Building Trust in Deep Learning System towards Automated Disease Detection

Zhan Wei Lim,[1] **Mong Li Lee,**[1] **Wynne Hsu,**[1] **Tien Yin Wong**[2]
[1]School of Computing [2]Duke-NUS Medical School
National University of Singapore

Abstract

Though deep learning systems have achieved high accuracy in detecting diseases from medical images, few such systems have been deployed in highly automated disease screening settings due to lack of trust in how well these systems can generalize to out-of-datasets. We propose to use uncertainty estimates of the deep learning system's prediction to know when to accept or to disregard its prediction. We evaluate the effectiveness of using such estimates in a real-life application for the screening of diabetic retinopathy. We also generate visual explanation of the deep learning system to convey the pixels in the image that influences its decision. Together, these reveal the deep learning system's competency and limits to the human, and in turn the human can know when to trust the deep learning system.

Introduction

Recent progress in deep learning has enabled many medical imaging applications to achieve high accuracy in detecting diseases such as tuberculosis, skin cancer and diabetic retinopathy (Ting et al. 2017; Gulshan, Peng, and others 2016). Deep learning systems can potentially replace human to look at these medical images to reduce waiting time and cost, expand coverage of screening program, and deliver better outcome in public. Despite many advantages, they are rarely deployed in highly automated situations. A major obstacle to adoption of automated disease detection models is that it is difficult for human to trust the deep learning systems.

In contrast to deep learning systems, typical diagnostic tests are often well-understood at every step, from chemical reactions, responses measurement, to the final interpretation. Protocols can be established to control for variables that are known to interfere with the test to ensure its accuracy. The inherent complexity and "black box" nature of deep learning systems mean that it is difficult for human to grasp the computation applied to arrive at its prediction. We are left to guess the variables that may cause a deep learning system to fail such as camera model and parameters, photo-taking techniques, patient population, etc. Thus, we do not know when a deep learning system may fail.

From a machine learning perspective, the failure of a neural network on images arising from real-world deployment can be attributed to overfitting. That is, the deep neural network has overfitted to the training data and unable to generalize to the true data distribution. To prevent overfitting, the validation set should be close to real-world data distribution by covering a wide range variations that may appear in the real-world. However, without knowledge of the variations that matter, it is hard to construct an all encompassing validation dataset. Even if the deep learning system is restricted to a domain it is familiar with, such as a particular patient profile and camera, there is no guarantee that there will be no domain shift during its deployment.

Building trust is a crucial step towards the deployment of deep learning system in automated disease detection and can contribute to better validation methods. Diabetic retinopathy (DR) is a medical condition of the eye that arises due to diabetes. Deep learning systems have shown high accuracy in predicting the severity of DR from retina fundus images on a range of validation datesets. In this paper, we describe how we can build user trust in a deep learning system for a real-life application, specifically the screening of DR.

To enhance trust in the deep learning system, there should be measures that let the user knows when the output of the system is reliable and when it is not. We propose to use uncertainty estimation as a measure to know when to trust the deep learning system output. A well-calibrated deep neural network may be uncertain about its output when it is applied to an image that comes from a different camera, or from a poorly taken photo, or rare phenotype of the disease that is not well represented in training data, etc.

We apply stochastic batch normalization (Atanov et al. 2018) to obtain uncertainty estimation on a deep neural network trained for detecting diabetic retinopathy. Unlike previous works that demonstrate out-of-dataset detection by artificially splitting a dataset by classes (CIFAR5) or generating new images by rotation (notMINIST), we observe that domain shift in real-world dataset is more subtle. The neural network may still generalize properly on a subset of a new dataset but make mistakes on another part of the dataset. We present novel analysis of uncertainty estimation and propose new uncertainty estimate on a real-world dataset. We show that uncertainty estimates can be a used to inform us about the reliability of the deep neural network on images.

We can use visual explanation for deep learning system as a proxy to understanding the inner working of the system. We define the qualitative criteria for visual explanation to enhance trust and apply Integrated Gradient to generate compelling visualization on a deep neural network trained to detect diabetic retinopathy. Intuitively, if the visual explanation of the deep learning system corresponds to how a human would explain his decision to another human, it makes the deep learning system trustable. We also propose a novel approach to sharpen visualization from deep neural network using the stochasticity of the stochastic batch normalization layers in the neural network.

