A VARIATIONAL DIRICHLET FRAMEWORK FOR OUT-OF-DISTRIBUTION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

With the recently rapid development in deep learning, deep neural networks have been widely adopted in many real-life applications. However, deep neural networks are known to have very little control over its uncertainty for test examples, which can potentially cause very harmful and annoying consequences in practical scenarios. In this paper, we are particularly interested in designing a higher-order uncertainty metric for deep neural networks and investigate its performance on the out-of-distribution detection task proposed by Hendrycks & Gimpel (2016). Our method is based on a variational inference framework where we interpret the output distribution $p(x)$ as a stochastic variable z lying on a simplex of multi-dimensional space and represent the higher-order uncertainty via the entropy of the latent distribution $p(z)$. Under the variational Bayesian framework with a given dataset D, we propose to adopt Dirichlet distribution as the approximate posterior $F_\theta(z|x)$ to approach the true posterior distribution $p(z|D)$ by maximizing the evidence lower bound of marginal likelihood. By identifying the over-concentration issue in the Dirichlet framework, we further design a log-scaling smoothing function to avert such issue and greatly increase the robustness of the entropy-based uncertainty measure. Through comprehensive experiments on various datasets and architectures, our proposed variational Dirichlet framework is observed to yield state-of-the-art results for out-of-distribution detection.

1 INTRODUCTION

Recently, deep neural networks (LeCun et al., 2015) have surged and replaced the traditional machine learning algorithms to demonstrate its potentials in many real-life applications like speech recognition (Hannun et al., 2014), image classification (Deng et al., 2009; He et al., 2016), and machine translation (Wu et al., 2016; Vaswani et al., 2017), reading comprehension (Rajpurkar et al., 2016), etc. However, unlike the traditional machine learning algorithms like Gaussian Process, etc, deep neural networks are very limited in their capability to measure their uncertainty over the unseen test cases and tend to produce over-confident predictions. Such over-confidence issue (Amodei et al., 2016; Zhang et al., 2016) is known to be harmful or offensive in real-life applications. Even worth, such models are prone to adversarial attacks and raise concerns in AI safety (Goodfellow et al., 2014; Moosavi-Dezfooli et al., 2016). Therefore, it is very essential to design a robust and accurate uncertainty metric in deep neural networks in order to better deploy them into real-world applications.

Recently, an out-of-distribution detection task has been proposed in Hendrycks & Gimpel (2016) as a benchmark to promote the uncertainty research in the deep learning community. In the baseline approach, a simple method using the highest softmax score is adopted as the indicator for the model's confidence to distinguish in- from out-of-distribution data. Later on, many follow-up algorithms (Liang et al., 2017; Lee et al., 2017; Shalev et al., 2018; DeVries & Taylor, 2018) have been proposed to achieve better performance on this benchmark. In ODIN (Liang et al., 2017), the authors follow the idea of temperature scaling and input perturbation (Pereyra et al., 2017; Hinton et al., 2015) to widen the distance between in- and out-of-distribution examples. Later on, adversarial training (Lee et al., 2017) is introduced to explicitly introduce boundary examples as negative training data to help increase the model's robustness. In DeVries & Taylor (2018), the authors proposed to directly output a real value between [0, 1] as the confidence measure. The most recent

paper (Shalev et al., 2018) leverages the semantic dense representation into the target labels to better separate the label space and uses the cosine similarity score as the confidence measure.

Though the above methods (Shalev et al., 2018; Liang et al., 2017; Hendrycks & Gimpel, 2016) achieve significant improvements on various datasets and neural architectures, they are still restricted to pursuing better representation for lower-order uncertainty (the uncertainty over the predicted class) in neural networks. Inspired by Subjective Logic (Jøsang, 2016; Yager & Liu, 2008; Sensoy et al., 2018), we are motivated to pursue an effective higher-order uncertainty metric (the uncertainty over a distribution) in deep neural networks. In order to leverage randomness in the output distribution $p(x)$, we first view the output $p(x)$ of the neural network as a latent variable z lying on the simplex of multi-dimensional space $\mathbb{S}_K$, and then investigate the higher-order uncertainty by studying the randomness of z. Specifically, the more random z is (e.g. higher entropy), the more uncertain the network is about its output distribution. Thus, the problem of finding uncertainty metric simply becomes inferring the true latent distribution $p(z)$ of the random variable z. Under a Bayesian framework with a given dataset D containing (x, y) pairs, the latent distribution $p(z)$ can be understood as the posterior distribution $p(z|D)$, which is approximated by a parameterized Dirichlet function $F_\theta(z|x)$ through maximization of the evidence lower bound. We use the approximated Dirichlet's entropy as the network's higher-order uncertainty for out-of-distribution detection. However, we observe an overwhelming over-concentration problem, which greatly hinders the detection accuracy of our framework. Therefore, we propose to smooth the Dirichlet distribution by calibrating the concentration parameters and lower model's over-confidence on unseen data. Combined with the input perturbation method (Liang et al., 2017; Krizhevsky & Hinton, 2009), our proposed Dirichlet framework can greatly widen the distance between in- and out-of-distribution data to achieve state-of-the-art performances on various datasets and architectures.

