Robust Adversarial Quantification via Conflict-aware Evidential Deep Learning

Charmaine Barker, Daniel Bethell, Simos Gerasimou

University of York

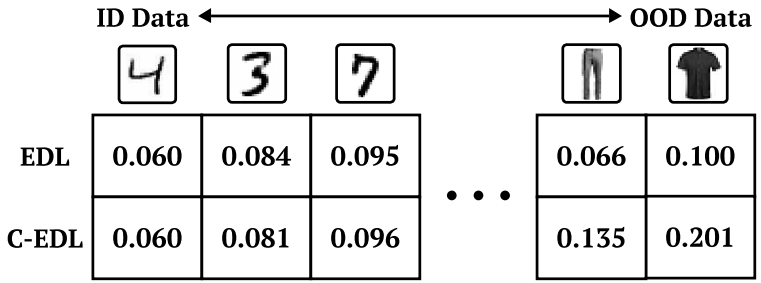
Cyprus University of Technology

Robust Uncertainty Quantification

Uncertainty quantification (UQ) estimates **how confident a model should be** in its predictions.

Evidential Deep Learning (EDL) enables **lightweight** UQ via a Dirichlet distribution over class probabilities.

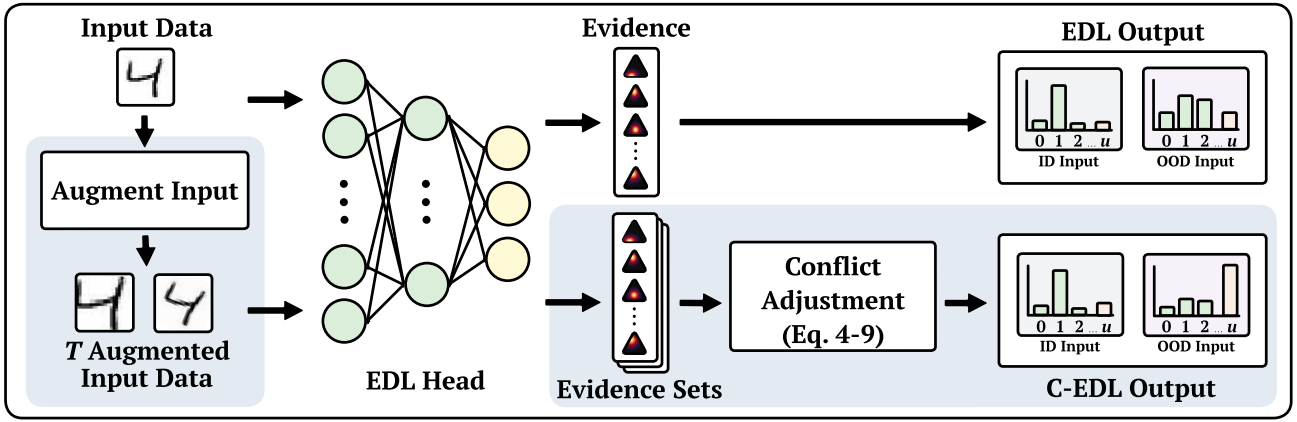
However, EDL can remain **overconfident** on **out-of-distribution** (OOD) and **adversarial inputs**.



Conflict-aware Evidential Deep Learning

C-EDL is a **lightweight post-hoc** extension to EDL that measures **prediction conflict** across **task-preserving transformed views** of the same input.

When this conflict is **high**, it reduces evidential strength and increases predictive uncertainty.



