Robustness of Bayesian Neural Networks to Gradient-Based Attacks

Ginevra Carbone, Matthew Wicker, Luca Laurenti, Andrea Patane, Luca Bortolussi, Guido Sanguinetti

Arxiv preprint: 2002.04359

Jiaru Zhang
8.13.2020

Background

- Adversarial attacks: Small, potentially imperceptible perturbations of test inputs can lead to misclassifications of NNs.
- Many attack strategies are based on identifying directions of high variability in the loss function by evaluating gradients.
- This paper shows a remarkable property of BNNs: The gradients of the expected loss function of a BNN vanish in a suitably defined large data limit.

Background

Bayesian Neural Networks and Adversarial Attacks

- The predictions of BNNs are obtained by

$$f\left(\mathbf{x}^*|D\right) = \langle f(\mathbf{x}^*, \mathbf{w}) \rangle_{p(\mathbf{w}|D)} \simeq \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{x}^*, \mathbf{w_i}) \qquad \mathbf{w_i} \sim p\left(\mathbf{w}|D\right)$$

- It can be seen as an ensemble of NNs.

- One of the most popular adversarial attack is Fast Gradient Sign Method (FGSM):

$$\tilde{\mathbf{x}} \simeq \mathbf{x} + \epsilon \, \mathrm{sgn}\left(\langle \nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{w}) \rangle_{p(\mathbf{w}|D)}\right) \simeq \mathbf{x} + \epsilon \, \mathrm{sgn}\left(\sum_{i=1}^{n} \nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{w}_i)\right)$$

Adversarial Robustness of BNN

- Key Theorem:

Theorem 1. *Let $f(\mathbf{x}, \mathbf{w})$ be a fully trained overparametrized BNN on a prediction problem with data manifold $\mathcal{M}_D \subset \mathbb{R}^d$ and posterior weight distribution $p(\mathbf{w}|D)$. Assuming $\mathcal{M}_D \in \mathcal{C}^\infty$ almost everywhere, in the large data limit we have a.e. on $\mathcal{M}_D$*

$$\left(\langle \nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{w}) \rangle_{p(\mathbf{w}|D)}\right) = \mathbf{0}. \tag{3}$$

Adversarial Robustness of BNN

- The Key Theorem:

Theorem 1. *Let $f(\mathbf{x}, \mathbf{w})$ be a fully trained overparametrized BNN on a prediction problem with data manifold $\mathcal{M}_D \subset \mathbb{R}^d$ and posterior weight distribution $p(\mathbf{w}|D)$. Assuming $\mathcal{M}_D \in \mathcal{C}^\infty$ almost everywhere, in the large data limit we have a.e. on $\mathcal{M}_D$*

$$\left(\langle \nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{w}) \rangle_{p(\mathbf{w}|D)}\right) = \mathbf{0}. \tag{3}$$

- Any gradient-based attack will be ineffective against a BNN in the limit.
- Necessary premise:
  - Fully trained BNN, i.e., it has enough expressive power to fit any function
  - large data limit, i.e., the training data are enough to represent the data manifold

Proof

- Lemma 1

Lemma 1. *Let* $f(\mathbf{x}, \mathbf{w})$ *be a fully trained overparametrized NN on a prediction problem with a.e. smooth data manifold* $\mathcal{M}_D \subset \mathbb{R}^d$*. Let* $\mathbf{x}^* \in \mathcal{M}_D$ *s.t.* $B_d(\mathbf{x}^*, \epsilon) \subset \mathcal{M}_D$*, with* $B_d(\mathbf{x}^*, \epsilon)$ *the d-dimensional ball centred at* $\mathbf{x}^*$ *of radius* ϵ *for some* $\epsilon > 0$*. Then* $f(\mathbf{x}, \mathbf{w})$ *is robust to gradient-based attacks at* $\mathbf{x}^*$ *of strength* $\leq \epsilon$ *(i.e. restricted in* $B_d(\mathbf{x}^*, \epsilon)$*).*

Proof

- Lemma 1

Lemma 1. *Let $f(\mathbf{x}, \mathbf{w})$ be a fully trained overparametrized NN on a prediction problem with a.e. smooth data manifold $\mathcal{M}_D \subset \mathbb{R}^d$. Let $\mathbf{x}^* \in \mathcal{M}_D$ s.t. $B_d(\mathbf{x}^*, \epsilon) \subset \mathcal{M}_D$, with $B_d(\mathbf{x}^*, \epsilon)$ the d-dimensional ball centred at $\mathbf{x}^*$ of radius ϵ for some $\epsilon > 0$. Then $f(\mathbf{x}, \mathbf{w})$ is robust to gradient-based attacks at $\mathbf{x}^*$ of strength $\leq \epsilon$ (i.e. restricted in $B_d(\mathbf{x}^*, \epsilon)$).*

- The key observation for proving Lemma 1 is:

- Over-parametrised NNs provably achieve zero loss on the whole data manifold, hence the function f would be locally constant at x^*.