The contributions of this paper are described as follows:

- We propose a variational Dirichlet algorithm for deep neural network classification problem and define a higher-order uncertainty measure.
- We identify the over-concentration issue in our Dirichlet framework and propose a smoothing method to alleviate it.
- Our method of learning higher-order uncertainty is able to achieve state-of-the-art performance on various datasets and architectures.

2 MODEL

The traditional classification networks only contain deterministic operation to output fixed probability distribution $p(x)$ for each x. Such architecture is only able to express label-wise confidence rather than distribution-wise confidence. In order to enable neural networks to output their confidence or uncertainty on the distribution level, we propose to view the output distribution $p(x)$ as a stochastic variable z, where each potential value of z is a particular distribution $p(x)$ over K classes. The random variable z lies in the K (number of classes) dimensional support space $\mathbb{S}_k$, where $\mathbb{S}_k = \{z| \sum_{i=1}^{K} z_i = 1; 0 \leq z_i \leq 1\}$. In order to measure the randomness of the deep neural network, we define a distribution $p(z)$ on the support space $\mathbb{S}_k$ to model the stochastic variable z, where $p(z)$ could be intuitively viewed as an underlying factory producing multinomial distributions. Then we can use statistical properties of $p(z)$ to quantitatively study the higher-order uncertainty of deep neural networks.

With the above problem definition, we can interpret the problem of finding the higher-order uncertainty measure as inferring the underlying higher-order distribution $p(z)$. Given the dataset D containing i.i.d. observations of tuples (x, y), we assume that $p(z) = \lim_{|D|\to\infty} p(z|D)$. Here we consider using Bayesian analysis, where we compute a true posterior distribution $p(z|D)$ through Bayes' rule $p(z|D) = \frac{p(D|z)p(z)}{\int_z p(D|z)p(z)dz}$. However, direct computation is impossible due to the intractable integral, therefore, we resort to the variational inference framework, where we optimize the parameters θ of some parameterized model $F_\theta(z|x)$ towards the true posterior distribution $p(z|D)$ via Kullback-Leibler divergence $D_{KL}(F_\theta(z|x)||p(z|D))$. This KL-divergence is minimized in practice by maximizing the evidence lower bound $\mathcal{L}$ of the marginal likelihood $p(y|x) = \int_z p(y|z, x)p(z|x)dz$ of the

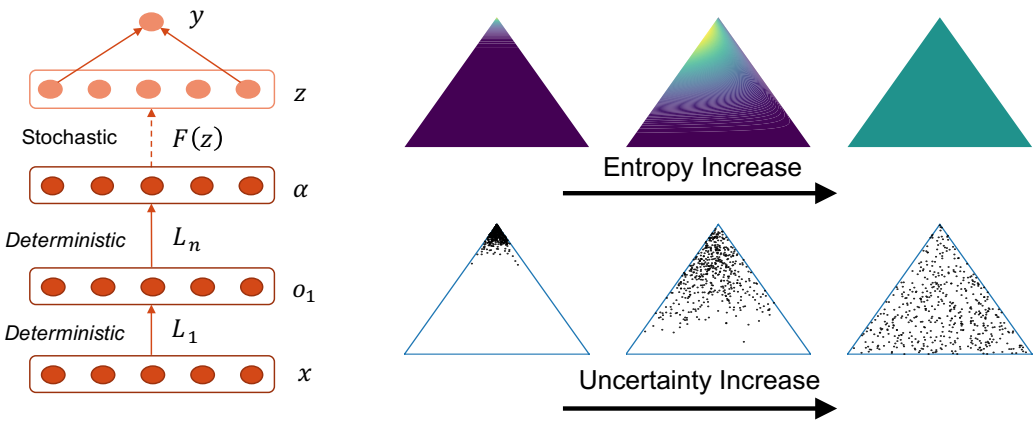


Figure 1: An intuitive explanation of higher-order distribution to measure model uncertainty.

data D as follows:

$$\mathcal{L} = \mathop{\mathbb{E}}_{z \sim F_\theta(z|x)} [\log p(y|z) - D_{KL}(F_\theta(z|x)||p(z|x))] \tag{1}$$

$\mathcal{L}$ is known to contain two parts: classification loss and KL loss, where $p(z|x)$ represents the prior distribution, which our approximated function needs to follow. The lower bound $\mathcal{L}$ plus $D_{KL}(F_\theta(z|x)||p(z|D))$ equals the (conditional) marginal log-likelihood $\sum_{(x,y)\in D} logp(y|x)$. Since the marginal likelihood is constant w.r.t. θ, maximizing $\mathcal{L}$ is equivalent to minimizing the $D_{KL}(F_\theta(z|x)||p(z|D))$.

