EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Medizinische Fakultät

AG Berens

Research Project

Modelling multiple sources of uncertainty

University of Tübingen
Institute of Ophthalmic Research
AG Berens

Yutong Wen, yutong.wen@student.uni-tuebingen.de, 1. August 2022

Bearbeitungszeitraum: von 8. April 2022 bis 1. August 2022

Betreuer/Gutachter: Prof. Philipp Berens, Universität Tübingen
Zweitgutachter: Dr. Lisa Koch, Universität Tübingen

Abstract

Predictive uncertainty estimation is essential to machine learning models, as it tells the user how confident the model is in the prediction. The user can then decide whether human intervention is needed, thus improving safety. There are three sources of uncertainty, namely uncertainty in the model, uncertainty in the data, and uncertainty in the distribution. Most previous work either conflates model uncertainty with distributional uncertainty or data uncertainty with distributional uncertainty. Recently, some Dirichlet Prior Network-based approaches can distinguish distributional uncertainty from other uncertainties, which means we can detect out-of-distribution data from the dataset. However, the problem of distinguishing between in-domain data with data uncertainty, in-domain data without data uncertainty, and out-of-distribution data remains unsolved. Therefore, in this work, we first reproduce the experiments from [MG18], using the Dirichlet Prior Network, and then train this model on a corrupted MNIST dataset we created to model the problem better. These experimental results show that the model can only distinguish between in-domain without data uncertainty and out-of-distribution data or in-domain with data uncertainty and out-of-distribution data. However, this model can not perfectly classify in-domain data with or without data uncertainty.

Keywords— Dirichlet Prior Network, Out-of-distribution, Uncertainty Estimation

Contents

1. Introduction **5**
- 1.1. Uncertainty . . . 5
- 1.2. Related work . . . 5
- 1.3. Contributions and Thesis Organisation . . . 6
  - 1.3.1. Contributions . . . 6
  - 1.3.2. Structure . . . 7

2. Prerequisites **8**
- 2.1. Dirichlet distribution . . . 8
  - 2.1.1. Definition . . . 8
  - 2.1.2. Properties . . . 9
- 2.2. Dirichlet Prior Network . . . 10
  - 2.2.1. Construction . . . 10
  - 2.2.2. Training . . . 11
- 2.3. Uncertainty Measures . . . 11
  - 2.3.1. Total Uncertainty (Entropy of y) . . . 11
  - 2.3.2. Expected data uncertainty . . . 12
  - 2.3.3. Entropy of μ . . . 12
  - 2.3.4. Mutual Information of y and μ . . . 12

3. Experiments **13**
- 3.1. Synthetic Experiments . . . 13
  - 3.1.1. Setup . . . 13
  - 3.1.2. Results . . . 14
- 3.2. MNIST Experiments . . . 16
  - 3.2.1. Reproduction with reverse KL-divergence loss . . . 16
  - 3.2.2. Corrupted MNIST . . . 17

4. Discussion and Conclusion **19**
- 4.1. Discussion . . . 19
  - 4.1.1. Comparison with previously published results . . . 19
  - 4.1.2. Problem . . . 19
  - 4.1.3. Further work . . . 21
- 4.2. Conclusion . . . 22

A. Appendix **23**
A.1. Reverse KL-divergence 23
A.1.1. KL Divergence between two Dirichlet Distribution 23
A.1.2. Derivative of KL-divergence 23
A.1.3. Definition of reverse KL-divergence loss 23
A.2. Derivations for Uncertainty Measures 24
A.2.1. Mutual Information 24
A.2.2. Expected data uncertainty 24

1. Introduction

Machine learning models can be used to solve many problems in real life. However, when these models are wrong, we do not receive any warning. This is why we need predictive uncertainty estimation. Based on this estimation, we can accept the prediction or reject it and take a human intervention.

In this chapter, we will first give some background information and some previous approaches to uncertainty estimation, and then we will talk about our contributions and the structure of this report.

1.1. Uncertainty

There are three different sources of predictive uncertainty: model uncertainty, distributional uncertainty, and data uncertainty.

Model uncertainty measures the uncertainty in estimating model parameters[MG18], and it can be reduced when we have enough data. On the opposite, **data uncertainty** is generated by the data itself, such as category overlap, which is irreducible. **Distributional uncertainty** arises from the dataset shift[MG18] between training and test data. We refer to the training data as in-domain data, and for test data that appear outside the training distribution, we refer to them as out-of-distribution data.

In this work, we focus on distributional uncertainty and data uncertainty, and we want to find a way to distinguish distributional uncertainty from other uncertainties.

1.2. Related work

In this section, we present some previous approaches to predictive uncertainty estimation and their problems. Most of the formulas are quoted from papers [MG18, NHL20]

Consider finite dataset $\mathcal{D} = \{x_i, y_i\}$, where x denotes images and y denotes class labels. $p(x, y)$ is the distribution over x and y. For an input x^*, the data uncertainty is captured by posterior distribution $P(y = w_c|x^*, \theta)$, where θ represents model parameters and the model uncertainty is captured by model posterior $p(\theta|D)$. The

predictive uncertainty is defined as follows:

$$P(y = w_c|x^*, \mathcal{D}) = \int P(y = w_c|x^*, \theta)P(\theta|\mathcal{D})d\theta \tag{1.1}$$

However, the true posterior for θ is intractable, so approximation approaches are needed, such as **Monte-Carlo dropout**[GG16] or **Deep Ensembles**[LPB17].

