DeepMind

A Closer Look at the Adversarial Robustness of Information Bottleneck Models

Iryna Korshunova*, David Stutz*[1], Alexander A. Alemi[2], Olivia Wiles[3], and Sven Gowal[3]

* Work done during an internship at DeepMind, [1] Max Planck Institute for Informatics, [2]Google Research, [3]DeepMind

- Information bottlenecks have been shown to significantly improve adversarial robustness of DNNs [1,2]
- We run a number of diagnostics to validate these claims
- Our analysis suggests that previous IB robustness results were influenced by gradient obfuscation

Information Bottlenecks

- The idea is to learn a compressed representation **Z** of an input **X** that is predictive of a target **Y** via the following **IB** objective:

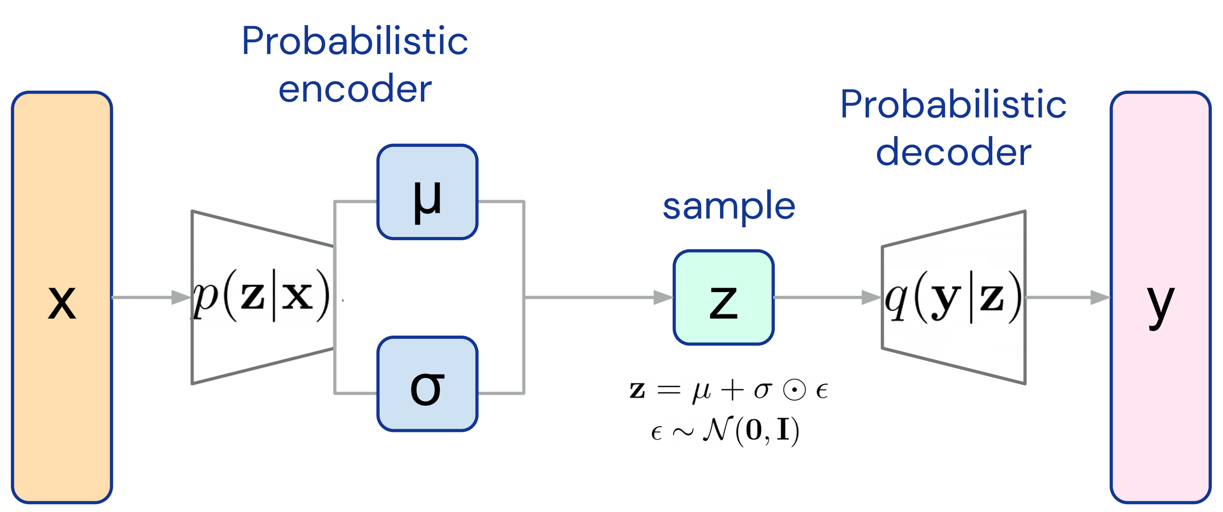
$$\min_{Z} -I(Z,Y) + \beta I(Z,X)$$

- The Variational Information Bottleneck (**VIB**) [1] makes the IB objective practical. Training a neural network with VIB is similar to that of a VAE:

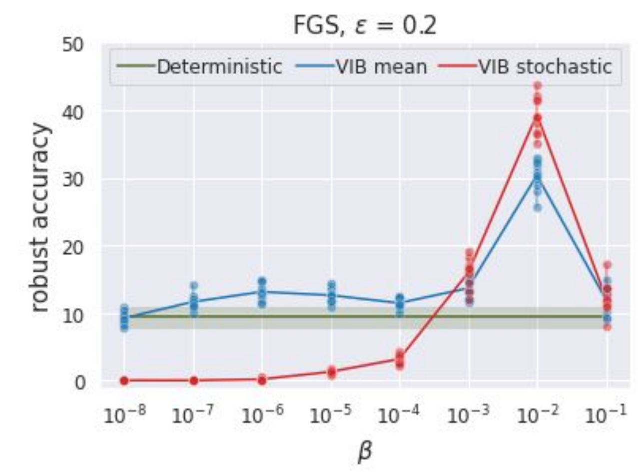
$$\min_{p(\mathbf{z}|\mathbf{x})} \mathbb{E}_{p(\mathbf{x},\mathbf{y})p(\mathbf{z}|\mathbf{x})}\left[-\log q(\mathbf{y}|\mathbf{z}) + \beta \log \frac{p(\mathbf{z}|\mathbf{x})}{q(\mathbf{z})}\right]$$

