Robustness Verification Problem for Tree Ensembles and Certified Robustness

A Review of Modern Literature

Isaac Bergl (22710992)

Introduction

In recent years, an important area of interest for researchers has been Machine Learning (ML), which have proven to be vastly superior to prior methods for data mining, decision making in unpredictable environments and classification tasks (Stoica et al., 2017). As such, it is crucial to have the safety and robustness of these systems formally guaranteed, for example for the safety of systems such as self-driving cars (Burton et al., 2017). One such requirement is defence against adversarial examples which are small, often undetectable to humans, changes to input data that causes misclassification in ML systems. This occurs in ML classifiers such as Deep Neural Networks (DNN) (Szegedy et al., 2013), and tree ensembles (James et al., 2013). It has been recognised in current literature that the robustness properties of tree ensemble methods are yet to be fully understood and explored in research (H. Chen et al., 2019). This paper aims to review modern literature on primarily tree ensemble classifiers regarding their robustness including current frameworks of robustness, the robustness verification problem as a formal guarantee of robustness and current methods of improving the robustness of these models.

Overview of Robustness Verification

Most of the current literature regarding adversarial examples is in the context of DNNs, however the basic idea of an adversarial example in ML is fundamental across all ML classifiers. In a paper by Dreossi et al. (2019), it is noted that in the current literature any definition of *robustness* is specific to the paper itself and attempts to formalise the term generally enough to encompass many different definitions of the term. Future papers would go on to adopt this framework (Gross et al., 2020).

Furthermore, Tsipras et al. (2018) suggests in their paper that there is a fundamental relationship between robustness and accuracy and that in order to ensure the robustness of a classifier a degree of accuracy for non-adversarial examples must be sacrificed.

Taking a game-theory perspective, with one player being a robust classifier and the other being an adversarial attacker, the paper by Pinot et al. (2020) concludes that no classifier can have a guaranteed robustness against all attacks, by finding a Nash-equilibrium in such a game. The paper discusses the importance of randomisation and tests randomized classifiers against adversarial trained systems, trained using state-of-the-art attack and finds that their method prevails.

How exactly input data is perturbed can be measured by an norm, and in this paper by Wang et al. (2020), it is noted that all other papers prior had only considered the L-infinity norm. The paper investigates how perturbation on different features are related and studies the robustness verification problem for tree ensembles with respect to a general L-n norm. It finds that for n=1 and infinity the problem is polynomial and builds upon past verifying algorithms by generalising to an L-n norm.

In the paper by Goodfellow et al. (2014) the idea of indistinguishability as an small amount of noise added to each pixel, an L-infinity bounded norm, is proposed. Wong et al. (2019) offers an alternative model of how to define a perturbational change. The paper uses the Wasserstein distance, being how much of the subjects of an image are distorted and created an algorithm to attack classifiers using this idea. This different interpretation of perturbation proved to be more successful for image recognition than previous norms, and the paper notes that research into metrics that reflect our human intuition of variance is incomplete yet potentially valuable to the field.

Robustness Verification for Tree Ensembles

Tree based models are becoming increasing powerful for many ML challenges, including classification problems (Ke et al., 2017). Many novel tree-based methods have been proposed over the last few years that are highly competitive with other ML techniques such as DNNs (T. Chen & Guestrin, 2016; Zhou & Feng, 2017; Zhu et al., 2019)**.**

However, these tree-based models are prone to adversarial attacks (Kantchelian et al., 2016). Many new attack methods have been created in recent years, for example in a paper by Cheng et al. (2018), where one such attack method is presented by formulating the attack as a real-valued optimisation problem and solved via an optimisation algorithm, and noted this algorithm is effective against Gradient Boosted Decision Trees (GBDT). In the paper by H. Chen et al. (2019) the weakness in tree-based models is shown by proposing an attack algorithm, and in a paper by Singh (2020) a black-box attack approach (where the inner workings of the targeted system are unknown) was used to produce adversarial examples for an arbitrary classifier. Brendel et al. (2017) presents an attack method, called the *Boundary Attack*, that takes a large initial perturbation and then gradually works back towards the original input until the change is unnoticeable. While these methods can give empirical evidence of a system's robustness, they are not methods for formal verification which requires a tight lower bound on the minimal adversarial perturbation (H. Chen et al., 2019).

The vulnerability of tree ensemble methods is important as it presents concerns for the safety and robustness of such systems as self-driving cars. It is thus important to discover formal methods for verifying robustness. Törnblom & Nadjm-Tehrani (2020) provides a method of verifying robustness of tree-based classifiers by ensuring their input-output mapping comply with some arbitrary requirement in their paper. However, this paper does not solve the formal robustness verification problem, as the minimal adversarial example is still not found.

