propagate, along with the image data, in a multi-stage Bayesian deep learning framework. Using computer-aided detection (CAD) of pulmonary nodules as an example, we demonstrate that Bayesian modeling in a multi-stage setting not only provides model confidence associated with nodule detection decisions, but, via propagation of uncertainty information between networks, also improves the overall detection/classification performance. This is important to build physician trust in the model's performance for applications such as cancer detection, where knowing that a result has high vs. low confidence can influence a follow-on treatment decision. To the best of our knowledge, none of the existing methods for lung lesion detection [1–5] utilize or produce uncertainty/confidence information associated with nodule detection decisions.

The problem of computer-aided detection (or CAD) of pulmonary nodules using low-dose CT scans has a number of unique challenges [6, 7]. First, each scan is very large, prohibiting them from being used as full size 3D images for deep learning due to computational resource constraints. Second, pulmonary nodules to be detected are much smaller than full CT scans and they have high variability in terms of shape, size, and texture properties. These challenges have motivated researchers to divide the original problem into simpler subproblems and use multi-stage learning algorithms, where each algorithm attempts to solve a simpler subproblem. More specifically for the CAD problem, it is common to first segment 2D slices to find regions of interest (narrowing the search space) followed by performing 2D or 3D nodule detection within each region of interest to improve detection results.

Motivated by the CAD problem, we consider a multi-stage deep learning architecture as shown in Figure 1, comprising two consecutive Bayesian convolutional neural networks (CNNs) cascaded in a way that the predictions of the segmentation network inform the predictions of the nodule detection network. The overall performance of such an architecture depends on the individual performance of each network as well as on the coupling between them. In this architecture, each network has a different local view of the full CT scan. The segmentation network operates on full 2D axial CT slices (see Figure 2a), whereas the detection network operates on small 3D volumes.

Second workshop on Bayesian Deep Learning (NIPS 2017), Long Beach, CA, USA.

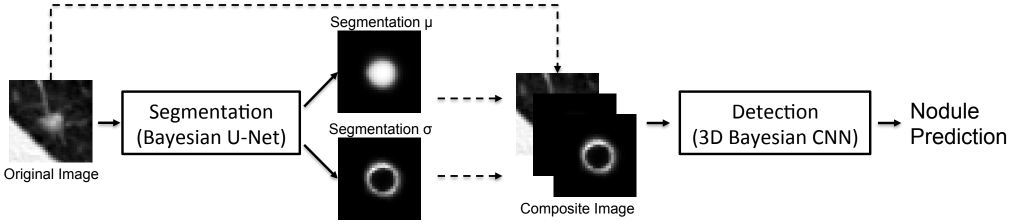


Figure 1: Two-Stage Bayesian CNN CAD architecture. Segmentation predictive mean and standard deviation maps are fused with the original image to form a 3-channel composite image, which is then fed into a 3D Bayesian CNN for final nodule detection.

Our goal here is to use segmentation predictions along with their uncertainties to improve nodule detection performance. To that end, segmentation predictions are incorporated as additional features for nodule detection by concatenating/fusing segmentation predictive mean and standard deviation with the original image to form a 3-channel composite image, as shown in Figure 1. This composite image is then fed into a Bayesian nodule detection network. The extra information provided by segmentation predictions along with associated uncertainties not only improves the detection performance for nodules that the segmentation network has high confidence of, but it also increases the overall model confidence of the nodule detection network. Since the segmentation predictions are based on 2D local context and as a result can produce false positives and lower confidence true positives, we also use the original 1-channel image to train a second Bayesian nodule detection network. Combining predictions from these two detection networks enables us to combine the strengths of both approaches, resulting in improved overall accuracy and higher model confidence. Although the specific application we consider here is pulmonary nodule detection, the approach of propagating and fusing uncertainty can be used in other multi-stage deep learning applications such as autonomous driving, where a first network detects and classifies objects in the scene followed by a second network that makes decisions as to which direction to steer the car.