$$P(y = w_c|x^*, \mathcal{D}) \approx \frac{1}{M}\sum_{k}^{M} P(y = w_c|x^*, \theta^k) \tag{1.2}$$

where θ^k is sampled from variational approximation $q(\theta) \approx p(\theta|\mathcal{D})$. Each $P(y = w_c|x^*, \theta^k)$, presents a categorical distribution μ, is in an ensemble $\{P(y = w_c|x^*, \theta^k)\}_k^M$

$$P(y = w_c|x^*, \theta^k) \sim p(\theta|\mathcal{D}) \equiv \mu^k = [p(y = w_1), .., p(y = w_m)] \sim P(\mu|x^*, \mathcal{D}) \tag{1.3}$$

The ensemble can be visualized on a simplex, and we can determine the source of uncertainty based on its properties: it is sharp at the corner for in-domain data and flat for out-of-distribution data. However, in practice, there is no guarantee that the distribution over the simplex would have these properties[MG19]. **Dirichlet Prior Network** from[MG19, MG18], on the other hand make this objective explicit. We will discuss this approach further in Chapter 2.

Non-Bayesian approaches, such as works from [MRKG17, LLLS17], use predictive posteriors from DNN to measure uncertainties. In these works, DNNs are trained to output high entropy values for out-of-distribution data. However, it is not possible to detect whether this corresponding input is out-of-distribution data or in-domain data from a region of high data uncertainty. Therefore, it is not possible to distinguish distributional uncertainty from other sources of uncertainties.

Hence, the aim of our work is to address the problem of distinguishing distributional uncertainty from others by extending the work done in [MG18]. The contributions of our work are summarized in Chapter 1.3.1.

1.3. Contributions and Thesis Organisation

1.3.1. Contributions

First, we reproduce the results of the paper[MG18] on misclassification detection and out-of-distribution detection on the MNIST dataset using FashionMNIST as out-of-distribution. Second, we applied this approach to the new dataset we created to investigate whether this model can distinguish between in-domain data without data uncertainty, out-of-distribution data, and in-domain data with data uncertainty. Finally, we discuss further work we need to do before applying this approach to the field of medical image analysis.

1.3.2. Structure

This report is structured as follows: Chapter 2 is about some basics of Dirichlet Prior Network and uncertainty measures, Chapter 3 is about the experiments and results, and Chapter 4 contains the conclusion and discussions of this work.

2. Prerequisites

In this chapter, we briefly introduce some basic knowledge related to **Dirichlet Prior Network** and uncertainty measurements. Most of the concepts and formulas in this report are quoted from papers [MG18, MG19].

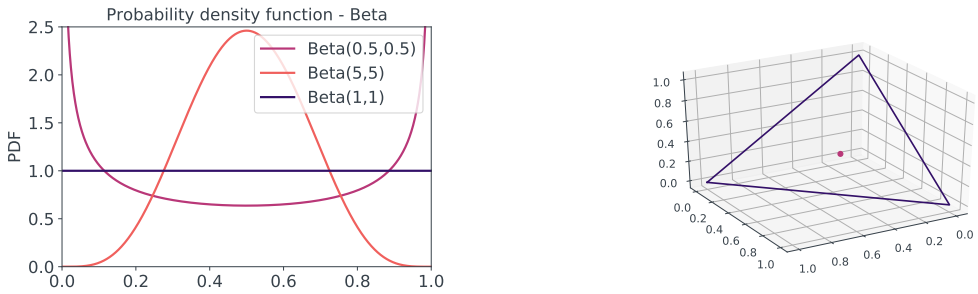
2.1. Dirichlet distribution

2.1.1. Definition

The **Dirichlet distribution**, denoted $\mathrm{Dir}(\mu|\alpha)$, is a multivariate probability distribution over the probability simplex. The simplex is a set of vectors whose components should sum to 1, and all components are non-negative. Thus, our μ has the following properties:

$$\sum_i \mu_i = 1 \qquad \mu_i \geq 0 \tag{2.1}$$

An easy way to interpret the simplex is to use a triangle, as shown in Fig. 2.1b. The μ corresponds to the barycentric coordinate of the point.



(a) Beta Distribution

(b) Simplex

Figure 2.1.: Distributions

The **Dirichlet distribution** is parameterized by vector $\alpha = (\alpha_1, ..\alpha_i, .., \alpha_c)$ where each α_i is a positive real number. It is a multivariate generalization of the beta distribution[Dir22], where the beta distribution $\mathcal{B}(\alpha, \beta)$ is an univariate probability distribution parameterized by α and β as shown in Fig. 2.1a.

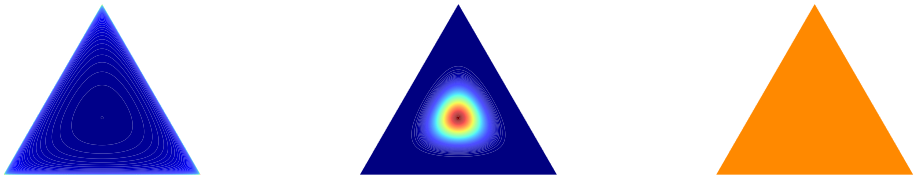
The probability density function of **Dirichlet distribution** is defined as follows:

$$\mathrm{Dir}(\mu|\alpha) = \frac{\Gamma(\alpha_0)}{\prod_c \Gamma(\alpha_c)} \prod_c \mu_c^{\alpha_c - 1} \tag{2.2}$$

where α is the concentration parameter of the distribution and $\alpha_0 = \sum_i \alpha_i$ is the precision.