In the paper by Kantchelian et al. (2016), it was shown that both Random Forest (RF) and are vulnerable to adversarial attacks. In the paper a Mixed-Integer Linear Programming (MLIP) method was proposed to find a minimal adversarial perturbation for a tree ensemble and showed it to be NP-complete. However this approach lacks scalability to large ensembles, as noted in a paper by Andriushchenko & Hein (2019) where the exact min-max robust loss for Boosted Decision Stumps, and for Boosted Trees the upper-bound for the robust loss.

Another approach is to use Satisfiability Modulo Theory (SMT) methods to solve the problem, as shown in the paper by Einziger et al. (2019) in the context of GBDT. This method however would later be shown to be slow compared to superior algorithms.

In this paper by H. Chen et al. (2019), previous work done in Robustness Verification was expanded upon using a different approach, noting that the existing MLIP and SMT approaches would be impractical for large tree ensembles, taking exponential time. First, they devised an algorithm for verifying a single decision tree in linear time, and for tree ensembles cast the problem as a max-clique searching problem in K-partite graphs. Using this concept, a multi-level verification algorithm was proposed and was reported to be hundreds of times faster than prior MLIP methods for ensembles in the size of the 100s. Similarly Hein & Andriushchenko (2017) devises a method for finding instance-specific lower bounds for the minimal adversarial example, however this method applies to kernel and Neural Network systems, not necessarily tree ensembles.

It is also important to mention the current work being done on DNNs. A MLIP approach to finding the minimal adversarial example as an optimisation problem was proposed in a paper by Tjeng et al. (2017), and while it's method was orders of magnitude faster than previous methods it proved too computationally expensive for large networks (Croce & Hein, 2020). In their paper Katz, Barrett et al. (2017) provides another method for verifying general neural networks using the simplex method. Alternatively, in the paper by Madry et al. (2017) proposes a Projected Gradient Descent method to find the minimal adversarial example. In the paper by Croce & Hein (2020) this was improved to be more time and space efficient, by reducing the number of resets the algorithm would have to do, as well as more adaptable to other classifiers. The minimal adversarial example was guaranteed with respect to the L-1, L-2, and L-infinity norms. It claims to be more quick, scalable and more effective that previous methods, and additionally is robust to gradient masking.

Improving Robustness

Many methods have been proposed in order to make these systems more robust. One approach to verifying a tree ensemble based classifier was presented in a paper by Sato et al. (2019) where, as they termed it, *violating inputs* are detected within a range. Using a separate process, any input within one of these ranges is filtered out before ever being passed into the classifier. The paper suggests using some separate software to process these inputs separately.

In the paper by Abbasi & Gagné (2017) an ensemble of specialists was used to detect adversarial examples, or as they call it *fooling instances*. The confusion matrix was used to define the speciality. The paper empirically confirmed this hypothesis and concluded it would indeed make a system more robust.

However, detection methods will not ever be perfect. In this paper by Carlini & Wagner (2017), ten methods of adversarial detection were bypassed using more sophisticated attacks. The paper gives credence to the idea that these detection methods are not complete and much like the zero sum game model previously discussed in the paper by Pinot et al. (2020) there will always be a new way to fool these systems.

In two papers by Ranzato & Zanella (2019, 2020a) a tool called *silva* is produced, which formally verifies tree ensembles such as GBDT and RF according to a set of pre-defined perturbations. The researchers noted that abstract interpretation can be utilised to formally verify DNNs and applied the same principles to tree ensemble classifiers. This tool was used in many future works by the authors including in a paper where a training method was proposed that makes tree ensembles more accurate, and more relevantly, more robust (Ranzato & Zanella, 2020b). The approach taken in the paper uses a genetic algorithm to maximise the robustness according to the framework proposed in their previous papers (Ranzato & Zanella, 2020a), and used *silva*.

However, in a paper by Gross et al. (2020) work done in the paper by Ranzato & Zanella (2019) was considered to be incomplete due to the fact the work is based on an abstract interpretation. Instead they opt to use the general framework for robustness proposed in the paper by Dreossi et al. (2019) and produce a method for formal verification of tree ensemble classifiers such as GBDT and RF against *randomised* attacks using MILP and SMT methods.

There are many proposed methods of increasing the robustness of ML classifiers. In a paper by Calzavara et al. (2020) an algorithm for constructing decision trees according to some threat model was designed. At every step of the decision tree's construction, a loss-function is minimised such that the tree is built to be both accurate and robust to specified perturbations.

