
The Limitations of Model Uncertainty in Adversarial Settings

Kathrin Grosse
CISPA Helmholtz Center
kathrin.grosse@cispa.saarland

David Pfaff
CISPA Helmholtz Center
david.pfaff@cispa.saarland

Michael T. Smith
Sheffield University
m.t.smith@sheffield.ac.uk

Michael Backes
CISPA Helmholtz Center
backes@cispa.saarland

Abstract

Machine learning models are vulnerable to adversarial examples: minor perturbations to input samples intended to deliberately cause misclassification. While an obvious security threat, adversarial examples yield as well insights about the applied model itself. We investigate adversarial examples in the context of Bayesian neural network's (BNN's) uncertainty measures. As these measures are highly non-smooth, we use a smooth Gaussian process classifier (GPC) as substitute. We show that both confidence and uncertainty can be unsuspecting even if the output is wrong. Intriguingly, we find subtle differences in the features influencing uncertainty and confidence for most tasks.

1 Introduction

Machine learning classifiers are used for various purposes in a variety of research and industry applications. However, these classifiers have been shown to be vulnerable to a number of different attacks [Lowd and Meek, 2005, Mei and Zhu, 2015, Goodfellow et al., 2015, Biggio and Roli, 2018]. *Adversarial examples*, or evasion attacks, present a direct threat to classification at test-time. The classifier outputs a wrong label given an input sample with original utility, however slightly modified by the attacker. In the area of computer vision, adversarial examples are often visually indistinguishable images which are misclassified by state-of-the-art models [Goodfellow et al., 2015, Moosavi-Dezfooli et al., 2016].

Adversarial examples and Bayesian uncertainty have been investigated before. Bekasov and Murray [since 2018] show the importance of priors in robustness. Gal and Smith [2018] propose an attack to sample garbage examples in the pockets of the uncertainty of BNN. BNN are further investigated by Louizos and Welling [2017], Rawat et al. [2017] and Liu et al. [2019]. All these authors test simple fast-gradient-sign adversarial examples on (variants of) BNN and find notable differences in model uncertainty for adversarial examples. Further, Li and Gal [2017] observe differences for high confidence adversarial examples. We show that adversarial examples exist that show no deviation for uncertainty measures, are visually similar to the original, and yet are misclassified by BNNs.

2 Experimental Setup and Results

We extend the notion of adversarial examples from a classifiers' output to Bayesian confidence and uncertainty. Here, we assume a binary classifier. Confidence is a real value between 0 (not at all confident) and 1 (very confident). Further, uncertainty is a real larger than 0, where larger

Table 1: Average size of perturbation and uncertainty change of HC/HCLU to benign data.

	Spam	FMNIST19	FMNIST57	MNIST19	MNIST38
$\ \delta_{HC}\ _2$	0.006±0.01	0.194±0.036	0.019±0.012	0.053±0.014	0.029±0.011
$\ \delta_{HCLU}\ _2$	0.008±0.006	0.194±0.036	0.019±0.013	0.053±0.014	0.03±0.012
diff. unc.	2.35	0.0	0.89	0.31	0.74

means more uncertain. We formalize the resulting optimization problem of high-confidence-low-uncertainty (HCLU) examples:

$$\begin{aligned} \min_{\delta} \quad & \|\delta\|_2 \\ \text{s.t.} \quad & \text{confidence}(f(x + \delta)) > 0.95, \\ \text{and} \quad & \text{uncertainty}(f(x + \delta)) \leq \text{uncertainty}(f(x)), \end{aligned}$$

where we minimize the perturbation δ using the L_2 norm. An extension to other norms (as by Carlini and Wagner [2017]), or a restriction to change only a subset of the features, is straight forward. In general, we minimize the change between benign sample x and its corresponding adversarial example $x + \delta$. The first constraint maximize the confidence of classifier f and yields misclassification (e.g., the initial confidence was small). The second constraint limits the uncertainty.

Yet, as Smith and Gal [2018] show in their work, the uncertainty estimates of BNNs are highly non-smooth, and the above problem potentially hard to solve. We thus decide to use a GPC as a substitute, since its decision surface is smoother.

