Learn To Be Uncertain: Leveraging Uncertain Labels In Chest X-rays With Bayesian Neural Networks

Hao-Yu Yang[1], Junling Yang[2], Yue Pan[1], Kunlin Cao[1],
Qi Song[1], Feng Gao[1]*, Youbing Yin[1]*
[1] {haoyuy, yuep, cao, song, fengg, yin} @curacloudcorp.com
[2] junlin.yang@yale.edu

Abstract

Communication of uncertainty is important for both radiology reports and deep neural networks (DNNs). For radiologists, conveying diagnostic uncertainty in the written report is a challenging and yet inevitable task. On the other hand, while deep learning models have shown compelling potentials in disease classification and lesion detection, applications of DNNs in the medical domain should provide a quantitative measurement of prediction confidence for risk management purposes. In this paper, we investigate the relationship between uncertainty in diagnostic chest x-ray radiology reports and uncertainty estimation of corresponding DNN models using Bayesian approaches. Two sampling methods, Bernoulli and Gaussian dropout have been tested. Our results show that the incorporation of uncertainty labels during model training results in higher predictive variance for uncertain cases at test time. The uncertain cases are inherently difficult to diagnose for human readers, which often needs a further psychical examination to confirm. Returning uncertain predictions on these cases will prevent the DNN model from making over-confident mistakes.

1. Introduction

Deep neural networks (DNNs) have achieved state-of-the-art performance across a wide range of computer vision tasks. Integrating DNNs into clinical workflows have drawn extensive interests and efforts [8]. However, there are a limited number of existing literatures discussing the predictive uncertainty of DNNs for such applications. Representing uncertainty is crucial to handling out-of-distribution samples, defending adversarial attacks, and managing risk, especially in the healthcare sectors where diagnostic reliably should be closely monitored. In this research, we investigate the correlation between uncertainty in the radiology report and uncertainty in DNNs. We quantify the DNN uncertainty by different sampling methods based on different dropout distributions.

1.1. Uncertainty in Radiology report

The radiology report is the primary means of communication between one radiologist and other physicians. Radiologists are constantly balancing between brevity and clarity to make sure that the intended level of confidence is conveyed accordingly to the readers [1]. Despite substantial advancement of the imaging technologies in recent centuries, radiologists still face a great deal of uncertainty in their daily work. A typical radiology report [5] can read:"*Cardiac size is top normal. Bibasilar opacities, larger on the left side, could be due to atelectasis but superimposed infection cannot be excluded.*" Words like "*could be*" and "*cannot be excluded*" indicates ambiguity but currently, there are no rigorous standards in determining the extent of uncertainty. Another source of radiography uncertainty comes from the physical limitation of the imaging modality. A representative case would be diagnosing cardiomegaly with chest x-rays. Cardiomegaly is the enlargement of the heart that may cause other complications like blood clots and heart failure. A chest x-ray screen is usually how cardiomegaly is detected. However, confirmation of cardiomegaly requires a blood test or an electrocardiogram because different patient positioning during the scan can effectively alter the appearance of the heart in an x-ray image. A DNN system that is being overconfident on images with inherent constraints could lead to erroneous and harmful conclusions.

1.2. Related Work

The current state-of-the-art methods for evaluating the uncertainty of DNNs are Bayesian-based methods that form a posterior distribution of the network parameters [2]. From a Bayesian viewpoint, training a DNN concerns finding the maximum a posterior (MAP) of the weight matrices $\mathbf{W}$ given the training dataset $\mathcal{D} = (\mathbf{X}, \mathbf{Y})$, where $\mathbf{X}$ denotes

*Corresponding authors.

the training inputs and $\mathbf{Y}$ denotes the corresponding labels. Given a new query input x, the posterior predictive distribution is

$$p(y \mid x, \mathcal{D}) = \int p(y \mid x, \mathbf{W}) p(\mathbf{W} \mid \mathcal{D}) \, d\mathbf{W}. \quad (1)$$

However, the posterior $p(\mathbf{W} \mid \mathcal{D})$ is often analytically intractable, and the alternative is to employ a variational distribution $q(\mathbf{W})$ to approximate $p(\mathbf{W} \mid \mathcal{D})$, such that

$$p(y \mid x, \mathcal{D}) = \int p(y \mid x, \mathbf{W}) q(\mathbf{W}) d\mathbf{W}. \quad (2)$$

$q(\mathbf{W})$ is usually optimized by minimizing the Kullblack-Leiber (KL) divergence between $q(\mathbf{W})$ and $p(\mathbf{W} \mid \mathcal{D})$. The posterior predictive distribution can now be approximated through Monte-Carlo sampling of the weights from $q(\mathbf{W})$,

$$p(y \mid x, \mathcal{D}) \simeq \frac{1}{T} \sum_{t=1}^{T} p(y \mid x, \mathbf{W}_t), \quad (3)$$

where T denotes the number of stochastic forward passes, which is equivalent to the number of Monte-Carlo samples. Numerous methods have been formulated for variational distribution [7] based on different sampling distribution, e.g., Bernoulli, Gaussian, and Spike-and-Slab dropout. Previous research [6] has explored using dropout-based Bayesian uncertainty measures for diagnosing diabetic retinopathy and their results established an informative interpretation of the source of uncertainty.

