View Reviews

Paper ID

2689

Paper Title

Bayesian Neural Networks for Uncertainty Estimation of Imaging Biomarkers

Reviewer #1

Questions

2. Please provide a summary of the paper (a few lines)

The paper investigate four different segmentation uncertainty generation methods for the follow up group analysis and disease classification. The results shows that integrating segmentation uncertainty can lead to higher classification accuracy.

3. Please list the major strengths of the paper; you should write about a novel formulation, an original way to use data, demonstration of clinical feasibility, a novel application, or anything else that is a strong aspect of this work (bulleted list)

+ A good idea of investigating effect of integrating segmentation uncertainty in follow-up task.
+ Comparison of different uncertainty generation method is applaudable.
+ Results clearly demonstrate the effectiveness of the proposed method.

4. Please list the major weaknesses of the paper (bulleted list).

- Authors state that the integration of uncertainty in group analysis and follow up task has not been done before. This is clearly false as uncertainty effect in group analysis is done in Bayesian QuickNAT paper [1], which authors cite but doesn't mention. Similarly, propagating uncertainty from one task to the follow up task has also been done before [2]. This makes the novelty of the proposed method questionable.

- It is not clear, why coefficient values of Beta_4, which is related to diabetes status, if reported in the Table:1 instead of coefficient values related to uncertainty measures like IoU or CV^-1, which would reflect the usefulness of them in the group analysis.

[1] Roy et al. "Bayesian QuickNAT: model uncertainty in deep whole-brain segmentation for structure-wise quality control." NeuroImage 2019
[2] Mehta et al. "Propagating Uncertainty Across Cascaded Medical Imaging Tasks for Improved Deep Learning Inference." UNSURE, MICCAI 2019.

5. Please rate the clarity and organization of this paper

Good

6. Please provide detailed and constructive comments for the authors. Please also refer to our Reviewer's guide on what makes a good review: https://miccai2020.org/en/REVIEWER-GUIDELINES.html

-- It would be good, if authors can clarify if segmentation network was also retrained 1000 times, similar to the process mentioned in sec:3.4
-- Although, results demonstrate clear usefulness of uncertainty in increasing diabetes classification accuracy, there is no statistical significance analysis done for the same.
-- There is no clear winner among different uncertainty generation methods. Some explanation for the same would be appreciated.

Reviewer #2

Questions

2. Please provide a summary of the paper (a few lines)

This paper proposes the use of uncertainty estimates for the biomarker analysis task. They specifically use four probabilistic methods to estimate the segmentation uncertainty and use the associated confidence measure for the group analysis and disease classification.

3. Please list the major strengths of the paper; you should write about a novel formulation, an original way to use data, demonstration of clinical feasibility, a novel application, or anything else that is a strong aspect of this work (bulleted list)

Trying to make use of the prediction uncertainty to improve the disease classification task is the only strength of the paper to me.

4. Please list the major weaknesses of the paper (bulleted list).

- Methods #3 and #4 don't look suitable for the task in hand
- It is not clear how the uncertainty maps are generated
- The method employed for making use of uncertainty estimations in the disease classification is basic

5. Please rate the clarity and organization of this paper

Good

6. Please provide detailed and constructive comments for the authors. Please also refer to our Reviewer's guide on what makes a good review: https://miccai2020.org/en/REVIEWER-GUIDELINES.html

Comparing the first two methods (MC dropout and Full Bayesian) with the last two methods in generating reliable uncertainty maps and segmentation predictions are not valid as the last two methods are not designed for that purpose (as the authors also admit in page 7). Instead, they could have used methods such as Model Ensembles or MC-Dropconnect or M-heads for that purpose.

It is also unclear how they generated the uncertainty maps. Is it by computing the standard deviation over the predicted samples? How many samples were generated in each case? For example, for the MC Dropout method, how many rounds of MC simulations were used?

In any case, averaging the predictions for one sample and considering it as the prediction and use it for the next steps does not sound right to me for methods #3 and #4 as they initially were developed and proposed to generate a set of diverse but plausible segmentations.

Performance of the probabilistic methods are evaluated using dice score. While dice score is commonly used in the literature for comparing a deterministic prediction with a unique ground truth, it does not tell much about the distribution of segmentations and the quality of the generated uncertainty maps.

Replace Figure 2 with tables as the numbers are very close and not readable. Adding statistical analysis can illustrate more about the significance of the differences.

