PRUV 2019: Assessing the generalization performance of Bayesian convolutional neural networks on a pneumonia-screening task

Tyler Lian
with Dr. Paul Bendich, Ph.D.

May 27 to July 7, 2019

1 Introduction

Soon after the advent of the medical X-ray came the first radiologists—doctors to read, label, and interpret the growing range of medical images now essential for modern diagnosis and treatment. It would seem today that computational algorithms are ready to take up the mantle. Much like how a doctor-in-training must first learn from case studies and previous experiences, machine learning algorithms are able to "learn" from inputted "training" images to perform as well if not better than human radiologists on important tasks, including the classification of medical images as normal or pathological ([1]).

Many of these successes can be attributed to convolutional neural networks (CNNs), which have been put forward as promising machine learning tools for the automated classification of images, including medical ones. Researchers, however, less often assess—among other crucial benchmarks of trustworthiness—the extent that these algorithms are able to generalize beyond the limited scope of their initial training data. It is true that, in machine learning, it is standard practice to set aside a subset of data (known as the test set) before training to fairly evaluate model performance on unseen data. This test set is assumed to be fully representative of the relevant data that the model has yet to encounter, but confounding factors in the available data may hinder the model's true generalization performance. To this end, in [2], Zech *et al.* trained convolutional neural networks (CNNs) on single hospital databases to classify chest X-rays as either normal or pneumonic; however, they also showed that these CNNs could not replicate the same level of performance when tested on X-rays from other hospital systems.

Unlike human doctors, who recognize and interpret medical images using their knowledge of anatomy and physiology, algorithms have no such insight into the biology of the human body. Accordingly, where physicians consistently see bones and organs, algorithmic models perceive only rows and columns of pixels with varying intensity, which the algorithms are free to leverage in unexpected

and arbitrary ways. An automated tool that produces accurate diagnoses for X-ray images from the Duke University Hospital, for example, may not be accurate for X-rays at the UNC Medical Center or even other X-rays at Duke taken using a different X-ray machine, at a different resolution, or with some other overlooked anomaly. This is very dangerous for patients, whose health and well-being depend on the reliability of these tools.

One potential solution is the Bayesian convolutional neural network (BCNN), a probabilistic variant of the standard CNN. This project seeks to assess whether the probabilistic uncertainty information incorporated in BCNN methods confers an improved generalizability. In an experimental set-up similar to [2], CNNs and BCNNs were evaluated on a pneumonia-screening task using chest X-rays from one of two different public hospital databases. The "internal" (test source same as training source) and "external" (test source different from training source) generalization performances of the different models were assessed.

This report first provides a basic overview of the mathematics underlying Bayesian neural networks and Bayesian convolutional neural networks before discussing the methods, results, and possible future directions of this project.

1.1 Bayesian neural networks and convolutional neural networks

Under certain conditions, standard (i.e., popular) neural network methods can be re-interpreted as familiar concepts in probability. From a probabilistic perspective, the standard neural network is a model $P(y|\mathbf{x}, \mathbf{w})$, which takes in an input $\mathbf{x} \in \mathbb{R}^p$ and assigns a probability to each possible class $y \in \mathcal{Y}$. In classification, $\mathcal{Y}$ is a finite set.

Given training data D composed of pairs of inputs and labels $(\mathbf{x}_i, y_i)$, the typical task of the neural network is to "learn" model parameters (or weights) $\mathbf{w}$ by gradient descent and backpropagation such that the categorical cross-entropy loss function $\mathcal{H}(D)$ is minimized. The optimal set of weights $\hat{w}$ that minimizes cross-entropy loss is equivalent to the maximum likelihood estimate (MLE). This is clear from the definition of $\mathcal{H}(D)$,

$$\mathcal{H}(D) = \sum_i y_i \log P(y|\mathbf{x}_i, \mathbf{w}) + (1 - y_i) \left(1 - \log P(y|\mathbf{x}_i, \mathbf{w})\right).$$

Point estimates like the MLE, however, lose important probabilistic information about the model parameters. In general, Bayesian neural networks put neural networks in a proper Bayesian framework by representing the set of weights $\mathbf{w}$ with an updateable distribution rather than single estimates.

Thus, given labelled image data D and a prior distribution over the weights $P(\mathbf{w})$, the task of the Bayesian neural network is to compute the posterior distribution for the model parameters $P(\mathbf{w}|D)$. By Bayes' theorem,

$$P(\mathbf{w}|D) \propto P(D|\mathbf{w})P(\mathbf{w}).$$

From there, the posterior predictive distribution can be derived as

$$P(y^*|\mathbf{x}^*, D) = \mathbb{E}[P(y^*|\mathbf{x}^*, \mathbf{w})] = \int P(y^*|\mathbf{x}^*, \mathbf{w}) P(\mathbf{w}|D)\, d\mathbf{w},$$

which returns the probability that an image $\mathbf{x}^*$ is in a class y^* given the training data D. By integrating over all possible network weights, the prediction incorporates information about uncertainty, similar to averaging via an ensemble method.