- The Conditional Entropy Bottleneck (**CEB**) [2] gives a tighter bound on the IB objective:

$$\min_{p(\mathbf{z}|\mathbf{x})} \mathbb{E}_{p(\mathbf{x},\mathbf{y})p(\mathbf{z}|\mathbf{x})}\left[-\log q(\mathbf{y}|\mathbf{z}) + e^{-\rho} \log \frac{p(\mathbf{z}|\mathbf{x})}{q(\mathbf{z}|\mathbf{y})}\right]$$

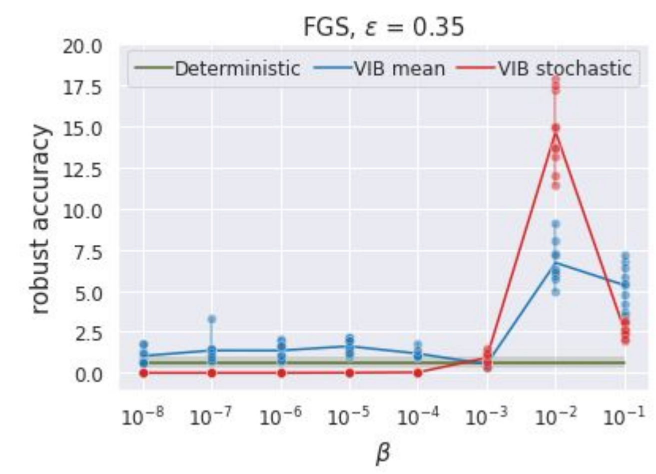


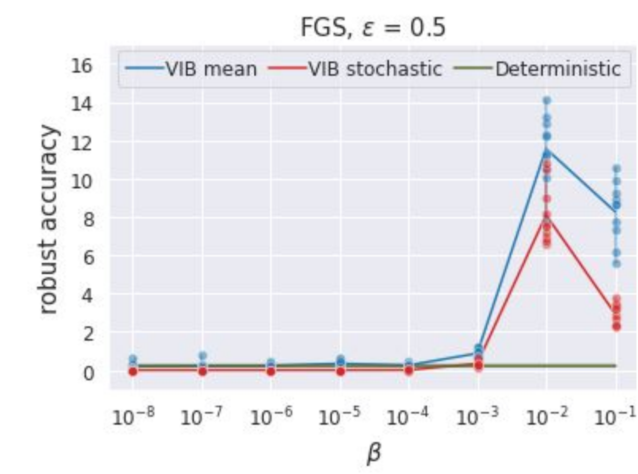
Experiments: MNIST



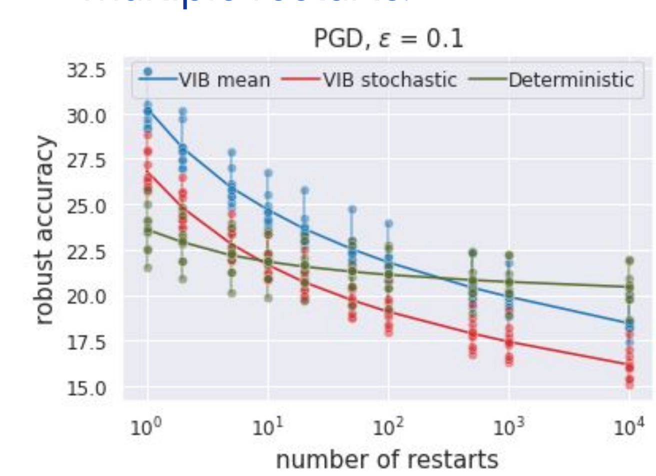
The robust accuracy of VIB models under a Fast Gradient Sign attack with different values of ε.

The results are similar to those of Alemi et al., 2017, which show an improved robustness in comparison to undefended deterministic models.