Related Works

Human's trust in a machine learning model is deeply related to our optimisim of generalizability of a model to unseen data. Generalization errors of machine learning models has been well studied in machine learning literature. A model's generalizability is often viewed as a trade-off between bias and variance by ways such as limiting the model's complexity (Friedman, Hastie, and Tibshirani 2001). Studies on generalization error are concerned with the error at dataset level, i.e., its error across the entire population of inputs. However, this work aims to quantify the generalizability of the model to each input example by means of uncertainty estimation, so that we can know when to trust the model.

Unlike state-of-the-art deep neural networks that are trained using maximum-likelihood regime, Bayesian probabilistic models are able to capture uncertainty over its parameters and thus give better uncertainty estimates of its predictions. However, Bayesian inference on deep neural networks is intractable due to its nonlinearity. Various methods for practical approximate inference of Bayesian neural network have been proposed (Louizos and Welling 2017; Hernández-Lobato et al. 2016). While these methods are principled, they have not been applied to computer vision tasks such as ImageNet (Deng et al. 2009) or on medical images.

Due to training difficulties and computational cost associated with Bayesian neural networks compared to recent non-Bayesian deep neural networks, another line of research estimates predictive uncertainty using non-Bayesian deep neural network. Some works interpret existing components of deep neural network as Bayesian probabilistic models to obtain uncertainty estimation without little modification to the training procedure and network architecture. (Gal and Ghahramani 2016) established the equivalence between neural network with dropout (Srivastava et al. 2014) applied before every weight layers and an approximation probabilistic deep Gaussian Process (Damianou and Lawrence 2013).

Another work (Atanov et al. 2018) interpret the mean and variance of mini-batch statistics used in batch normalization (Ioffe and Szegedy 2015) as random variables since they depend on stochastic shuffling of training examples into mini-batches during training. Thus, the neural network with batch normalization layers can be viewed as a probabilistic model during training. During inference, the stochasticity due to dropout units and batch normalization layers is removed by averaging the predictions and normalizing with long term means and variances respectively. For both works, uncertainty estimations are obtained by Monte-Carlo evaluation of the network with the same randomness as during training.

Ensemble of deep neural network and adversarial training has been proposed to get predictive uncertainty estimation (Lakshminarayanan, Pritzel, and Blundell 2017). Ensembles have been used to boost predictive accuracy and produce well-calibrated models. The authors proposed to use scoring rules to evaluate their models for calibration and showed that their models are less confident on out-of-dataset inputs.

Feature visualization methods have been proposed to help human interpret the inner working of deep neural network. Feature attributions point to parts of the input that influenced the neural network's output. Feature attributions may be achieve by tracking the difference in the neural network output when the input image is perturbed (Zeiler and Fergus 2014). Another approach to feature attributions is to backpropagate the prediction scores through the network to the input. However, it is hard to evaluate the strength of each method without an objective metric. Recent works have focused on establishing desirable properties (Sundararajan, Taly, and Yan 2017) of attribution methods and analysing the failure cases.

Uncertainty Estimation

It is difficult to predict all variables that may adversely performance of deep learning systems and control for them. A basic quality control step is to exclude images of poor quality. We can train a deep neural network to predict and exclude images of poor quality e.g., retina fundus images that are too dark or have cataract that can interfere with disease detection. This is implemented in deep learning systems for referable DR (Ting et al. 2017; Krause et al. 2018).

As deep learning systems are data-driven, another approach may be to standardise the data acquisition, such as using the same camera model, ensure same camera parameters, taking the same view of retina fundus, and using the system for the same population it was trained on, etc. However, this may reduce the applicability of such systems, and it may be impractical to replicate exactly how the data arise. Even if we can control the data, there may be variables such as occurrence of rare phenotypes that the deep learning system have not seen before in its training and validation data. Deep neural networks are also known to be susceptible to adversarial attacks where imperceptible image perturbation can arbitrarily change its output. Hence, we cannot be sure whether there are more variables that can affect the system.

We propose to use uncertainty estimation for deep learning system's predictions to inform us when to trust the predictions. The deep learning system should give high confidence predictions when the predictions are likely to be correct and low confidence when the system is unsure.

However, it is not straightforward to evaluate uncertainty estimations because there is no ground truth for uncertainty of a prediction. For classification problems, metrics such as average classification does not take into account the uncertainty of the predictions. Proper scoring rules (Gneiting and

Raftery 2007) can be used to evaluate the quality of predictive uncertainty. Many common loss functions used in neural network such as softmax cross entropy are proper scoring rules (Lakshminarayanan, Pritzel, and Blundell 2017).