In order to optimize $F_\theta(z|x)$ to approximate the true posterior $p(z|D)$, we adopt the widely used Dirichlet distribution $F_\theta(z|x) = Dir(z|\alpha)$ with parameters θ since its support space exactly conforms to $\mathbb{S}_k$. The probability density function (pdf) of Dirichlet distribution over all possible values of the K-dimensional stochastic variable z can be written as:

$$Dir(z|\alpha) = \begin{cases} \frac{1}{B(\alpha)} \prod_{i=1}^{K} z_i^{\alpha_i - 1} & for \quad z \in \mathbb{S}_k \\ 0 & otherwise, \end{cases} \tag{2}$$

where α is the concentration parameter of the Dirichlet distribution and $B(\alpha) = \frac{\prod_i^K \Gamma(\alpha_i)}{\Gamma(\sum_i^k \alpha_i)}$ is the normalization factor. In order to choose an appropriate prior distribution, we clip the maximum value of α to one and obtain α_c (like [1, 1, 3] $\rightarrow$ [1, 1, 1]), then we compute its KL-divergence with the non-informative Dirichlet distribution $Dir(z|U) = Dir(1, 1, \cdots, 1)$. It not only regularizes the model's concentration but also maintains network's belief about the maximum dimension. Therefore, we re-write the parameterized lower bound $\mathcal{L}$ as follows:

$$\mathcal{L} = \sum_{(x,y)\in D} [\mathop{\mathbb{E}}_{z \sim Dir(z|\alpha)} [\log z_y] - D_{KL}(Dir(z|\alpha_c)||Dir(z|U))] \tag{3}$$

Fortunately, both the classification loss and KL loss adopt closed-form solutions:

$$\mathop{\mathbb{E}}_{z \sim Dir(z|\alpha)} [\log z_y] = \psi(\alpha_y) - \psi(\alpha_0) \tag{4}$$

$$D_{KL}(Dir(z|\alpha_c)||Dir(z|U)) = \log \frac{1}{B(\alpha_c)\Gamma(k)} + \sum_{i=1}^{k} (\alpha_{ci} - 1)(\psi(\alpha_{ci}) - \psi(\alpha_{c0})) \tag{5}$$

where α_{c0} is the sum of clipped concentration α_c over K dimensions, Γ denotes the gamma function, ψ denotes the digamma function. Here we parameterize the poster Dirichlet with a deep neural network $f_\theta(x)$, which takes x as input to obtain a real-valued K-dimensional vector. Since α is required to be larger than zero, we adopt an exponential transform to project the output into the positive real domain $\mathbb{R}^+$ by $\alpha = \exp(f_\theta(x))$. We write the derivative of $\mathcal{L}$ w.r.t to parameters θ

based on the chain-rule: $\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}}{\partial \alpha} \odot \alpha \cdot \frac{\partial f_\theta(x)}{\partial \theta}$, where $\odot$ is the Hardamard product and $\frac{\partial f_\theta(x)}{\partial \theta}$ is the Jacobian matrix. We first re-weight the two terms in $\mathcal{L}$ and then propose to use mini-batch gradient descent to optimize the network parameters θ as follows:

$$\frac{\partial \mathcal{L}}{\partial \alpha} = \sum_{x,y \in B(x,y)} [\frac{\partial[\psi(\alpha_y) - \psi(\alpha_0)]}{\partial \alpha} + \eta \frac{\partial D_{KL}(F(z|\alpha_c)||Dir(z|U))}{\partial \alpha}] \tag{6}$$

where $B(x, y)$ denotes the mini-batch in dataset D, and η is the balancing factor between KL regularizer and log-likelihood terms. During inference, we use the maximum α's index as the model prediction class $\hat{y} = \arg\max_i f_{\theta,i}(x)_i$, where $f_{\theta,i}(x)$ denotes the i_{th} dimension of network output.