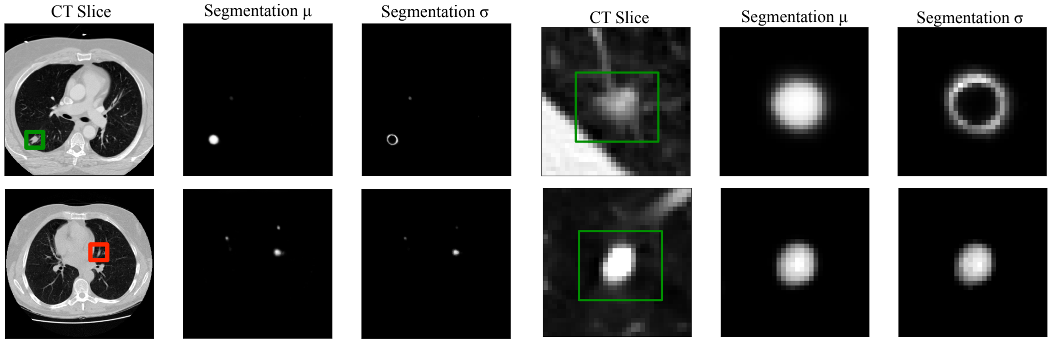
2 Two-Stage Bayesian CNN CAD Architecture & Uncertainty Fusion

As described in the previous section and shown in Figure 1, we have two Bayesian CNNs in our CAD architecture. Each network has a different local view of the full CT scan. For 2D segmentation (1st stage of the CAD architecture), we employ a 10-layer U-net architecture [8] as a base network and add dropout with probability of 0.5 after each convolutional layer. Using stochastic dropout at test time [9, 10], we obtain 50 Monte Carlo (MC) samples approximating the segmentation prediction probability distribution [11]. We note that, to our knowledge, Bayesian U-Net has not been employed for medical image segmentation before.

Segmentation alone is not sufficient for nodule detection as it results in a large number of false positives (see Section 4 for numerical results). To perform nodule detection with reduced false positive rates, we employ a 3D CNN (2nd stage of the CAD architecture) with three convolutional and two fully connected layers. This detection network operates on small 3D nodule candidates that are extracted based on stacked segmentation predictions from the 2D axial slices. The reason we use 3D volumes for nodule detection is that suspicious nodules have unique 3D features that differentiate them from normal lung lesions. Similar to the approach for segmentation, we perform approximate Bayesian inference via stochastic dropout at test time to obtain 50 MC samples approximating the nodule prediction probability distribution.

Since the segmentation network is Bayesian, we can approximate segmentation predictive mean (μ) and standard deviation (σ) via MC samples as shown in Figure 2. The top rows in Figure 2a and 2b show example cases where the segmentation network is able to accurately identify true nodule pixels as well as most of the normal lung tissue with high confidence, with the exception of nodule borders where the predictive uncertainty is high. On images for which the segmentation network performs well with high confidence, these segmentation predictions help improve the performance of the 3D nodule detection network, especially since the segmentation network is able to see larger context in full 2D slices whereas the detection network has 3D local context.

Segmentation prediction statistics are used as additional features for nodule detection by concatenating segmentation predictive mean and standard deviation with the original image to form a 3-channel composite image, as shown in Figure 1. We call this step 'Uncertainty Fusion'. Our goal is to make nodule detection easier for cases such as shown in Figure 2 top row. Structural similarities between nodules and normal lung features in 2D means that the segmentation network that only uses 2D context can produce false positives, as well as lower confidence true positives. We show such examples cases in the bottom row of Figure 2. In Figure 2a bottom row, there are false positives with high uncertainty, whereas in Figure 2b bottom row, there is a true detection with high uncertainty. To further improve performance, since there are difficult cases for which there is high overall uncertainty such as the ones shown in Figure 2b bottom row, we employ a second nodule detection network, trained using the original 1-channel image, without propagation of segmentation predictive uncertainty. As shown in the next section, combining predictions from these two detection networks enables us to combine the strengths of both approaches.



(a) Full size examples

(b) Zoomed examples (different than (a))

Figure 2: Example segmentation results. Left: Original axial CT slices. Green boxes and red box depict true positives and false positive after segmentation, respectively. Center: Segmentation mean μ. Right: Segmentation uncertainty represented by standard deviation σ. The top row in (b) indicates an ideal case with confident segmentation predictions for true nodule location and uncertain predictions for the edges whereas the bottom row indicates that the model is overall uncertain.

