Are Uncertainty Quantification Capabilities of Evidential Deep Learning a Mirage?

Maohao Shen[1*] J. Jon Ryu[1*] Soumya Ghosh[2] Yuheng Bu[3] Prasanna Sattigeri[2] Subhro Das[2] Gregory W. Wornell[1]

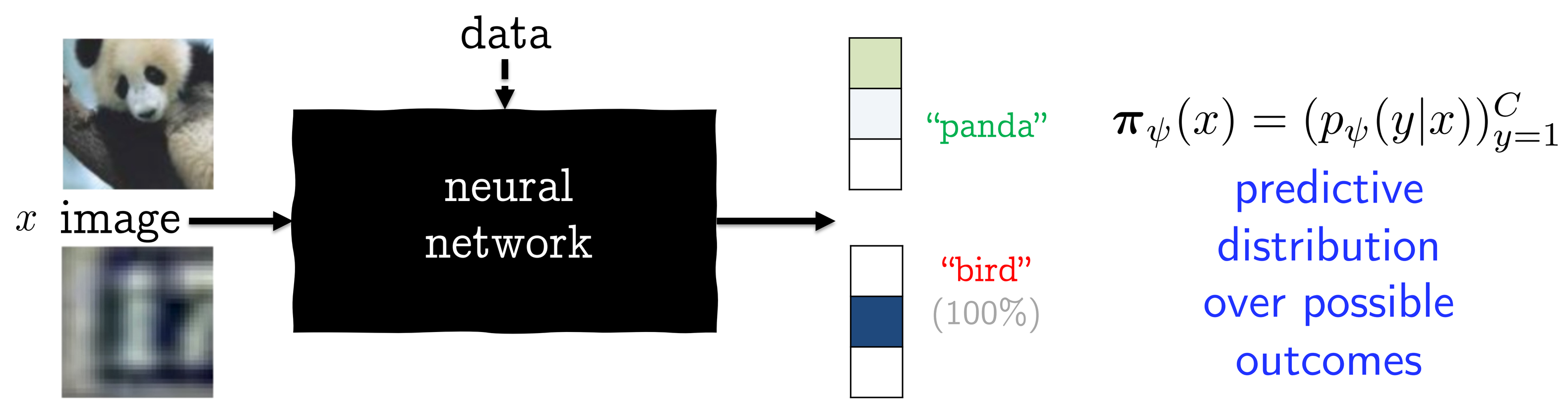
[1]Massachusetts Institute of Technology
[2]MIT-IBM Watson AI Lab
[3]University of Florida

