

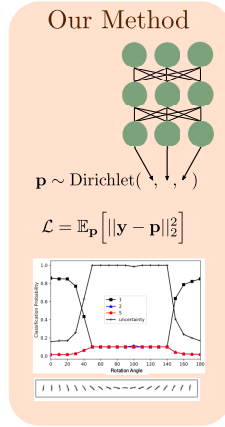
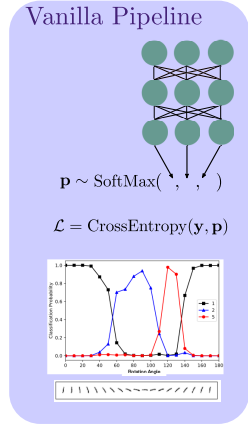
Evidential Deep Learning to Quantify Classification Uncertainty

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada

M. Sensoy¹, L. Kaplan², M. Kandemir³ | ¹Özyeğin University, ²US Army Research Labs ³Bosch Center for Artificial Intelligence



1 Motivation



2 The Subjective Logic Interpretation

- Consider a frame of K mutually exclusive singletons (e.g., class labels)
- Assign a belief mass $b_k \geq 0$ on each singleton $k = 1, \dots, K$ with and define an uncertainty score $u \geq 0$ such that

$$u + \sum_{k=1}^K b_k = 1.$$

- Let $e_k \geq 0$ be the evidence derived for the k^{th} singleton, then the belief b_k and the uncertainty u are computed as

$$b_k = \frac{e_k}{S} \quad \text{and} \quad u = \frac{K}{S},$$

where $S = \sum_{i=1}^K (e_i + 1)$.

- This way, a subjective opinion can be derived easily from a Dirichlet distribution with parameters α_k such that

$$b_k = (\alpha_k - 1)/S.$$

3 The Loss Design

As our method provides a distribution on class probabilities for a given input, we need to minimize the Bayes risk with respect to a loss:

$$\begin{aligned} \mathcal{L}_i(\Theta) &= \int \|\mathbf{y}_i - \mathbf{p}_i\|_2^2 \frac{1}{B(\boldsymbol{\alpha}_i)} \prod_{i=1}^K p_{ij}^{\alpha_{ij}-1} d\mathbf{p}_i \\ &= \sum_{j=1}^K \mathbb{E} \left[y_{ij}^2 - 2y_{ij}p_{ij} + p_{ij}^2 \right] \\ &= \sum_{j=1}^K \left(y_{ij}^2 - 2y_{ij}\mathbb{E}[p_{ij}] + \mathbb{E}[p_{ij}^2] \right). \end{aligned}$$

Regularize the loss against unjustified evidence prediction with an absolutely uncertain predictor:

$$\begin{aligned} \mathcal{L}(\Theta) &= \sum_{i=1}^N \mathcal{L}_i(\Theta) \\ &+ \lambda_t \sum_{i=1}^N KL[D(\mathbf{p}_i | \tilde{\boldsymbol{\alpha}}_i) \parallel D(\mathbf{p}_i | (1, \dots, 1))]. \end{aligned}$$

4 Theoretical Properties

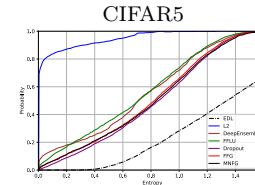
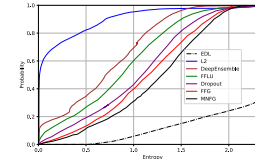
Our loss can be expressed in the following easily interpretable form

$$\begin{aligned} \mathcal{L}_i(\Theta) &= \sum_{j=1}^K (y_{ij} - \mathbb{E}[p_{ij}])^2 + \text{Var}(p_{ij}) \\ &= \sum_{j=1}^K \underbrace{(y_{ij} - \alpha_{ij}/S_i)^2}_{\mathcal{L}_{ij}^{err}} + \underbrace{\frac{\alpha_{ij}(S_i - \alpha_{ij})}{S_i^2(S_i + 1)}}_{\mathcal{L}_{ij}^{var}} \end{aligned}$$

which satisfies the following three propositions.

5 Experiments

Detection of Out-of-Distribution Samples notMNIST

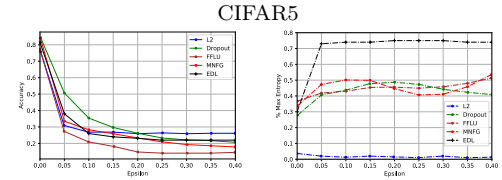
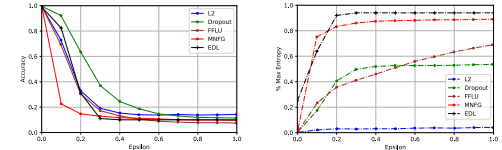


6 Take Homes

- Replace the SoftMax-generated class probabilities with a Dirichlet distribution.
- Minimize Bayes risk in addition to the empirical risk.
- Draw links to and get inspiration from opinion modeling.
- Outperform state of the art in detection of out-of-distribution samples and white-box attacks without any security-specific design.

- Proposition 1.** For any $\alpha_{ij} \geq 1$, the inequality $\mathcal{L}_{ij}^{var} < \mathcal{L}_{ij}^{err}$ is satisfied.
i.e. The loss prioritizes data fit over variance estimation.
- Proposition 2.** For a given sample i with the correct label j , \mathcal{L}_i^{err} decreases when new evidence is added to α_{ij} and increases when evidence is removed from α_{ij} .
i.e. The loss has a tendency to fit to the data.
- Proposition 3.** For a given sample i with the correct class label j , \mathcal{L}_i^{err} decreases when some evidence is removed from the biggest Dirichlet parameter α_{il} such that $l \neq j$.
i.e. The loss performs learned loss attenuation.

Detection of White-Box Adversarial Attacks MNIST



Paper



Code