Maximizing the Representation Gap between In-domain & OOD examples

Jay Nandy [1] **Wynne Hsu** [1] **Mong Li Lee** [1]

Abstract

Among the existing uncertainty estimations approaches, only Dirichlet Prior Network (DPN) distinctly models different uncertainty types. However, in this paper, we show that for in-domain examples with high data uncertainties among multiple classes, a DPN also produces almost indistinguishable representations from the out-of-distribution (OOD) examples, compromising their OOD detection performance. We address this shortcoming by proposing a new loss function for DPN models that maximizes the *representation gaps* between the in-domain and OOD examples. Experimental results suggest that our proposed technique consistently improves OOD detection performance by solving this issue.

1. Introduction

Predictive uncertainties of a classification model can arise from three different sources (Gal, 2016; Malinin & Gales, 2018): *Model or epistemic uncertainty* captures the uncertainty to estimate the model parameters, conditioned on training data (Gal, 2016). *Data or aleatoric uncertainty* arises from the complexities of the underlying distribution, such as class overlap, label noise, homoscedastic, and heteroscedastic noise (Gal, 2016). *Distributional uncertainty* arises due to the distributional mismatch between the training and test examples (Candela et al., 2009). That is, the test data is *out-of-distribution (OOD)*.

Recently notable progress has been made for predictive uncertainty estimation using both Bayesian (Hernandez-Lobato & Adams, 2015; Gal, 2016; Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017; Hein et al., 2019; Meinke & Hein, 2020) and non-Bayesian neural networks (Lee et al., 2018a; Hendrycks et al., 2019; Liang et al., 2018; Lee et al., 2018b). However, none of these approaches robustly determine the source of predictive uncertainty (Malinin & Gales, 2018). In particular, the presence of high data uncertainty among multiple classes leads the non-Bayesian classifiers to produce uniform categorical predictions for in-domain examples, making them indistinguishable from the OOD examples.

Dirichlet Prior Network (DPN) separately models different uncertainty types by producing sharp Dirichlet distributions for in-domain examples, and flat Dirichlet distributions for OOD examples (Malinin & Gales, 2018; 2019). However, we show that for *in-domain examples with high data uncertainties*, the existing loss function for DPN leads to much flatter distributions.

In this paper, we propose an alternative approach for a DPN classifier that instead produces sharp multi-modal Dirichlet, that uniformly spreads the density at each corner of the simplex, for OOD examples to maximize the *"representation gap"* from in-domain examples. We propose a new loss function that separately models the mean and the precision of the output Dirichlet distributions by introducing a novel *explicit precision regularizer* along with the soft-max cross-entropy loss. Experimental results demonstrate that our proposed approach consistently improves OOD detection performance by addressing this issue.

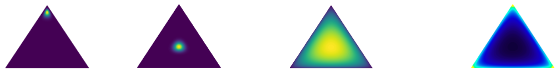
2. Dirichlet Prior Network

A DPN classification model directly parametrizes a Dirichlet distribution as the prior to the predictive categorical distribution over a simplex (Malinin & Gales, 2018; 2019). It attempts to produce a sharp Dirichlet in one corner of a simplex when it predicts confidently for the in-domain examples (Fig 1a), inducing a sharp uni-modal categorical distribution over the class labels. For in-domain examples with high data uncertainty, it attempts to produce a sharp distribution in the middle of the simplex (Fig 1b), inducing a multi-modal categorical distribution over the class labels.

Finally, for OOD examples, an existing DPN model attempts to produce a flat Dirichlet distribution to indicate high-order distributional uncertainty (see Fig 1c). However, in section 3, we demonstrate that in the case of higher data uncertainty among multiple classes, an existing DPN model also produces flatter Dirichlet distribution for in-domain examples, leading to indistinguishable representations from the OOD examples. Hence, we propose to produce sharp multi-

[1] National University of Singapore, Singapore. Correspondence to: Jay Nandy <jaynandy@comp.nus.edu.sg>.

Presented at the ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning. Copyright 2020 by the author(s).



(a) Confident (b) Data (c) Distributional (Existing) (d) Distributional (Proposed)

Figure 1. Desired outputs of a DPN under different predictive uncertainty types.

modal Dirichlet distributions, that uniformly spreads the densities at each corner of the simplex, for OOD examples (see Fig 1(d)). It increases their "representation gap" from in-domain examples, leading to improve the OOD detection performance. Note that, both Dirichlet distributions in Fig 1c) and 1d) induces uniform predictive categorical distribution over the class labels for OOD examples. As we see that, compared to Fig 1a or Fig 1b, the probability densities of both Dirichlet distribution in Fig 1c and Fig 1d are more *diverse* over the simplex. We compute this *"diversity"* using measures, such as *mutual information ($\mathcal{MI}$)* to distinguish the OOD examples (Malinin & Gales, 2018; 2019).