For all but the simplest networks, the derivation of the posterior distribution $P(\mathbf{w}|D)$ is intractable. Although Bayesian approaches to neural network learning have their origins in landmark papers published almost thirty years ago (e.g., [3], [4], [5]), slow and unwieldy computational steps made these methods impractical for processing large amounts of data and held them back from more mainstream attention.

The past decade has seen a resurgence of interest in Bayesian neural networks. In [6], Graves introduced a stochastic approach to earlier variational inference techniques that could be more easily implemented with standard gradient descent. In the absence of an analytic form of the posterior, a stochastic optimizer is instead used to find parameters θ such that a new variational distribution $q(\mathbf{w}|\theta)$ of arbitrary family (e.g., Gaussian) resembles the posterior $P(\mathbf{w}|D)$ as much as possible. The Kullback-Liebler (KL) divergence quantifies the difference between the two probability distributions.

Since then, various approaches have been further developed to improve the viability of Bayesian neural networks. A *local reparameterization trick* reduces the variability of stochastic gradients via more efficient sampling ([7]). This trick has been generalized to be compatible with a standard backpropagation scheme that uses gradient descent ([8]). Another improvement is the expansion of variational techniques to more complicated variational families ([9], [10], [11]). Alternatives to variational inference approaches have also been proposed, including probabilistic backpropagation ([12], [13]), stochastic gradient Langevin dynamics ([14]), and deep ensemble methods ([15]).

One of the most prolific techniques in practical application, however, has been Monte Carlo Dropout ([16]) and its convolutional neural network extension ([17]), which has been used for disease detection in images ([18]), image segmentation ([19]), including for medical images ([20]), and autonomous driving ([21]).

Monte Carlo Dropout (MC-DO) has been able to capture more widespread interest because its implementation is simple. It employs the already widely used Dropout regularizer ([22]) to approximate a Bayesian procedure. Standard CNNs can be easily modified to be Bayesian by inserting Dropout layers after each layer of convolution. During testing, predictions are made by averaging the model's outputs over T runs for each image, as follows:

$$P(y^*|\mathbf{x}^*, D) \approx \int P(y^*|\mathbf{x}^*, \theta) q(\theta|D)\, d\theta = \frac{1}{T} \sum_{t=1}^{T} P(y^*|\mathbf{x}^*, \hat{\theta}_t),$$

where $\hat{\theta}_t$ represents a subset of the trained network weights on the t-th run, during which a proportion of the weights in each layer are randomly "dropped" to zero.

Similarly, the variance of the T outputs approximates the variance of the predictive posterior distribution, which is used as a measure of uncertainty, i.e. how sure is the model of its prediction given a specific image.

This project focuses on MC-DO because of its relative popularity in applications. It also considers a recent modification of MC-DO that makes use of newer Batch Normalization techniques in lieu of Dropout ([23]). This method, referred to in this report as MC-BN, is implemented in a manner analogous to MC-DO. Other BCNN methods, like multiplicative normalizing flows ([10]), are left for further research.

2 Methods

2.1 Datasets and Image Preprocessing

Chest X-rays were obtained from two different hospital systems: the NIH Clinical Center ($n = 112,120$, inpatient and outpatient 1992-2015) and Stanford Medical Center ($n = 224,316$, inpatient and outpatient 2002-2017). Images were pre-annotated with zero (absence of other pathologies) or one or more diagnostic labels extracted from radiological reports using natural language processing, as described in [24] and [25]. In the NIH database, bounding boxes were also included for about 1000 images, given as pixel coordinates in a separate list. Both databases were anonymized and are publicly available to researchers.

The Stanford database included labels for 14 common radiographic observations (including pathologies), whereas the the NIH database included labels for 14 lung pathologies. Only the seven labels the two databases had in common (Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Pneumonia, Pneumothorax) were considered during training. Further, in the Stanford database, the presence of each of the 14 observations were labeled as either positive, negative, or uncertain. For the sake of simplicity, all uncertain labels were interpreted as positive, as a conservative assumption.

Since the NIH database contained only frontal-view X-rays, lateral-view images were removed from the Stanford data. All remaining images were resized to 224 by 224 pixels for the sake of computational speed and limited resource capacity. In addition, images were rendered in RGB format and normalized by channel in order to take advantage of transfer learning on the ImageNet database. The two databases, after all changes described, are summarized in Table 1.

Both datasets were split into train and test subsets; images from the same patient were ensured to appear in only one of the two subsets. During training only, datasets were augmented by random horizontal flips.