Intra-class Conflict

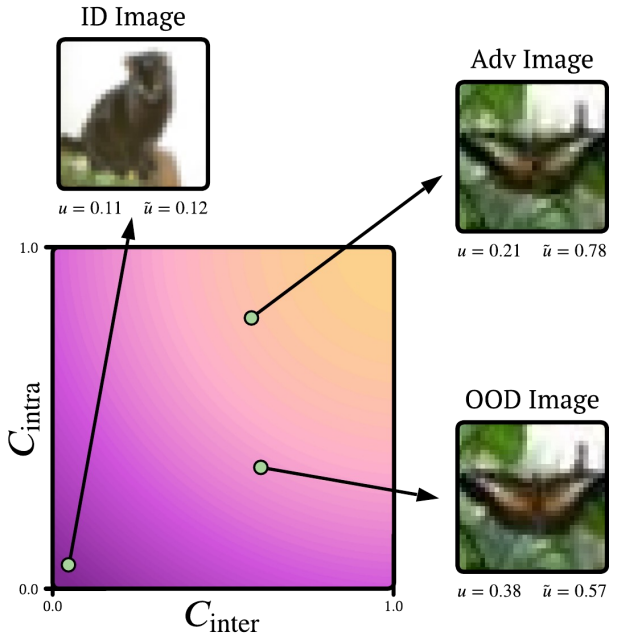
Measures how much evidence for each class **varies across transformed views**.

$$C_{\text{intra}} = \frac{1}{K}\sum_{k=1}^{K}\frac{\sigma(\{\alpha_k^{(t)}\}_{t=1}^{T})}{\mu(\{\alpha_k^{(t)}\}_{t=1}^{T}) + \epsilon}$$

High C_{intra} means **unstable class-wise evidence**.

Where:

- K: number of classes
- T: number of transformed views
- $\alpha_k^{(t)}$: Dirichlet parameter for class k for transformed view t
- $\sigma(\cdot)$: standard deviation across the T transformed views
- $\mu(\cdot)$: mean across the T transformed views
- ϵ: small positive constant for numerical stability



Inter-class Conflict

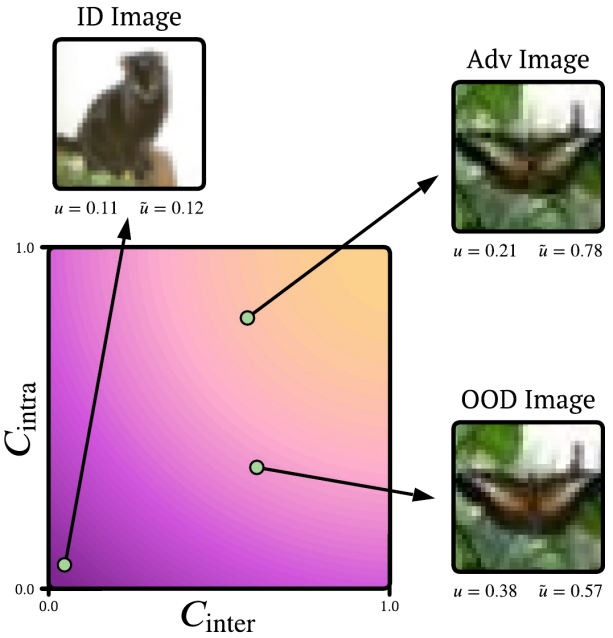
Measures contradiction between competing classes **within a view**.

$$C_{\text{inter}} = \frac{1}{T}\sum_{t=1}^{T}\left(1 - \exp\left(-\beta \sum_{k=1}^{K}\sum_{j=k+1}^{K} \left(\frac{\min(\alpha_k^{(t)}, \alpha_j^{(t)})}{\max(\alpha_k^{(t)}, \alpha_j^{(t)})} \times \frac{\min(\alpha_k^{(t)}, \alpha_j^{(t)})}{\sum_{k=1}^{K} \alpha_k^{(t)}} \times 2\right)^2\right)\right)$$

High C_{inter} means **multiple classes are strongly and similarly supported**.

Where:

- j: index of a competing class
- β: scaling parameter controlling the sharpness of the penalty



Combined Conflict & Adjustment

We combine C_{intra} and C_{inter} into C.