Setup. We use spam [Lichman, 2013] data and two sub-tasks of each MNIST [LeCun et al., 1998] and Fashion MNIST [Xiao et al., 2017]. On Fashion MNIST, we train on trousers vs. ankle boots (one vs. nine) and sandals vs. sneakers (five vs. seven). On MNIST, we train on the one vs. nine and three vs. eight sub-tasks. We choose small tasks to fit scalability of the GPC substitute, which is further optimized on only 750 samples. Our experiments are implemented in Python. For deep neural networks (DNN) and BNN, we use Tensorflow [Abadi et al., 2016], and GPy [GPy, 2012] for GPC. The attack on GPC is implemented using the optimization routines of SciPy [Jones et al., 2001] and L-BFGS-B [Zhu et al., 1997]. Our attack is publicly available in the adversarial robustness toolbox [Nicolae et al., 2018], which we also used for Carlini and Wagner [2017]’s L_2 attack.

Properties of HCLU examples. We show the HCLU adversarial examples with the smallest δ in the fifth row of Figure 1. These examples are still adversarial: we see in the figure that almost all are visually similar to their benign origin. To verify for all depicted examples that they were altered in the crafting process, we plot the original sample beneath the examples. The success rates on GPC are 100%, where however often the specified confidence of 0.95 is barely not met, and the resulting confidence is around 0.948. Table 1 shows the statistics of the adversarial perturbations per feature measures using the L_2 -norm. We observe very small changes on spam, large changes on Fashion MNIST19 and descent changes for all other datasets.

We can also craft examples that only maximize confidence by removing the second constraint. Surprisingly, the perturbation δ_{hc} is barely different from the original examples, as visible in Table 1: only on the spam data and two MNIST tasks, a difference is observable. The pictures in the fourth row of Figure 1 reveal that albeit looking very similar, the examples are actually different. Some features are generally changed, whereas others seem only correlated with uncertainty. We conclude that slightly different features are learned for confidence and uncertainty, respectively. Consequently, for all datasets except Fashion MNIST ankle boot vs trousers, the observed uncertainty is indeed lower when targeted. Concerning the unchanged Fashion MNIST task, we observe that the change in uncertainty for those is examples is only 1% from the original value. In this case, confidence and uncertainty seem to rely on the same features.

Transferability of misclassification. We investigate how many HCLU are misclassified on a GPC (trained on a different subset of the data), a DNN, and the BNN. The accuracies are plotted in Figure 2. Dark colors denote benign accuracy, lighter colors the accuracy on HCLU.