2.1.2. Properties

Concentration parameter



(a) $\alpha = (0.9, 0.9, 0.9)$ (b) $\alpha = (10, 10, 10)$ (c) $\alpha = (1, 1, 1)$

Figure 2.2.: Effect of concentration parameter on the shape of the distribution.

As shown in Fig. 2.2, the concentration parameter affects the shape of the distribution. As shown in Fig. 2.2a, if all α_i are less than 0.9, we have a U-shaped distribution, which can also be seen in the beta distribution of Fig. 2.1a. When all α_i are greater than 1, we have a mode in the distribution, illustrated in Fig. 2.2b. When all the values are equal to 1, we obtain a flat distribution as shown in Fig. 2.2c.

The expected value of the Dirichlet distribution

$$\mathbb{E}(\mu_c) = \frac{\alpha_c}{\sum_c \alpha_c} \tag{2.3}$$

In Chapter 2.2, we use the expected value of the Dirichlet distribution to calculate the final class label.

2.2. Dirichlet Prior Network

2.2.1. Construction

A **Dirichlet prior network** is an approach to analysis the source of the uncertainty in the prediction. It parametrizes the Dirichlet Distribution as prior over simplex[MG18].

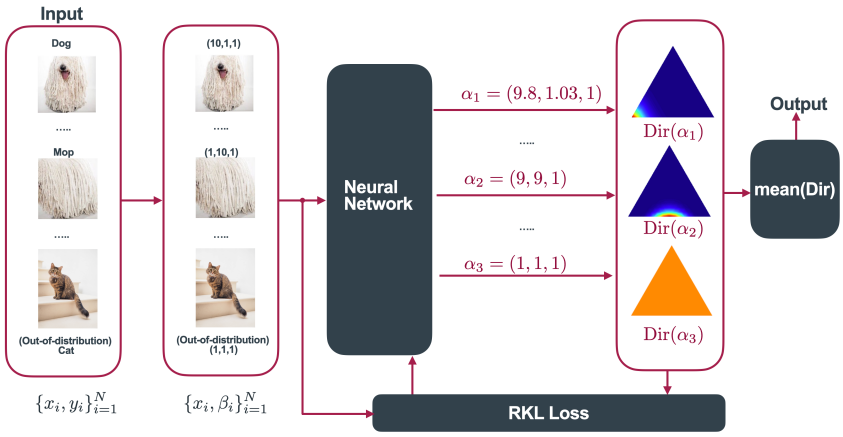


Figure 2.3.: Overview of the Dirichlet Prior Network architecture.
This is an example of using the Dirichlet Prior Network. The input in-domain data contains three classes: dog ($y_i = 0$), mop ($y_i = 1$), and brush ($y_i = 2$). The out-of-distribution data are photos of cats. We first convert the input class labels into concentration parameters: dog$(10,1,1)$, mop$(1,10,1)$, brush$(1,1,10)$, and cat$(1,1,1)$. Then we take these inputs into the neural network to predict the concentration parameters of the Dirichlet distribution. After that, we get the output category labels based on the distribution expectation. Figures of Dog and mop are from [Aga18]. The figure of the cat is from [Ale].

The structure of the Dirichlet Prior Network is shown in Fig. 2.3. For each label y_i in the input data $\{x_i, y_i\}_{i=1}^N$, we first convert it to the corresponding concentration parameter β_i. For out-of-distribution data, we set all β_i^c in β_i to 1. We can then use this concentration to generate a flat uniform Dirichlet distribution as the target distribution for these data. For in-domain data, β_i is set as follows:

$$\beta_i^c = \begin{cases} 1 & y_i \neq c \\ K & y_i = c \end{cases}$$

We can then use this concentration parameter to generate a sharp Dirichlet distribution focused on one of the simplex corners as the target distribution for these data. We then feed these converted inputs $\{x_i, \beta_i\}_{i=1}^N$ into the Neural Network to predict the concentration parameters α of the Dirichlet distribution.

$$\alpha = f(x; \hat{\theta}), \qquad p(\mu|x, \hat{\theta}) = \mathrm{Dir}(\mu|\alpha) \tag{2.4}$$

When the model is confident in its predictions, the Dirichlet distribution Dir(α) should be a sharp distribution at the corner of the simplex illustrated in Fig. 2.4a. For in-domain inputs with high data uncertainty, Dir(α) should be a sharp distribution at the center of the simplex as shown in Fig. 2.4b. For out-of-distribution inputs, Dir(α) should be a flat distribution over the simplex illustrated in Fig. 2.4c. When we have only two overlapping classes, Dir(α) should be a sharp distribution at the edge of both classes, as shown in Fig. 2.4d.

(a) Confident (b) All classes overlap (c) Out-of-distribution (d) Two classes overlap

Figure 2.4.: Different types of predictive uncertainty of DPN.

The posterior over class labels can then be calculated as the expected value of the Dirichlet distribution:

$$p(y = w_c|x^*, \hat{\theta}) = \mathbb{E}_{p(\mu|x^*;\hat{\theta})}[P(y = w_c|\mu)] = \frac{\alpha_c}{\sum_c \alpha_c} \tag{2.5}$$

2.2.2. Training

In this work, we trained the **Dirichlet Prior Network** based on works of [MG19]. It is trained to minimize the reverse KL-divergence between the model and target distribution. The loss function is defined as follows:

$$\mathcal{L}(\theta, \mathcal{D}; \beta_{\text{in}}, \beta_{\text{out}}) = \mathcal{L}^{RKL}(\theta, \mathcal{D}_{in}; \beta_{\text{in}}) + \mathcal{L}^{RKL}(\theta, \mathcal{D}_{out}; \beta_{\text{out}}) \tag{2.6}$$

Mathematical formulas and calculations related to the loss function are available in Appendix.