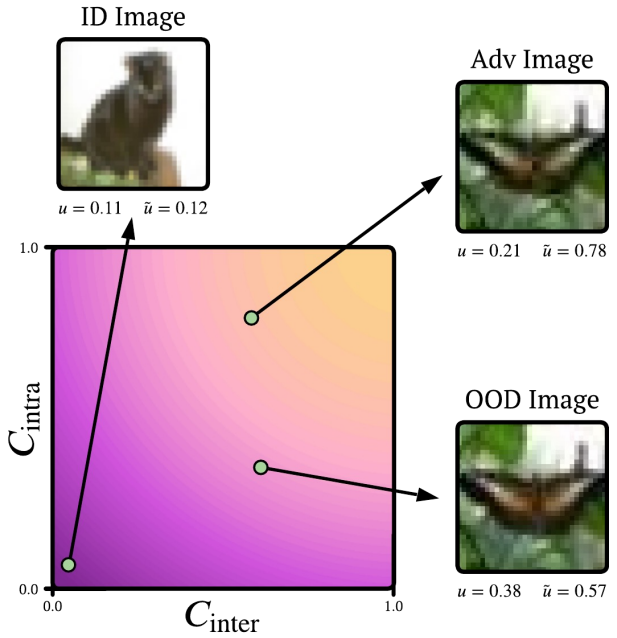
$$C = C_{\text{inter}} + C_{\text{intra}} - C_{\text{inter}}C_{\text{intra}} - \lambda(C_{\text{inter}} - C_{\text{intra}})^2$$

Evidence is then scaled using C to reduce overconfident predictions.

$$\tilde{\alpha}_k = \bar{\alpha}_k \times \exp(-\delta C)$$

Where:

- λ: asymmetric penalization weighting
- $\overline{\alpha_k}$: aggregated Dirichlet parameter for class k across transformed views
- $\widetilde{\alpha_k}$: conflict-adjusted Dirichlet parameter for class k
- δ: adjustment sensitivity hyperparameter



Results

C-EDL consistently **improves abstention** under OOD shift and adversarial attack across six dataset pairs, outperforming seven alternative techniques.

MNIST → FMNIST

Method	ID Acc (%)	ID Cov (%)	OOD Cov (%)	Adv Cov (%)
EDL	**99.96**	**96.61**	2.52	52.21
R-EDL	**99.96**	96.27	2.34	49.05
DA-EDL	99.89	95.16	2.98	28.74
C-EDL	**99.96**	94.18	**2.00**	**15.51**

Here,

EDL: Baseline Evidential Deep Learning[1]

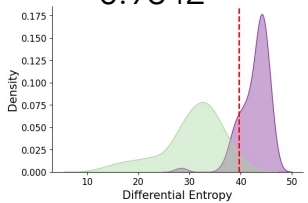
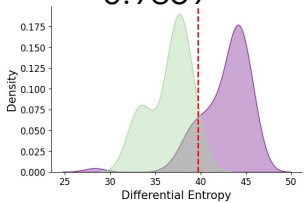
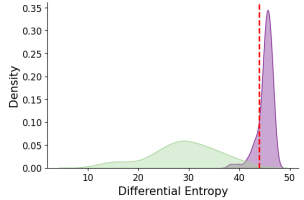
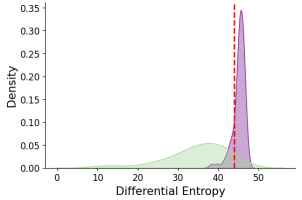
R-EDL: Relaxed Evidential Deep Learning[2]

DA-EDL: Density Aware Evidential Deep Learning[3]

Results

C-EDL consistently **improves abstention** under OOD shift and adversarial attack across six dataset pairs, outperforming seven alternative techniques.

MNIST → FMNIST

Method	ID-OOD AUROC	ID-Adv AUROC
EDL	0.9842 	0.9659 
C-EDL	0.9984 	0.9830 

References

1. Sensoy, M., Kaplan, L. and Kandemir, M., 2018. Evidential deep learning to quantify classification uncertainty. Advances in neural information processing systems, 31.

2. Chen, M., Gao, J. and Xu, C., 2024. R-edl: Relaxing nonessential settings of evidential deep learning. In The Twelfth International Conference on Learning Representations.

3. Yoon, T. and Kim, H., 2024, July. Uncertainty Estimation by Density Aware Evidential Deep Learning. In International Conference on Machine Learning (pp. 57217-57243). PMLR.

Full Paper

GitHub

Thank you!

UNIVERSITY of York