Towards Maximizing the Representation Gap between In-Domain & Out-of-Distribution Examples

NUS National University of Singapore
School of Computing

Jay Nandy, Wynne Hsu, Mong Li Lee
{jaynandy, whsu, leeml}@comp.nus.edu.sg

NEURAL INFORMATION PROCESSING SYSTEMS

1. Introduction

- Deep learning based models produce wrong predictions without any warning
- This raises questions about their reliability for sensitive real-world applications
- Determining source of uncertainty can allow manual intervention in an informed way, enhancing the reliability of DNN based models

2. Types of Predictive Uncertainty

Model or Epistemic uncertainty

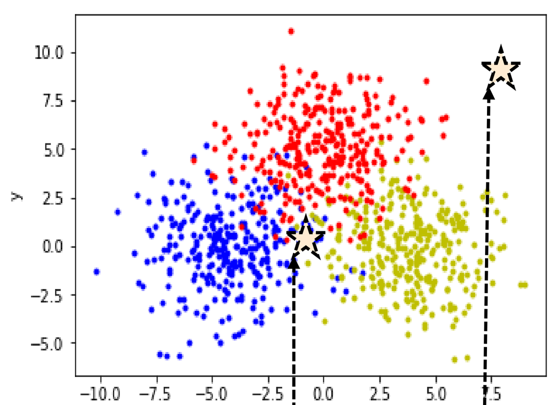
- Uncertainty in estimating network parameters
- Reducible with enough training data

Data or Aleatoric uncertainty

- Arises due to the natural complexities of the underlying distribution, such as class overlap, label noise, homoscedastic and heteroscedastic noise

Distributional uncertainty

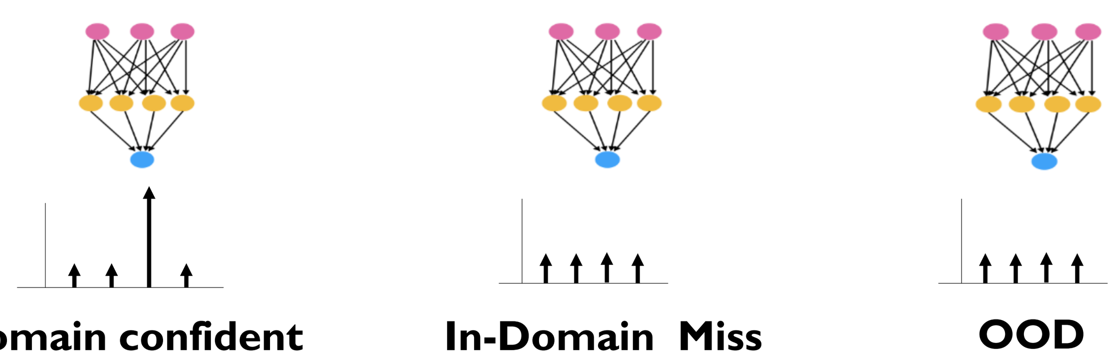
- Distributional mismatch between the training and test examples during inference
- Test data is out-of-distribution (OOD)