Equation (6) is unclear to me. It is not clear how this formulation gives higher importance to the samples with more confident predictions. Adding a couple of sentences on that or correcting the possible typos can help to clarify it.

Reviewer #3

Questions

2. Please provide a summary of the paper (a few lines)

The paper proposes the use of image segmentation to extract metrics that can serve as biomarkers for predicting diseases. However, to combat uncertainties within predictions that may negatively affect biomarker analysis, the paper propagates segmentation uncertainty to aid in better group analysis and predictions. Four varieties of BNNs are used for segmentation and experiments show the effect of uncertainties in statistical inference.

3. Please list the major strengths of the paper; you should write about a novel formulation, an original way to use data, demonstration of clinical feasibility, a novel application, or anything else that is a strong aspect of this work (bulleted list)

The paper compares a variety of methods for segmentation and shows that uncertainty can be used to improve accuracy in the subsequent statistical analysis.

4. Please list the major weaknesses of the paper (bulleted list).

The differences between aleatoric and epistemic uncertainties are seriously misunderstood in the paper, and no justification has been shown for comparing both the uncertainties (more details in comments).

5. Please rate the clarity and organization of this paper

Good

6. Please provide detailed and constructive comments for the authors. Please also refer to our Reviewer's guide on what makes a good review: https://miccai2020.org/en/REVIEWER-GUIDELINES.html

“image segmentation [7,10,11] have been developed that do not only provide the mode” - A BNN usually predicts the parameters of the posterior distribution. In most cases like the Gaussian one of the parameters happens to be the mean (and the mode) but in general the NN outputs the parameters of the distribution and not necessarily the mode. Please change the mode to parameters of an underlying distribution.

Out of four of the networks, the first two use epistemic uncertainty (stochasticity in model parameters) and the other two use aleatoric uncertainty (output is a probability distribution). Is the comparison of models with these kinds of uncertainty justified? In fact, Figure 3 shows that the 3rd and 4th models are overconfident in their uncertainty estimation. This is indeed a problem that occurs due to miscalibration in neural networks and their uncertainty [1] [2] and comparing epistemic with aleatoric uncertainties is not justified since they have semantically different meanings. How do these uncertainties affect the final prediction? Fig 2(a) seems to show that MC Dropout performs the best in terms of segmentation, but all four methods have similar results on prediction of liver volume, suggesting that the confidence value predicted by the NN is not informative.
The paper has completely missed out on the point that aleatoric uncertainties need to be calibrated, especially in the context of classification (segmentation) where the outputs are parameters of a categorical distribution.

“The probabilistic and hierarchical models were designed to learn annotations from multiple raters, while we only have annotations from a single rater, which may explain the lower stochasticity of these models in our experiments.” This is partially correct, but misleading. If that is the case, one might ask how do MC dropout and Full Bayesian output more meaningful uncertainties. The authors have misinterpreted the sources of uncertainty and how they should (not) be compared.

The use of CV-1 as a confidence metric is fine, but using it in group analysis sounds a bit problematic to me. This is because CV-1 is nothing but (mean volume from different outputs) / (std. volume by outputs). Neural nets can make systematic mistakes, and if the std is more or less similar for different images, then the regression coefficient \beta_5 can ideally be equal to std and all other coefficients can be zero leading to a good fit in terms of MSE error but zeroes out all other \beta which provides no information at all. However, in a more realistic scenario, the use of C = CV-1 can still interfere with the actual values of other \beta because this input feature is equal to the volume of the liver multiplied with some value and the output variable is also the volume.
This may also explain why IoU performs better than CV-1 in all experiments. Overall, it seems like CV-1 is a bad design choice given the inputs and outputs of the group analysis.

To check this, the paper can also show the coefficient \beta_5 for the “Variable” method and analyse if there are significant differences in the values assigned to them for both confidences.

Also, is volume supposed to linearly depend with the confidence value? I expected the “Instance” method to work better but it does not. Is there more analysis as to why this happens? If there are definite trends (like bigger volumes have more uncertainties) then the performance of “Variable” can be justified.

What is the “ground truth” for the volume used in group analysis? Is it computed from the expert segmentation or is present independently? This is not clear from the paper.

[1] Guo, Chuan, et al. "On calibration of modern neural networks." Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017.

[2] Jena, R., & Awate, S. P. (2019, June). A Bayesian Neural Net to Segment Images with Uncertainty Estimates and Good Calibration. In International Conference on Information Processing in Medical Imaging (pp. 3-15). Springer, Cham.