2.3. Uncertainty Measures

In this section, we present several measures to quantify uncertainty.

2.3.1. Total Uncertainty (Entropy of y)

The first measure is the entropy of the prediction posterior over the class label, which represents the uncertainty contained in the entire distribution. We use this metric to

measure the total uncertainty of the predictions.

$$\mathcal{H}[P(y|x^*;\mathcal{D})] = -\sum_c P(y = w_c|x^*,\mathcal{D}) \ln(P(y = w_c|x^*,\mathcal{D}))$$

2.3.2. Expected data uncertainty

Data uncertainty is measured by the entropy of the true data distribution. However, we do not know the true distribution, so we can only get an approximate data uncertainty from the Dirichlet Prior network. The expected data uncertainty is the average of these approximate data uncertainties. It is defined as follows:

$$\mathbb{E}_{p(\mu|x^*;\mathcal{D})}[\mathcal{H}(P(y|\mu)]$$

2.3.3. Entropy of μ

This is a measure of the entropy of the Dirichlet distribution of the output of the forward network. This metric is maximum when the Dirichlet distribution is flat, i.e., all categorical distributions are equally likely.

$$\mathcal{H}(P(\mu|x^*,\mathcal{D})) = -\int P(\mu|x^*,\mathcal{D}) \ln P(\mu|x^*,\mathcal{D})$$

2.3.4. Mutual Information of y and μ

The mutual information of y and μ is calculated by the difference between the entropy of y and the expected data uncertainty. It expresses distributional uncertainty.

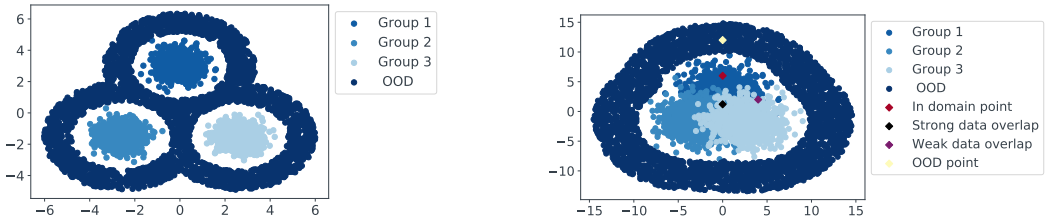
$$\mathcal{I}(y,\mu|x^*,\hat{\theta}) = \mathcal{H}[P(y|x^*;\mathcal{D})] - \mathbb{E}_{p(\mu|x^*;\mathcal{D})}[\mathcal{H}(P(y|\mu)]$$

The derivations for the uncertainty measures are in the Appendix.

3. Experiments

To study how to distinguish distributional uncertainty from other uncertainties, in this Chapter, we first reproduce the synthetic and MNIST experiments from [MG18] and add a measure, expected data uncertainty, to quantify the uncertainty. We then replicate the experiments proposed in [MG18] on the corrupted MNIST dataset we created, which would allow us to better model high data uncertainty to answer the remaining questions in [MG18] regarding the behavior of this Dirichlet Prior Network approach to distinguish between data uncertainty and distributional uncertainty.

3.1. Synthetic Experiments



(a) Distinct Dataset

(b) Dataset with strong class overlap

Figure 3.1.: Synthetic datasets for the preliminary experiments.
In (a), the in-domain data consists of three non-overlapping classes, and the out-of-distribution data is shown in dark blue. In (b), the in-domain data have a large amount of class overlap, and the four colored data points are samples from different regions of the dataset.

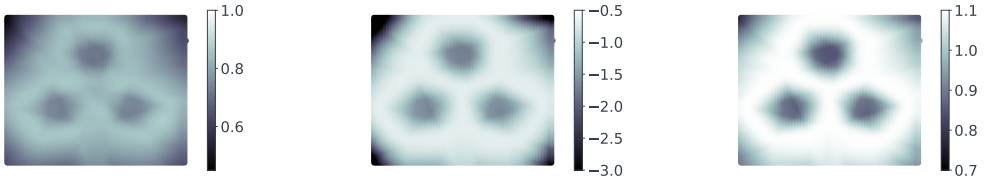
3.1.1. Setup

We train the Dirichlet Prior Networks with 50 neurons in 1 hidden layer on in domain and out-of-distribution data introduced in Fig. 3.1. The in-domain data contains 3 Gaussian distributed groups with equal mean and variance. Each group of the in-domain data includes 1000 samples, and the size of the in-domain samples equals the out-of-distribution samples. Now we consider two cases, the first one, as illustrated in Fig. 3.1a, is based on a dataset containing 3 distinct classes, while the

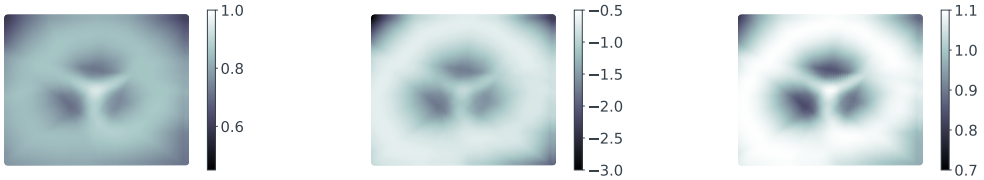
second case (Fig. 3.1b) uses a dataset with significant class overlap. In this experiment, we used the same way of sampling out-of-distribution data as [MG19], where the out domain data surrounds the training data. The target concentration follows the setting in Chapter 2.2 with $K = 4$, and we train the Dirichlet Prior Network with the loss of reverse KL-divergence.