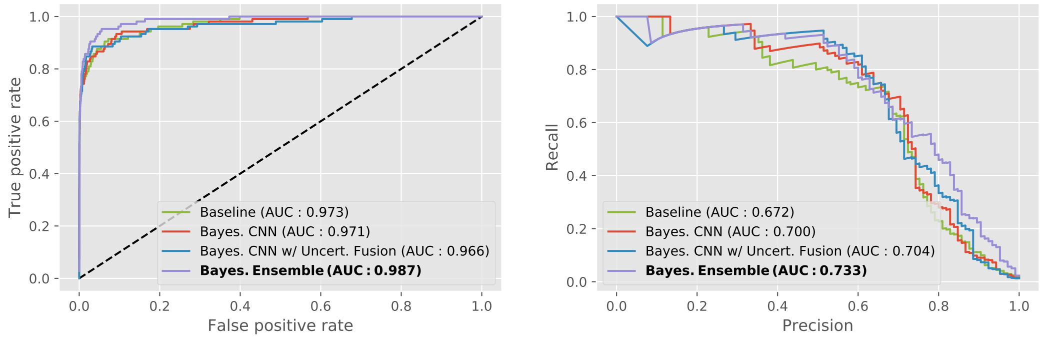
3 Experiments & Discussion

To evaluate the false positive reduction performance provided by the 3D nodule detection network, we use the LUNA16 dataset comprising 3D CT scans of 888 patients [7]. Among 888 patients, 601 have nodules with a total number of 1186 nodules. We split the dataset into an $80\%/10\%/10\%$ train, validation, and test split. After segmentation, we get a recall and precision of $\sim$92% and $\sim$0.9% on the test data, respectively. Figure 3 shows the performance of Bayesian 3D nodule detection networks on the test data. Note that for these figures the total number of positives is limited to those that are extracted by the 2D segmentation network, as our goal is to compare the performance of the 3D nodule detection networks. If the segmentation network misses a nodule, that nodule can no longer be recovered by the detection network since the candidates for the detection network are generated based on stacked segmentation predictions. We present both ROC and Precision-Recall curves because the candidate set is extremely imbalanced ($\sim$0.9% positive samples). As a performance benchmark, we also show results for a non-Bayesian 1-channel 3D CNN, called 'Baseline'. We refer to the Bayesian nodule detection network that fuses uncertainty (via 3-channel composite image) as 'Bayesian CNN with Uncertainty Fusion'. The Bayesian nodule detection network that only uses the original 1-channel image is referred to as 'Bayesian CNN'. Both detection networks have exactly the same architectures.

Figures 3 and 4 show that the Bayesian CNN with Uncertainty Fusion has comparable AUC with the the 1-channel Bayesian CNN, but has higher overall confidence (lower average predictive uncertainty) due to the extra information provided by the segmentation network. We compute average predictive uncertainty in Figure 4 via averaging standard deviations of predictions over all candidates. Example results on the validation set revealed that the Bayesian CNN with Uncertainty Fusion is able to make high confidence predictions for positive candidates similar to what is shown at the top row of Figure 2b, whereas it struggles for candidates similar to the bottom row. This is due to similar candidates that are negatives as shown at the bottom row of Figure 2a. In contrast, the 1-channel 3D CNN struggles with some of the small nodules that the 3-channel 3D CNN easily identifies thanks to the uncertainty information provided by the segmentation network.

To assess the dependence between predictions of the two Bayesian CNNs, we computed Spearman's rank corrrelation coefficient on the validation set resulting in $\rho = 0.346$ ($p < 10^{-5}$), which shows weak dependence between the predictions, confirming our manual inspection. Based on this information, we combined the two predictions via convex combination to create an ensembled prediction. We selected a weight of 0.5 on each prediction, because it maximized the AUC on the validation set. As shown in Figure 3, this ensembling approach resulted in significant improvement relative to either network operating independently. Further, denoting the test predictions of the two networks as y_1^* and y_2^*, we can show that $y_1^*|x^*, \mathcal{D}$ and $y_2^*|x^*, \mathcal{D}$ are two independent random variables, where x^* and $\mathcal{D}$ are the test input and the training data, respectively. Therefore, we can calculate the predictive variance of the ensembling approach for each ensemble prediction. Figure 4 shows that the ensembling approach results in a slight increase in model uncertainty compared to the 3-channel Bayesian CNN (due to the large difference between the uncertainties of the two models), but this uncertainty is still much lower than that of the 1-channel Bayesian CNN.

The performance of 3D Bayesian CNNs are comparable or better than the non-Bayesian baseline model. More importantly, unlike the non-Bayesian baseline, they provide model uncertainty/confidence information which could be extremely valuable for medical applications. We note that the performance improvement is more pronounced for the Precision-Recall curves since the positive class is extremely underrepresented in the data. We also show Brier scores [12] (weighted due to heavy class imbalance) of test predictions in Table 1 showing that the Bayesian CNN with Uncertainty Fusion provides better performance than the 1-channel Bayesian CNN in terms of Brier score, and that ensembling the two methods provides the best performance.