Proof

- Corollary 1

Corollary 1. *Let $f(\mathbf{x}, \mathbf{w})$ be a fully trained overparametrized NN on a prediction problem with data manifold $\mathcal{M}_D \subset \mathbb{R}^d$ smooth a.e. (where the measure is given by the data distribution $p(D)$). If f is vulnerable to gradient-based attacks at $x^* \in \mathcal{M}_D$ in the infinite data limit, then a.s.* $\dim(\mathcal{M}_D) < d$ *in a neighbourhood of x^*.*

- It had been already empirically noticed that adversarial perturbations often arise in directions which are normal to the data manifold.

- A consequence of Corollary 1 is:

$$\nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{w}) = \nabla_{\perp \mathbf{x}} L(\mathbf{x}, \mathbf{w})$$

Proof

- Recall: we want to prove $\nabla_{\perp \mathbf{x}} \langle L(\mathbf{x}, \mathbf{w}) \rangle_{p(\mathbf{w}|D)} = 0.$

- We only need to prove the following symmetry:

Lemma 2. *Let $f(\mathbf{x}, \mathbf{w})$ be a fully trained overparametrized NN on a prediction problem on data manifold $\mathcal{M}_D \subset \mathbb{R}^d$ a.e. smooth. Let $\hat{\mathbf{x}} \in \mathcal{M}_D$ to be attacked and let the normal gradient at $\hat{\mathbf{x}}$ be $\mathbf{v}_{\mathbf{w}}(\hat{\mathbf{x}}) = \nabla_{\perp \hat{\mathbf{x}}} L(\mathbf{x}, \mathbf{w})$ be different from zero. Then, in the infinite data limit and for almost all $\hat{\mathbf{x}}$, there exists a set of weights $\mathbf{w}'$ such that*

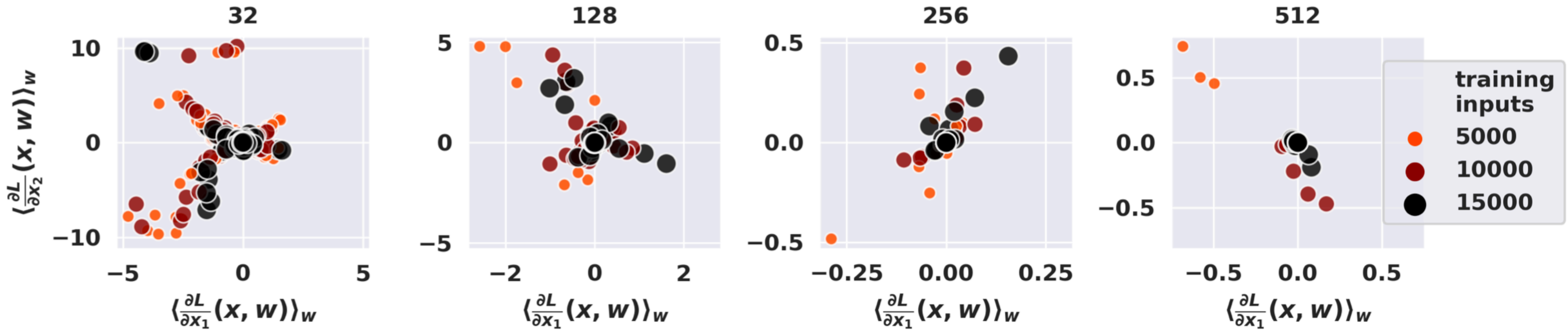
$$f(\mathbf{x}, \mathbf{w}') = f(\mathbf{x}, \mathbf{w}) \text{ a.e. in } \mathcal{M}_D, \tag{4}$$

$$\nabla_{\perp \hat{\mathbf{x}}} L(\hat{\mathbf{x}}, \mathbf{w}') = -\mathbf{v}_{\mathbf{w}}(\hat{\mathbf{x}}). \tag{5}$$

- The proof of this lemma rests on constructing a function satisfying (4) and (5).

