香港中文大學
The Chinese University of Hong Kong

之江实验室
ZHEJIANG LAB

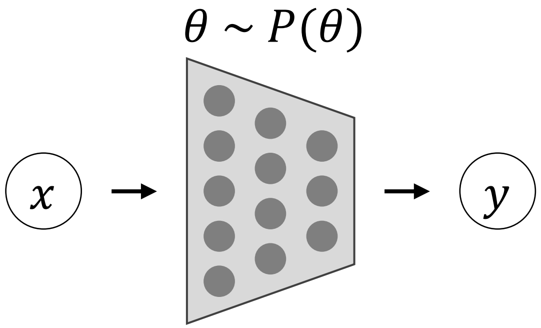
Uncertainty Estimation by Fisher Information-based Evidential Deep Learning

Danruo Deng[1], Guangyong Chen[2*], Yang Yu[1], Furui Liu[2], Pheng-Ann Heng[1]

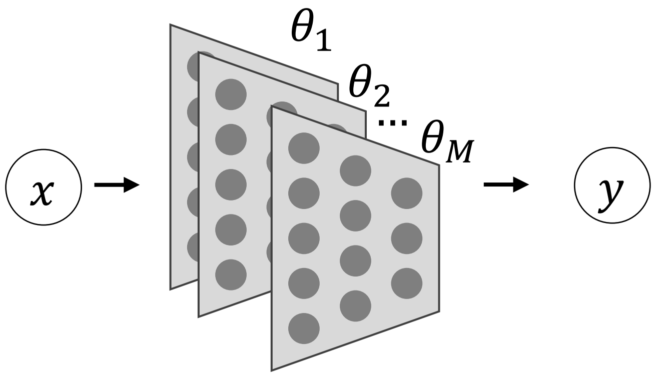
[1] The Chinese University of Hong Kong, [2] Zhejiang Lab

Background

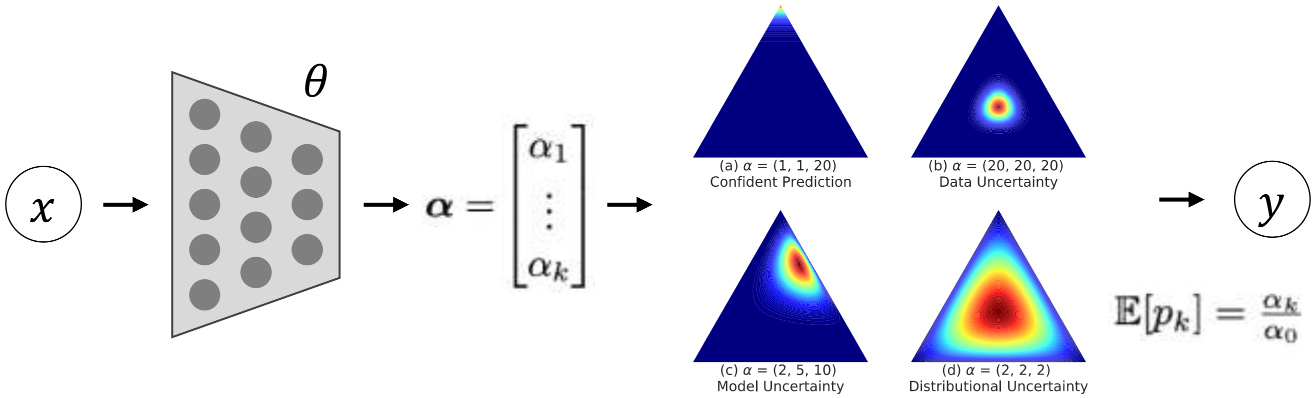
Bayesian Neural Networks (BNNs):



Ensemble Methods:



Evidential Neural Networks [1]:



$$\mathbb{E}[p_k] = \frac{\alpha_k}{\alpha_0} = \frac{\alpha_k}{\sum_{i=1}^{K} \alpha_i}$$

$$\boldsymbol{p} \sim Dir(\boldsymbol{\alpha})$$

$$Dir(\boldsymbol{p}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} p_k^{\alpha_k - 1}, \alpha_0 = \sum_{k=1}^{K} \alpha_k$$

- Predictions with *softmax* is **over-confidence** and **cannot distinguish between different uncertainties.**
- BNNs and Ensemble methods are computationally expensive, and also **cannot distinguish between distributional uncertainty and other uncertainties.**
- Evidential neural networks **quantify different types of uncertainty** by modeling the output as the evidence use to obtain concentration parameters of a Dirichlet distribution.

[1] Sensoy et al., Evidential deep learning to quantify classification uncertainty. NeurIPS 2018.

Motivation: EDL underestimates data uncertainty

Limitation: EDL **cannot** distinguish samples with **different data uncertainties**.