full paper

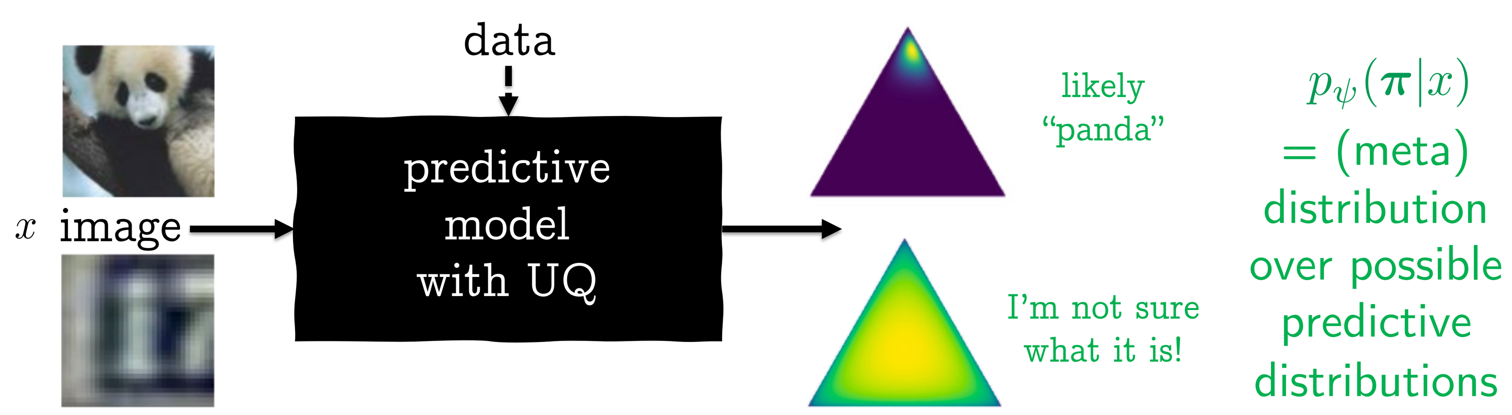
NEURAL INFORMATION PROCESSING SYSTEMS 2024

Background

Single prediction can be unreliable...



We wish to quantify confidence/uncertainty!

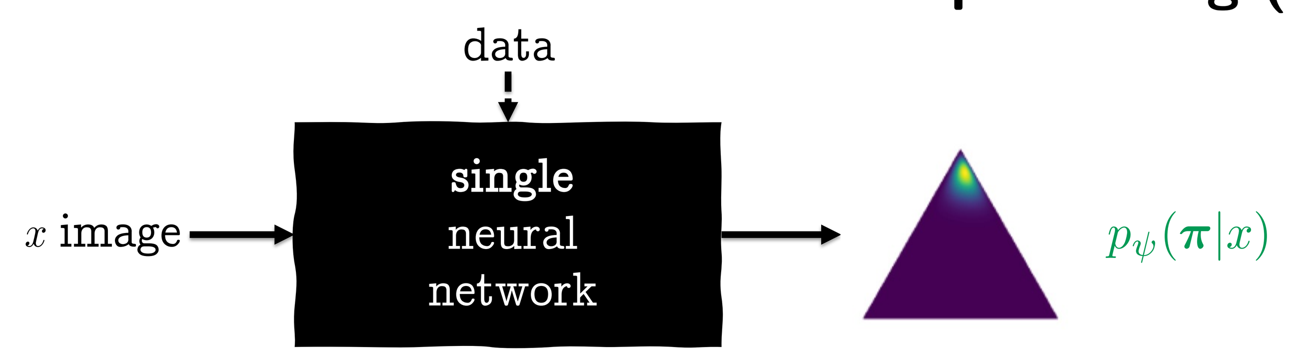


Classical approaches to estimate/induce $p_\psi(\pi|x)$

- **Bayesian methods**: variational inference, MCMC, Monte Carlo Dropout, ...
- **Frequentist methods**: jackknife, bootstrap, ...
- **Ensemble methods**

However, these methods are **computationally inefficient** in general!

An alternative: Evidential Deep Learning (EDL)



EDL aims to quantify uncertainty with **a single neural network** $p_\psi(\pi|x)$

- Various EDL objectives have been proposed from different motivations for different settings (i.e., prediction for discrete, continuous, count outcomes)
- Empirical successes shown for downstream tasks (e.g., OOD detection)

However, recent works reported **suspicious behaviors** (e.g., non-vanishing epistemic uncertainty) + **unifying theoretical understanding is lacking**

Goal: Demystifying EDL Methods

We answer to these questions in this paper:

Q1. What do EDL methods learn as uncertainty?
Q2. Why are the EDL methods empirically successful?
Q3. How can we make EDL methods more reliable?

Answers: **A1.** made-up target / **A2.** ∵ EDL ≈ EBM OOD detector / **A3.** bring back external stochasticity

References

[7] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," in NeurIPS 2018.
[8] A. Malinin and M. Gales, "Reverse KL-divergence training of prior networks: Improved uncertainty and adversarial robustness," in NeurIPS 2019.
[10] M. Sensoy et al., "Evidential deep learning to quantify classification uncertainty," in NeurIPS 2018.
[12] W. Chen et al., "A variational Dirichlet framework for out-of-distribution detection," arXiv 2018.
[13] T. Joo et al., "Being Bayesian about categorical probability," in ICML 2020.
[15] B. Charpentier et al., "Posterior network: Uncertainty estimation without OOD samples via density-based pseudo-counts," in NeurIPS 2020.
[20] B. Charpentier et al., "Natural posterior network: Deep Bayesian uncertainty for exponential family distributions," in ICLR 2022.