	NIH Clinical Center	Stanford Univ. Hospital
No. X-rays (train; test)	112,120 (86,524; 25,596)	191,229 (165,630; 25,599)
No. unique patients	30,805	64,734
Age, mean (SD), years	46.9 (16.6)	60.7 (17.8)
Female (%)	48,780 (43.5%)	78,969 (41.3%)
Pneumonia (%)	1,353 (1.2%)	4,683 (2.4%)
Atelectasis (%)	11,535 (10.3%)	29,795 (15.6%)
Cardiomegaly (%)	2,772 (0.2%)	23,451 (12.3%)
Consolidation (%)	4,667 (4.2%)	13,015 (6.1%)
Edema (%)	2,303 (2.1%)	49,717 (26.0%)
Effusion (%)	13,307 (11.9%)	76,963 (40.2%)
Pneumothorax (%)	5,298 (4.7%)	17,700 (9.3%)

Table 1: Summary of the NIH and Stanford chest X-ray databases, after initial preprocessing. Age data was available for: 112,077/112,120 NIH, 191,226/191,229 Stanford. Sex data was available for: 191,228/191,229 Stanford.

2.2 Models and Model Training

The performances of three neural network architectures were studied for this pneumonia-screening task. The Keras API with a TensorFlow backend was used for the models and model training ([26]). Two BCNNs, one implementing MC-DO and the other MC-BN, were both 44-layer models derived from VGG-19 architecture ([27]). One standard CNN, a 121-layer DenseNet model ([28]), was considered as well for comparison.

All models were designed to be multi-class classifiers, i.e. labels were not mutually exclusive. As a form of class balancing, models were designed to classify seven labels (Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Pneumonia, Pneumothorax), but ultimately only their performance on pneumonia-screening task was evaluated. All models were initialized with weights pre-trained on ImageNet.

Models used the Adam optimizer with weight decay to minimize a binary cross-entropy loss. The learning rate was decreased by a factor of 10 on plateau. Training stopped when the loss failed to improve after 3 epochs to prevent overfitting. Further details, including model hyperparameters, can found in the code on this project's GitHub repository: `https://github.com/tylerwlian/bcnn-generalize`.

2.3 Experimental Design

Each model was trained on the training set of one of the two hospital databases and then evaluated on the test sets from both hospitals. The "internal" test refers to the test set whose hospital source is the same as the training set, and

the "external" test refers to the other.

AUCs, the areas under the receiver-operator characteristic (ROC) curve, were reported for the internal and external tests as the primary measure of good performance for a given trained model. In broad terms, a higher AUC indicates a better ability to distinguish between true positive and true negative cases. Delong's test for ROC curves was used to determine whether differences in AUC were statistically significant, assessed at a significance level of 0.05.

AUC is preferred to accuracy as a measure because it is independent of the chosen binary threshold; however, accuracy, sensitivity, and specificity were also reported. The threshold was chosen to guarantee a sensitivity of at least 0.95, since false positives (taking unneeded precautionary measures) would be preferred to false negatives (not detecting the pneumonia) in a clinical setting. This choice of threshold is not meant to compete with state-of-the-art results, but rather to assess the basic techniques that underlie today's most popular deep learning tools.

Further exploration can be done with the BCNN models. Uncertainty was quantified in the MC-DO and MC-BN models by the predictive variance, computed over 50 runs of the model. The predictive variance of the images was plotted as a histogram, to compare the distributions of uncertainty for the internal and external databases.

2.4 Class Activation Maps

For each trained model, gradient-weighted class activation maps (Grad-CAM) were used as a quick method to visualize the regions of the image that were most relevant to the model's predictive outcome ([29]). For the BCNN models, maps were created by averaging Grad-CAM heatmaps over 100 runs, an ad-hoc procedure meant to take into account the probabilistic nature of the BCNNs. Bounding boxes were also superimposed on the NIH images when applicable.

3 Results

3.1 Databases

The difference in the proportion of pneumonia between the two databases is significant at a 0.05 level (P-value < 0.00001), as is the difference for all the pathology labels. As demonstrated in [2], any between-database differences may be exploited by the deep learning models, disregarding the visual presentation of pathology.

3.2 Internal and External Performance

The standard CNN was able to reproduce the results of [2]. External testing significantly suffered relative to internal testing, as summarized in Table 2.

Train	NIH		STF	
Test	NIH	STF	NIH	STF
AUC (CI)	0.643* (0.619-0.667)	0.558 (0.535-0.580)	0.663 (0.639-0.687)	0.724* (0.702-0.745)
Acc	0.164	0.107	0.127	0.178
TPR	0.952	0.951	0.952	0.951
TNR	0.150	0.087	0.112	0.160
PPV	0.021	0.025	0.020	0.027
NPV	0.994	0.986	0.992	0.993

Table 2: Various measures of internal and external model performance for the standard CNN (DenseNet). The confidence level is 0.95. Legend: TPR=true positive rate (sensitivity), TNR=true negative rate (specificity), PPV=positive predictive value (precision), NPV=negative predictive value.