A *Dirichlet distribution* is parameterized using its *concentration parameters*, $\boldsymbol{\alpha} = \{\alpha_1, \cdots, \alpha_K\}$, as: $Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_{c=1}^K \Gamma(\alpha_c)} \prod_{c=1}^K \mu_c^{\alpha_c - 1}, \quad \alpha_c > 0.$ Here, $\alpha_0 = \sum_{c=1}^K \alpha_c$ is the *precision* of the Dirichlet. A larger α_0 produces a sharper uni-modal Dirichlet distribution (Fig 1a). However, as we uniformly decrease $\alpha_c < 1 \ \forall c$, we obtain a sharp multi-modal Dirichlet with the densities uniformly distributed at each corner of the simplex (Fig 1d).

Given an input $\boldsymbol{x}^*$, a DPN, $f_{\boldsymbol{\theta}}$ produces $\boldsymbol{\alpha}$ for each class, i.e, $\boldsymbol{\alpha} = f_{\boldsymbol{\theta}}(\boldsymbol{x}^*)$. The expected categorical distribution over class labels, ω_c, is given by the mean of the Dirichlet: $p(\omega_c|\boldsymbol{x}^*;\boldsymbol{\theta}) = \int p(\omega_c|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{x}^*;\boldsymbol{\theta})d\boldsymbol{\mu} = \frac{\alpha_c}{\alpha_0}$, where $p(\boldsymbol{\mu}|\boldsymbol{x}^*;\boldsymbol{\theta}) = Dir(\boldsymbol{\mu}|\boldsymbol{\alpha})$ and $\boldsymbol{\theta}$ is the model parameters.

Uncertainty Measures. For a given input $\boldsymbol{x}^*$, we can measure the *total predictive uncertainty* by computing the maximum probability ($maxP$) from its predictive categorical distribution i.e. $maxP = \max_c p(y = \omega_c|\boldsymbol{x}^*, \boldsymbol{\theta})$.

$maxP$ behaves similarly for any classification models. We get a higher $maxP$ score for in-domain confident examples. However, it produces lower scores for both data and distributional uncertainty, making it difficult to distinguish the OOD examples. Hence, it also indicates the *limitation* of the non-Bayesian models (Malinin & Gales, 2018).

In contrast, a DPN efficiently captures the *distributional uncertainty* by computing the *mutual information ($\mathcal{MI}$)* between class-labels, y and categorical $\boldsymbol{\mu}$ as: $\mathcal{MI}[y, \boldsymbol{\mu}|\boldsymbol{x}, \boldsymbol{\theta}] = \mathcal{H}[\mathbb{E}_{p(\mu|\boldsymbol{x},\boldsymbol{\theta})} P(y|\boldsymbol{\mu})] - \mathbb{E}_{p(\mu|\boldsymbol{x},\boldsymbol{\theta})} \mathcal{H}[P(y|\boldsymbol{\mu})]$. It produces lower scores for in-domain examples (when the density is concentrated in a single model as in Fig 1a or 1b) and high scores for OOD examples (Fig. 1c or 1d). We can also use the *precision*, α_0 as a distributional uncertainty measure as a DPN explicitly models to produce higher α_0 values for in-domain examples. Existing DPN models also applied differential entropy ($D.Ent$), that produces a low score for a sharp Dirichlet, to distinguish the OOD examples. However, our proposed solution behaves differently to produce a sharp multi-modal Dirichlet for an OOD example and a sharp uni-modal Dirichlet for in-domain confident predictions. Hence, we cannot detect OOD examples based on the sharpness of the output Dirichlet. (see more details in Appendix B).

Construction. A standard DNN, with the softmax cross-entropy loss, can be viewed as a DPN, such that $\alpha_c = e^{z_c(\boldsymbol{x}^*)}$; $z_c(\boldsymbol{x}^*)$ is the logit output for class, c for input $\boldsymbol{x}^*$. Here, the categorical posterior for class label ω_c is:

$$p(\omega_c|\boldsymbol{x}^*;\boldsymbol{\theta}) \ = \frac{\alpha_c}{\alpha_0} = \frac{e^{z_c(\boldsymbol{x}^*)}}{\sum_{c=1}^K e^{z_c(\boldsymbol{x}^*)}} \tag{1}$$

However, since the mean of the Dirichlet is *insensitive* to any arbitrary scaling of α_c, the precision, α_0, of the output Dirichlet degrades under cross-entropy loss.

Existing loss functions. Malinin & Gales (2018) propose the *forward KL (FKL) loss* to explicitly minimizes the KL divergence between the model and the given target Dirichlet. Malinin & Gales (2019) propose the *reverse KL (RKL)* loss, that reverses the terms in the KL divergence, to address the previous limitations of FKL, and improve the scalability of a DPN model for classification tasks with a larger number of classes. The RKL loss is given as follows:

$$\begin{aligned}\mathcal{L}^{rkl}(\boldsymbol{\theta};\gamma,\boldsymbol{\beta}^y,\boldsymbol{\beta}^{out}) &= \mathbb{E}_{P_{in}}\mathrm{KL}[p(\boldsymbol{\mu}|\boldsymbol{x},\boldsymbol{\theta})||Dir(\boldsymbol{\mu}|\boldsymbol{\beta}^y)] \\ &+ \gamma \cdot \mathbb{E}_{P_{out}}\mathrm{KL}[p(\boldsymbol{\mu}|\boldsymbol{x},\boldsymbol{\theta})||Dir(\boldsymbol{\mu}|\boldsymbol{\beta}^{out})]\end{aligned} \tag{2}$$

β^y and β^{out} are their *hand-crafted target concentration parameters*. P_{in} and P_{out} are the distribution for the in-domain and OOD training examples.

