LiBRe: A Practical Bayesian Approach to Adversarial Detection

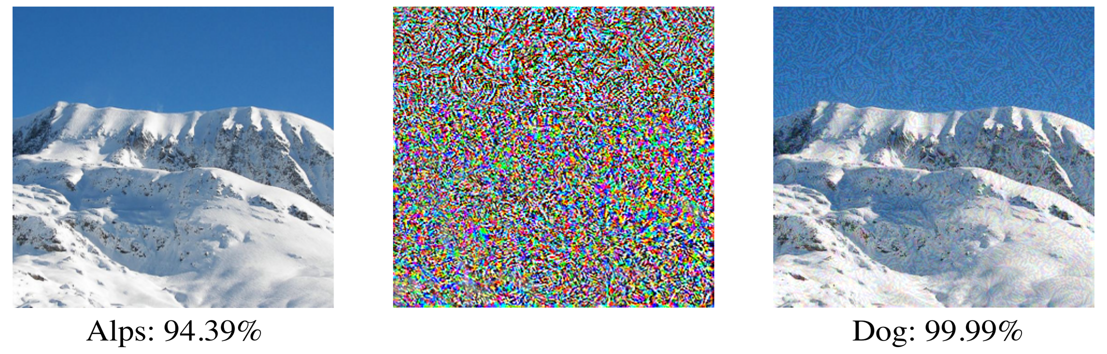
Zhijie Deng[1], Xiao Yang[1], Shizhen Xu[2], Hang Su[1], Jun Zhu[1]

[1] Dept. of Comp. Sci. and Tech., BNRist Center, Institute for AI, Tsinghua-Bosch Joint ML Center, THBI Lab, Tsinghua University [2] RealAI

CVPR VIRTUAL JUNE 19-25

Motivation and introduction:

- DNNs are vulnerable against **adversarial examples**, which are generated by adding human-imperceptible perturbations upon clean examples to deliberately cause misclassification.



Dong et al., 2018

- *Current* ***defenses*** to adversarial examples
  - **Adversarial training** methods are effective, yet cause added training overheads and undermine the predictive performance on clean data.
  - **Adversarial detection** methods detect the adversarial examples ahead of decision making, yet are usually developed for specific tasks or attacks, thus lack the flexibility to effectively *generalize* to other tasks or attacks.

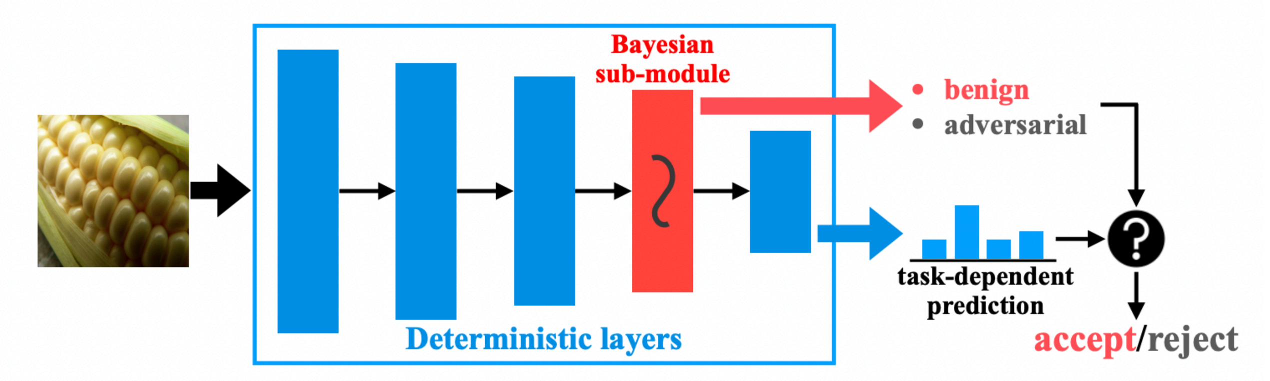
★ Key insight: think of adversarial examples as a special kind of out-of-distribution (OOD) data, and proceed in a **Bayesian** way.

- **Bayesian neural networks** (BNNs) are as **flexible** as DNNs for data fitting in various tasks, and the **epistemic** uncertainty yielded by them suffices for detecting **heterogeneous** OOD/adversarial data in principle.
- Yet, current BNN methods may be less effective in predictive performance, hard to implement, and expensive to train.