3 Uncertainty Measure

After optimization, we obtain a parametric Dirichlet function $F_\theta(z|\alpha)$ and compute its entropy E as the higher-order uncertainty measure. Formally, we write the such metric as follows:

$$E(\alpha) = -C(\alpha) = \log B(\alpha) + (\alpha_0 - K)\psi(\alpha_0) - \sum_i^k (\alpha_i - 1)\psi(\alpha_i) \tag{7}$$

where α is computed via the deep neural network f_θ, and α_0 is the sum of α over K dimensions. Here we use negative of entropy as the confidence score $C(\alpha)$. By investigating the magnitude distribution of concentration parameter α for in-distribution test cases, we can see that α is either adopting the prior $\alpha = 1.0$ or adopting a very large value $\alpha \gg 1.0$. In order words, the Dirichlet distribution is heavily concentrated at a corner of the simplex to produce very similar distributions $p(x)$ (see the first right hand Dirichlet in Figure 1). Specifically, the model holds very strong confidence about what the form of output distribution $p(x)$ on unseen test cases even though it cannot correctly classify them. Such over-concentration issue makes the model very prone to out-of-distribution noise and leads to compromised detection accuracy. Therefore, we propose to smooth the concentration parameters α to decrease model's confidence for better separating in- and out-of-distribution data.

Concentration smoothing In order to smooth the concentration parameters α, we adopt the log-scaling function $\hat{\alpha} = \log(\alpha + 1)$ to calibrate the concentration α and obtain the smoothed confidence score $C(\hat{\alpha})$. By plotting the histogram of concentration magnitude before and after log-scaling in Figure 2, we can observe a very strong smoothing effect in the Dirichlet distribution, where the Dirichlet is less concentrated on certain dimensions. We also demonstrate the histogram of confidence score for in- and out-of-distribution samples before and after log-scaling in Figure 2, we can see the two distributions are better separated. In the experimental section, we also compare different smoothing functions to discuss its effect.

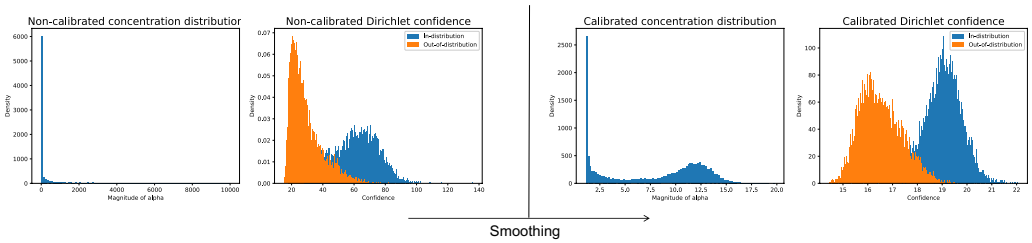


Figure 2: Concentration and confidence distribution before and after smoothing for CIFAR10 under VGG13 architecture with iSUN as out-of-distribution dataset.

Input Perturbation Inspired by fast gradient sign method (Goodfellow et al., 2014), we propose to add perturbation in the data before feeding into neural networks:

$$\hat{x} = x - \epsilon * sign(\nabla_x[\psi(\alpha_0) - \psi(\alpha_y)]) \tag{8}$$

where the parameter ϵ denotes the magnitude of the perturbation, and (x, y) denotes the input-label data pair. Here, similar to Liang et al. (2017) our goal is also to improve the entropy score of any

given input by adding belief to its own prediction. Here we make a more practical assumption that we have no access to any form of out-of-distribution data. Therefore, we stick to a rule-of-thumb value $\epsilon = 0.01$ throughout our experiments.

Detection For each input x, we first use input perturbation to obtain $\hat{x}$, then we feed it into neural network $f_\theta(\hat{x})$ to compute the concentration α, finally we use log-scaling to calibrate α and compute $C(\hat{\alpha})$. Specifically, we compare the confidence $C(\hat{\alpha})$ to the threshold δ and say that the data x follows in-distribution if the confidence score $C(\hat{\alpha})$ is above the threshold and that the data x follows out-of-distribution, otherwise.

4 EXPERIMENTS

In order to evaluate our variational Dirichlet method on out-of-distribution detection, we follow the previous paper (Hendrycks & Gimpel, 2016; Liang et al., 2017) to replicate their experimental setup. Throughout our experiments, a neural network is trained on some in-distribution datasets to distinguish against the out-of-distribution examples represented by images from a variety of unrelated datasets. For each sample fed into the neural network, we will calculate the Dirichlet entropy based on the output concentration α, which will be used to predict which distribution the samples come from. Finally, several different evaluation metrics are used to measure and compare how well different detection methods can separate the two distributions.

4.1 IN-DISTRIBUTION AND OUT-OF-DISTRIBUTION DATASET

These datasets are all available in Github [1].