To preserve the accuracy of the deep neural network for detecting DR, we want to keep the network architecture and training regime as close as possible to the best models that we have. Thus, we use stochastic batch normalization layers (Atanov et al. 2018) for uncertainty estimation. Stochastic batch normalization can be applied to existing trained neural network. We iterate through the training examples to collect means and variances of mini-batches statistics. At inference time, we use these statistics to simulate sampling a random mini-batch. Due to our choice of network architecture (Residual network, (He et al. 2016)), dropout unit is not a default component. Hence, we did not use Monte-Carlo dropout to obtain uncertainty estimations.

In our experiments, we show that prediction error are correlated with high estimated uncertainty. We also show that uncertainty estimates can be used to exclude a small number of high uncertainty predictions to improve performance on the rest of the dataset.

Visual Explanations

Visual explanation, either via features visualization that look at features of the network, or features attributions that look at what the network sees in an image, are attempts to project the concepts detected by deep neural network to a visual representation that human can understand. It may be impossible to fully grasp the computations undertaken by deep neural networks. Hence, model interpretability is often a misnomer because the visualization may not represent the inner working of the deep neural networks. Despite the incompatibility between what a human can understand and what the network actually understands, trust can still be derived by seeing that the deep neural network picks out visual cues that human typically picks to justify its decision.

For a visual explanation to enhance trust, it has to be specific and relevant. It should only highlight the parts of image that is most relevant to how human justify its decision. For example, in visual explanation of DR detection, the visual explanation should only highlight the lesions in the image but not the vessels.

Sharp visualization is necessary for the visual explanation to pinpoint pixels of interest (specificity). Inspired by SmoothGrad (Smilkov et al. 2017), we propose stochastic batch norm-SmoothGrad. The key idea is to sample images similar to the target image (by adding Gaussian noise to each pixel) and take the average of the visualization generated. Instead of sampling similar input images, we use the stochasticity in stochastic batch normalization to sample activations within the deep neural network.

Experiments

We trained our deep neural network classifier on retinal images of patients with diabetes who participated in the Singapore national DR screening program (SiDRP) between 2010 and 2013 (SiDRP) and test it on both retina images collected from the same program between 2014 and 2015. We also test the deep neural network on test set of Kaggle DR dataset. Kaggle DR dataset can be considered "out-of-dataset" because it comprises of mainly Caucasian patients while SiDRP consists of mainly patients of Chinese, Malay, Indian ethnicity. Each image belongs to one of the five levels of DR severity: level 0 No DR, level 1 Mild DR, level 2 Moderate DR, level 3 Severe DR, level 4 Proliferative DR.

Dataset	Model	Brier loss	F1 score
SiDRP14-15	SBN	0.0113	0.492
	BN	0.0104	0.538
Kaggle	SBN	0.104	0.667
	BN	0.110	0.643

Table 1: Brier loss and F1 scores of models.

We trained a Residual Network with 50 layers to give four binary outputs: Mild DR or worse, Moderate DR or worse, Severe DR or worse, and Proliferative DR. The neural network weights are initialized with a pre-trained ImageNet model. In our experiments, we use the output for Moderate DR or worse unless otherwise stated.

For each image in the test set, we obtain 4 rotations at multiples of 90 degrees of the image For stochastic batch normalization (SBN) network, we ran the neural network 3 times for each rotation, for a total of 12 runs per image. For each inference, we sample a new mean and variance at normalize the activations at every stochastic batch normalization layers. From these runs, we obtain the mean and standard deviation. We estimate the uncertainty of the neural network output by the standard deviation and the entropy of the mean predictive distribution. We also perform inference on the same neural network using batch normalization (BN).

Performance of Models

We establish the performance of the models on each dataset. We report F1 score and Brier loss, which is a proper scoring function that measures the probability calibration of a model. Performance measures for SBN are obtained using the means of 12 runs.

Table 1 shows the results. We observe that the Brier loss

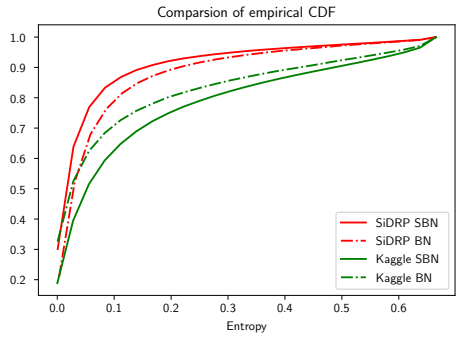

Figure 1: Entropy CDF on SiDRP14-15 and Kaggle.