In-domain example with Data uncertainty

OOD example leading to distributional uncertainty

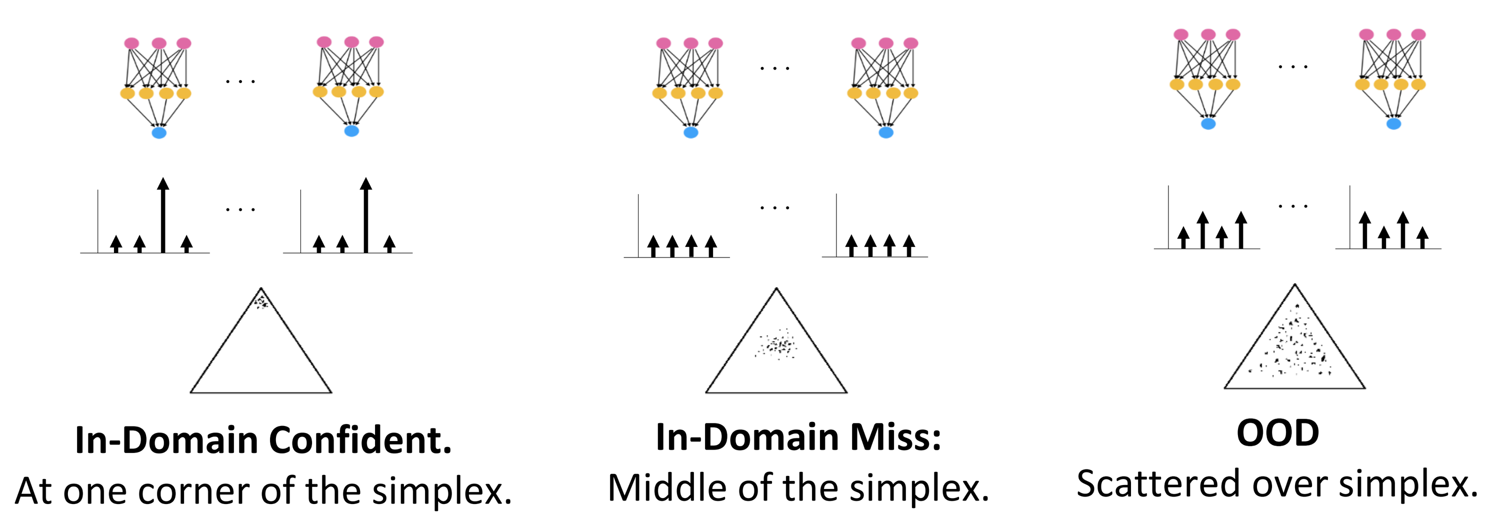
3. Existing Approaches: Non-Bayesian



Limitation:

- In the presence of high data uncertainty among multiple classes, existing OOD detectors produce similar representation for both in-domain and OOD examples.
- Compromise their performance for OOD detection

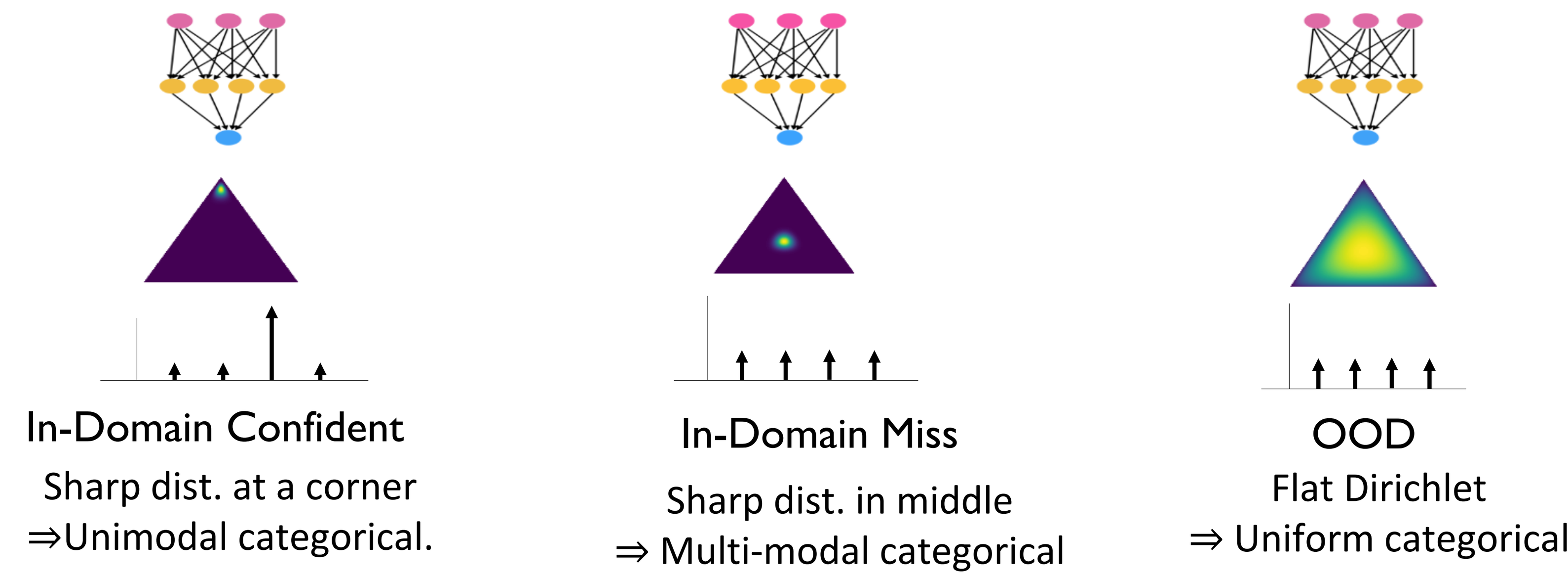
4. Existing Approaches: Bayesian



Limitation:

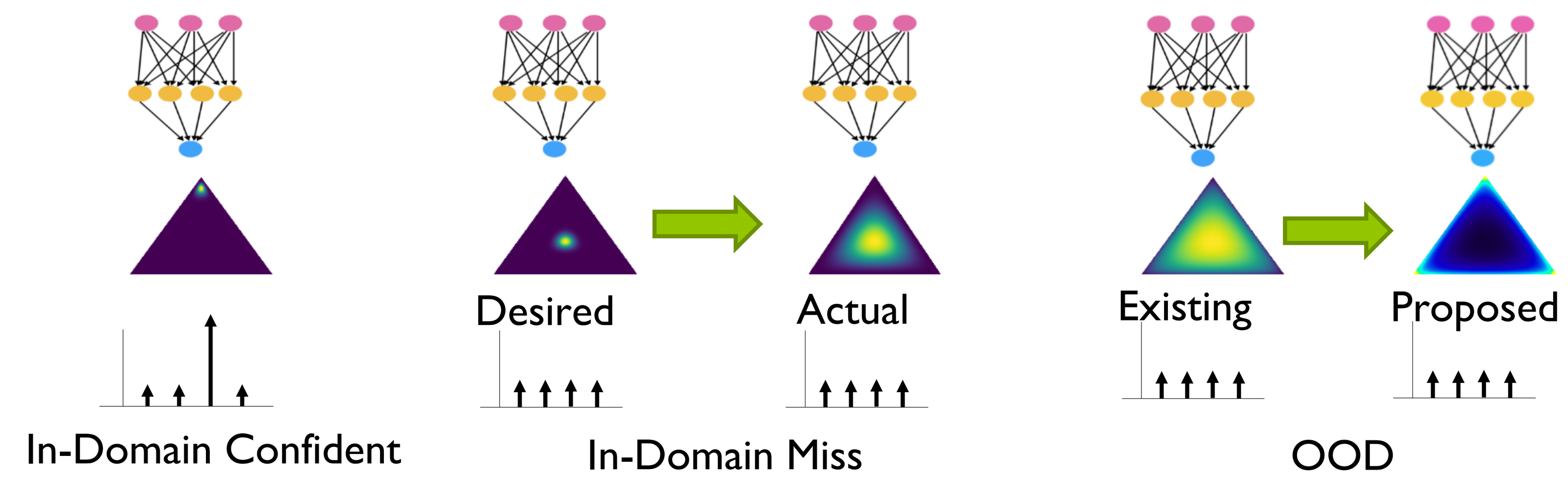
- Computationally expensive to produce the ensemble
- Difficult to control this desired behavior

5. Existing Dirichlet Prior Network (DPN)



Emulating the behavior of Bayesian (ensemble) approaches [Malinin & Gales, 2018; 2019]
Parameterize a prior Dirichlet distribution to the categorical over a simplex

6. Proposed Representation for OOD



- Maximize representation gap by producing sharp multi-modal Dirichlet for OODs.
- We show that existing RKL loss cannot produce this representation.
- We propose a novel loss function for DPN to address this limitation.

7. Dirichlet Distribution

Parameterized using the concentration parameters, $\boldsymbol{\alpha} = \{\alpha_1, \cdots, \alpha_K\}$

$$Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma\alpha_0}{\prod_{c=1}^{K}\Gamma\alpha_c}\prod_{c=1}^{K}\mu_c^{\alpha_c - 1}, \qquad where\ \alpha_c > 0\ \forall\ c$$

$\alpha_0 = \sum\alpha_c$ denotes the precision of the Dirichlet distribution.

- Larger α_0 with at least one $\alpha_c > 0$ produces a sharp unimodal Dirichlet
- $\alpha_c < 1\ \forall\ c$ produces sharp multi-modal Dirichlet

Visualization of Dirichlet distributions for different concentration parameters

8. A standard DNN with soft-max is a DPN

- Concentration parameters of the Dirichlet is exponential of logits: $\alpha_c = \exp(z_c(\boldsymbol{x}^*))$
- *Categorical posterior, obtained from soft-max, is the mean of the Dirichet distribution*

$$p\left(\frac{y = \omega_c}{\boldsymbol{x}^*}\right) = \frac{\alpha_c}{\alpha_0} = \frac{\exp(z_c(\boldsymbol{x}^*))}{\sum_c \exp(z_c(\boldsymbol{x}^*))}$$

- **Limitation:** Cannot control the individual α_c to produce desired Dirichlet distributions.