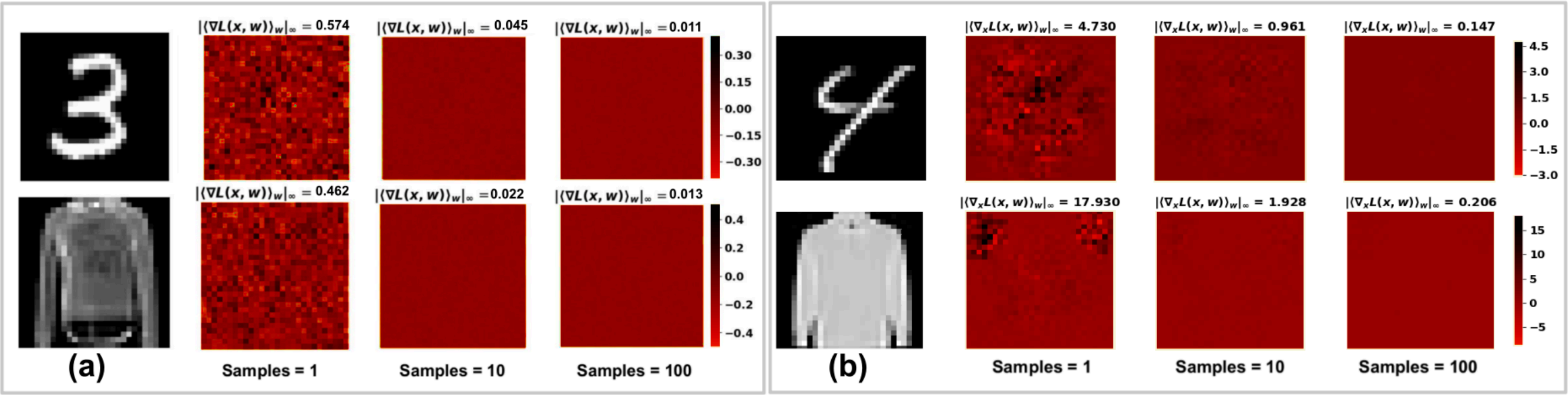
Experiments

- The magnitude of the expectation of the gradient shrinks as we increase the network's parameters and the number of training inputs.



Experiments

- The expected loss gradients of BNNs vanish when increasing the number of samples.



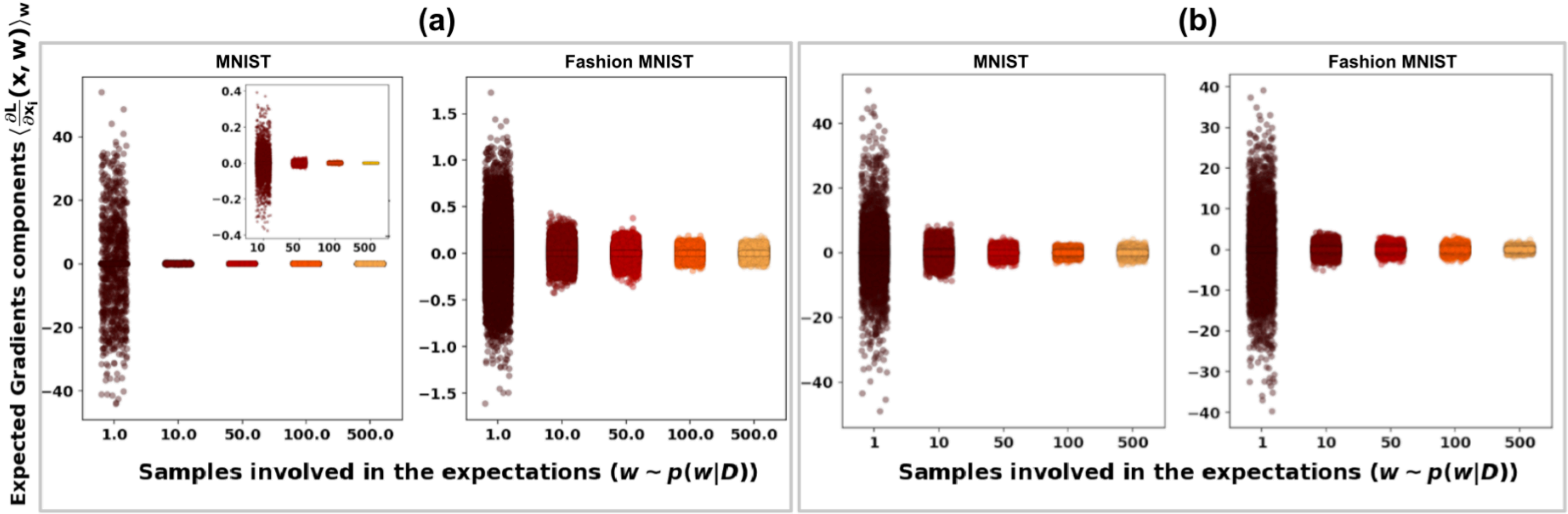
Experiments

- The random attack outperforms the gradient-based attacks.
- The vanishing behaviour of the gradient makes FGSM and PGD attacks ineffective.