3. Proposed Methodology

In this section, we first demonstrate that the RKL loss function tends to produce flatter Dirichlet distributions for in-domain misclassified examples, compared to the confidently predicted examples. We can decompose the RKL loss using *reverse cross entropy*, $\mathbb{E}_{Pr(\boldsymbol{\mu}|\boldsymbol{x},\boldsymbol{\theta})}[-\ln Dir(\boldsymbol{\mu}|\overline{\boldsymbol{\beta}})]$ and $D.Ent$, $\mathcal{H}[p(\boldsymbol{\mu}|\boldsymbol{x},\boldsymbol{\theta})]$ (Malinin & Gales, 2019).

$$\begin{aligned}&\mathbb{E}_{\tilde{P}_T(\boldsymbol{x},y)}\mathrm{KL}\big[\,p(\boldsymbol{\mu}|\boldsymbol{x},\boldsymbol{\theta}) \;||\; Dir(\boldsymbol{\mu}|\boldsymbol{\beta})\,\big] \\ &= \mathbb{E}_{\tilde{P}_T(\boldsymbol{x})}\Big[\mathbb{E}_{P(\boldsymbol{\mu}|\boldsymbol{x},\boldsymbol{\theta})}[-\ln Dir(\boldsymbol{\mu}|\overline{\boldsymbol{\beta}})] - \mathcal{H}\big[p(\boldsymbol{\mu}|\boldsymbol{x},\boldsymbol{\theta})\big]\Big]\end{aligned} \tag{3}$$

where, ψ is the digamma function. $\boldsymbol{\beta} = \{\beta_1^{(c)}, \cdots, \beta_K^{(c)}\}$ represents their hand-crafted target concentration parameters. $\overline{\boldsymbol{\beta}}$ represents the concentration parameter of the expected target Dirichlet with respect to the empirical training distribution, $\tilde{P}_T$. We can replace $\tilde{P}_T$ with the empirical distribution of in-domain training examples, $\tilde{P}_{in}$ or OOD training examples, $\tilde{P}_{out}$ for our analysis.

Minimizing $-\mathcal{H}[p(\boldsymbol{\mu}|\boldsymbol{x},\boldsymbol{\theta})]$ always leads to a flatter distribution. Hence, we rely only on $\mathbb{E}_{P(\boldsymbol{\mu}|\boldsymbol{x},\boldsymbol{\theta})}[-\ln Dir(\boldsymbol{\mu}|\overline{\boldsymbol{\beta}})]$ to produce sharper distributions.

Malinin & Gales (2019) choose the target concentration value for in-domain examples as: $(\beta+1)$ for the correct class and 1 for the incorrect classes. Thus, we get:

$$\begin{aligned}
&\mathbb{E}_{\tilde{P}_T(\boldsymbol{x},y)}\ \mathrm{KL}\big[\, p(\boldsymbol{\mu}|\boldsymbol{x},\boldsymbol{\theta}) \,||\, Dir(\boldsymbol{\mu}|\boldsymbol{\beta})\,\big]\\
&= \mathbb{E}_{\tilde{P}_T(\boldsymbol{x})}\Big[\sum_c\sum_k \tilde{p}(\omega_c|\boldsymbol{x})(\beta_k^{(c)}-1)\big[\psi(\alpha_0)-\psi(\alpha_k)\big]\Big]\\
&= \mathbb{E}_{\tilde{P}_T(\boldsymbol{x})}\Big[\beta\sum_c \tilde{p}(\omega_c|\boldsymbol{x})\big[\psi(\alpha_0)-\psi(\alpha_c)\big]\Big]
\end{aligned} \tag{4}$$

We can see in Eqn. 3, the reverse cross-entropy term maximizes $\psi(\alpha_c)$ for each class c with the factor, $\beta\tilde{pr}(\omega_c|\boldsymbol{x})$, while minimizes $\psi(\alpha_0)$ with the factor, β. Hence, for an in-domain example with confident predictions, it produces a sharp Dirichlet with a large concentration value for the correct class and very small concentration parameters ($<< 1$) for the incorrect classes. However, for an input with high data uncertainty, β is distributed among multiple classes. It leads to smaller concentration parameters (≥ 1) for all overlapping classes, producing a much flatter and diverse Dirichlet distributions.