BCNNs, however, also demonstrated similar trends. Both MC-DO and MC-BN models performed significantly worse on external versus internal test sets, for both hospital training sets. The results are summarized in Table 3.

3.3 Model Uncertainty

With predictive variance as a measure of uncertainty, the histograms in Figure 1 appear to provide further evidence that the BCNN models had trouble generalizing to the external database. The MC-DO models seemed to be more certain of internal classifications relative to external ones. This trend was less clear for MC-BN models.

One advantage of the BCNN was that uncertainty information may be helpful for anticipating model errors. For the MC-DO models, correct classifications were associated with lower uncertainty for both hospital databases, as seen in Figure 2.

3.4 Class Activation Maps

For a few NIH X-rays positive for pneumonia, Grad-CAM heatmaps were created with the trained CNN and MC-DO models. (Maps for MC-BN could not be reliably generated; more troubleshooting is needed.) Maps for four pneumonia-positive X-rays are presented, two from the NIH database and two from the Stanford database; bounding box information was available for the NIH images only. Since this sample of images is small, more research is required before more conclusive statements can be made about model trends or behavior.

For a given X-ray image, the areas of "high importance" were not consistent from model to model. Sometimes, important regions were at or near the labelled bounding boxes (for the NIH images), but at other times they were concentrated in the corners of the image, away from the anatomical chest altogether. This

Train	NIH		STF	
Test	NIH	STF	NIH	STF
AUC (CI)	0.606* (0.582-0.630)	0.533 (0.511-0.555)	0.641 (0.617-0.666)	0.697* (0.675-0.719)
Acc	0.153	0.080	0.148	0.125
TPR	0.952	0.951	0.952	0.951
TNR	0.138	0.059	0.133	0.105
PPV	0.020	0.024	0.020	0.025
NPV	0.993	0.980	0.993	0.989

(a) MC-DO

Train	NIH		STF	
Test	NIH	STF	NIH	STF
AUC (CI)	0.631* (0.607-0.655)	0.549 (0.527-0.571)	0.593 (0.568-0.618)	0.669* (0.647-0.691)
Acc	0.182	0.100	0.101	0.159
TPR	0.952	0.951	0.952	0.951
TNR	0.168	0.079	0.085	0.139
PPV	0.021	0.025	0.019	0.026
NPV	0.995	0.985	0.989	0.991

(b) MC-BN

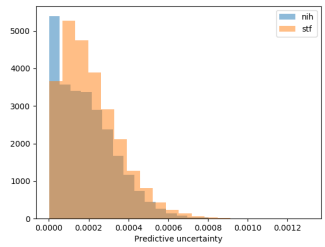
Table 3: Various measures of internal and external model performance for the BCNN models: MC-DO and MC-BN. The confidence level is 0.95. Legend: TPR=true positive rate (sensitivity), TNR=true negative rate (specificity), PPV=positive predictive value (precision), NPV=negative predictive value.

suggests that model predictions may have been influenced by confounding factors present in the images, extraneous information that is separate from the disease pathology itself.

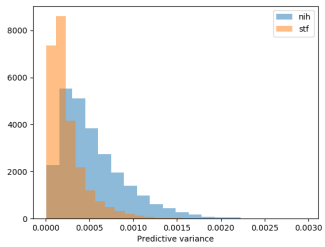
MC-DO models seemed to localize better than CNN models. Further, since the Stanford database contained about twice the training data compared to the NIH database, it also makes sense that the Stanford-trained models seemed to localize better than the NIH-trained models. Increased localization, however, did not generally correspond to gains in model quality, in the sense that important regions still did not consistently align with provided bounding boxes or the chest area.

4 Discussion

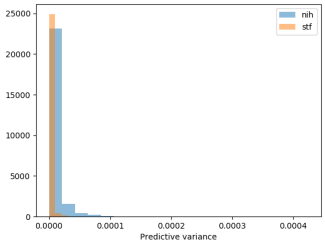
Since they incorporate probabilistic information about the data and the model, BCNNs were hypothesized to be more robust than CNNs across different databases; however, in this project, BCNNs were shown to suffer from the same issues of



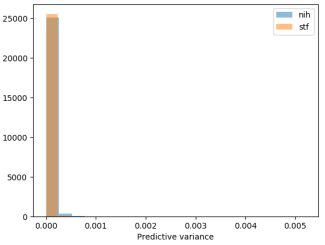
(a) NIH-trained MC-DO



(b) Stanford-trained MC-DO



(c) NIH-trained MC-BN



(d) Stanford-trained MC-BN

Figure 1: Histograms of predictive uncertainty for the NIH-trained and Stanford-trained BCNN models.

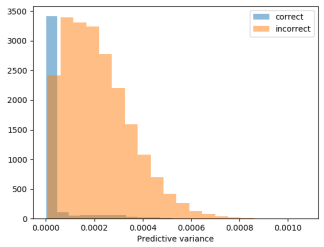
generalizability when trained on one of two public chest X-rays databases and evaluated on both. Predictive BCNN models were significantly worse when X-ray images differed in hospital source from the training data, a discrepancy that would not be expected from a human radiologist given X-rays from two different hospitals.