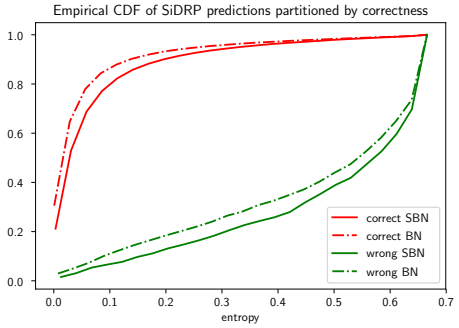

Figure 2: Entropy CDF on SiDRP partition by correctness.

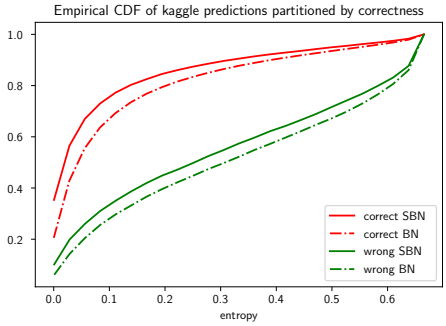

Figure 3: Entropy CDF on Kaggle partition by correctness.

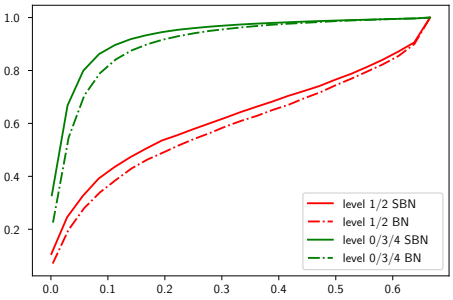

Figure 4: Entropy CDF on SiDRP14-15 by DR levels.

on SiDRP14-15 is much lower than Kaggle because of its similarity to the training dataset.

Uncertainty of Datasets

Previous works on uncertainty estimation use true out-of-dataset evaluation by training the network on 5 of the 10 CIFAR10 classes and evaluate on the other 5 classes or evaluating network trained on MNIST on nonMNIST data. However, the difference between examples in datasets that arise from real-world applications is usually less drastic.

Figure 1 shows the entropy CDF on SiDRP and Kaggle datasets. On an empirical CDF plot, the closer to the bottom and to the right means there are more predictions with higher entropy, and thus the classifier is more uncertain. Both deep

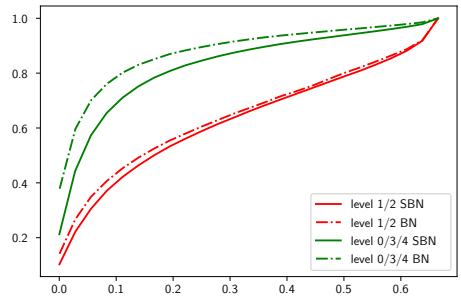

Figure 5: Entropy CDF on Kaggle by DR levels.

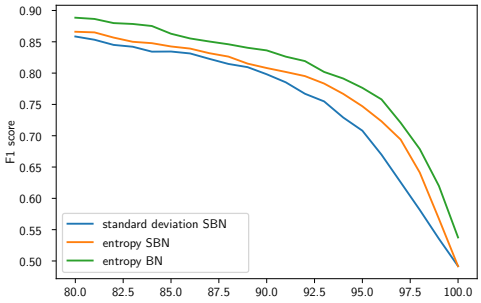

Figure 6: F1 scores on SiDRP14-15 as higher uncertainty predictions are included.

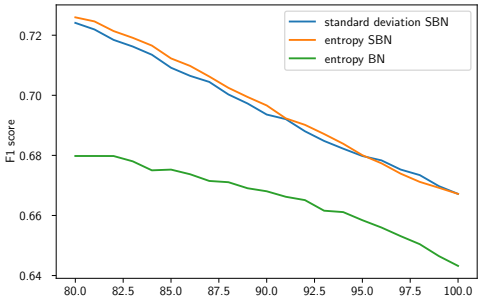

Figure 7: F1 scores on Kaggle as higher uncertainty predictions are included.

neural networks with BN and SBN are more confident on SiDRP than on Kaggle. This is expected as Kaggle is considered "out of dataset". Between SBN and BN predictions, SBN is slightly more confident than BN on SiDRP, but BN is more confident than SBN on Kaggle. This suggests that BN is over-confident on Kaggle compared to SBN.