(a) Dir_1:$\{0.01, 0.01, 101.98\}$ $D.Ent = -199.1$, $\mathcal{MI} = 8e\text{-}4$ (b) Dir_2:$\{34, 34, 34\}$ $D.Ent = -3.45$, $\mathcal{MI} = 9.7e\text{-}3$

Figure 2. Dirichlet distributions with the same precision but different concentration parameters.

For example, let us consider two Dirichlet, Dir_1 and Dir_2, with the same precision, $\alpha_0 = 102$, but different concentration parameters of $\{0.01, 0.01, 101.98\}$ (Fig. 2a) and $\{34, 34, 34\}$ (Fig. 2b). We can see that Dir_1 produces much lower $D.Ent$ and $\mathcal{MI}$ scores for than Dir_2. It respectively indicates that Dir_2 is more flatter and diverse than Dir_1, even with the same precision. The differences of these scores become more significant in higher dimensions. As we consider the same example for a classification task with 100 classes, $D.Ent$ and $\mathcal{MI}$ would respectively become $-9.9e3$ and 0.02 for Dir_1 and -370.5 and 0.23 for Dir_2. Further, we empirically show that DPN models also leads to lower precision values along with flatter and diverse Dirichlet distributions for misclassified examples (Table 2).

This behavior is *not* desirable: since the reverse KL-divergence also trains the DPN to produce flat Dirichlet distributions for OOD examples, it often leads to producing indistinguishable distributions for in-domain misclassified examples and OOD examples, in the boundary cases. Hence, we can instead produce sharp, multi-modal Dirichlet distributions for distributional uncertainties (Figure 1d) to ensure that the OOD examples always remain distinguishable from the in-domain examples.

For OOD training examples, we should choose identical values for target concentration parameters, $(\tau+1)$ with $\tau > -1$, for all classes. Using $(\tau+1)$ for $\beta_k^{(c)}$ in Eq. 3, we get the RKL loss for OOD examples as:

$$\mathbb{E}_{P_T(\boldsymbol{x})}\Big[\tau K\psi(\alpha_0) - \sum_c \tau\,\psi(\alpha_c) - \mathcal{H}\big[p(\boldsymbol{\mu}|\boldsymbol{x},\boldsymbol{\theta})\big]\Big] \tag{5}$$

Malinin & Gales (2019) choose $\tau = 0$, leading to minimize $-\mathcal{H}[pr(\boldsymbol{\mu}|\boldsymbol{x},\boldsymbol{\theta})]$. Hence, the DPN produces flat Dirichlet distributions for OOD examples. We investigate the other choices for τ: Choosing $\tau > 0$ leads to minimizing the precision, α_0 (i.e. $\sum_{c=1}^K \alpha_c$) of the output Dirichlet while maximizing individual concentration parameters, α_c (see Eq. 5). In contrast, choosing $\tau \in (-1, 0)$ leads to maximize the α_0 while minimizing α_c's. In other words, either choice of τ may lead to an output Dirichlet with *uncontrolled concentration parameter values* for an OOD example.

Explicit Precision Regularization. We now present our proposed loss function for DPN that *separately models* the mean and precision of the output Dirichlet, providing greater control over the desired outputs. We use the *soft-max cross-entropy loss* to models the mean of the output Dirichlet, as shown in Eq. 1), along with a novel *explicit precision regularizer*, to model the precision.

We propose to use a bounded approximation of the precision value, i.e $\frac{1}{K}\sum_{c=1}^K \mathrm{sigmoid}(z_c(\boldsymbol{x}))$ as our regularizer. For an in-domain confident example, we design the loss to produce a sharp uni-modal Dirichlet with the mode in the corner of the correct class: $\mathcal{L}_{in}(\boldsymbol{\theta};\lambda_{in}) :=$

$$\mathbb{E}_{P_{in}(\boldsymbol{x},y)}\Big[-\log p(y|\boldsymbol{x},\boldsymbol{\theta}) - \frac{\lambda_{in}}{K}\sum_{c=1}^K \mathrm{sigmoid}(z_c(\boldsymbol{x}))\Big] \tag{6}$$

For OOD training examples, we design the loss function to produce smaller precision, with uniform concentration values, for the output Dirichlet that induces a uniform categorical posterior over all class labels: $\mathcal{L}_{out}(\boldsymbol{\theta};\lambda_{out}) :=$

$$\mathbb{E}_{P_{out}(\boldsymbol{x},y)}\Big[\mathcal{H}_{ce}(\mathcal{U};p(y|\boldsymbol{x},\boldsymbol{\theta})) - \frac{\lambda_{out}}{K}\sum_{c=1}^K \mathrm{sigmoid}(z_c(\boldsymbol{x}))\Big] \tag{7}$$

where, $\mathcal{H}_{ce}$ is the cross-entropy function. $\mathcal{U}$ is the uniform distribution over the class labels. λ_{in} and λ_{out} are hyper-parameters that control the precision of the output distributions. We train the DPN in a multi-task fashion:

$$\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta};\gamma,\lambda_{in},\lambda_{out}) = \mathcal{L}_{in}(\boldsymbol{\theta},\lambda_{in}) + \gamma\mathcal{L}_{out}(\boldsymbol{\theta},\lambda_{out}) \tag{8}$$

where $\gamma > 0$ balances between the loss values for in-domain examples and OOD examples.