Further, visualization of each model's areas of "importance" suggests possible confounding factors. This possibility, along with the generalization issue, should be troubling for any researcher interested in implementing these algorithms in a real-world hospital setting.

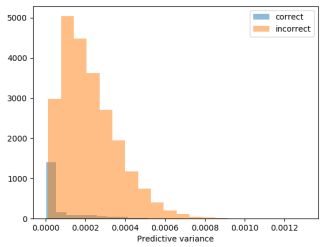
Despite these weaknesses, Bayesian methods may still offer clinicians and researchers advantages that standard deep learning methods cannot. Deep learning models are notoriously uninterpretable; however, BCNNs naturally incorporate a rigorous quantification of uncertainty in its classification process, which may lend it some quality of "trustworthiness" if wielded responsibly.

4.1 Future Directions

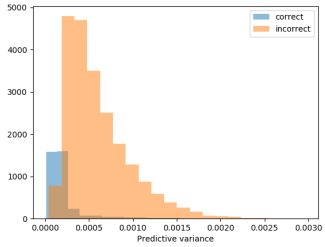
This project still leaves much to explore in this new yet fast-growing field. First, it should be noted that compromises had to be made in this project for the sake of limited computing resources and runtime. Aggressive downsampling of the images to 224 by 224 pixels facilitated a less resource-intensive training



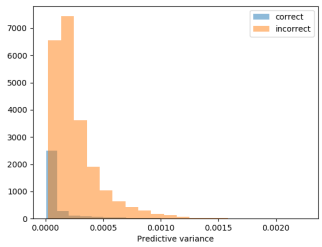
(a) NIH-trained, NIH images



(b) NIH-trained, STF images



(c) STF-trained, NIH images



(d) STF-trained, STF images

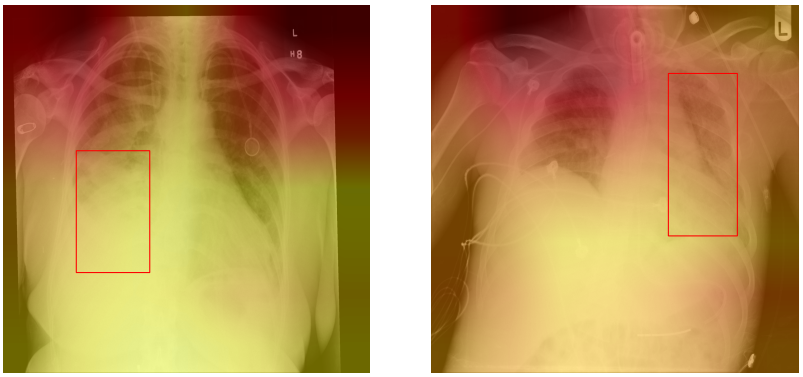
Figure 2: Histograms of predictive uncertainty, separated by correct or incorrect classification, for the NIH-trained and Stanford-trained MC-DO models.

process, but at the cost of image detail and clarity, perhaps exacerbating the models' reliance on confounding factors. The addition of a third database would have also solidified the results of this project. Unfortunately, the database in consideration, the MIMIC-CXR database from the Beth Israel Hospital in Boston ([30]), proved to be too large to incorporate in the limited timeframe of the project.

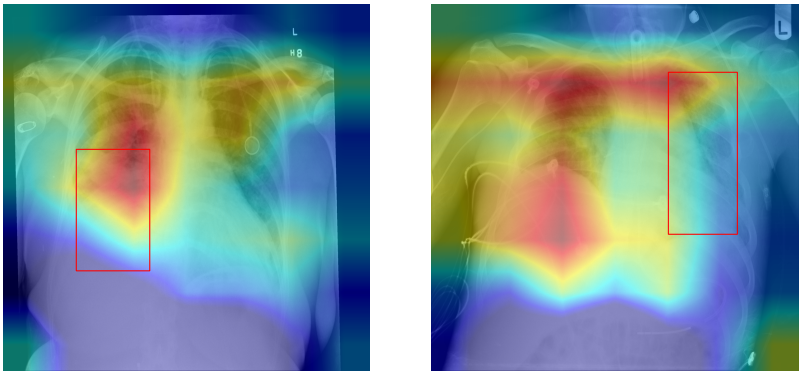
This project only studied MC-DO and MC-BN BCNN architectures, which are popular precisely because they are "cheap" methods for achieving Bayesian outputs; however, other more intensive methods may be less prone to this issue of generalizability, though at the moment this would come at the expense of computing speed.