- CIFAR10/100 (in-distribution): The CIFAR-10 and CIFAR100 dataset (Krizhevsky & Hinton, 2009) consists of RGB images of 32×32 pixels. Each image is classified into 10/100 classes, such as dog, cat, automobile, or ship. The training split for both datasets is comprised of 50,000 images, while the test split is comprised of 10,000 images.
- SVHN (in-distribution): The Street View Housing Numbers (SVHN) dataset (Netzer et al., 2011) consists of colored housing number pictures ranging from 0 to 9. Images are also with a resolution of 32×32. The official training split is comprised of 73,257 images, and the test split is comprised of 26,032 images.
- TinyImageNet (out-of-distribution): The TinyImageNet dataset2 is a subset of the ImageNet dataset (Deng et al., 2009). The test set for TinyImageNet contains 10,000 images from 200 different classes for creating the out-of-distribution dataset, it contains the original images, downsampled to 32×32 pixels.
- LSUN (out-of-distribution): The Large-scale Scene UNderstanding dataset (LSUN) (Yu et al., 2015) has a test set consisting of 10,000 images from 10 different scene classes, such as bedroom, church, kitchen, and tower. We downsample LSUN's original image and create 32×32 images as an out-of-distribution dataset.
- iSUN (out-of-distribution): The iSUN dataset (Xu et al., 2015) is a subset of the SUN dataset, containing 8,925 images. All images are downsampled to 32×32 pixels.

Before reporting the out-of-distribution detection results, we first measure the classification accuracy of our proposed method on the two in-distribution datasets in Table 1, from which we can observe that our proposed algorithm has minimum impact on the classification accuracy.

4.2 TRAINING DETAILS

In order to make fair comparisons with other out-of-distribution detectors, we follow the same setting of Liang et al. (2017); Zagoruyko & Komodakis (2016); DeVries & Taylor (2018); Shalev et al. (2018) to separately train WideResNet (Zagoruyko & Komodakis, 2016) (depth=16 and widening factor=8 for SVHN, depth=28 and widening factor=10 for CIFAR100), VGG13 (Simonyan & Zisserman, 2014), and ResNet18 (He et al., 2016) models on the in-distribution datasets. All models are

[1] `https://github.com/ShiyuLiang/odin-pytorch`

Dataset	Cross-Entropy			Ours		
	VGG13	WideResNet	ResNet-18	VGG	WideResNet	ResNet-18
CIFAR10	93.2	-	94.4	93.8 (+0.6)	-	94.8 (+0.4)
CIFAR100	-	81.4	-	-	81.8 (+0.4)	-
SVHN	96.8	96.7	-	95.9 (-0.9)	96.1 (-0.6)	-

Table 1: Classification accuracy of Dirichlet framework on various datasets and architectures.

trained using stochastic gradient descent with Nesterov momentum of 0.9, and weight decay with 5e-4. We train all models for 200 epochs with 128 batch size. We initialize the learning with 0.1 and reduced by a factor of 5 at 60th, 120th and 180th epochs. we cut off the gradient norm by 1 to prevent from potential gradient exploding error. We save the model after the classification accuracy on validation set converges and use the saved model for out-of-distribution detection.

4.3 Experimental Results

We measure the quality of out-of-distribution detection using the established metrics for this task (Hendrycks & Gimpel, 2016; Liang et al., 2017; Shalev et al., 2018).

- FPR at 95% TPR (lower is better): Measures the false positive rate (FPR) when the true positive rate (TPR) is equal to 95%.
- Detection Error (lower is better): Measures the minimum possible misclassification probability defined by $\min_\delta\{0.5P_{in}(f(x) \leq \delta) + 0.5P_{out}(f(x) > \delta)\}$.
- AUROC (larger is better): Measures the Area Under the Receiver Operating Characteristic curve. The Receiver Operating Characteristic (ROC) curve plots the relationship between TPR and FPR.
- AUPR (larger is better): Measures the Area Under the Precision-Recall (PR) curve, where AUPR-In refers to using in-distribution as positive class and AUPR-Out refers to using out-of-distribution as positive class.

We report our VGG13's performance in Table 2 and ResNet/WideResNet's performance in Table 3, where we list the performance of Baseline Hendrycks & Gimpel (2016), ODIN Liang et al. (2017), Bayesian Neural Network (Gal, 2016)[2], Semantic-Representation (Shalev et al., 2018) and Learning-Confidence (DeVries & Taylor, 2018). The results in both tables have both shown remarkable improvement brought by our proposed variational Dirichlet framework. For CIFAR10/100 dataset, the achieved improvements are very significant, however, the FPR score on CIFAR100 is still unsatisfactory with nearly half of the out-of-distribution samples being wrongly detected. For SVHN dataset, the current algorithms already achieve close-to-perfect results, therefore, the improvements brought by our algorithm is comparatively minor. In order to individually study the effectiveness of our proposed methods (entropy-based uncertainty measure, concentration smoothing, and input perturbation), we design a series of ablation experiments in Table 4. 1) Our variational Dirichlet achieves higher detection accuracy than Evidential Dirichlet Sensoy et al. (2018), which confirms our entropy-based uncertainty. 2) We also observe that concentration smoothing is playing a more important role than input perturbation.