3.1.2. Results

Qualitative Analysis



(a) Data Uncertainty (b) Entropy of μ (c) Entropy of y



(d) Data Uncertainty (e) Entropy of μ (f) Entropy of y

Figure 3.2.: Uncertainty measures for synthetic experiments.

Fig. 3.2 presents three different uncertainty measures, namely, the entropy of y, the entropy of μ and expected data uncertainty. The first row of the figure shows the results on the test set with distinct datasets, and the second row shows the uncertainty measures for the test set with significant class overlap.

As shown in the first row of Fig. 3.2, all groups are perfectly distinct, and we can see that all of the uncertainty measures have similar behavior, with lower values for the in-domain data and higher values for out-of-distribution data.

As shown in the second row of Fig. 3.2, we can see that we have higher values in both the class overlap region and the out-of-distribution region. For the entropy of y shown in Fig. 3.2f, the color received in the class overlap region and the out-of-distribution region are almost identical, reflecting the fact that we have similar total uncertainty in the out-of-distribution region and the high class overlap region. As shown in Fig. 3.2d, the class overlap region is darker than the out-of-distribution region, which means that we have relatively high expected data uncertainty in the class overlap

region and relatively low expected data uncertainty in the out-of-distribution region, and they are both higher than the expected data uncertainty in the in-domain area. For the entropy of μ shown in Fig. 3.2e, it is clear that we have the darkest colors in out-of-distribution regions, darker colors in areas with class overlap, and lightest colors in regions with in-domain data.

In conclusion, the Dirichlet Prior Network with reverse KL-divergence loss reflects the structure of the synthetic experiments dataset.

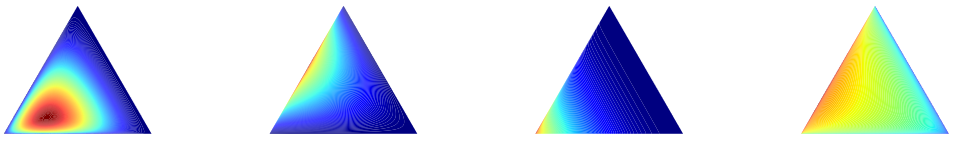
Quantitative Analysis

We now focus on some specific points in Fig. 3.1b, which are samples from the high class overlap region, weak class overlap region, in-domain region, and out-of-distribution region. The quantitative results are shown below:

Points	**Entropy of y**	**Entropy of μ**	**Data Uncertainty**
In domain	0.9071	-1.4869	0.7523
High class overlap	1.0439	-1.0974	0.8985
weak class overlap	1.0402	-1.0914	0.8739
Out-of-distribution	1.0951	-0.7129	0.8441

Table 3.1.: Uncertainty measures of points

The results in Table 3.1 are consistent with our previous observations. The point from the out-of-distribution region has high entropy of μ, and the point from the significant class overlap region has high expected data uncertainty, so that we can distinguish well between out-of-distribution data and high class overlap data by the entropy of μ and the expected data uncertainty. We visualize the results of the forward network of these points as shown in Fig. 3.3.



(a) High class overlap (b) weak class overlap (c) In domain point (d) Out-of-Distribution

Figure 3.3.: Dirichlet distribution of these points

As shown in Fig. 3.3a, when the point is located in the region of strong class overlap, we can receive a uni-mode in the Dirichlet distribution in the center toward the lower left. When the point lies in the overlap between two classes, our mode is placed at the edge between these classes, as shown in Fig. 3.3b. When the point is

sampled from in domain region, we can receive a sharp distribution on the corner of the simplex, and when the point is out-of-distribution, we can get a flat Dirichlet distribution illustrated in Fig. 3.3d.

3.2. MNIST Experiments

This experiment aims to test whether this model clearly distinguishes out-of-distribution data from in-domain data with or without data uncertainty. In this section, we first reproduce some of the experiments from [MG18] using MNIST and FashionMNIST datasets. We then train the Dirichlet Prior network on the corrupted MNIST dataset we created.

3.2.1. Reproduction with reverse KL-divergence loss

In-domain misclassification detection

We used the same setup from [MG18] in this experiment. We run the experiments on MNIST and FashionMNIST with reverse KL-divergence loss. The images in both datasets were reduced to 28×28 pixels. FashionMNIST is used as out-of-distributed data in both training and test datasets. The goal of this experiment is to detect whether an in-domain prediction is incorrect. Therefore, we choose incorrect classification as the positive example and evaluate their performance using the area under ROC (AUROC) and Precision-Recall (AUPR). In this experiment, instead of reproducing the results of the max probability as the paper suggested, we focus more on expected data uncertainty. The results are displayed in the first row in Table 3.2, where **Ent.y** represents the entropy of y, **Ent.**μ represents the entropy of μ, **M.I.** stands for the mutual information, and **D.U.** stands for the expected data uncertainty.