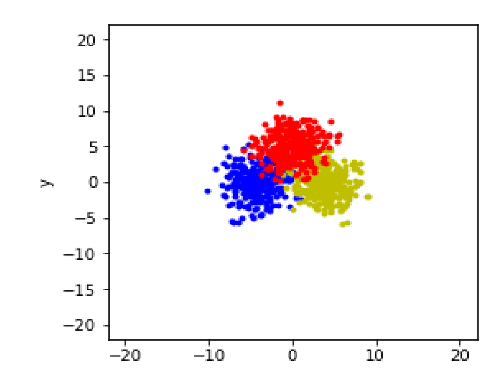
9. Proposed Loss Function

Proposed a novel regularizer to control the concentration parameters, α_c
Train using both in-domain and OOD training examples in multi-task fashion

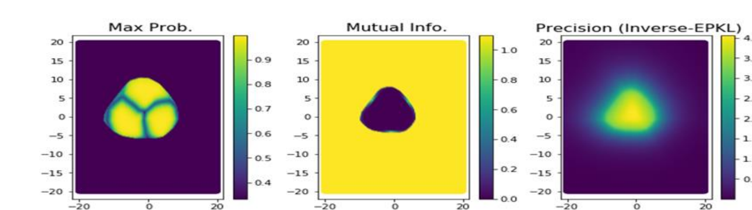
- In-domain: $\mathbb{E}_{p_{in}(x,y)}\left[-\log p(y|x,\boldsymbol{\theta}) - \frac{\lambda_{in}}{K}\sum sigmoid(z_c(x))\right]$ where $\lambda_{in} > 0$
- OOD: $\mathbb{E}_{p_{in}(x,y)}\left[\mathcal{H}_{ce}(\mathcal{U}; p(y|x,\boldsymbol{\theta})) - \frac{\lambda_{out}}{K}\sum sigmoid(z_c(x))\right]$ where $\lambda_{out} < \lambda_{in}$

$\lambda_{out} > 0$ $\qquad$ $\lambda_{out} < 0$ (Proposed)

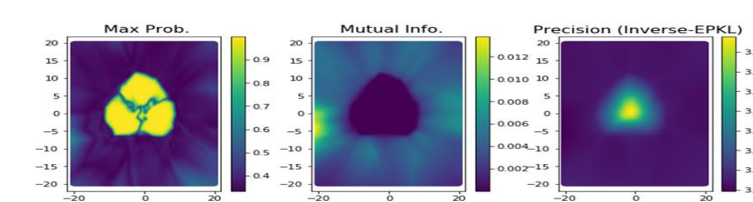
10. Results on Synthetic Dataset



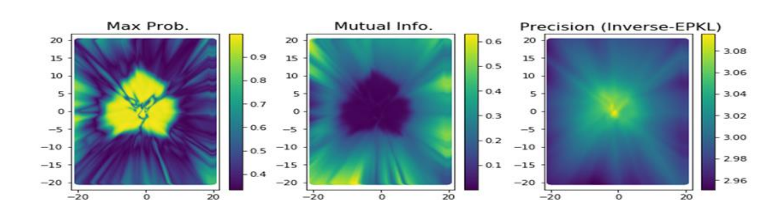
In-Domain Training Data



DPN$^-$ using $\lambda_{in} > 0$ & $\lambda_{out} < 0$ (Proposed)



DPN$^+$ using $\lambda_{in} > \lambda_{out} > 0$



DPN$_{\text{rev}}$ using RKL [Malinin & Gales, 2019]