4.4 Effects on Smoothing functions and Input Perturbation

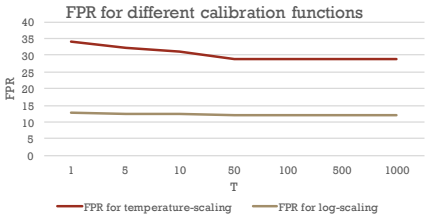
Here we investigate the impact of different smoothing functions and different input perturbation magnitude on the out-of-distribution detection accuracy. For smoothing functions, we mainly consider the following function forms: 1) temperature-scaling $\hat{\alpha} = \frac{\alpha}{T}$, 2) log-scaling $\hat{\alpha} = \log(\frac{\alpha}{T} + 1)$ with $T > 1$. Both smoothing functions are aimed at decreasing the larger concentration value while maintaining the smaller values. We gradually increase the hyper-parameter T to calibrate α and draw the FPR curve in Figure 3. For both smoothing modules, we have observed that larger T generally converges to better detection performance, however, the gain brought to log-scaling is not quite remarkable. Besides, we also gradually increase the input perturbation ϵ and draw the FPR

[2]We use the variational ratio as uncertainty measure to perform out-of-distribution detection, specifically, we forward Bayesian deep network 100 times for each input sample for Monte-Carlo estimation.

Model	OOD-Dataset	Method	FPR (TPR=0.95)	Detection Error	AUROC	AUPR In	AUPR Out
VGG13 CIFAR-10	iSUN	Baseline	43.8	11.4	94	95.5	91.5
		ODIN	22.4	10.2	95.8	96.3	94.9
		BNN	56.4	14.6	91.2	93.6	84.4
		Confidence	16.3	8.5	97.5	98	96.9
		Ours	**10.9**	**6.9**	**98.0**	**98.4**	**97.7**
	LSUN-resized	Baseline	41.9	11.5	94	95.1	92.2
		ODIN	20.2	9.8	95.9	95.8	95.8
		BNN	52.4	14.2	91.3	93.8	86.7
		Confidence	16.4	8.3	97.5	97.8	97.2
		Ours	**9.9**	**6.6**	**98.1**	**98.4**	**97.9**
	Tiny-ImageNet	Baseline	43.8	12	93.5	94.6	91.7
		ODIN	24.3	11.3	95.7	95.9	95.9
		BNN	53.7	16.9	90.2	91.9	82.6
		Confidence	18.4	9.4	97	97.3	96.9
		Ours	**13.8**	**7.9**	**97.5**	**97.8**	**97.3**
VGG13 SVHN	iSUN	Baseline	10	6	98	99.3	93.7
		ODIN	1.6	2.95	99.5	99.8	98.8
		Confidence	0.9	2.3	99.7	99.9	98.9
		Ours	**0.5**	**1.8**	**99.7**	**99.9**	**99.6**
	LSUN-resized	Baseline	9.4	5.7	98.1	99.3	94.3
		ODIN	1.4	2.6	99.6	99.7	99.1
		Confidence	1	2.3	99.7	99.9	99
		Ours	**0.8**	**2.2**	**99.8**	**99.9**	**99.2**
	Tiny-ImageNet	Baseline	11.4	6.2	97.8	99.2	93.7
		ODIN	2.3	3.4	99.3	99.7	98.6
		Confidence	1.5	2.8	99.5	99.8	98.7
		Ours	**1.4**	**2.3**	**99.7**	**99.8**	**99.2**

Table 2: Experimental Results on VGG13 architecture, where Confidence refers to Learning Confidence algorithm (DeVries & Taylor, 2018), BNN refers to Bayesian Neural Network (Gal, 2016).

curve in Figure 3, from which we can observe that a moderate perturbation of $T = 0.01$ can generally achieve promising results. In conclusion, log-scaling is more effective than temperature-scaling calibration, since there is no held-out dataset to finetune the hyper-parameters, we simply stick to $\hat{\alpha} = \log(\alpha + 1)$ throughout our experiments.



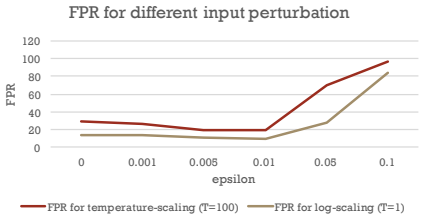


Figure 3: Impact of different smoothing hyperparameter T and different input perturbation magnitudes on out-of-distribution detection accuracy. The network architecture is VGG13 with in-distribution CIFAR10 dataset and out-of-distribution iSUN dataset.