The expected data uncertainty yields the best results. For the remaining measures, we have similar conclusions to the paper: the entropy of y performs better than the entropy of μ, and the difference between them is much more significant in AUPR than in AUROC.

	AUROC				AUPR			
Detections	**Ent. y**	**Ent.**μ	**M.I.**	**D.U.**	**Ent. y**	**Ent.**μ	**M.I.**	**D.U.**
Misclassification	92.3	89.7	89.9	92.4	33.9	29.2	29.7	34.3
OOD	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.9
Misclass. vs OOD	100.0	100.0	100.0	99.3	100.0	100.0	100.0	100.0

Table 3.2.: Results of reproduction

Out-of-distribution detection

In this experiment, we still use FashionMNIST as out-of-distribution data in the training and test datasets. The purpose of this experiment is to detect whether the data is out-of-distribution. Therefore, we choose out-of-distribution samples as the positive example and use the area under ROC (AUROC) and Precision-Recall (AUPR) to evaluate their performance. The results are shown in the last two rows in Table 3.2. The second row of the table shows the results of out-of-distribution detection for all data in the MNIST and FashionMNIST datasets. In the last row of the table, we perform out-of-distribution detection for all data in the FashionMNIST dataset and misclassified data in the MNIST dataset, which have higher data uncertainty. From these results, we can conclude that Dirichlet Prior Network can determine the out-of-distribution data using the uncertainty measures on MNIST and FashionMNIST datasets.

We can conclude from the above two experiments that the entropy of μ is not as helpful as we saw in the synthetic experiments. The authors in [MG18] explained that this is due to the low data uncertainty of MNIST. Therefore, in the next section, we run experiments on corrupted MNIST, which has higher data uncertainty than MNIST.

(a) Training set (b) FashionMNIST (c) Test set 1 (d) Test set 2

Figure 3.4.: Training and test sets

3.2.2. Corrupted MNIST

In this experiment, we want to test whether the model can distinguish between the in-domain dataset with or without data uncertainty and the out-of-distribution dataset given uncertainty measures.

	Dataset	**Size**	**Erased regions**%
Training in domain	MNIST	60000	50
Training OOD	FashionMNIST	60000	0
Test set 1	MNIST	10000	0
Test set 2	MNIST	10000	100
Test set OOD	FashionMNIST	10000	0

Table 3.3.: Training and evaluation datasets

Setup

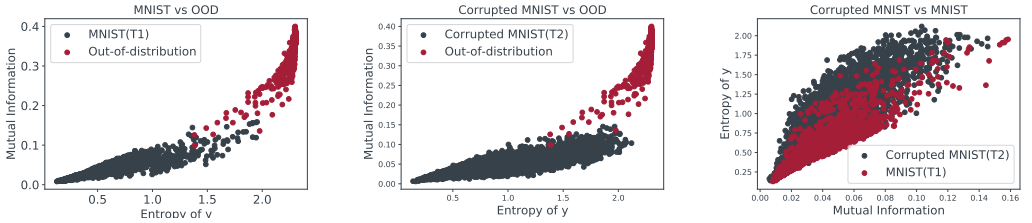
The dataset information can be found in Table 3.3. For the training dataset, we partially erased half of the digits, and the location of the erasure was utterly random. The test set 1 is the original MNIST test set, while for test set 2, we partially erase all the digits, and the erasure position is also random. We repeat the two experiments in Chapter 3.2.1.

Results

		AUROC				AUPR			
Detections	**Sets**	**Ent. y**	**Ent.μ**	**M.I.**	**D.U.**	**Ent. y**	**Ent.μ**	**M.I.**	**D.U.**
Misclassification	T1	92.9	87.3	80.2	93.4	25.6	17.9	11.9	25.8
	T2	86.7	77.7	71.9	87.0	45.4	35.7	29.7	45.8
OOD	T1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.9
	T2	100.0	100.0	100.0	99.9	100.0	100.0	100.0	99.8
In-domain	T1,T2	72.0	67.5	59.2	72.7	73.8	68.0	60.0	74.5

Table 3.4.: Results

Table 3.4 illustrates that DPN can classify both in-domain and out-of-distribution data using all uncertainty measures for our new dataset. For the detection of misclassification, the expected uncertainty of data performs better than the others. Test set 1 (T1 MNIST) has better AUROC results and worse AUPR results than test set 2 (T2 Corrupted MNIST) for all uncertainty measures, which is expected because the test set 2 has higher data uncertainty than test set 1, and AUPR is more sensitive to imbalance classes. For in-domain classification between Test set 1 and Test set 2, we performed poorly for the distinction between in-domain with or without uncertainty with all the uncertainty measures.



(a) MNIST vs OOD (b) Corrupted MNISTvs OOD (c) MNIST vs Corrupted MNIST

Figure 3.5.: Visualization of classification between different test sets.

As shown in the first two plots of Fig. 3.5, out-of-distribution and in-domain data can be classified by uncertainty measures. However, for the corrupted MNIST dataset, it is not easy to classify it and the MNIST dataset perfectly. In conclusion, we can use this Dirichlet Prior Network with reverse KL-divergence loss to classify the out-of-distribution and in-domain data.

4. Discussion and Conclusion

4.1. Discussion

This section discusses the following two questions: why our results are partially different from the paper and why we have difficulties estimating data uncertainty.

4.1.1. Comparison with previously published results

If we compare our results with those in [MG18], we observe a different behavior of differential entropy in the synthetic experiments. We used the reverse KL divergence loss, while the authors used the KL divergence loss in that paper.