(a)

EDL	0.63	0.43	0.72	0.63	0.14
$\mathcal{I}$-EDL	1.54	1.49	0.98	0.82	0.47

(b)

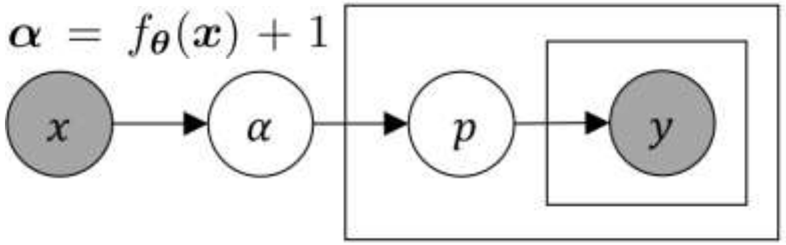
EDL	0.23	0.12	0.15	0.35	0.09
$\mathcal{I}$-EDL	1.48	1.44	1.36	1.35	1.28

Figure 1. Data uncertainty for (a) digit "4" in MNIST, (b) "horse" in CIFAR10. $\mathcal{I}$-EDL has the ability to distinguish between hard samples (orange) and easy samples (green), but EDL cannot.

Analysis: For samples with **high data uncertainty** but annotated with **one-hot vectors**, the learning process of evidence for those mislabeled classes is **over-penalized and remains hindered**.

Graphical model representation of EDL：

$$\boldsymbol{\alpha} = f_{\boldsymbol{\theta}}(\boldsymbol{x}) + 1$$



$$\boldsymbol{p} \sim Dir(\boldsymbol{\alpha}) \quad \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{p}, \sigma^2 \boldsymbol{I})$$

Objective function:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{P}} \mathbb{E}_{\boldsymbol{p}\sim Dir(\boldsymbol{\alpha})} \left[(\boldsymbol{y}-\boldsymbol{p})^T (\boldsymbol{y}-\boldsymbol{p}) \right]$$

$$=$$

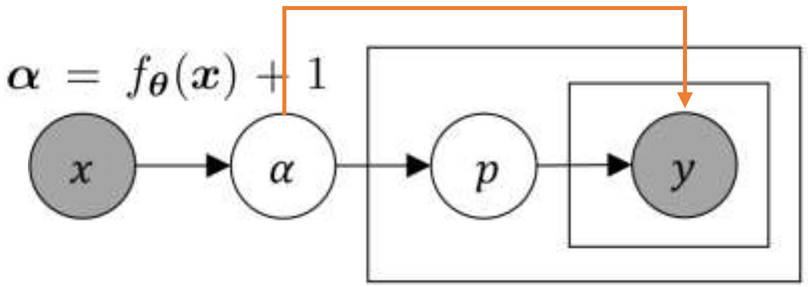
$$\max_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{P}} \left[\log \mathbb{E}_{\boldsymbol{p}\sim Dir(\boldsymbol{\alpha})} [\mathcal{N}(\boldsymbol{y}|\boldsymbol{p}, \sigma^2 \boldsymbol{I})] \right]$$

Methods: Fisher Information-based Evidential Deep Learning (I-EDL)

Key Idea: A certain class label with **higher evidence** is allowed to have a **larger variance**, so that **the evidence for missing labels can be preserved** while maximizing the likelihood of the observed labels.

Method: Use Fisher Information Matrix (FIM) to measure the amount of information that **the observed class probabilities p** carry about the **concentration parameters α** of a Dirichlet distribution that models p.

Graphical model representation of **I-EDL**：



$\boldsymbol{p} \sim Dir(\boldsymbol{\alpha})$ $\quad \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{p}, \sigma^2 \mathcal{I}(\boldsymbol{\alpha})^{-1})$

Objective function:

$$\max_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{P}} \left[\log \mathbb{E}_{\boldsymbol{p}\sim Dir(\boldsymbol{\alpha})} [\mathcal{N}(\boldsymbol{y}|\boldsymbol{p}, \sigma^2 \mathcal{I}(\boldsymbol{\alpha})^{-1})]\right]$$

$$\begin{aligned}\mathcal{I}(\boldsymbol{\alpha}) &= \mathbb{E}_{Dir(\boldsymbol{p}|\boldsymbol{\alpha})}\left[\frac{\partial \ell}{\partial \boldsymbol{\alpha}} \frac{\partial \ell}{\partial \boldsymbol{\alpha}^T}\right] \\ &= \mathrm{diag}([\psi^{(1)}(\alpha_1), \cdots, \psi^{(1)}(\alpha_K)]) - \psi^{(1)}(\alpha_0)\mathbf{1}\mathbf{1}^T\end{aligned}$$

- $\psi^{(1)}(\cdot)$denotes the trigamma function, defined as $\psi^{(1)}(x) = d\psi(x)/dx$= $d^2 \ln \Gamma(x)/dx^2$.
- Since $\psi^{(1)}(x)$ is a monotonically decreasing function when x > 0, the class label with higher evidence corresponds to less Fisher information.