By choosing $\lambda_{in} > 0$ for in-domain examples, our regularizer imposes the network to maximize $\mathrm{sigmoid}(z_c(\boldsymbol{x}))$ for

	OOD test sets	TIM (Li et al., 2017)			STL-10 (Coates et al., 2011)			LSUN (Yu et al., 2015)		
		$maxP$	$\mathcal{MI}$	α_0	$maxP$	$\mathcal{MI}$	α_0	$maxP$	$\mathcal{MI}$	α_0
C10	Baseline	88.9$_{\pm 0.0}$	-	-	75.9$_{\pm 0.0}$	-	-	90.3$_{\pm 0.0}$	-	-
	MCDP	88.7$_{\pm 0.1}$	88.1$_{\pm 0.1}$	-	76.2$_{\pm 0.0}$	76.0$_{\pm 0.0}$	-	90.6$_{\pm 0.0}$	90.2$_{\pm 0.0}$	-
	OE	98.2$_{\pm 0.1}$	-	-	81.4$_{\pm 1.2}$	-	-	98.4$_{\pm 0.3}$	-	-
	DPN$_{rev}$	97.5$_{\pm 0.5}$	97.8$_{\pm 0.4}$	97.8$_{\pm 0.4}$	81.6$_{\pm 1.7}$	82.2$_{\pm 1.7}$	82.2$_{\pm 1.6}$	98.5$_{\pm 0.4}$	98.7$_{\pm 0.3}$	98.7$_{\pm 0.3}$
	DPN^{+}	98.0$_{\pm 0.2}$	98.0$_{\pm 0.2}$	98.0$_{\pm 0.2}$	81.6$_{\pm 1.4}$	81.8$_{\pm 1.2}$	81.8$_{\pm 1.2}$	98.2$_{\pm 0.3}$	98.3$_{\pm 0.4}$	98.3$_{\pm 0.4}$
	DPN^{-}	**99.0**$_{\pm \mathbf{0.1}}$	**99.0**$_{\pm \mathbf{0.1}}$	97.7$_{\pm 0.9}$	84.7$_{\pm 0.4}$	**85.3**$_{\pm \mathbf{0.5}}$	84.9$_{\pm 0.5}$	99.2$_{\pm 0.1}$	**99.3**$_{\pm \mathbf{0.0}}$	98.1$_{\pm 0.1}$
C100	Baseline	68.8$_{\pm 0.2}$	-	-	69.6$_{\pm 0.0}$	-	-	72.5$_{\pm 0.0}$	-	-
	MCDP	69.7$_{\pm 0.3}$	70.6$_{\pm 0.3}$	-	70.7$_{\pm 0.1}$	71.6$_{\pm 0.2}$	-	74.5$_{\pm 0.1}$	75.9$_{\pm 0.2}$	-
	OE	89.5$_{\pm 1.0}$	-	-	91.2$_{\pm 0.7}$	-	-	92.2$_{\pm 0.9}$	-	-
	DPN$_{rev}$	81.2$_{\pm 0.2}$	83.8$_{\pm 0.1}$	83.8$_{\pm 0.1}$	87.2$_{\pm 0.1}$	89.3$_{\pm 0.1}$	89.3$_{\pm 0.1}$	86.7$_{\pm 0.0}$	89.3$_{\pm 0.1}$	89.3$_{\pm 0.1}$
	DPN^{+}	85.9$_{\pm 0.3}$	92.2$_{\pm 0.1}$	92.2$_{\pm 0.1}$	89.1$_{\pm 0.2}$	95.0$_{\pm 0.0}$	95.0$_{\pm 0.0}$	90.3$_{\pm 0.3}$	95.0$_{\pm 0.1}$	95.0$_{\pm 0.1}$
	DPN^{-}	89.2$_{\pm 0.1}$	**94.5**$_{\pm \mathbf{0.1}}$	**94.5**$_{\pm \mathbf{0.1}}$	92.8$_{\pm 0.1}$	**96.8**$_{\pm \mathbf{0.1}}$	**96.8**$_{\pm \mathbf{0.1}}$	92.8$_{\pm 0.1}$	**96.5**$_{\pm \mathbf{0.1}}$	**96.5**$_{\pm \mathbf{0.1}}$

Table 1. AUROC scores for OOD detection (mean ± s.d of 3 models). Higher scores are better.