(a) KL (b) RKL

Figure 4.1.: Target distribution using KL and RKL

The difference between KL divergence and reverse KL-divergence is that when there is a large amount of data uncertainty, the target distribution will be multi-modal with lower precision, as shown in Fig. 4.1a if we use KL-divergence as a loss function. On the contrary, when we use reverse KL-divergence, the target distribution will be uni-modal with high precision, as illustrated in Fig. 4.1b. In short, the reverse KL-divergence is better suited to handle data with high data uncertainty. The properties and analysis of reverse KL-divergence can be found in the paper[MG19].

4.1.2. Problem

Why does our model not distinguish well between data with and without class overlap? To answer this question, we first look at five ideal Dirichlet distribution examples illustrated in Fig. 4.2, with the following setup:

For **in-domain data**, we build a sharp Dirichlet distribution concentrated in the upper corner of the simplex with a concentration parameter of 10, as shown in Fig.

Data	Dirichlet distribution
(a)In-domain	Dir(1,1,10)
(b)High data uncertainty	Dir(10,10,10)
(c)Perfect class overlap	Dir(1,10,10)
(d)Low data uncertainty	Dir(1,2,10)
(e)Out-of-distribution	Dir(1,1,1)

Table 4.1.: Setup

4.2a.

For data point with **high data uncertainty**, we build a sharp Dirichlet distribution concentrated in the center of the simplex with equal concentration parameters (10) and high precision (30), as shown in Fig. 4.2b.

For data point in the **perfect overlap region of two classes** (class A and B), perfect here means that the point is equally likely to belong to class A and B. We build a sharp Dirichlet distribution between two corners of the simplex with concentration parameters (1,10,10), as shown in Fig. 4.2c.

For data point with **low data uncertainty** (about 17% may belong to class B and 83% to class A), we construct a sharp Dirichlet distribution concentrated between two corners, but near the upper corner of the simplex, with concentration parameters (1,2,10), as shown in Fig. 4.2e.

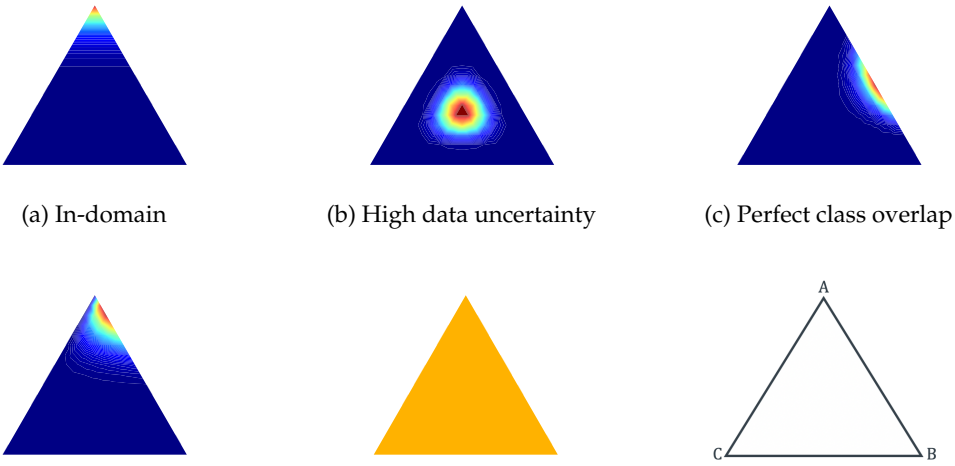
For **out-of-distribution** data, we build a flat Dirichlet distribution with concentration parameters (1,1,1), as shown in Fig. 4.2f.

We calculate the uncertainty measures directly on these example distributions, and the results as shown below:

	Ent.y	**Ent.μ**	**D.U.**	**M.I.**
In-domain	0.5661	-2.9823	0.4957	0.0703
Low data uncertainty	0.6871	-2.6140	0.6194	0.0677
Perfect class overlap	0.8516	-2.8941	0.8082	0.0433
High data uncertainty	1.0986	-2.2691	1.0660	0.0326
Out-of-distribution	1.0986	-0.6931	0.8333	0.2653

Table 4.2.: Uncertainty measures

From the results in the table, we can see the huge difference between the data from the high data uncertainty region, Out-of-distribution region, and in-domain region. However, the measurements for in-domain data are relatively similar to those from low data uncertainty regions. Back to our corrupted MNIST dataset test set 2, we calculate the number of some special outputs(concentration parameter α)



(a) In-domain (b) High data uncertainty (c) Perfect class overlap

(d) Low data uncertainty (e) Out-of-distribution (f) Classes

Figure 4.2.: Visualization of Dirichlet Distributions

of the forward network. For these special outputs, they have multiple components α_i greater than 10. The reason we choose 10 is that: For in-domain data points, the largest component of the target distribution has a value of 100, and the rest have a value of 1. We have a total of 10 classes, so for the data point with multiple components α_i greater than 10, we regard this data point as having relatively high data uncertainty. The corrupted MNIST test dataset has 10000 images, and we only receive about 2500 kinds of outputs, which means only about 2500 images have relatively high data uncertainty. Most of them are only two to three class overlaps, and we have 10 classes. Therefore, we have a relatively low class overlap in this dataset, so in this case, it is difficult to perfectly distinguish whether the data comes from a class overlap region or a non-class overlap region. Theoretically, if we have a dataset with a large amount of data uncertainty or if our dataset has a relatively low label space dimension, we can distinguish well between the in-domain data with data uncertainty, in-domain data without data uncertainty, and the out-of-distribution data. However, in real life, most clinical image datasets are unlikely to have as perfect data uncertainty as the toy example in Fig. 4.2b. Hence, it is challenging to get good results when we apply this model to medical diagnosis.