Dataset/Method	**Rand**	**FGSM**	**PGD**
MNIST/HMC	**0.850**	0.960	0.970
MNIST/VI	0.956	**0.936**	0.938
Fashion/HMC	**0.812**	0.848	0.826
Fashion/VI	**0.744**	0.834	0.916

Experiments

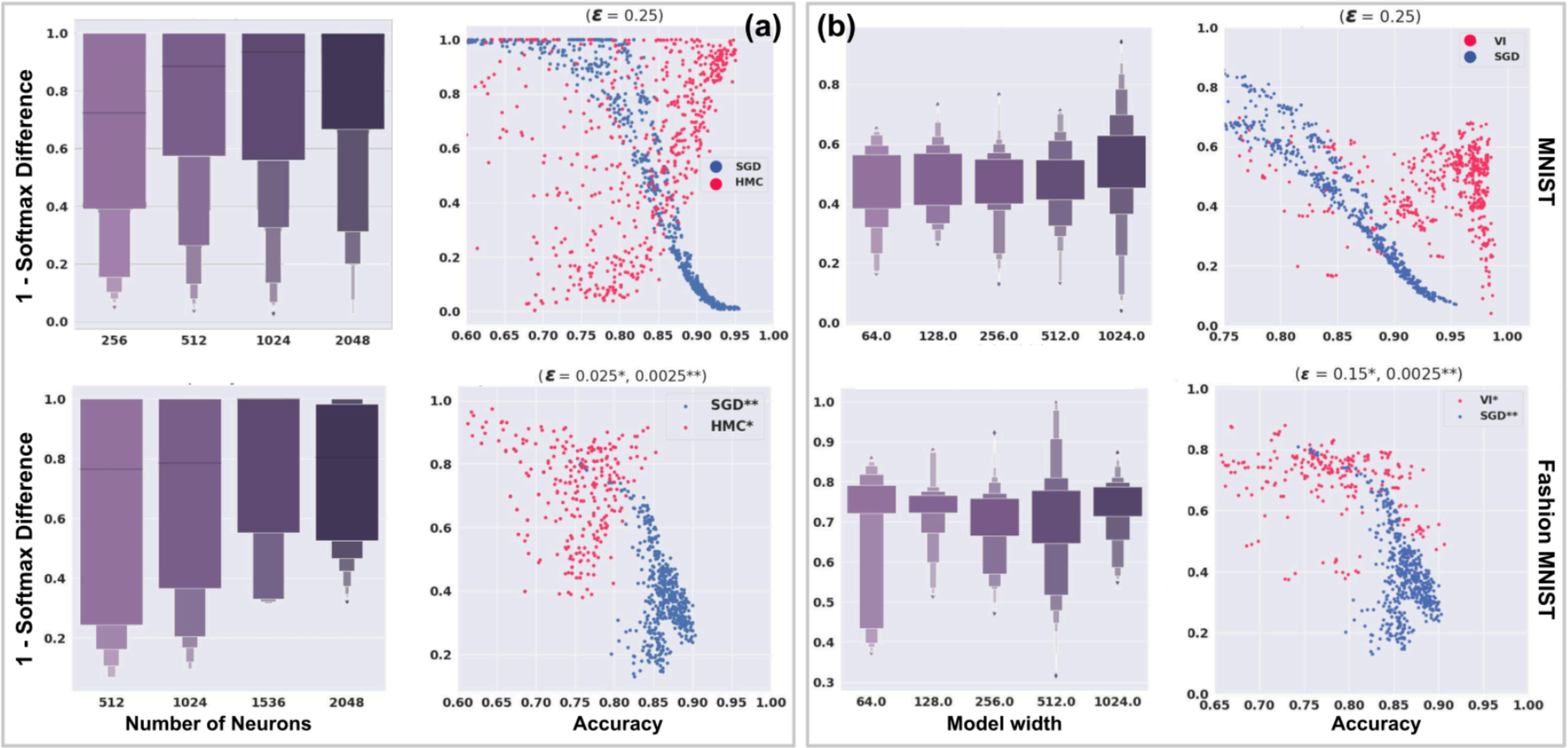
- The random attack outperforms the gradient-based attacks.
- The vanishing behaviour of the gradient makes FGSM and PGD attacks ineffective.