Unifying EDL Objectives: A New Taxonomy

Method (name of loss)	likelihood	$D(\cdot,\cdot)$	prior α_0	γ_{ood}	$\alpha_\psi(x)$ parameterization
FPriorNet (F-KL loss) [7]	categorical	fwd. KL	$=\mathbb{1}_C$	>0	direct
RPriorNet (R-KL loss) [8]	categorical	rev. KL	$=\mathbb{1}_C$	>0	direct
EDL (MSE loss) [10]	Gaussian	rev. KL	$=\mathbb{1}_C$	$=0$	direct
Belief Matching (VI loss) [12, 13]	categorical	rev. KL	$\in\mathbb{R}^C_{>0}$	$=0$	direct
PostNet (UCE loss) [15]	categorical	rev. KL	$=\mathbb{1}_C$	$=0$	density w/ single flow
NatPN (UCE loss) [20]	categorical	rev. KL	$=\mathbb{1}_C$	$=0$	density w/ multiple flows

Q. What's the common principle behind these objectives?

A. $$\mathcal{L}(\psi) := \underbrace{\mathbb{E}_{p(x,y)}[D(p^{(\nu)}(\pi|y), p_\psi(\pi|x))]}_{\text{in-distribution objective}} + \gamma_{ood}\underbrace{\mathbb{E}_{p_{ood}(x)}[D(p(\pi), p_\psi(\pi|x))]}_{\text{OOD objective}}$$

"fixed" uncertainty target — EDL model — prior

EDL aims to fit "fixed" uncertainty target!

Theorem 5.1. *For any prior $p(\pi)$ and likelihood $p(y|\pi)$, we have*

$$\min_\psi \mathbb{E}_{p(x,y)}[D(p_\psi(\pi|x) \,\|\, p_\nu(\pi|y))] \equiv \min_\psi \mathbb{E}_{p(x)}[D(p_\psi(\pi|x) \,\|\, p^\star(\pi|x))],$$

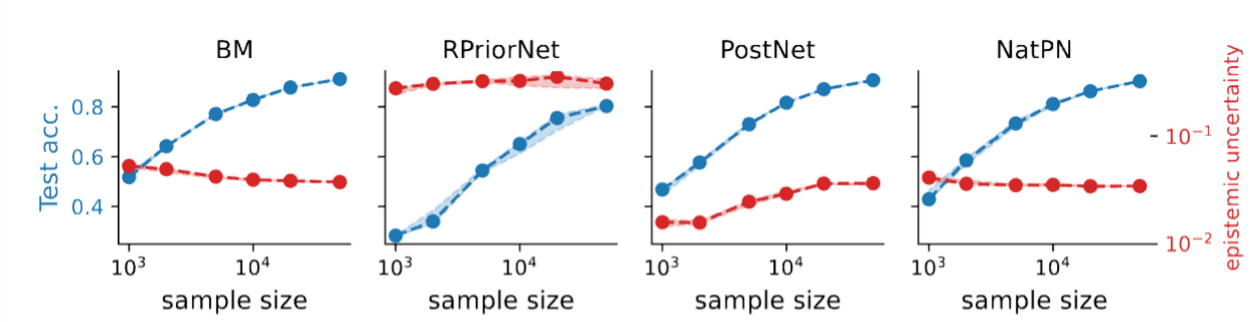
where $$p^\star(\pi|x) := \frac{p(\pi)\exp(\nu\mathbb{E}_{p(y|x)}[\log p(y|\pi)])}{\int p(\pi)\exp(\nu\mathbb{E}_{p(y|x)}[\log p(y|\pi)])\,d\pi}.$$

in-distribution objective with reverse KL div.