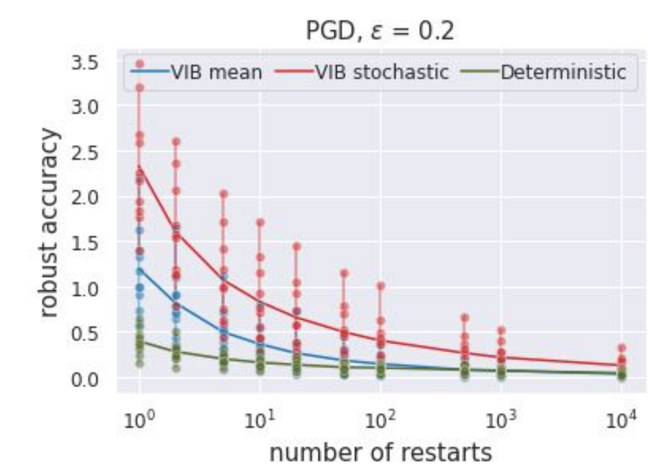




The robust accuracy dramatically decreases as we use a PGD attack with multiple restarts:



The robust accuracy of VIB models keeps decreasing as we do more restarts.

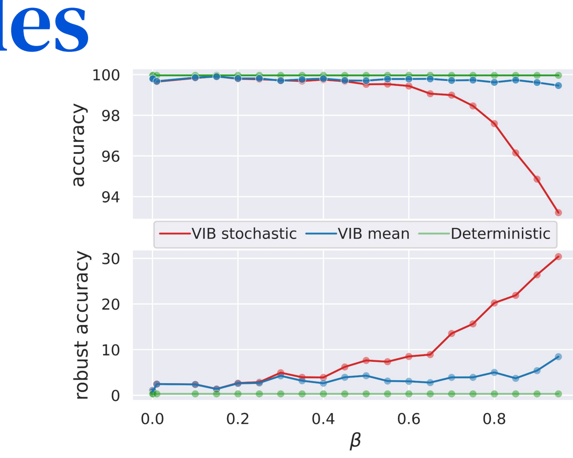


With enough random restarts, the robust accuracy goes to zero.

Toy Examples

$$x_1 = \begin{cases} +y & \text{w.p. } 0.95 \\ -y & \text{w.p. } 0.05 \end{cases}$$

$$x_2 \ldots x_{101} \overset{i.i.d}{\sim} \mathcal{N}(0.3y, 1)$$



This example from Tsipras et al. (2019) motivates the use of IB models for adversarial robustness.

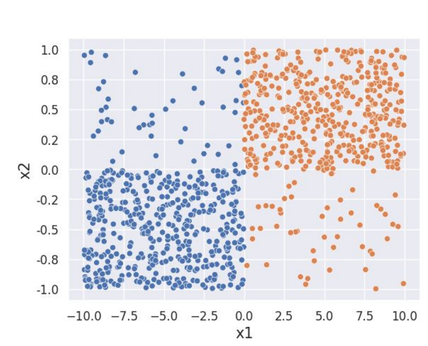
The following example illustrates a failure mode of VIB models:

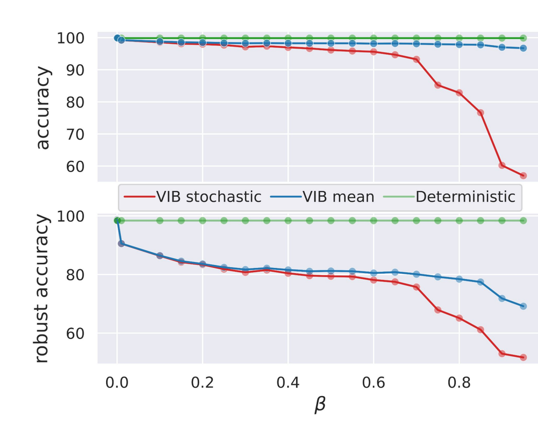
$$p(x_1|y=1) = \mathcal{U}(0,10)$$

$$p(x_2|y=1) = \begin{cases} \mathcal{U}(0,1) & \text{w.p. } 0.9 \\ \mathcal{U}(-1,0) & \text{w.p. } 0.1 \end{cases}$$