Experiments

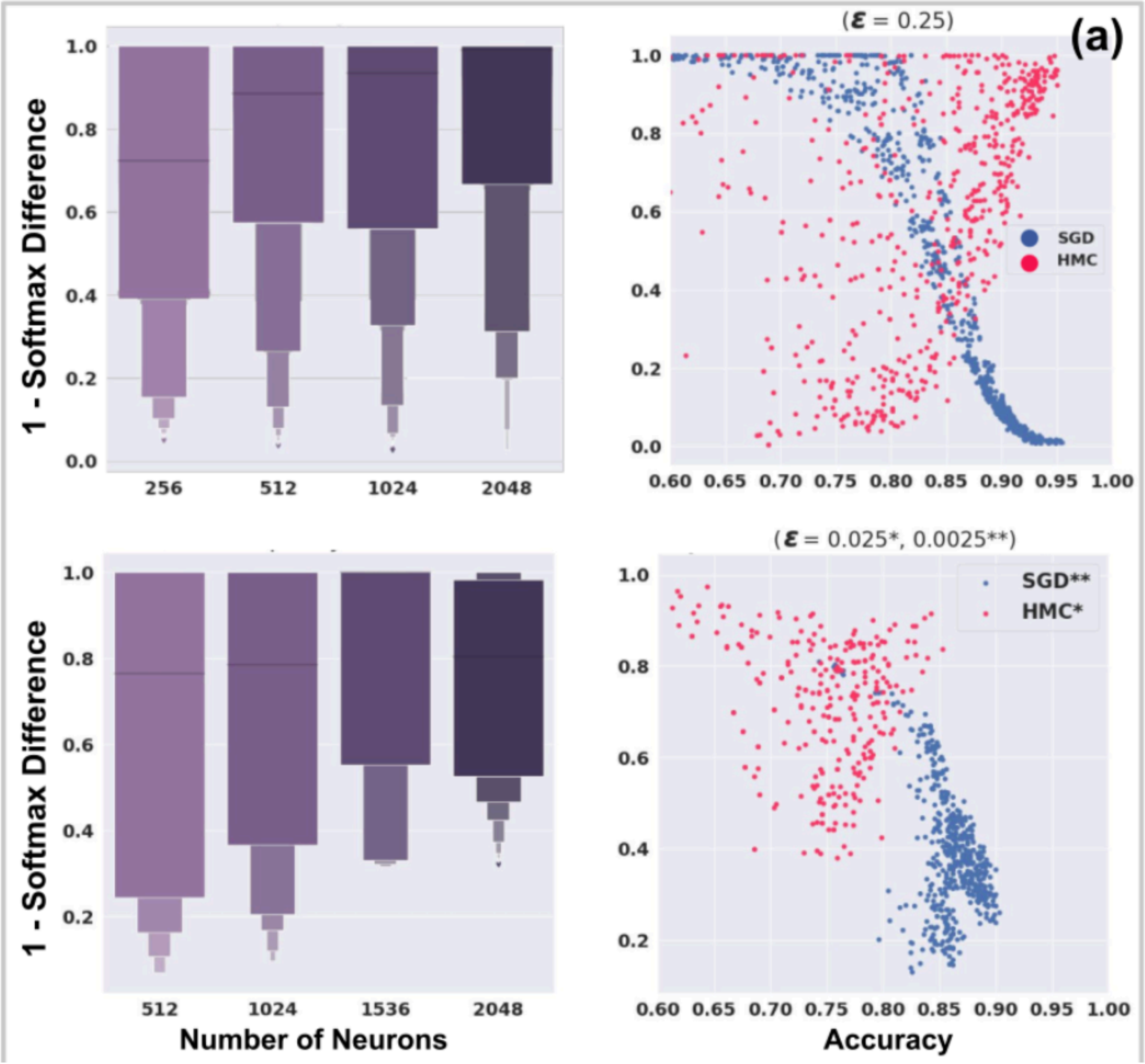
Robustness Accuracy Analysis

- 1000 different NNs and BNNs are tested in this experiment.
- Metric: 1 - the average difference in the softmax prediction.
- The larger it is, the more robust the model is.