Uncertainty by Prediction Correctness

To further evaluate uncertainty estimations, we partition each dataset into a set of correct predictions, and another set of incorrect predictions. We expect correct predictions to

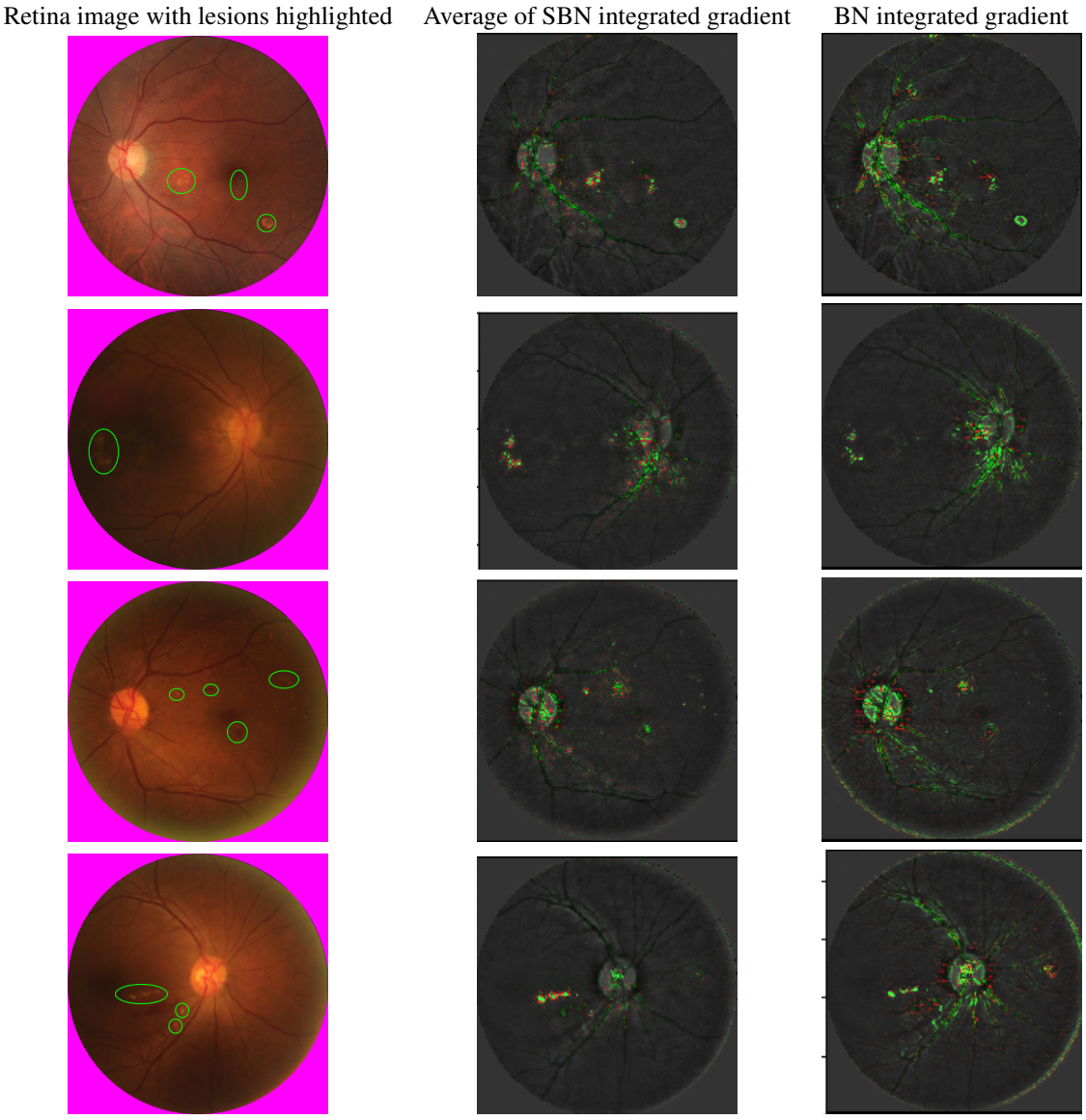

Figure 8: Visual explanation using integrated gradient.

have higher confidence (lower uncertainty) while incorrect predictions should have lower confidence.

Indeed, this is the case for both SiDRP14-15 and Kaggle dataset as shown in Figures 2 and 3 respectively. The curves for incorrect predictions are lower and to right than the correct predictions. The difference between the curves of SBN and SB follows that of the overall dataset, i.e. SBN is more certain than BN on SiDRP14-15 and less certain on Kaggle.