2. Method

2.1. Data

We trained and experimented on two large publicly available chest x-ray datasets, CheXpert [4] and Chest X-ray-14 (CXR-14) [9]. The CheXpert dataset contained 224,316 chest radiographs from 65,240 patients while the CXR-14 dataset contained 108,948 chest radiographs from 32,717 patients. Both datasets used rule-based Natural Language Processing (NLP) labeler to extract 14 common mentions from raw radiology reports. The main difference between the two datasets lies in the inclusion of uncertainty label and the disease categories. The CheXpert labeler extracted uncertain findings of the disease that are denoted as u, while the CXR-14 omitted such mentions. For the CheXpert labels, some diseases are more likely to be marked as uncertain than others. For example, uncertain labels constitute 15.66% and 12.78% of all atelectasis and consolidation respectively. The two datasets have 7 disease labels in common: Atelectasis, Cardiomegaly, Effusion, Pneumonia, Pneumothorax, Consolidation, and Edema. Our experiments are conducted on the 7 overlapping diseases.

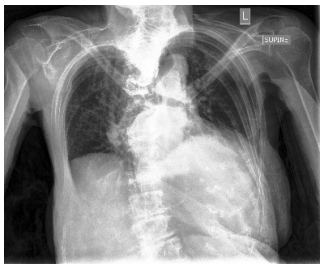

Figure 1. Female patient diagnosed with lung opacity and fracture and suspected of cardiomegaly, edema, atelectasis and pleural effusion.

2.2. Deep Bayesian Network

In deep learning frameworks, a straightforward approach to performing sampling is multiplying the feature maps $\mathcal{F}$ with a sampling matrix $\hat{M}$ where each element is drawn from some distribution M,

$$\hat{W} = \mathcal{F} \odot \hat{M}, \quad (4)$$

where $\odot$ denotes element-wise multiplication. The difference between each sampling method is, therefore, the distribution that the elements in $\hat{M}$ are generated from.

Bernoulli and Gaussian Dropout We describe two methods for generating the sampling matrix. The Bernoulli dropout samples each entity in the mask independently from a Bernoulli distribution. The probability of a connection being dropped is therefore $\mathrm{Bern}(1-p)$. In Gaussian dropout, each element is sampled from a Gaussian distribution with mean μ and variance σ^2. Element (i, j, k) of the sampling matrix is therefore $\hat{M}_{i,j,k} \sim \mathcal{N}(\mu, \sigma)$.

2.3. Uncertainty Labels

We compare different approaches to utilizing the uncertainty label from the CheXpert dataset in model training. The two main methods we tried were binary mapping and separate label classification.

Binary mapping In this setting, the u labels are mapped to either positive, negative, or simply ignored during training. In the CheXpert paper [4], these methods served as baseline methods for performance evaluation. In our experiments, we have ignored samples with the u labels.

Uncertainty classification With uncertainty classification, the u labels are treated as a separate class from positive and negative labels. During inference time, the softmax function is restrained to only positive and negative class predictions.

3. Experiments

Two independent DenseNet-121 [3] were trained for each of the aforementioned datasets. We use 50 Monte-Carlo samples during test time. For both datasets, we use only the frontal-view images. Since there were no uncertain labels in the official CheXpert validation set, we split

the training set into 80% training and 20% validation. The learning rate was set at 1×10^{-4}. The Bernoulli dropout rate was set at 0.5, while the mean and standard deviation of the Gaussian dropout were 0 and 1, respectively.

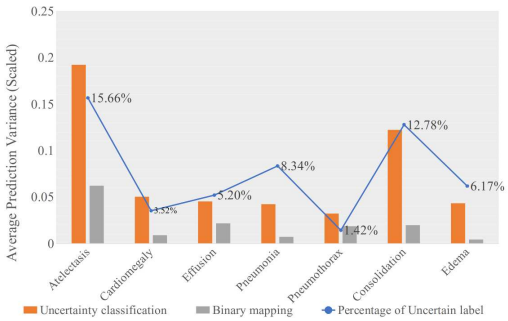

Figure 2. Predictive variance across 7 diseases

3.1. Uncertain radiologist diagnosis

We compared the results of using binary mapping for uncertainty labels and as a separate class label. Fig 2 showcase the prediction variances of DenseNet-121 trained on the CheXpert dataset. The line plot indicates the percentage of total uncertainty label for a disease category. The two bar plots represent the average prediction variance of 50 Monte-Carlo samples scaled by 10^4 for easier visualization. We observe that disease classes with a higher percentage of uncertain label tend to have prediction with higher uncertainty as well. Furthermore, treating the uncertain label as a separate class during training will lead to low confidence predictions of uncertain cases in test time.