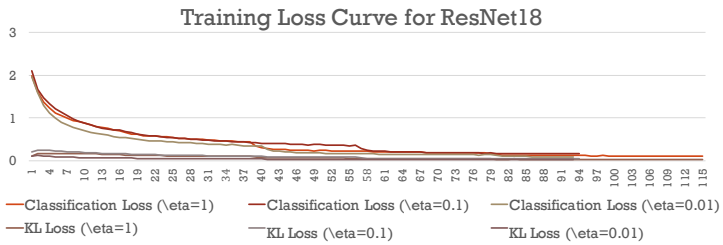
4.5 Effects on KL-divergence

Here we investigate the impact of KL-divergence in terms of both classification accuracy and detection errors. By gradually increasing the weight of KL loss (increasing the balancing factor η from 0 to 10), we plot their training loss curve in Figure 4. With a too strong KL regularization, the model's classification accuracy will decrease significantly. As long as η is within a rational range, the classification accuracy will become stable. For detection error, we can see from Figure 4 that adopting either too large value or too small value can lead to compromised performance. For the very small

Model	OOD-Dataset	Method	FPR (TPR=0.95)	Detection Error	AUROC	AUPR In	AUPR Out
ResNet18 CIFAR-10	iSUN	Baseline	52.6	13.6	92.4	94.6	88.9
		ODIN	22.7	9.6	96.3	97.2	95
		Semantic	21.5	9.2	96.3	97.1	94.3
		Ours	**13.8**	**8.4**	**97.1**	**97.4**	**96.8**
	LSUN-resized	Baseline	50.2	12.3	93.1	94.8	90.8
		ODIN	17.9	8.4	96.9	97.5	96.3
		Semantic	23	14	96	96.7	94.8
		Ours	**12.1**	**6.3**	**97.4**	**97.4**	**97.5**
	Tiny-ImageNet	Baseline	59	15.1	91.1	93.2	88.1
		ODIN	32.1	11.2	94.9	95.8	93.6
		Semantic	32.1	13.1	93.2	94.2	90.6
		Ours	**18.4**	**9.9**	**95.9**	**95.8**	**96.1**
WideResNet CIFAR100	iSUN	Baseline	82.7	43.9	72.8	74.2	69.2
		ODIN	57.3	31.1	85.6	85.9	84.8
		Ours	**44.7**	**25.3**	**88.4**	**88.2**	**88.1**
	LSUN-resized	Baseline	82.2	43.6	73.9	75.7	70.1
		ODIN	56.5	30.8	86	86.2	84.9
		Ours	**45.9**	27.3	87.9	88.1	87.9
	Tiny-ImageNet	Baseline	79.2	42.1	72.2	70.4	70.8
		ODIN	55.9	30.4	84.0	82.8	84.4
		Ours	**51.2**	**28.6**	**86.1**	**84.9**	**86.1**
WideResNet SVHN	iSUN	Baseline	9.6	5.9	98	99.3	93.4
		ODIN	1.1	2.7	99.6	99.8	99.1
		Ours	**0.8**	**1.8**	**99.8**	**99.9**	**99.5**
	LSUN-resized	Baseline	9.5	5.8	98	99.3	94
		ODIN	1.5	2.9	99.6	99.8	99
		Ours	**0.6**	**1.5**	**99.8**	**99.9**	**99.5**
	Tiny-ImageNet	Baseline	10.6	6.1	97.8	99.2	93.6
		ODIN	2.1	3.2	99.5	99.8	98.8
		Ours	**1.8**	**2.8**	**99.8**	**99.8**	**99.5**

Table 3: Experimental results for ResNet architecture, where Semantic refers to multiple semantic representation algorithm (Shalev et al., 2018)

value $\eta \to 0$, the variational Dirichlet framework degrades into a marginal log-likelihood, where the concentration parameters are becoming very erratic and untrustworthy uncertainty measure without any regularization. For larger $\eta > 1$, the boundary for in- and out-of-distribution becomes fuzzier leading to worse detection performance. Without further finetuning the hyper-parameter η on the held-out dataset, we simply stick to $\eta = 0.01$ throughout our experiments.



\eta	Classification accuracy	FPR at (TPR=0.95)
0	93.8	0.45
0.001	94.8	0.26
0.01	94.4	0.13
0.1	94.1	0.28
1	93.4	0.29
10	88.7	0.42

Figure 4: The training loss curve under ResNet18 on CIFAR10 dataset for different η is demonstrated on the left side, the accuracy and out-of-distirbution detection results on the right side.