The uncertainty information given by the BCNNs is also worthy of further analysis. Alternatives to the predictive variance could be considered ([21]). In addition, a simulated referral-like system was proposed in [18] as a possible approach to safely integrating model uncertainty into routine clinical flow; further development and application of this idea would be interesting.

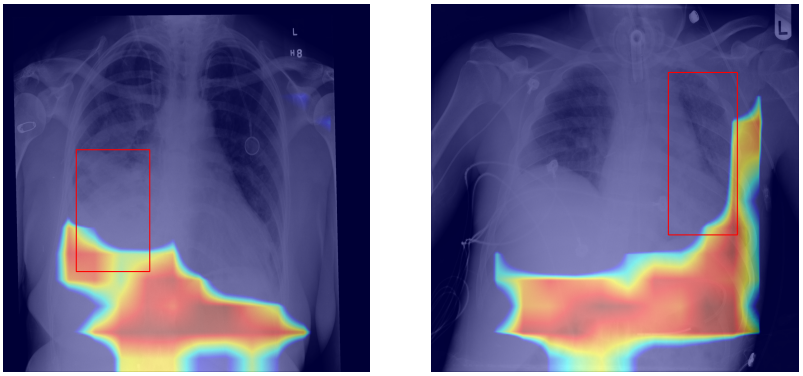
Future directions of this project, however, will likely focus more on theoretical perspectives. The validity of the MC-DO model has been questioned before (notably in [31]), and so the strength, limitations, and mathematical rigor of the model is worth investigating. Techniques for the removal of confounding factors in deep learning models have been proposed (e.g., [32]); generalizing these tech-



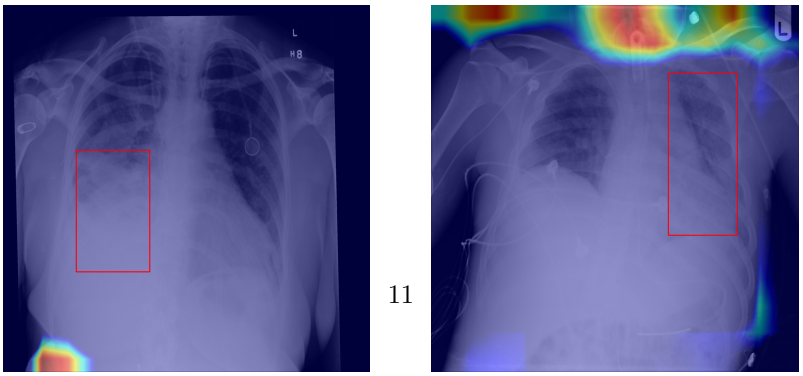
(a) NIH-trained CNN



(b) STF-trained CNN

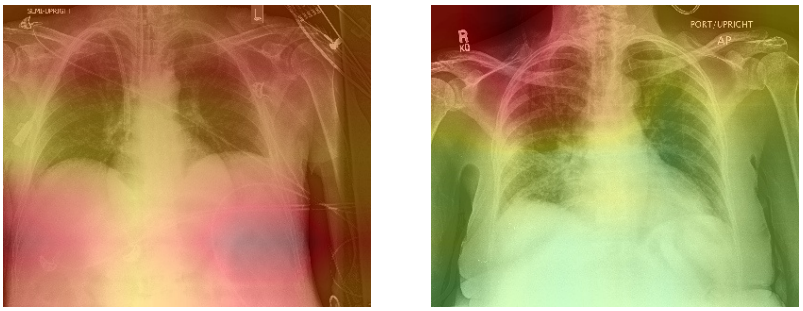


(c) NIH-trained MC-DO

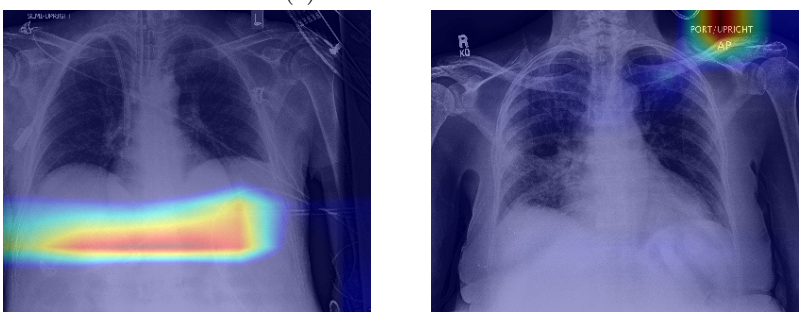


(d) STF-trained MC-DO

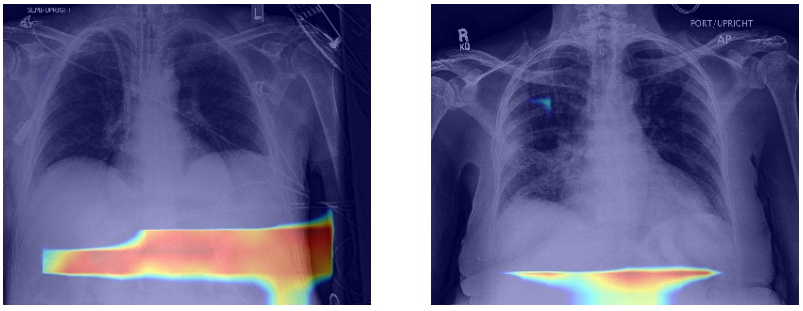
Figure 3: Grad-CAM heatmaps for two pneumonia-positive X-rays from the NIH database, with bounding boxes.