		$maxP$	$\mathcal{MI}$	α_0	Acc.
C10	Baseline	93.3$_{\pm 0.1}$	-	-	94.1
	MCDP	93.6$_{\pm 0.2}$	93.2$_{\pm 0.1}$	-	94.2
	OE	92.0$_{\pm 0.0}$	-	-	94.2
	DPN$_{rev}$	89.6$_{\pm 0.1}$	88.7$_{\pm 0.2}$	88.7$_{\pm 0.2}$	90.6
	DPN^{+}	92.2$_{\pm 0.3}$	90.3$_{\pm 0.1}$	90.3$_{\pm 0.1}$	94.0
	DPN^{-}	92.6$_{\pm 0.1}$	89.9$_{\pm 0.0}$	89.9$_{\pm 0.0}$	94.4
C100	Baseline	86.8$_{\pm 0.1}$	-	-	72.3
	MCDP	87.2$_{\pm 0.0}$	83.3$_{\pm 0.3}$	-	72.7
	OE	86.9$_{\pm 0.0}$	-	-	71.6
	DPN$_{rev}$	79.3$_{\pm 0.1}$	73.5$_{\pm 0.1}$	73.1$_{\pm 0.1}$	71.1
	DPN^{+}	86.4$_{\pm 0.1}$	81.2$_{\pm 0.0}$	81.3$_{\pm 0.0}$	72.1
	DPN^{-}	86.4$_{\pm 0.1}$	82.3$_{\pm 0.0}$	82.3$_{\pm 0.0}$	72.3

Table 2. AUROC scores for misclassified image detection (mean ± s.d. of 3 models). Higher scores are better.

all classes. However, for confidently predicted examples, the cross-entropy loss ensures to maximize the logit value of the correct class. In contrast, in the presence of high data uncertainty, the cross-entropy loss produces multiple smaller modes for the overlapping classes. Hence, as before, it leads to producing a flatter distribution for misclassified examples (see Fig 2). Now, by choosing $\lambda_{in} > \lambda_{out} > 0$, we also enforce the network to produce a flatter distribution with $\alpha_c = \exp z_c(\boldsymbol{x}*) \geq 1$ for an OOD example $\boldsymbol{x}*$. Hence, it leads to an indistinguishable representation for the OOD example as an in-domain example with high data uncertainty, as in the case of the RKL loss (Eq. 3).

However, now we can address this problem by choosing $\lambda_{out} < 0$. It enforces the DPN to produce uniform negative values for $z_c(\boldsymbol{x}^*)$, leading to $\alpha_c < 1\ \forall c$, for an OOD example to produce a sharp multi-modal Dirichlet with uniform weights at each corner of the simplex (Fig 1d).

Note that, the choice of $\lambda_{in} = 0$, $\lambda_{out} = 0$ reduces the proposed loss to the non-Bayesian OE model (Hendrycks et al., 2019). However, it now fails to produce desirable Dirichlet distributions to indicate different uncertainty types.

4. Experimental Study

We carry out experiments on CIFAR-10 and CIFAR-100 datasets (Krizhevsky & Hinton, 2009) using VGG-16 (Simonyan & Zisserman, 2015). We train the C10 classifiers using CIFAR-10 training images as in-domain data and CIFAR-100 training images as the OOD data. For the C100 classifiers, we use CIFAR-100 training images as in-domain and CIFAR-10 training images as OOD. We study the performance our DPN models using $\lambda_{in} > 0$ and both $\lambda_{out} > 0$ and $\lambda_{out} < 0$, denoted as DPN^{+} and DPN^{-} respectively. (See Appendix A for more details.)

We evaluate the performances of our models for OOD detection and in-domain misclassification detection using *area under the receiver operating characteristic curve (AUROC)* metric (Manning & Schütze, 1999; Hendrycks & Gimpel, 2017). We compare the performance of our models with the standard DNN as Baseline (Hendrycks & Gimpel, 2017), Bayesian MCDP (Gal & Ghahramani, 2016), non-Bayesian OE (Hendrycks et al., 2019) and DPN$_{rev}$ (Malinin & Gales, 2019). For MCDP, we measure the $maxP$ and MI. For the other models, $\mathcal{MI}$ and α_0 are not defined.

Tables 1 shows the OOD detection performance of C10 and C100 classifiers. We observe that our DPN^{-} models consistently outperform the other models using $\mathcal{MI}$ measure. It produces sharp multi-modal Dirichlet distributions, with uniform weights at each corner of the simplex, for OOD examples, leading to higher $\mathcal{MI}$ scores and maximize the representation gap from in-domain examples.

In Table 2, we observe that our DPN^{-} models achieve comparable performance with the existing models for misclassification detection. While the Baseline and MCDP models often achieve higher scores for misclassification detection task, their poor performance for OOD detection makes it difficult for real-world applications.

We can see that the DPN models achieve higher scores even for the $\mathcal{MI}$ and α_0 measures. In previous studies (Malinin, 2019; Malinin & Gales, 2019), we can also observe similar results using $D.Ent$ as an uncertainty measure. This supports our assertion that *in the presence of data uncertainty, DPN models tend to produce flatter and diverse Dirichlet distributions, compared to the confident predictions.* Hence, we produce sharp multi-modal Dirichlet, with uniform densities at each corner of the simplex, for OOD examples to keep them distinguishable from the in-domain examples.