★ The solution: **LiBRe** -- **Lightweight Bayesian Refinement:**

Given a **pre-trained task-dependent** DNN

1. LiBRe converts its last **few layers** (e.g. the last ResBlock) to be *Bayesian.*
2. LiBRe **inherits** the **pre-trained** parameters.
3. LiBRe launches **several**-round adversarial detection-oriented **fine-tuning**.



Lightweight Bayesian Refinement:

- A **BNN** is specified by a parameter prior $p(w)$ and a data likelihood $p(D|w)$. We concern the posterior $p(w|D)$. $D=\{D_i\}_{i=1}^{n}$.
  - **Variational BNNs** have shown promise recently. They use a variational $q(w|\theta)$ to approximate $p(w|D)$ by maximizing **ELBO**:

$$\max_{\theta} \mathrm{E}_{q(w|\theta)} \sum_i \log p(D_i|w) - KL(q(w|\theta)||p(w)).$$

  - Predict by $p(D'|D) \approx \mathrm{E}_{q(w|\theta)} p(D'|w) \approx \frac{1}{T}\sum_{t=1}^{T} p\left(D'\middle|w^{(t)}\right), w^{(t)} \sim q(w|\theta)$.
  - Quantifying **epistemic** uncertainty by softmax variance is not universal (e.g. regression), so we adopt the **predictive variance of hidden feature**:

$$Unc = \frac{1}{T-1}\left(\sum_{t=1}^{T}\left\|z^{(t)}\right\|_2^2 - T\left\|\frac{1}{T}\sum_{t=1}^{T} z^{(t)}\right\|_2^2\right) \quad (z^{(t)} \text{ is the hidden feature under } w^{(t)}).$$

- **Partial** Bayesian treatment: **Few-lAyer Deep Ensemble (FADE)**

$$q(w|\theta) = \frac{1}{C}\sum_{c=1}^{C}\delta\left(w_b - w_b^{(c)}\right)\delta\left(w_{-b} - w_{-b}^{(0)}\right).$$

  – w_b: parameters of **tiny Bayesian sub-module**; w_{-b}: the deterministic ones.

  – **FADE** conjoins the **expressiveness** of *deep ensemble* [Lakshminarayanan et al., 2017] and the **efficiency** of *last-layer Bayesian learning* [Kristiadi et al., 2020].

  – A mixture of deltas is a singular approximating distribution, so we indeed relax $q(w|\theta)$ as ***a mixture of Gaussians with small variance*** to estimate $KL(q(w|\theta)||p(w))$.

- ELBO maximization by **stochastic variational inference (SVI)**

$$\max_{\theta} \mathcal{L} = \frac{1}{|\mathcal{B}|}\sum_{\mathcal{B}_i} \log p\left(\mathcal{B}_i\middle|w_b^{(c)}, w_{-b}^{(0)}\right), c \sim \{1,2,\dots,C\}, \mathcal{B} \subset D.$$

  – **Exemplar reparameterization** for variance reduction:

$$\max_{\theta} \mathcal{L}^* = \frac{1}{|\mathcal{B}|}\sum_{\mathcal{B}_i} \log p\left(\mathcal{B}_i\middle|w_b^{(c_i)}, w_{-b}^{(0)}\right), c_i \sim \{1,2,\dots,C\}\ \forall i = 1,\dots,|\mathcal{B}|.$$

- Adversarial example **free uncertainty correction**

$$\max_{\theta} \mathcal{R} = \frac{1}{|\mathcal{B}|}\sum_{\mathcal{B}_i} \min\left(\left\|\tilde{z}_i^{(c_{i,1})} - \tilde{z}_i^{(c_{i,2})}\right\|_2^2, \gamma\right).$$

  – $\tilde{z}_i^{(c_{i,j})}$ refers to the feature of i^{th} training instances with **uniform** input perturbations under parameter sample $w^{(c_{i,j})} = \{w_b^{(c_{i,j})}, w_{-b}^{(0)}\}$.