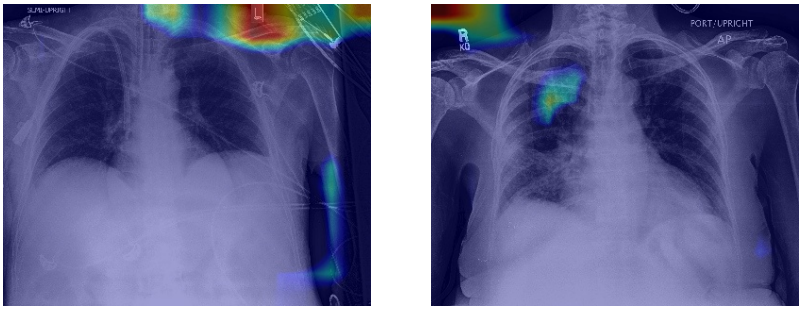
(a) NIH-trained CNN



(b) STF-trained CNN



(c) NIH-trained MC-DO



(d) STF-trained MC-DO

Figure 4: Grad-CAM heatmaps for two pneumonia-positive X-rays from the Stanford database.

niques to Bayesian deep learning methods could be an exciting step forward in regards to the issue of true generalization performance raised in this project.

References

[1] Syed Muhammad Anwar et al. "Medical Image Analysis using Convolutional Neural Networks: A Review". In: *Journal of Medical Systems* 42.11 (Nov. 2018), p. 226. ISSN: 0148-5598. DOI: 10.1007/s10916-018-1088-1. URL: http://www.ncbi.nlm.nih.gov/pubmed/30298337%20http://link.springer.com/10.1007/s10916-018-1088-1.

[2] John R. Zech et al. "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study". In: *PLOS Medicine* 15.11 (Nov. 2018). Ed. by Aziz Sheikh, e1002683. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1002683. URL: http://dx.plos.org/10.1371/journal.pmed.1002683.

[3] David J C Mackay'. *A Practical Bayesian Framework for Backpropagation Networks*. Tech. rep. URL: https://www.mitpressjournals.org/doi/pdf/10.1162/neco.1992.4.3.448.

[4] Radford M Neal. *BAYESIAN LEARNING FOR NEURAL NETWORKS*. Tech. rep. 1995. URL: https://pdfs.semanticscholar.org/db86/9fa192a3222ae4f2d766674a378e47013b1b.pdf.

[5] Geoffrey E. Hinton and Drew van Camp. "Keeping the Neural Networks Simple by Minimizing the Description Length of the Weights". In: *Proceedings of the Sixth Annual Conference on Computational Learning Theory*. COLT '93. Santa Cruz, California, USA: ACM, 1993, pp. 5–13. ISBN: 0-89791-611-5. DOI: 10.1145/168304.168306. URL: http://doi.acm.org/10.1145/168304.168306.

[6] Alex Graves. "Practical Variational Inference for Neural Networks". In: *Advances in Neural Information Processing Systems 24*. Ed. by J. Shawe-Taylor et al. Curran Associates, Inc., 2011, pp. 2348–2356. URL: http://papers.nips.cc/paper/4329-practical-variational-inference-for-neural-networks.pdf.

[7] Diederik P. Kingma, Tim Salimans, and Max Welling. "Variational Dropout and the Local Reparameterization Trick". In: (June 2015). arXiv: 1506.02557. URL: http://arxiv.org/abs/1506.02557.

[8] Charles Blundell et al. *Weight Uncertainty in Neural Networks Daan Wierstra*. Tech. rep. arXiv: 1505.05424v2. URL: https://arxiv.org/pdf/1505.05424.pdf.

[9] Christos Louizos and Max Welling. "Structured and Efficient Variational Deep Learning with Matrix Gaussian Posteriors". In: (Mar. 2016). arXiv: 1603.04733. URL: http://arxiv.org/abs/1603.04733.

[10] Christos Louizos and Max Welling. "Multiplicative Normalizing Flows for Variational Bayesian Neural Networks". In: (Mar. 2017). arXiv: 1703.01961. URL: http://arxiv.org/abs/1703.01961.

[11] Nick Pawlowski et al. *Implicit Weight Uncertainty in Neural Networks*. Tech. rep. arXiv: 1711.01297v2. URL: https://arxiv.org/pdf/1711.01297.pdf.