(a) ROC Curves

(b) Precision-Recall Curves

Figure 3: 3D nodule detection results and comparisons on test data for nodules extracted by the 2D segmentation network.

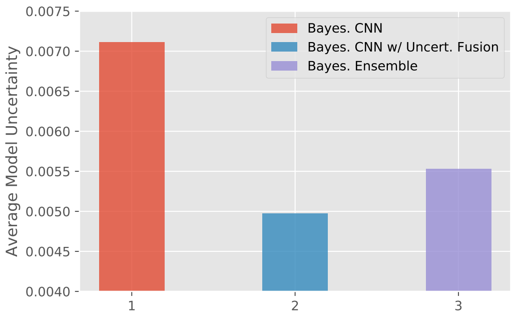


Figure 4: Average uncertainty of Bayesian CNN models.

Table 1: Brier Scores of 3D CNN Models

Model	Brier Score
Baseline	0.1045
Bayes. CNN	0.1214
Bayes. CNN w/ Uncert. Fusion	0.1010
Bayes. Ensemble	**0.0948**

4 Conclusion

Focusing on a CAD problem, we proposed a method to propagate and fuse uncertainty information in a multi-stage Bayesian convolutional neural network (CNN) architecture. Our experiments showed that propagating and fusing uncertainty improved the overall performance in terms of both final prediction accuracy and model confidence. Although the specific problem we considered in this paper was CAD, the proposed method could be used in a variety of applications for which the knowledge of uncertainty is important and informs subsequent decision making processes in the overall system. Example applications include medical image based diagnostics, autonomous driving, and robotics.

Our results pave the way for several interesting research questions that we will investigate as future work. Although we performed ensembling to combine the predictions from two detection networks, we will explore whether we can integrate this ensembling process within a single CNN architecture and have a single network learn how to best combine the information from different channels. We will also consider adaptive ensembling approaches, where predictions are combined by using the information about each prediction's uncertainty. In addition, we will investigate other approximate Bayesian inference techniques such as variational dropout [13] and assess how close the estimated uncertainties via different approximations are to true uncertainties. This information could be crucial for certain applications such as medicine, where overestimation of uncertainty may be preferable to underestimation. Finally, we will investigate whether we can separately model the uncertainty due to lack of training data and the uncertainty due to inherent noise in the data [14], which may have different implications for subsequent decision making processes.

References

[1] P.-P. Ypsilantis and G. Montana. Recurrent convolutional networks for pulmonary nodule detection in CT imaging. *arXiv:1609.09143*, 2016.

[2] A. A. A. Setio et al. Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks. *IEEE Trans. Medical Imaging*, 35(5):1160–1169, May 2016.

[3] H. R. Roth, L. Lu, J. Liu, J. Yao, A. Seff, K. Cherry, L. Kim, and R. M. Summers. Improving computer-aided detection using convolutional neural networks and random view aggregation. *IEEE Trans. Medical Imaging*, 35(5):1170–1181, May 2016.

[4] W. Zhu, C. Liu, W. Fan, and X. Xie. DeepLung: 3D deep convolutional nets for automated pulmonary nodule detection and classification. *arXiv:1709.05538*, 2017.

[5] J. Ding, A. Li, Z. Hu, and L. Wangy. Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks. *arXiv:1706.04303*, 2017.

[6] B. v. Ginneken et al. Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study. *Medical Image Analysis*, 14:707–722, 2010.

[7] A. A. A. Setio et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. *Medical Image Analysis*, 42:1–13, 2017.

[8] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[9] Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In *Proc. Int. Conf. Machine Learning (ICML)*, 2016.

[10] Y. Gal and Z. Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. In *Proc. Int. Conf. Learning Representations (ICLR)*, 2016.

[11] A. Kendall, V. Badrinarayanan, and R. Cipolla. Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In *Proc. British Machine Vision Conference (BMVC)*, 2017.

[12] M. H. DeGroot and S. E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society Series D (The Statistician)*, 32(1/2), 1983.

[13] D. P. Kingma, T. Salimans, and M. Welling. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[14] A. Kendall and Y. Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *arXiv:1703.04977*, 2017.