3.2. Out of distribution samples

To study the effects of uncertainty label on out-of-distribution samples, we cross-validated the two models on the validation set of its counterparts. Namely, we predicted on the CXR-14 dataset validation set using the model trained on CheXpert and vice versa. Fig 3 illustrate a patient with a positive label of effusion and suspected of atelectasis, consolidation, and cardiomegaly. Effusion was correctly predicted for both models. However, the CheXpert model trained on uncertain labels gave predictions that were less confident but closely resembled the ground truth of suspecting cardiomegaly and consolidation.

To evaluate the uncertainty estimation, we calculate the entropy of the predictive distribution. Given a query image x and corresponding prediction$y$, the entropy can be calculated as

$$H(y \mid x) = -\int p(y \mid x) \log p(y \mid x)\, dy. \qquad (5)$$

A robust DNN model should return higher entropy on out-of-distribution samples even if the imaging modality is the same. This is due to the fact that data collected from another imaging center may have different imaging protocols and the human user should proceed these cases with caution. Our experiment results are shown in Table 1. Both models produced higher predictive entropy on dataset that was not seen.

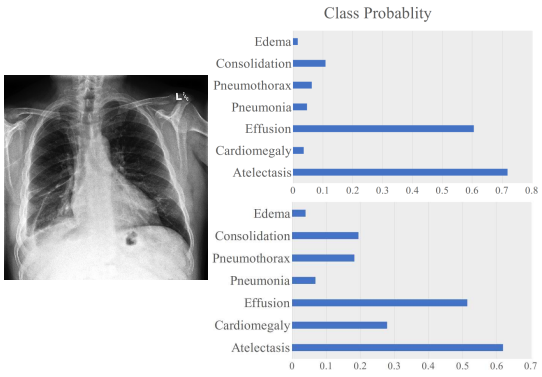

Figure 3. Left: Input image. Right upper: CXR-14 model. Right lower: CheXpert model

Table 1. Average predictive entropy across 7 common diseases. Higher entropy indicates uncertainty predictions.

Model	CXR-14 model	CheXpert model
Bernoulli dropout on CXR-14	0.182	0.531
Bernoulli dropout on CheXpert	0.428	0.308
Gaussian dropout on CXR-14	0.377	0.552
Gaussian dropout on CheXpert	0.440	0.379

4. Discussion and Future Work

In this work, we have established a connection between radiology uncertainty and DNN uncertainty. We demonstrate that incorporating uncertainty labels as a separate class during training enables the model to produce fewer confidence predictions on ambiguous cases as opposed to models trained with binary labels. Being able to produce an uncertain prediction on inconclusive cases holds significant clinical value as these cases often require physical examination or biopsy to confirm. By incorporating uncertainty information, typical over-confident mistakes of DNNs can be avoided. We acknowledge some limitations of our work that can be addressed in future work. First, the CheXpert official validation set provided ground truth labels by having multiple radiologists manually labeling the images and therefore does not contain uncertain labels. There is limited information on the quality and extent of the uncertain labels in the training set. Second, the current analysis is conducted on extracted labels from the original radiology report. The extraction process is not perfect compared to radiologist-annotated ground truth. With the release of the free-form radiology report from the CheXpert dataset in the future, a direct comparison of uncertainty from the radiology report and DNNs can be made. Our work can be further extended to the quantification of confidence in radiology vocabularies such as "possibly", "suggestive of", "consistent with", "not entirely excluded", etc.

References

[1] Michael A. Bruno, Jonelle Petscavage-Thomas, and Hani H. Abujudeh. Communicating uncertainty in the radiology report. *American Journal of Roentgenology*, abs/1901.07042, 2017.

[2] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *arXiv e-prints*, page arXiv:1506.02142, June 2015.

[3] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, July 2017.

[4] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *CoRR*, abs/1901.07031, 2019.

[5] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *CoRR*, abs/1901.07042, 2019.

[6] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, 7(1):17816, 2017.

[7] Patrick McClure and Nikolaus Kriegeskorte. Representation of uncertainty in deep neural networks through sampling. *CoRR*, abs/1611.01639, 2016.

[8] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *CoRR*, abs/1711.05225, 2017.

[9] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *CoRR*, abs/1705.02315, 2017.