[12] José Miguel Hernández-Lobato and Ryan P. Adams. "Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks". In: (Feb. 2015). arXiv: 1502.05336. URL: http://arxiv.org/abs/1502.05336.

[13] Shengyang Sun, Changyou Chen, and Lawrence Carin. *Learning Structured Weight Uncertainty in Bayesian Neural Networks*. Tech. rep. URL: http://proceedings.mlr.press/v54/sun17b/sun17b.pdf.

[14] Max Welling, D Bren, and Yee Whye Teh. *Bayesian Learning via Stochastic Gradient Langevin Dynamics*. Tech. rep. 2011. URL: https://www.stats.ox.ac.uk/%7B~%7Dteh/research/compstats/WelTeh2011a.pdf.

[15] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell Deepmind. *Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles*. Tech. rep. arXiv: 1612.01474v3. URL: https://arxiv.org/pdf/1612.01474.pdf.

[16] Yarin Gal and Zg201@cam Ac Uk. *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning Zoubin Ghahramani*. Tech. rep. 2016. arXiv: 1506.02142v6. URL: http://yarin.co..

[17] Yarin Gal and Zoubin Ghahramani. *BAYESIAN CONVOLUTIONAL NEURAL NETWORKS WITH BERNOULLI APPROXIMATE VARIATIONAL INFERENCE*. Tech. rep. arXiv: 1506.02158v6. URL: https://arxiv.org/pdf/1506.02158.pdf.

[18] Christian Leibig et al. "Leveraging uncertainty information from deep neural networks for disease detection". In: *Scientific Reports* 7.1 (Dec. 2017), p. 17816. ISSN: 2045-2322. DOI: 10.1038/s41598-017-17876-z. URL: http://www.nature.com/articles/s41598-017-17876-z.

[19] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. *Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding*. Tech. rep. arXiv: 1511.02680v2. URL: https://arxiv.org/pdf/1511.02680.pdf.

[20] Yongchan Kwon et al. *Uncertainty quantification using Bayesian neural networks in classification: Application to ischemic stroke lesion segmentation*. Tech. rep. URL: https://pdfs.semanticscholar.org/146f/8844a380191a3f883c3584df3d7a6a56a999.pdf.

[21] Rhiannon Michelmore, Marta Kwiatkowska, and Yarin Gal. *Evaluating Uncertainty Quantification in End-to-End Autonomous Driving Control*. Tech. rep. arXiv: 1811.06817v1. URL: https://arxiv.org/pdf/1811.06817.pdf.

[22] Nitish Srivastava et al. *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*. Tech. rep. 2014, pp. 1929–1958. URL: http://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf?utm%7B%5C_%7Dcontent=buffer79b43%7B%5C&%7Dutm%7B%5C_%7Dmedium=social%7B%5C&%7Dutm%7B%5C_%7Dsource=twitter.com%7B%5C&%7Dutm%7B%5C_%7Dcampaign=buffer.

[23] Andrei Atanov et al. "Uncertainty Estimation via Stochastic Batch Normalization". In: (Feb. 2018). arXiv: 1802.04893. URL: http://arxiv.org/abs/1802.04893.

[24] Xiaosong Wang et al. "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases". In: *CoRR* abs/1705.02315 (2017). arXiv: 1705.02315. URL: http://arxiv.org/abs/1705.02315.

[25] Jeremy Irvin et al. "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison". In: (Jan. 2019).

[26] François Chollet et al. *Keras*. https://keras.io. 2015.

[27] Karen Simonyan and Andrew Zisserman. *VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION*. Tech. rep. 2015. arXiv: 1409.1556v6. URL: http://www.robots.ox.ac.uk/.

[28] Gao Huang et al. *Densely Connected Convolutional Networks*. Tech. rep. arXiv: 1608.06993v5. URL: https://github.com/liuzhuang13/DenseNet..

[29] Ramprasaath R Selvaraju et al. *Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization*. Tech. rep. arXiv: 1610.02391v1. URL: http://gradcam.cloudcv.org.

[30] Alistair E. W. Johnson et al. "MIMIC-CXR: A large publicly available database of labeled chest radiographs". In: *CoRR* abs/1901.07042 (2019). arXiv: 1901.07042. URL: http://arxiv.org/abs/1901.07042.

[31] Ian Osband and Google Deepmind. *Risk versus Uncertainty in Deep Learning: Bayes, Bootstrap and the Dangers of Dropout*. Tech. rep. arXiv: 1602.04621. URL: http://bayesiandeeplearning.org/2016/papers/BDL%7B%5C_%7D4.pdf.

[32] Haohan Wang, Zhenglin Wu, and Eric P Xing. *Removing Confounding Factors Associated Weights in Deep Neural Networks Improves the Prediction Accuracy for Healthcare Applications*. Tech. rep. arXiv: 1803.07276v3. URL: https://arxiv.org/pdf/1803.07276.pdf.