5. Conclusion

The existing DPN models often produce indistinguishable representations for the in-domain examples with high data uncertainty among multiple classes and OOD examples. We propose a novel loss function to maximize the representation gap between in-domain and OOD examples. Our experimental results show that we can consistently improve the OOD detection performance by solving this issue.

Acknowledgment: This research is supported by the National Research Foundation Singapore under its AI Singapore Programme (Award Number: AISG-RP-2018-008).

References

Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset shift in machine learning*. The MIT Press, 2009.

Coates, A., Ng, A., and Lee, H. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, 2011.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

Gal, Y. *Uncertainty in deep learning*. PhD thesis, PhD thesis, University of Cambridge, 2016.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016.

Hein, M., Andriushchenko, M., and Bitterwolf, J. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *CVPR*, 2019.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ICLR*, 2017.

Hendrycks, D., Mazeika, M., and Dietterich, T. Deep anomaly detection with outlier exposure. In *ICLR*, 2019.

Hernandez-Lobato, J. M. and Adams, R. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *ICML*, 2015.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. *Citeseer*, 2009. URL http://www.cs.toronto.edu/~kriz/cifar.html.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.

Lee, K., Lee, H., Lee, K., and Shin, J. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *ICLR*, 2018a.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018b.

Li, F.-F., Karpathy, A., and Johnson, J. Tiny imagenet visual recognition challenge. *Stanford University CS231N*, 2017. URL https://tiny-imagenet.herokuapp.com.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018.

Malinin, A. Uncertainty estimation in deep learning with application to spoken language assessment. In *Doctoral thesis*, 2019.

Malinin, A. and Gales, M. Predictive uncertainty estimation via prior networks. In *NeurIPS*, 2018.

Malinin, A. and Gales, M. Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. In *NeurIPS*, 2019.

Manning, C. D. and Schütze, H. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

Meinke, A. and Hein, M. Towards neural networks that provably know when they don't know. In *ICLR*, 2020.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

Yu, F., Zhang, Y., Song, S., Seff, A., and Xiao, J. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv*, 2015.

A. Experimental Details

We use the VGG-16 network (Simonyan & Zisserman, 2015) for C10, C100 and TIM classification tasks. For C10, we use CIFAR-10 training images ($50,000$ images) as our in-domain training data and CIFAR-100 training images ($50,000$ images) as our OOD training data.

For C-100, we use CIFAR-100 training images ($50,000$ images) as our in-domain training data and CIFAR-10 training images ($50,000$ images) as our OOD training data.

We keep the test examples separately during the training phase of our models. In Table 3, we present the details of our experimental settings.

For the OOD detection task, we treat the OOD examples as the positive class and in-domain examples as the negative class. For the experiments on misclassification detection, we consider the misclassified examples as the positive class and correctly classified examples as the negative class.

Hyper-parameters. Similar to (Malinin & Gales, 2018; 2019; Hendrycks et al., 2019), we do not need to tune any hyper-parameters during testing. In other words, the OOD test examples remain unknown to our DPN classifiers, as in a real-world scenario.

We set $\gamma = 0.5$ for our loss function in Eqn. 8, as applied in (Hendrycks et al., 2019). We train two different DPN models

Classifier	Input	#Classes	Training Datasets		Test Datasets	
			In-Domain	OOD	In-Domain	OOD
C10	32×32	10	CIFAR-10 Training Set (50,000 images)	CIFAR-100 Training Set (50,000 images)	CIFAR-10 Test Images (10,000 Images)	TIM, STL-10, LSUN etc.
C100	32×32	100	CIFAR-100 Training Set (50,000 images)	CIFAR-10 Training Set (50,000 images)	CIFAR-100 Test Set (10,000 Images)	TIM, STL-10, LSUN etc.

Table 3. Details of Training and Test Datasets used for C10 and C100 classifiers.

for each classification task, using both positive and negative values for λ_{out} to analyze the effect of both flat Dirichlet distributions and sharp Dirichlet distributions at each corner of the simplex respectively for the OOD examples.

The hyper-parameters λ_{in} and λ_{out} controls the precision of the output Dirichlet distributions. For our experiments, we always choose these hyper-parameters as follows: DPN$^+$ is trained with positive $\lambda_{out} = \frac{1}{\#class} + 0.5$ and $\lambda_{in} = 1.5$. DPN$^-$ is trained with negative $\lambda_{out} = \frac{1}{\#class} - 0.5$, and $\lambda_{in} = 0.5$.

A.1. Competitive Systems

We compare the performance of our models with standard DNN as baseline model (Hendrycks & Gimpel, 2017), the Bayesian framework, monte-carlo dropout (MCDP) (Gal & Ghahramani, 2016), DPN$_{fwd}$ and DPN$_{rev}$ using the loss function proposed in (Malinin & Gales, 2018) and (Malinin & Gales, 2019), non-Bayesian frameworks such as outlier exposure (OE) (Hendrycks et al., 2019). We use the same architecture as our DPN models for the other competitive models.