$$p(x_1|y=-1) = \mathcal{U}(-10,0)$$

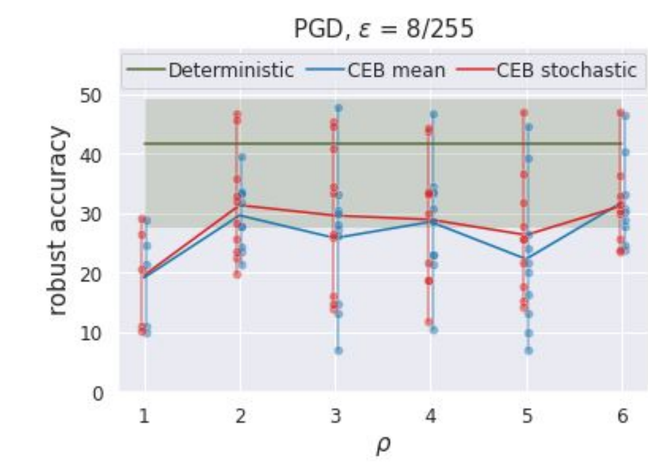
$$p(x_2|y=-1) = \begin{cases} \mathcal{U}(-1,0) & \text{w.p. } 0.9 \\ \mathcal{U}(0,1) & \text{w.p. } 0.1 \end{cases}$$

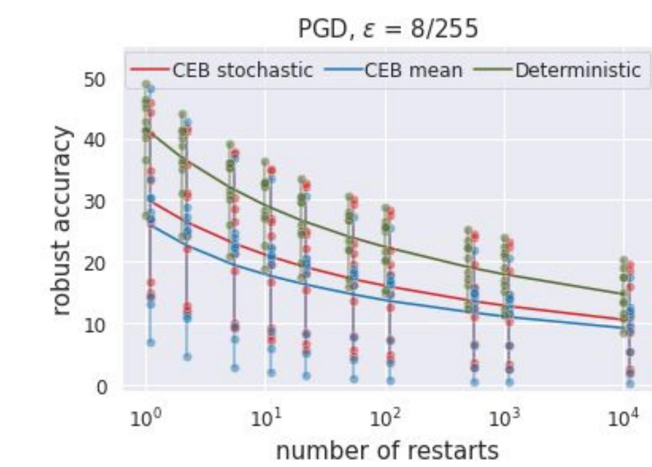


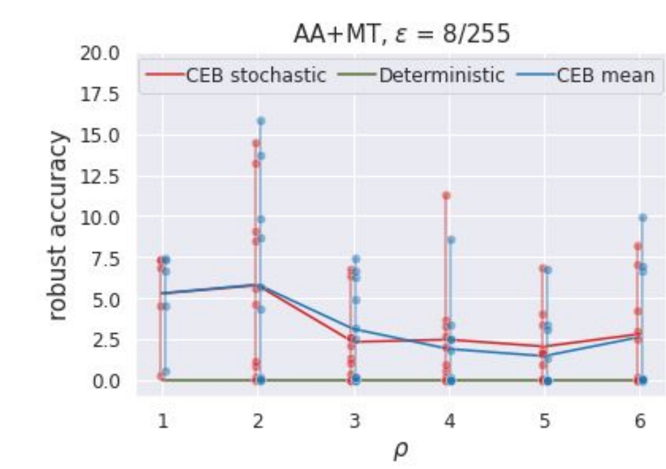


Experiments: CIFAR-10

For CEB models, we also observe a decline in the robust accuracy as we perform more restarts.