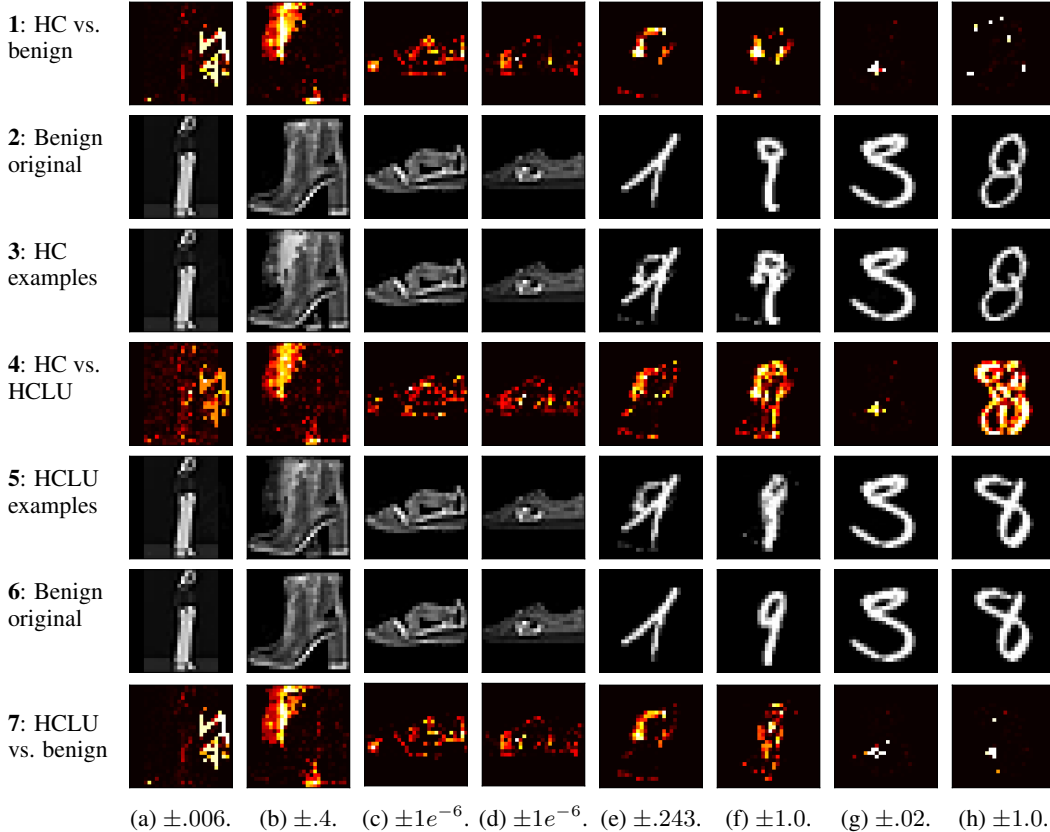


Figure 1: Comparison of HCLU, HC, and benign originals. Grey are examples and samples, red plots (row one, four and seven) show differences between images. HC vs. HCLU differences are scaled according to column label. For Fashion MNIST, these scales are also used for the comparisons in row one and seven. We plot all differences using a logarithmic spectrum that allows to see small changes. In the spectrum, colors go from black over red to yellow to white (strongest change). Benign originals are the samples started with to craft the upper HCLU or lower HC example. Figures with differences are best seen in color.

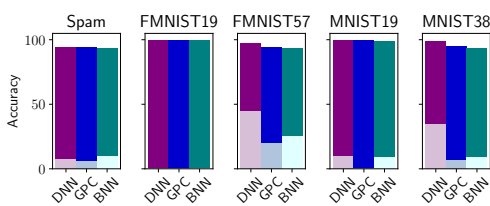


Figure 2: Transferability of optimized GPC adversarial examples. Dark color denotes accuracy on benign data, lighter color on HCLU examples.

Bayesian decision boundaries, Papernot et al. [2016] showed that adversarial examples often fool several models. In this experiment, we are interested whether behavior *differs* between benign and adversarial data. We chose Carlini and Wagner’s L_2 attack as a baseline: L_2 , as in our case, allows the best optimization. We further configure the attack as to increase transferability of the examples (corresponding to higher “confidence” on the target DNN)¹.

¹We set $\kappa = 0.7$. The attack definition, in a nutshell, defines $\kappa > 0$ to encourage the solver to find a confidently classified example. For details see Carlini and Wagner [2017].

The accuracy on these L_2 adversarial examples on DNN, GPC, and BNN is higher than on HCLU: the average accuracy is between 50% and in some cases higher than 90%. Yet, the accuracy under the transferred attack is always lower than the accuracy observed using clean test data.

We depict the results concerning Bayesian confidence and uncertainty in Figure 3, where we distinguish correctly classified (gray shades) and wrongly classified (red shades) benign and adversarial data. We measure the mean (confidence, left plots) and variance (uncertainty, right plots) of the sampled posteriors and bin them using 25 bins between 0.0 and 1.0. As large fractions of the histograms are empty, we plot only the relevant parts. To outline overall trends, we plot the normalized bins of correct and wrongly classified data stacked on top of each other.

In general, the BNN is more confident on benign data/Carlini and Wagner examples that are correctly classified. This observation holds across all data sets. For HCLU, this trend is reversed: the BNN is confident on many misclassified examples. Intriguingly, the BNN outputs low confidence on some HCLU examples which are not misclassified, or correctly assigned to their original class. Analogously, the uncertainty measures are similar between benign data and Carlini and Wagner’s attack. Uncertainty is generally low for correctly classified data and high for wrongly classified data. This observations are again reversed for HCLU examples: here uncertainty is often low if an example is wrongly classified.