what EDL objectives set as the "optimal" meta distribution

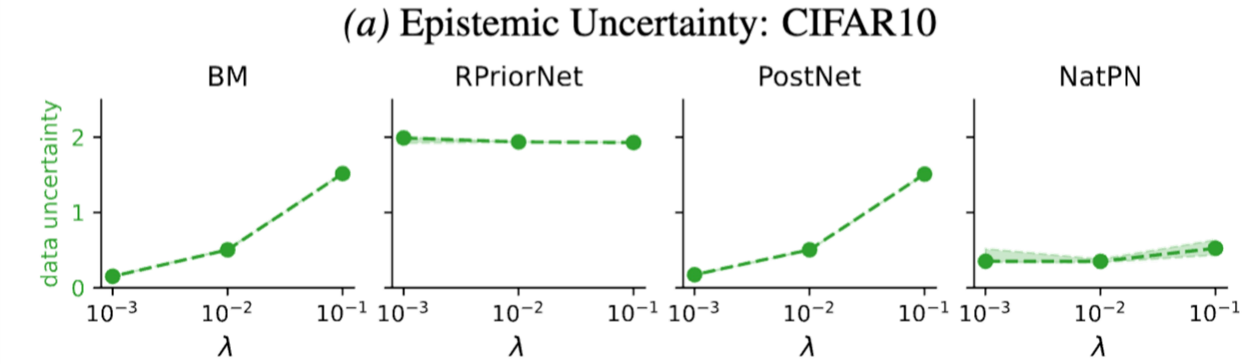
The Unifying View Demystifies EDL

Uncertainties learned by EDL exhibits spurious behaviors



Learned epistemic uncertainty is non-vanishing as data size grows

(a) Epistemic Uncertainty: CIFAR10



Learned aleatoric uncertainty is sensitive to regularization parameter

(b) Aleatoric Uncertainty: CIFAR10

EDL Methods ≈ EBM-based OOD detector

$$\mathcal{L}(\psi) := \underbrace{\mathbb{E}_{p(x,y)}[D(p^{(\nu)}(\pi|y), p_\psi(\pi|x))]}_{\text{in-distribution objective}} + \gamma_{ood}\underbrace{\mathbb{E}_{p_{ood}(x)}[D(p(\pi), p_\psi(\pi|x))]}_{\text{OOD objective}}$$

$$\approx$$

$$-\mathbb{E}_{p(x,y)}[\log p_\psi(y|x)] + \tau\{\mathbb{E}_{p(x)}[\max(0, E_\phi(x) - m_{id})^2]\} + \mathbb{E}_{p_{ood}(x)}[\max(0, m_{ood} - E_\phi(x))^2]\}$$

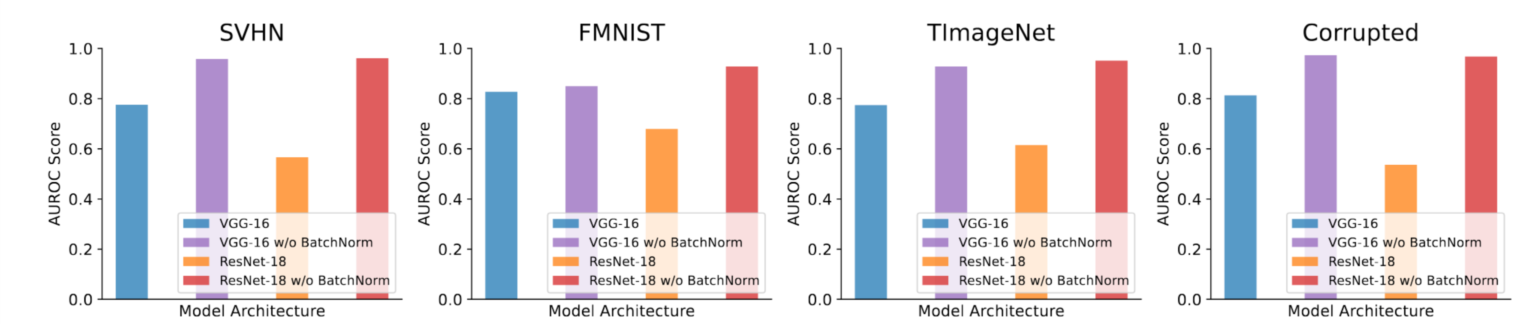
- This resemblance explains EDL methods' empirical success on OOD detection