5 Related Work

The novelty/anomaly detection problem (Pimentel et al., 2014) has already a long-standing research topic in traditional machine learning community, the previous works (Vincent & Bengio, 2003;

Model	OOD-Dataset	Method	FPR (TPR=0.95)	Detection Error	AUROC	AUPR In	AUPR Out
	iSUN	E-Dirichlet	39.9	13.2	93.7	95.2	92.0
		Dirichlet	34.0	11.4	94.8	96.0	93.3
		Dirichlet +Smooth	14.4	7.9	97.4	97.9	96.8
		Dirichlet +Smooth +Perturbation	**10.9**	**6.9**	**98.0**	**98.4**	**97.7**
	LSUN-resized	E-Dirichlet	40.0	12.8	93.9	95.0	92.9
		Dirichlet	32.5	11.3	94.9	95.8	94.0
		Dirichlet +Smooth	13.8	7.7	97.5	97.8	97.2
		Dirichlet +Smooth +Perturbation	**9.9**	**6.6**	**98.1**	**98.4**	**97.9**
	Tiny-ImageNet	E-Dirichlet	47.1	14.2	92.5	93.8	91.2
		Dirichlet	35.1	12.1	94.3	95.1	93.4
		Dirichlet +Smooth	18.9	9.1	96.7	97.1	96.1
		Dirichlet +Smooth +Perturbation	**13.8**	**7.9**	**97.5**	**97.8**	**97.2**

Table 4: Ablation experiments for VGG13 architecture to investigate the impact of our proposed smoothing technique, where E-Dirichlet refers to Evidential Dirichlet (Sensoy et al., 2018).

Ghoting et al., 2008; Schlegl et al., 2017) have been mainly focused on low-dimensional and specific tasks. Their methods though achieving very promising performance, are known to be unreliable in high-dimensional space. Recently, more research works about detecting an anomaly in deep learning have been proposed. Akcay et al. (2018) combines the ability of deep networks to extract a progressively rich representation of data with the one-class objective of creating a tight envelope around normal data. Lee et al. (2017); Chalapathy et al. (2018) propose to explicitly generate adversarial samples to enhancing model's robustness.

Another line of research is Bayesian Networks (Gal & Ghahramani, 2016; 2015; Gal, 2016; Kingma et al., 2015), which are powerful in providing stochasticity in deep neural networks by assuming the weights are stochastic. However, Bayesian Neural Networks' uncertainty measure like variational ratio and mutual information rely on Monte-Carlo estimation, where the networks have to perform forward passes many times, which greatly reduces the detection speed. Instead of modeling the weight uncertainty in deep neural networks, our method directly models randomness of output neuron with Dirichlet distribution, which arrives in closed-form uncertainty measure and bypasses the cost of Monte-Carlo sampling.

6 Conclusion

In this paper, we aim at finding an effective way for deep neural networks to express their uncertainty over their output distribution. Our variational Dirichlet framework is empirically demonstrated to yield better results, but its detection accuracy on a more challenging setup like CIFAR100 is still very compromised. We conjecture that better prior Dirichlet distribution or smoothing function could help further improve the performance. In the future work, we plan to apply our method to broader applications like natural language processing tasks or speech recognition tasks.

References

Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. *arXiv preprint arXiv:1805.06725*, 2018.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Raghavendra Chalapathy, Aditya Krishna Menon, and Sanjay Chawla. Anomaly detection using one-class neural networks. *arXiv preprint arXiv:1802.06360*, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255. Ieee, 2009.

Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

Yarin Gal. Uncertainty in deep learning. *University of Cambridge*, 2016.

Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059, 2016.

Amol Ghoting, Srinivasan Parthasarathy, and Matthew Eric Otey. Fast mining of distance-based outliers in high-dimensional datasets. *Data Mining and Knowledge Discovery*, 16(3):349–364, 2008.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014. URL `http://arxiv.org/abs/1412.6572`.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Audun Jøsang. *Subjective Logic - A Formalism for Reasoning Under Uncertainty*. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer, 2016. ISBN 978-3-319-42335-7. doi: 10.1007/978-3-319-42337-1. URL `https://doi.org/10.1007/978-3-319-42337-1`.

Diederik P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems*, pp. 2575–2583, 2015.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2574–2582, 2016.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 5, 2011.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pp. 146–157. Springer, 2017.

Murat Sensoy, Melih Kandemir, and Lance Kaplan. Evidential deep learning to quantify classification uncertainty. *arXiv preprint arXiv:1806.01768*, 2018.

Gabi Shalev, Yossi Adi, and Joseph Keshet. Out-of-distribution detection using multiple semantic label representations. *arXiv preprint arXiv:1808.06664*, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Pascal Vincent and Yoshua Bengio. Manifold parzen windows. In *Advances in neural information processing systems*, pp. 849–856, 2003.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

Ronald R Yager and Liping Liu. *Classic works of the Dempster-Shafer theory of belief functions*, volume 219. Springer, 2008.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*, 2016. URL `http://www.bmva.org/bmvc/2016/papers/paper087/index.html`.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.