Methods: Fisher Information-based Evidential Deep Learning (I-EDL)

$$\max_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{P}}\left[\log \mathbb{E}_{\boldsymbol{p}\sim Dir(\boldsymbol{\alpha})}[\mathcal{N}(\boldsymbol{y}|\boldsymbol{p},\sigma^2\mathcal{I}(\boldsymbol{\alpha})^{-1})]\right]$$

Jensen's inequality

$$\begin{aligned}\min_{\boldsymbol{\theta}} \quad & \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{P}}\mathbb{E}_{\boldsymbol{p}\sim Dir(\boldsymbol{\alpha})}\left[-\log p(\boldsymbol{y}|\boldsymbol{p},\boldsymbol{\alpha},\sigma^2)\right] \\ \text{s.t.} \quad & \boldsymbol{\alpha} = f_{\boldsymbol{\theta}}(\boldsymbol{x}) + 1 \\ & \mathcal{I}(\boldsymbol{\alpha}) = \mathbb{E}_{Dir(\boldsymbol{p}|\boldsymbol{\alpha})}\left[-\frac{\partial^2 \log Dir(\boldsymbol{p}|\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}\boldsymbol{\alpha}^T}\right] \\ & p(\boldsymbol{y}|\boldsymbol{p},\boldsymbol{\alpha},\sigma^2) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{p},\sigma^2\mathcal{I}(\boldsymbol{\alpha})^{-1})\end{aligned}$$

PAC-Bayesian Bound

$$\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_i(\boldsymbol{\theta}) + \frac{1}{\lambda}D_{\mathrm{KL}}(Dir(\boldsymbol{p}_i|\boldsymbol{\alpha}_i)\|Dir(\boldsymbol{p}_i|\boldsymbol{\mu}_i)).$$

where

$$\begin{aligned}\mathcal{L}_i(\boldsymbol{\theta}) &= \mathbb{E}_{Dir(\boldsymbol{p}_i|\boldsymbol{\alpha}_i)}\left[-\log\mathcal{N}(\boldsymbol{y}_i|\boldsymbol{p}_i,\sigma^2\mathcal{I}(\boldsymbol{\alpha}_i)^{-1})\right] \\ &\propto \underbrace{\mathbb{E}\left[(\boldsymbol{y}_i-\boldsymbol{p}_i)^T\mathcal{I}(\boldsymbol{\alpha}_i)(\boldsymbol{y}_i-\boldsymbol{p}_i)\right]}_{\mathcal{L}_i^{\mathcal{I}\text{-MSE}}} - \sigma^2\underbrace{\log|\mathcal{I}(\boldsymbol{\alpha}_i)|}_{\mathcal{L}_i^{|\mathcal{I}|}}\end{aligned}$$

Table 1. Given a sample $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, the difference in loss function between $\mathcal{I}$-EDL and EDL are marked in blue.

	EDL	**$\mathcal{I}$-EDL**
MSE	$\sum_{j=1}^{K}(y_{ij}-\frac{\alpha_{ij}}{\alpha_{i0}})^2 + \sum_{j=1}^{K}\frac{\alpha_{ij}(\alpha_{i0}-\alpha_{ij})}{\alpha_{i0}^2(\alpha_{i0}+1)}$	$\sum_{j=1}^{K}(y_{ij}-\frac{\alpha_{ij}}{\alpha_{i0}})^2\psi^{(1)}(\alpha_{ij}) + \sum_{j=1}^{K}\frac{\alpha_{ij}(\alpha_{i0}-\alpha_{ij})}{\alpha_{i0}^2(\alpha_{i0}+1)}\psi^{(1)}(\alpha_{ij})$
KL	$D_{\mathrm{KL}}(Dir(\hat{\boldsymbol{\alpha}}_i)\|Dir(\mathbf{1}))$	$D_{\mathrm{KL}}(Dir(\hat{\boldsymbol{\alpha}}_i)\|Dir(\mathbf{1}))$
$\mathcal{I}$	-	$-\log|\mathcal{I}(\boldsymbol{\alpha}_i)|$

Takeaways

- Standard neural network for classification with Softmax is **over-confidence** and **cannot distinguish between different uncertainties.**
- Although Evidential Deep Learning (EDL) models different types of uncertainties, it still **cannot distinguish between samples of different data uncertainties.**
- we propose a novel and simple method, ***Fisher Information-based Evidential Deep Learning (I-EDL)***, to alleviate the over-penalization of the mislabeled classes by considering importance weights with different classes.
- Extensive experiments on various **image classification, confidence evaluation and OOD detection tasks** demonstrate **the effectiveness of our approach** in achieving **high classification and uncertainty quantification.**

Thanks for your listening.