Other empirical pitfalls about EDL methods (see right column)

- EDL methods' performance is inherently sensitive to model architecture
- EDL methods' auxiliary techniques, such as density param., are not robust

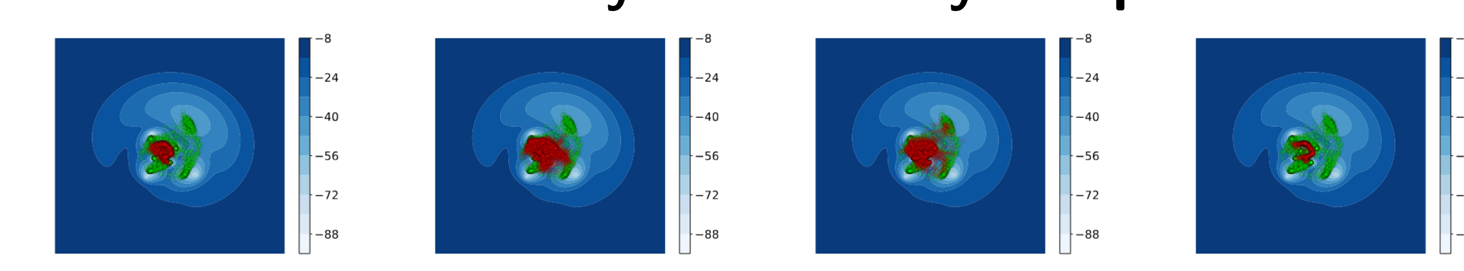
Learned Uncertainties in EDL are Fragile

EDL methods may be sensitive to model architecture



OOD Detection Performance of RPriorNet with Different Model Architecture. RPriorNet is sensitive to the choice of model architecture.

EDL with "flow density model" may not perform well



(a) SVHN (b) FMNIST (c) TImageNet (d) Corrupted

Visualization of Epistemic Uncertainty in PostNet's 2D Latent Feature Space. PostNet leverages a flow-based density estimation model, which outputs low (high) uncertainty for in-distribution (OOD) regions as expected. However, given that the input data is high-dimensional, PostNet suffers from the feature collapse issue that mapping **OOD data** to the same region as **ID data** in the latent space, making them indistinguishable.

How Can We Improve EDL?