11. Results on Benchmark Datasets

	OOD	Tiny [29]				STL-10 [32]				LSUN [33]			
		Max.P	MI	α_0	D.Ent	Max.P	MI	α_0	D.Ent	Max.P	MI	α_0	D.Ent
C10	Baseline	88.9±0.0	-	-	-	75.9±0.0	-	-	-	90.3±0.0	-	-	-
	MCDP	88.7±0.1	88.1±0.1	-	-	76.2±0.0	76.0±0.0	-	-	90.6±0.0	90.2±0.0	-	-
	DE	88.9±NA	87.8±NA	-	-	76.0±NA	75.6±NA	-	-	90.3±NA	89.7±NA	-	-
	OE	98.2±0.1	-	-	-	81.4±1.2	-	-	-	98.4±0.3	-	-	-
	DPN$_{rev}$	97.5±0.5	97.8±0.4	97.8±0.4	97.7±0.4	81.6±1.7	82.2±1.7	82.2±1.6	81.9±1.7	98.5±0.4	98.7±0.3	98.7±0.3	98.7±0.3
	DPN$^+$	98.0±0.2	98.0±0.2	98.0±0.2	98.0±0.2	81.6±1.4	81.8±1.2	81.8±1.2	81.8±1.2	98.2±0.3	98.3±0.4	98.3±0.4	98.3±0.4
	DPN$^-$	**99.0**±0.1	**99.0**±0.1	97.7±0.1	6.0±0.3	84.7±0.4	**85.3**±0.5	84.9±0.5	34.6±0.4	99.2±0.1	**99.3**±0.0	98.1±0.1	5.0±0.2

	OOD	Tiny [29]				STL-10 [32]				LSUN [33]			
		Max.P	MI	α_0	D.Ent	Max.P	MI	α_0	D.Ent	Max.P	MI	α_0	D.Ent
C100	Baseline	68.8±0.2	-	-	-	69.6±0.0	-	-	-	72.5±0.0	-	-	-
	MCDP	69.7±0.3	70.6±0.3	-	-	70.7±0.1	71.6±0.2	-	-	74.5±0.1	75.9±0.2	-	-
	DE	68.9±NA	69.6±NA	-	-	69.6±NA	70.2±NA	-	-	72.6±NA	73.4±NA	-	-
	OE	89.5±1.0	-	-	-	91.2±0.7	-	-	-	92.2±0.9	-	-	-
	DPN$_{rev}$	81.2±0.2	83.8±0.1	83.8±0.1	83.5±0.1	87.2±0.1	89.3±0.1	89.3±0.1	89.0±0.1	86.7±0.0	89.3±0.1	89.3±0.1	88.9±0.1
	DPN$^+$	85.9±0.3	92.2±0.1	92.2±0.1	92.3±0.1	89.1±0.2	95.0±0.0	95.0±0.0	94.8±0.0	90.3±0.3	95.0±0.1	95.0±0.1	95.0±0.1
	DPN$^-$	89.2±0.1	**94.5**±0.1	**94.5**±0.1	38.1±0.5	92.8±0.1	**96.8**±0.1	**96.8**±0.1	25.4±0.4	92.8±0.1	**96.5**±0.1	**96.5**±0.1	31.5±0.4

	OOD	CIFAR-10 [28]				CIFAR-100 [28]				Textures [34]			
		Max.P	MI	α_0	D.Ent	Max.P	MI	α_0	D.Ent	Max.P	MI	α_0	D.Ent
TIM	Baseline	76.9±0.2	-	-	-	73.6±0.2	-	-	-	70.9±0.2	-	-	-
	MCDP	77.4±0.1	77.5±0.2	-	-	74.0±0.2	73.6±0.2	-	-	70.3±0.2	63.6±0.2	-	-
	DE	76.9±NA	77.7±NA	-	-	73.7±NA	75.3±NA	-	-	71.1±NA	76.2±NA	-	-
	OE	91.3±0.4	-	-	-	89.5±0.5	-	-	-	95.8±0.3	-	-	-
	DPN$_{rev}$	85.4±0.7	82.8±1.4	81.9±1.6	85.6±0.9	84.2±0.8	82.5±1.4	81.7±1.6	85.0±0.9	90.9±0.3	91.2±0.6	90.6±0.6	92.6±0.3
	DPN$^+$	99.2±0.0	99.7±0.0	99.7±0.0	99.6±0.0	98.8±0.0	99.5±0.0	99.5±0.0	99.4±0.0	96.5±0.1	98.4±0.0	98.4±0.0	98.2±0.0
	DPN$^-$	99.7±0.0	**99.9**±0.0	**99.9**±0.0	3.5±0.1	98.7±0.1	**99.6**±0.0	**99.6**±0.0	7.5±0.2	95.8±0.1	**98.7**±0.1	**98.7**±0.1	19.3±0.4

AUROC scores for OOD detection (Higher scores are better)

Please refer to our paper for additional details and more experimental results:

ArXiv Paper Link

GitHub Code Link