4.1.3. Further work

If we want to apply this approach in medical image analysis, we first need to find baselines for out-of-distribution detection and misclassification detection and find suitable out-of-distribution datasets. Then, we also need to consider whether the size of the in-domain and out-of-distribution data has an impact on the results.

4.2. Conclusion

In this work, we first reproduce the experiments from [MG18] using Dirichlet Prior Network and reverse KL-divergence loss. Then we train the model on a corrupted MNIST dataset and evaluate it by misclassification and out-of-distribution detection. The experimental results show that Dirichlet Prior Networks can detect out-of-distribution data from datasets with varying degrees of data uncertainty, which indicates that it is accurate in estimating distributional uncertainty. The measures of expected data uncertainty and total uncertainty outperform other measures regarding in-domain misclassification detection. However, the model can not perfectly distinguish between in-domain data and in-domain data with some data uncertainty, as illustrated in the last corrupted MNIST experiment.

The difference between out-of-distribution and in-domain data of medical images is not as large as MNIST and FashionMNIST. Thus, the problem of selecting the out-of-distribution data for applying this approach to medical image analysis remains open.

A. Appendix

A.1. Reverse KL-divergence

A.1.1. KL Divergence between two Dirichlet Distribution

$$KL(P||Q) = \int p(x) \log \frac{p(x)}{q(x)} dx \tag{A.1}$$

$$= \int p(x)(\log p(x) - \log q(x)) dx \tag{A.2}$$

$$= \int p(x)(\underbrace{\log\Gamma(\alpha_0) - \log\Gamma(\beta_0) + \sum_c^C \log\Gamma(\beta_c) - \sum_c^C \log\Gamma(\alpha_c)}_{A} + \sum_c^C ((\alpha_c - 1) - \beta_c - 1) \log x_c) dx \tag{A.3}$$

$$= A \int p(x) dx + \sum_c^C ((\alpha_c - 1) - \beta_c - 1) \int p(x) \log x_c dx \tag{A.4}$$

$$= A + \sum_c^C ((\alpha_c - 1) - \beta_c - 1)(F(\alpha_c) - F(\alpha_0)) \tag{A.5}$$

$$= \log\Gamma(\alpha_0) - \log\Gamma(\beta_0) + \sum_c^C \log\Gamma(\beta_c) - \sum_c^C \log\Gamma(\alpha_c) + \sum_c^C ((\alpha_c - 1) - \beta_c - 1)(F(\alpha_c) - F(\alpha_0)) \tag{A.6}$$

A.1.2. Derivative of KL-divergence

$$\frac{\partial KL(P||Q)}{\partial \beta_k} = -F(\beta_k + \sum_{i \neq k} \beta_i) + F(\beta_k) + F(\alpha_0) - F(\alpha_k) \tag{A.7}$$

A.1.3. Definition of reverse KL-divergence loss

$$\mathcal{L}^{RKL}(y, x, \theta; \beta) = \sum_c \mathcal{I}(y = w_c) \times KL[p(\mu|x; \theta)||p(\mu|\beta^c)] \tag{A.8}$$

A.2. Derivations for Uncertainty Measures

A.2.1. Mutual Information

$$\begin{aligned}\mathcal{I}(y, \mu|x^*, \hat{\theta}) &= \mathcal{H}[P(y|x^*; \mathcal{D})] - \mathbb{E}_{p(\mu|x^*;\mathcal{D})}[\mathcal{H}(P(y|\mu)] \\ &= -\sum_c \frac{\alpha_c}{\alpha_0}(\ln \frac{\alpha_c}{\alpha_0} - F(\alpha_c + 1) + F(\alpha_0 + 1))\end{aligned}$$

A.2.2. Expected data uncertainty

$$\mathbb{E}_{p(\mu|x^*;\mathcal{D})}[\mathcal{H}(P(y|\mu)] = \frac{\alpha_c}{\alpha_0}(F(\alpha_c + 1) - F(\alpha_0 + 1))$$

Bibliography

[Aga18] Agata Gri. Puppies or food? 12 pics that will make you question reality, 2018.

[Ale] Alexander London. Unsplash-cat.

[Dir22] Dirichlet distribution. Dirichlet distribution — Wikipedia, the free encyclopedia, 2022. [Online; accessed 6-July-2022].

[GG16] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1050–1059. JMLR.org, 2016.

[LLLS17] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *CoRR*, abs/1711.09325, 2017.

[LPB17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6402–6413, 2017.

[MG18] Andrey Malinin and Mark J. F. Gales. Predictive uncertainty estimation via prior networks. *CoRR*, abs/1802.10501, 2018.

[MG19] Andrey Malinin and Mark J. F. Gales. Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14520–14531, 2019.

[MRKG17] Andrey Malinin, Anton Ragni, Kate Knill, and Mark J. F. Gales. Incorporating uncertainty into deep learning for spoken language assessment.

In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 45–50. Association for Computational Linguistics, 2017.

[NHL20] Jay Nandy, Wynne Hsu, and Mong-Li Lee. Towards maximizing the representation gap between in-domain \& out-of-distribution examples. *CoRR*, abs/2010.10474, 2020.