Fundamental cause: Ignorance of model stochasticity

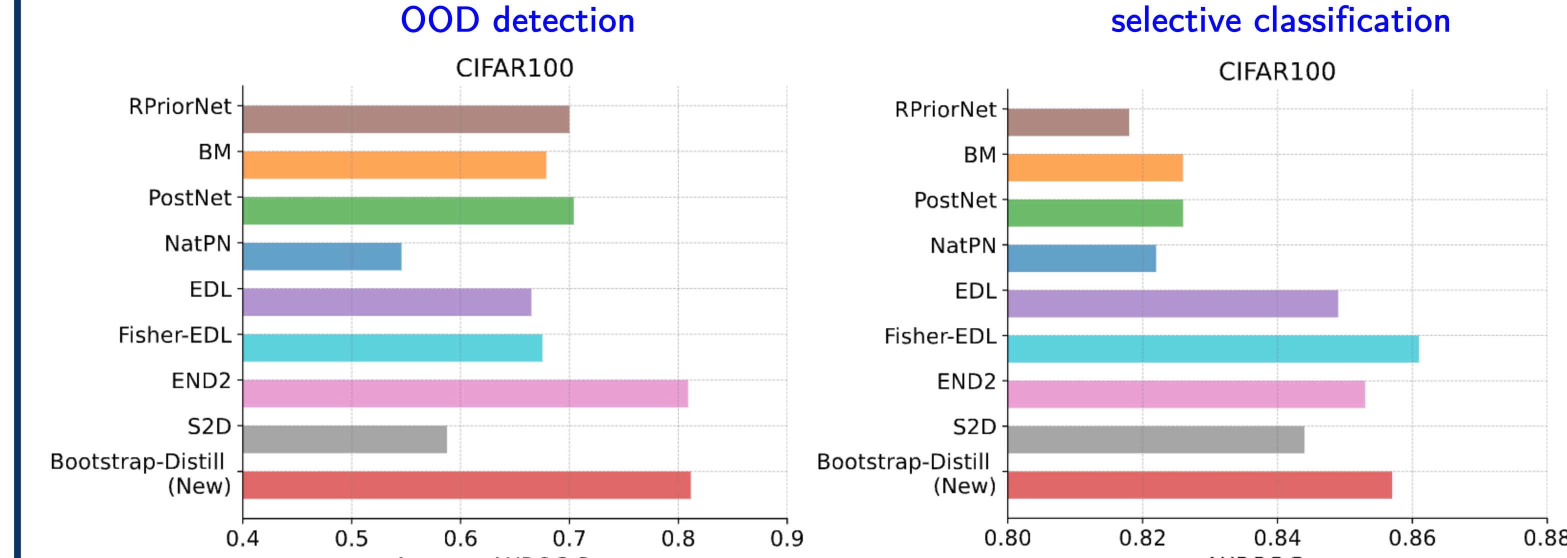
EDL methods assume no model randomness by setting $p(\psi|\mathcal{D}) \leftarrow \delta(\psi - \psi^\star)$

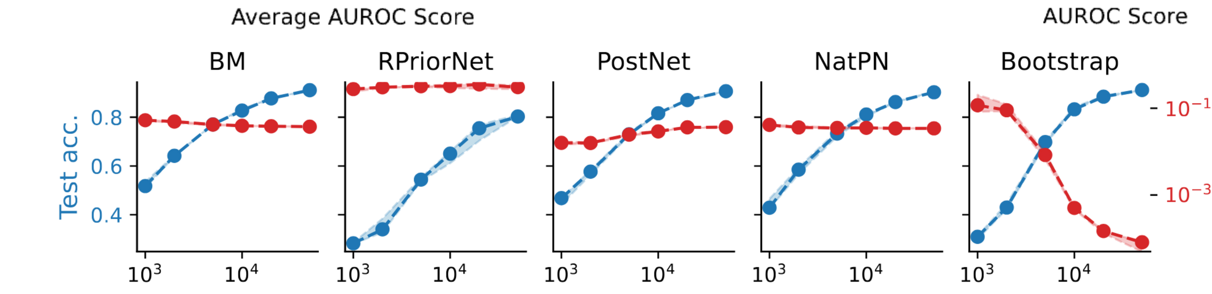
- Posterior $p(\pi|x,\mathcal{D}) \triangleq \int p(\pi|x,\psi)p(\psi|\mathcal{D})d\psi$ becomes degenerate
- The only benefit is computational efficiency
- EDL methods has to fit model $p_\psi(\pi|x)$ to an artificial uncertainty target

Use EDL to distill uncertainty from model stochasticity

$$\mathcal{L}(\theta) = D(\underbrace{p(\pi|x,\mathcal{D})}_{\text{posterior distribution induced from model stochasticity}} \,\|\, \underbrace{p_\theta(\pi|x)}_{\text{a single EDL model}})$$

- A new proposal: Bootstrap-Distill EDL
  - Train multiple models with different random subsamples (**bootstrap**)
  - A single network distills the model uncertainty via an EDL objective
- Bootstrap-Distill shows superior UQ performance!





Epistemic Uncertainty and Test Accuracy v.s. Number of Training Data. Compared to other EDL methods, our proposed Bootstrap Distillation method can faithfully quantify epistemic uncertainty, i.e., the uncertainty is monotonically decreasing with increasing number of data.