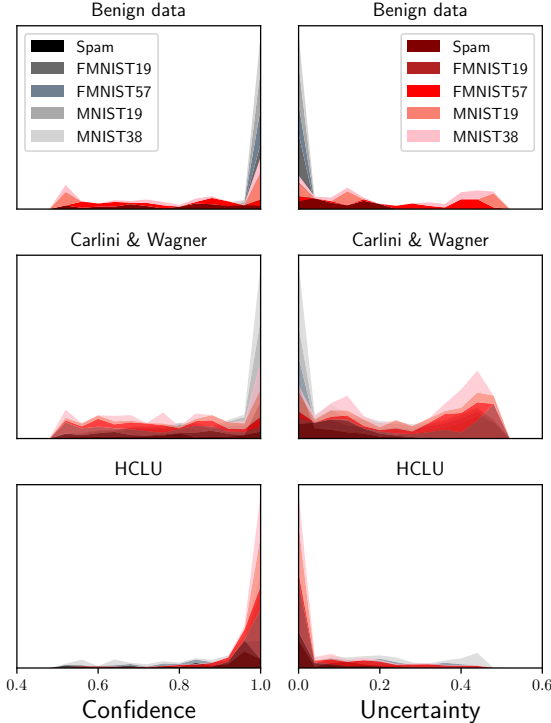


Figure 3: Transferability of HCLU examples (bottom) to Bayesian Neural Networks. We consider Carlini & Wagner’s L_2 attack as a comparison (middle). Benign data is also depicted as a baseline (top). Correctly classified data is plotted in gray shades, misclassified data in red shades. Figure is best seen in color.

3 Conclusion

In this paper, we studied the vulnerability of machine learning models providing Bayesian model uncertainty. We introduced a technique to craft HCLU adversarial examples, which achieve both high confidence and low uncertainty on a Gaussian process classifier. Also Bayesian neural networks misclassify HCLU. We further found that HCLU adversarial examples are misclassified with low uncertainty and high confidence, in contrast to high-confidence adversarial examples which are derived from non-Bayesian models such as DNN.

Our work has several implications. Firstly, different Bayesian models (GPC, BNN) learn similar confidence and uncertainty features, as our transferability study shows. This implies that such measures cannot be used as a defense. Our results show as well that there is a subtle difference in the features learned for uncertainty and confidence, respectively.

Acknowledgments

This work was supported by the German Federal Ministry of Education and Research (BMBF) through funding for the Center for IT-Security, Privacy and Accountability (CISPA) (FKZ: 16KIS0753). This work has further been supported by the Engineering and Physical Research Council (EPSRC) Research Project EP/N014162/1.

References

- Daniel Lowd and Christopher Meek. Good word attacks on statistical spam filters. In *CEAS 2005 - Second Conference on Email and Anti-Spam, July 21-22, 2005, Stanford University, California, USA*, 2005. URL <http://www.ceas.cc/papers-2005/125.pdf>.
- Shike Mei and Xiaojin Zhu. Using machine teaching to identify optimal training-set attacks on machine learners. In *AAAI*, pages 2871–2877, 2015.
- Ian J Goodfellow et al. Explaining and harnessing adversarial examples. In *Proceedings of the 2015 International Conference on Learning Representations*, 2015.
- Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.
- Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- Artur Bekasov and Iain Murray. Bayesian adversarial spheres: Bayesian inference and adversarial examples in a noiseless setting. *Bayesian Deep Learning at NeurIPS 2018*, since 2018.
- Yarin Gal and Lewis Smith. Idealised bayesian neural networks cannot have adversarial examples: Theoretical and empirical study. *CoRR*, abs/1806.00667, 2018. URL <http://arxiv.org/abs/1806.00667>.
- Christos Louizos and Max Welling. Multiplicative normalizing flows for variational bayesian neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 2218–2227, 2017.
- A. Rawat, M. Wistuba, and M.-I. Nicolae. Adversarial Phenomenon in the Eyes of Bayesian Deep Learning. *ArXiv e-prints*, November 2017.
- Xuanqing Liu, Yao Li, Chongruo Wu, and Cho-Jui Hsieh. Adv-BNN: Improved adversarial defense through robust bayesian neural network. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=rk4Qso0cKm>.
- Yingzhen Li and Yarin Gal. Dropout inference in bayesian neural networks with alpha-divergences. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 2052–2061, 2017.
- Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017.
- Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. pages 560–569, 2018.
- M. Lichman. UCI machine learning repository, 2013. URL <http://archive.ics.uci.edu/ml>.
- Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.
- Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.
- Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.
- GPy. GPy: A gaussian process framework in python. <http://github.com/SheffieldML/GPy>, 2012.
- Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001. URL <http://www.scipy.org/>.

Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.

Maria-Irina Nicolae, Mathieu Sinn, Tran Ngoc Minh, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Ian M. Molloy, and Benjamin Edwards. Adversarial robustness toolbox v0.2.2. *CoRR*, abs/1807.01069, 2018. URL <http://arxiv.org/abs/1807.01069>.

Nicolas Papernot, Patrick McDaniel, and Ian J. Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *CoRR*, abs/1605.07277, 2016. URL <http://arxiv.org/abs/1605.07277>.