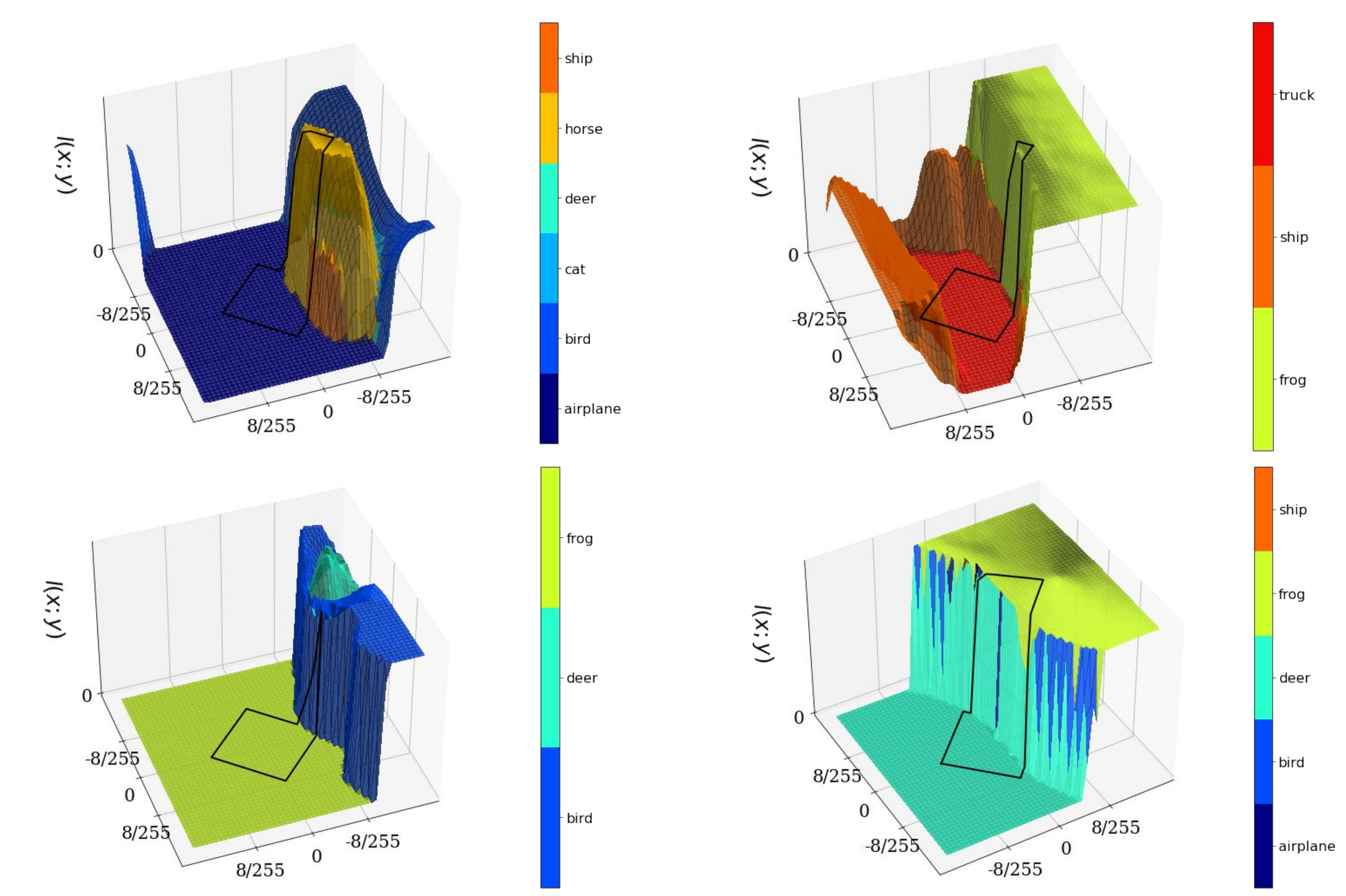






Under our strongest attack, an ensemble of AutoAttack [3] and Multi-targeted [4], the performance of CEB models greatly varies across random seeds.

Loss Surfaces of CEB Models



The flatness of these landscapes explains why gradient-based attacks with cross-entropy loss are not as effective.

References

[1] A. Alemi et al., "Deep variational information bottleneck," 2017
[2] I. Fischer and A. Alemi, "CEB improves model robustness," 2020
[3] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," 2020
[4] S. Gowal et al., "An alternative surrogate loss for PGD-based adversarial testing," 2019
[5] D. Tsipras et al., "Robustness may be at odds with accuracy," 2019