- Efficient training by **refining pre-trained DNNs**; efficient inference by **parallel computing**

Results:

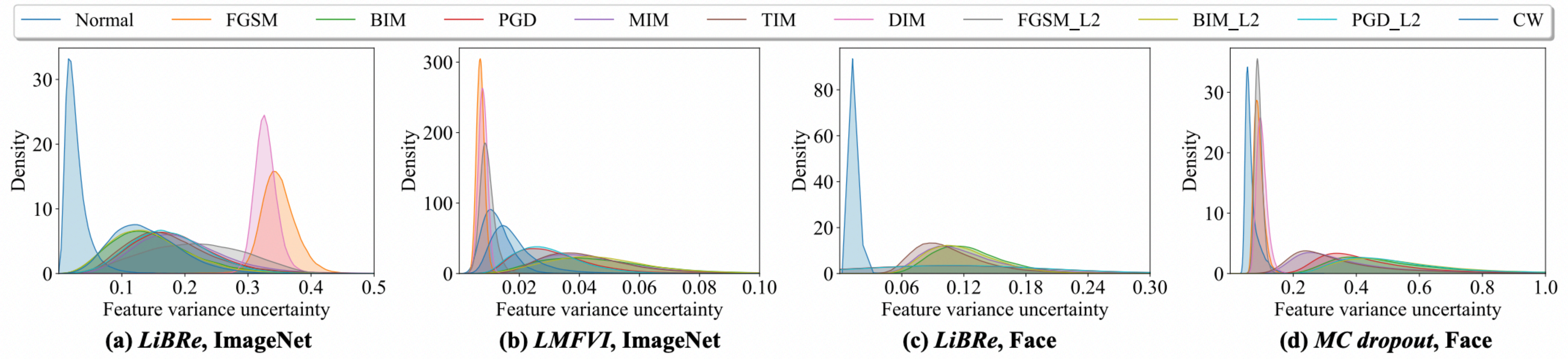
- We perform Bayesian fine-tuning for **only 6** epochs on ImageNet.
- LiBRe preserves **non-degraded accuracy** while demonstrating **near-perfect capacity of detecting adversarial examples**.

Method	Prediction accuracy ↑		AUROC of adversarial detection under *model transfer* ↑			
	TOP1	TOP5	PGD	MIM	TIM	DIM
MAP	76.13%	92.86%	-	-	-	-
MC dropout [17]	74.86%	92.33%	0.660	0.723	0.695	0.605
LMFVI	76.06%	92.92%	0.125	0.200	0.510	0.018
MFVI	75.24%	92.58%	0.241	0.205	0.504	0.150
LiBRe	**76.19%**	**92.98%**	**1.000**	**1.000**	**0.982**	**1.000**

Table 1: Left: comparison on accuracy. Right: comparison on AUROC of adversarial detection under *model transfer*. (ImageNet)

Method	FGSM	BIM	C&W	PGD	MIM	TIM	DIM	FGSM-ℓ_2	BIM-ℓ_2	PGD-ℓ_2
KD [14]	0.639	1.000	0.999	1.000	1.000	0.999	0.624	0.633	1.000	1.000
LID [39]	0.846	0.999	0.999	0.999	0.997	0.999	0.762	0.846	0.999	0.999
MC dropout [17]	0.607	1.000	0.980	1.000	1.000	0.999	0.628	0.577	0.999	0.999
LMFVI	0.029	0.992	0.738	0.943	0.996	0.997	0.021	0.251	0.993	0.946
MFVI	0.102	1.000	0.780	0.992	1.000	0.999	0.298	0.358	0.952	0.935
LiBRe	**1.000**	0.984	0.985	0.994	0.996	0.994	**1.000**	**0.995**	0.983	0.993

Table 2: Comparison on AUROC of adversarial detection for *regular attacks* ↑. (ImageNet)



- LiBRe can be easily applied to **face recognition** & **object detection**.

Conclusion

- Empowered by the **task and attack agnostic modeling** under **Bayes principle**, LiBRe can endow **a variety of** pre-trained task dependent DNNs with the ability of **defending heterogeneous adversarial attacks at a low cost**.
- We build the **FADE** variational and adopt the **pretraining & fine-tuning** workflow to boost the **effectiveness** and **efficiency**.
- We provide a novel insight to realise **adversarial detection-oriented uncertainty quantification** *without* inefficiently crafting adversarial examples.