MCDP models: For MCDP, we use the standard DNN (Baseline) model with randomly dropping the nodes during inference time. The predictive categorical distributions are obtained by averaging the outputs for 10 iterations.

DPN$_{rev}$ models: The DPN$_{rev}$ models are trained only using the ADAM optimizer (Kingma & Ba, 2014). Note that, We could not use the SGD optimizer to train these models due to the complex RKL loss. In contrast, we have not encountered such a problem for other OOD detection models.

For example, we use SGD to train the other models for the C10 classification task. We also observe that DPN$_{rev}$ achieves lower classification accuracies than the other classifiers (Table 2). For C100, we choose the ADAM optimizer for all models. We find that all the OOD detection models achieve similar classification accuracy in this case (Table 2).

We choose the same set of hyper-parameters to produce the *target hand-crafted concentration parameters* for DPN$_{rev}$ models (see Eq. 2) as suggested by the authors (Malinin & Gales, 2019; Malinin, 2019): for in-domain training examples are set to 100 for the correct class and 1 for the incorrect classes. For OOD training examples, we choose the concentration parameters as 1 for all classes. We set $\gamma = 0.5$, similar to our DPN$^+$ and DPN$^-$ models and non-Bayesian OE models (Hendrycks et al., 2019).

A.2. Description of the OOD Test Datasets

We use a wide-range of OOD dataset for our experiments, as described in the following. For C-10 and C-100 classifiers, these input test images are resized to 32×32.

TinyImageNet (TIM) (Li et al., 2017). This is a subset of Imagenet dataset (Deng et al., 2009). We use the validation set, that contains $10,000$ test images from 200 different image classes for our evaluation during test time.

STL-10 contains $8,000$ images of natural images from 10 different classes (Coates et al., 2011).

LSUN (Yu et al., 2015). The Large-scale Scene UNderstanding dataset (LSUN) contains images of 10 different scene categories. We use its validation set, containing $10,000$ images, as an unknown OOD test set.

B. Uncertainty Measures

In this section, we present the expressions to compute different uncertainty measures of a Dirichlet distribution.

Mutual Information of a Dirichlet distribution: The mutual information of the labels, y and the categorical, $\boldsymbol{\mu}$ of a DPN is computed as:

$$\begin{aligned} &\mathcal{MI}[y, \boldsymbol{\mu}|\boldsymbol{x}^*, \boldsymbol{\theta}] \\ =&\mathcal{H}[\mathbb{E}_{p(\mu|\boldsymbol{x},\boldsymbol{\theta})} P(y|\boldsymbol{\mu})] - \mathbb{E}_{p(\mu|\boldsymbol{x},\boldsymbol{\theta})} \mathcal{H}[P(y|\boldsymbol{\mu})] \\ =&\sum_{c=1}^{K} \frac{\alpha_c}{\alpha_0} \big[\psi(\alpha_c + 1) - \psi(\alpha_0 + 1) - \ln \frac{\alpha_c}{\alpha_0}\big] \end{aligned} \tag{9}$$

Precision, α_0 (Inverse-EPKL): The expected pairwise KL divergence (EPKL) measures the expected KL-divergence between pairs of independent *categorical distributions* samples from the output Dirichlet distribution. EPKL is simplified to $\frac{K-1}{\alpha_0}$ for a Dirichlet distribution.

Since the DPN models also produce smaller precision values for OOD examples, we can directly use the precision, α_0, or the inverse of EPKL as a distributional uncertainty measure.

Note that, both EPKL and precision (or inverse-EPKL) leads to the same OOD detection performance as they produce the same relative uncertainty scores (in reverse order) for a given set of input examples.

Differential Entropy of a Dirichlet distribution: Differential entropy ($D.Ent$), that measures the sharpness of a Dirichlet distribution. It is also used as a distributional uncertainty measure in previous studies (Malinin & Gales, 2018; 2019). However, unlike other DPN models, our DPN$^-$ behaves differently to produce sharp multi-modal Dirichlet distributions for OOD examples and sharp uni-modal distribution for in-domain confident predictions. Hence, we cannot directly detect OOD examples using $D.Ent$ measure. Differential Entropy of a Dirichlet distribution can be calculated as follows:

$$\begin{aligned}\mathcal{H}[p(\boldsymbol{\mu}|\boldsymbol{x}^*,\boldsymbol{\theta})] &= -\int p(\boldsymbol{\mu}|\boldsymbol{x}^*,\boldsymbol{\theta}) \ln p(\boldsymbol{\mu}|\boldsymbol{x}^*,\boldsymbol{\theta}) d\boldsymbol{\mu} \\ &= \sum_{c=1}^{K} \ln \Gamma(\alpha_c) - \ln \Gamma(\alpha_0) - \sum_{c=1}^{K} (\alpha_c - 1)(\psi(\alpha_c) - \psi(\alpha_0))\end{aligned} \tag{10}$$

Note that, α_c is a function of x^*. Γ and ψ denotes the Gamma and digamma functions respectively.