Uncertainty by DR Level

While we only look at the decision of Moderate DR or worse, the ground truth DR level should affect the uncertainty estimations. We expect the uncertainty of predictions to be higher for ground truth DR levels near the decision boundary (DR level 1, 2) than DR levels further away (DR level 0, 3, 4).

Figures 4 and 5 show the entropy CDF by DR levels for the SiDRP14-15 and Kaggle dataset respectively. We observe that our uncertainty estimations are consistent with finer-grain ground truth even though it is not explicitly modeled by the output.

F1 Scores by Excluding High Uncertainty

In a highly automated disease screening setting, we may use uncertainty estimates to flag predictions where the deep neural network is uncertain for human attention. In this experiment, we sort the predictions by uncertainty as measured by standard deviation of probability output for SBN model and entropy for both BN and SBN models. We want to look at the F1 score of subset of predictions with low uncertainty.

Figures 6 and 7 plot the F1 score on predictions starting from the 80 percent least uncertain predictions to 100 percent (full test set). For both SiDRP114-15 and all measures of uncertainty, the exclusion of high uncertain predictions can significantly improve compared to random exclusion.

The curves of F1 scores for SiDRP14-15 drop sharply near the right end of x axis. Hence, by excluding a small percentage of high uncertain predictions, the F1 score can be improved. Entropy of SBN is a better measure of uncertainty for exclusion as it maintains higher F1 score than standard deviation of SBN. The curve for entropy of BN is higher than SBN because of higher F1 score for BN. For the Kaggle test set, the SBN model has better F1 score, but the curves for entropy and standard deviation of SBN give similar results.

Visual Explanation

For the visual explanation to be specific, the visualization should be at the same resolution as the input image. We use Integrated Gradient (Sundararajan, Taly, and Yan 2017) that satisfies our criteria to generate visualization of the neural network. We smooth the visualization to focus on important features using average for five attributions on the same image using stochastic batch normalization. This achieved an effect similar to SmoothGrad to give sharper visualizations.

In Figure 8 where regions containing lesions in the original retina image are highlighted, we see that integrated gradient of BN highlights more of the vessels resulting in a visualization that is less specific compared to visualizations from the average of the SBN integrated gradient.

Conclusion

We have shown that uncertainty estimation can be helpful in enhancing trust in the deep learning systems, as a step towards deploying a deep learning system for the real-life screening of diabetic retinopathy. Predictions that the deep learning system is unsure of can be flagged for human attention. By excluding high uncertainty predictions, we show that accuracy can be improved. We proposed to use stochastic batch normalization to obtain uncertainty estimations and showed that its probabilities outputs are better calibrated than regular batch normalization on an external validation dataset. Stochastic batch normalization may be used to sharpen visual explanation, thus making it easier to convey human-interpretable concepts behind the decision of the deep learning system. Using uncertainty estimations, human may know when to trust the deep learning system and why it is trustable through visual explanation.

Acknowledgements

This work is supported by research grant R-252-000-A19-490. We thank the anonymous reviewers for their feedback.

References

Atanov, A.; Ashukha, A.; Molchanov, D.; Neklyudov, K.; and Vetrov, D. 2018. Uncertainty estimation via stochastic batch normalization. *arXiv:1802.04893*.

Damianou, A., and Lawrence, N. 2013. Deep Gaussian processes. In *Artificial Intelligence and Statistics*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2001. *The elements of statistical learning*, volume 1.

Gal, Y., and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*.

Gneiting, T., and Raftery, A. E. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102(477):359–378.

Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M. C.; Wu, D.; Marayanawamy, A.; et al. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama* 316(22).

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*.

Hernández-Lobato, J. M.; Li, Y.; Rowland, M.; Hernández-Lobato, D.; et al. 2016. Black-box α-divergence minimization. *International Conference on Machine Learning*.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*.

Krause, J.; Gulshan, V.; Rahimy, E.; Karth, P.; Widner, K.; et al. 2018. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology* 125(8).

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*.

Louizos, C., and Welling, M. 2017. Multiplicative normalizing flows for variational bayesian neural networks. *arXiv:1703.01961*.

Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; and Wattenberg, M. 2017. Smoothgrad: removing noise by adding noise. *arXiv:1706.03825*.

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).

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. *CoRR* abs/1703.01365.

Ting, D. S. W.; Cheung, C. Y.-L.; Lim, G.; et al. 2017. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *Jama* 318(22).

Zeiler, M. D., and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*. Springer.