Experiments

Robustness Accuracy Analysis

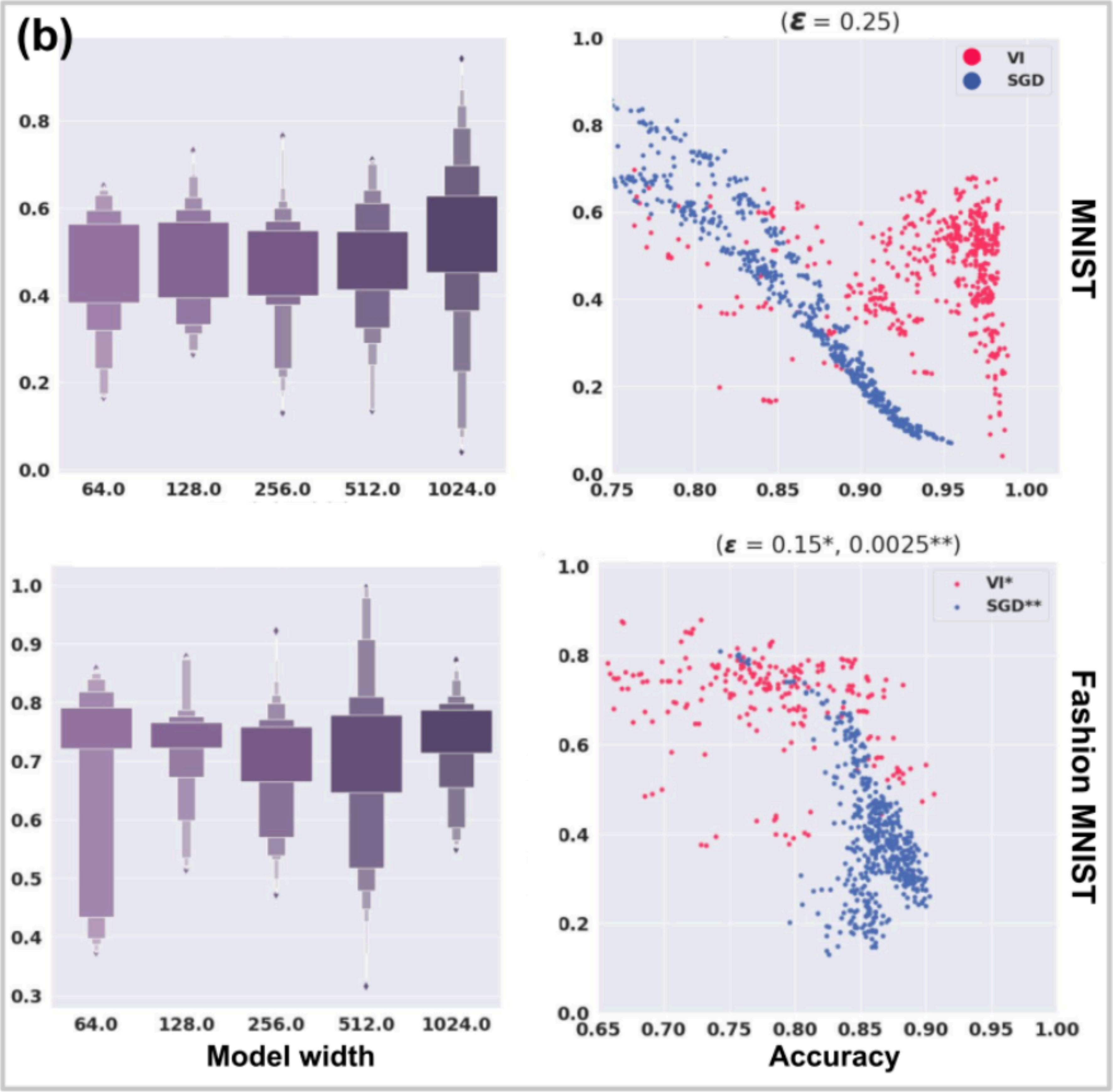


- With the increase of model size and accuracy, the robustness of BNNs increase.
- This trend is fully reversed for normal NNs trained with SGD.

Experiments

Robustness Accuracy Analysis

- This trend is less obvious on BNNs trained with VI.



Conclusion

- This paper shows that BNNs can evade a broad class of adversarial attacks.
- It also has some limitations:
  - Performing Bayesian inference in large non-linear models is extremely challenging.
  - Theoretical results hold in a thermodynamic limit which is never realized in practice.
  - We have focused on two attack strategies which directly utilize gradients.