A method of increasing the robustness of any ML classifier is to use some separate algorithm to pre-process any input data. For example, this paper by Salman et al. (2020) randomised smoothing is used to remove any small (and potentially adversarial) inputs before being parsed to the classifier, and shows it to be robust with respect to the L-p norm. The paper shows the effectiveness of such an approach by applying it to image classification API's from *Google*, *Azure*, *AWS* and *ClarifAI*. Similarly Cohen et al. (2019) applies randomised smoothing using Gaussian noise to input data and a tight robustness is guaranteed for the L-2 norm, showing similar results.

In the paper by Lecuyer et al. (2019), certified defence against adversarial examples was proposed that claims to be both scalable to large datasets and can be applied to arbitrary ML models. It provides a norm bounded robustness guarantee utilising the idea from cryptography – differential privacy, and they present their defence called *PixelDP*.

Conclusion

In this report, I have presented the current literature on ML model verification, particularly that of tree ensemble methods. As ML systems continue to replace prior systems in the real world the safety of such systems as tree ensembles should be better understood in research. Good work has been done in formalising frameworks and producing verification methods, but it is yet to be as well understood as DNNs and since the studies on verification of tree ensemble methods are relatively rare, while the tree ensemble methods have been paramount, more research work needs to be conducted in this area.

BIBLIOGRAPHY

Abbasi, M., & Gagné, C. (2017). Robustness to adversarial examples through an ensemble of specialists. *arXiv preprint arXiv:1702.06856*.

Andriushchenko, M., & Hein, M. (2019). *Provably robust boosted decision stumps and trees against adversarial attacks.* Paper presented at the Advances in neural information processing systems.

Brendel, W., Rauber, J., & Bethge, M. (2017). Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*.

Burton, S., Gauerhof, L., & Heinzemann, C. (2017). *Making the case for safety of machine learning in highly automated driving.* Paper presented at the International Conference on Computer Safety, Reliability, and Security.

Calzavara, S., Lucchese, C., Tolomei, G., Abebe, S. A., & Orlando, S. (2020). Treant: training evasion-aware decision trees. *Data Mining and Knowledge Discovery, 34*(5), 1390-1420.

Carlini, N., & Wagner, D. (2017). *Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods*. Paper presented at the Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, Dallas, Texas, USA. https://doi.org/10.1145/3128572.3140444

Chen, H., Zhang, H., Boning, D., & Hsieh, C.-J. (2019). Robust decision trees against adversarial examples. *arXiv preprint arXiv:1902.10660*.

Chen, H., Zhang, H., Si, S., Li, Y., Boning, D., & Hsieh, C.-J. (2019). *Robustness verification of tree-based models.* Paper presented at the Advances in neural information processing systems.

Chen, T., & Guestrin, C. (2016). *Xgboost: A scalable tree boosting system.* Paper presented at the Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.

Cheng, M., Le, T., Chen, P.-Y., Yi, J., Zhang, H., & Hsieh, C.-J. (2018). Query-efficient hard-label black-box attack: An optimization-based approach. *arXiv preprint arXiv:1807.04457*.

Cohen, J. M., Rosenfeld, E., & Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*.

Croce, F., & Hein, M. (2020). *Minimally distorted adversarial examples with a fast adaptive boundary attack.* Paper presented at the International Conference on Machine Learning.

Dreossi, T., Ghosh, S., Sangiovanni-Vincentelli, A., & Seshia, S. A. (2019). A formalization of robustness for deep neural networks. *arXiv preprint arXiv:1903.10033*.

Einziger, G., Goldstein, M., Sa'ar, Y., & Segall, I. (2019). *Verifying robustness of gradient boosted models.* Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Gross, D., Jansen, N., Pérez, G. A., & Raaijmakers, S. (2020). Robustness Verification for Classifier Ensembles. *arXiv preprint arXiv:2005.05587*.

Hein, M., & Andriushchenko, M. (2017). *Formal guarantees on the robustness of a classifier against adversarial manipulation.* Paper presented at the Advances in neural information processing systems.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112): Springer.

Kantchelian, A., Tygar, J. D., & Joseph, A. (2016). *Evasion and hardening of tree ensemble classifiers.* Paper presented at the International Conference on Machine Learning.

Katz, G., Barrett, C., Dill, D. L., Julian, K., & Kochenderfer, M. J. (2017). *Reluplex: An efficient SMT solver for verifying deep neural networks.* Paper presented at the International Conference on Computer Aided Verification.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., . . . Liu, T.-Y. (2017). *Lightgbm: A highly efficient gradient boosting decision tree.* Paper presented at the Advances in neural information processing systems.

Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., & Jana, S. (2019). *Certified robustness to adversarial examples with differential privacy.* Paper presented at the 2019 IEEE Symposium on Security and Privacy (SP).

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Pinot, R., Ettedgui, R., Rizk, G., Chevaleyre, Y., & Atif, J. (2020). Randomization matters. How to defend against strong adversarial attacks. *arXiv preprint arXiv:2002.11565*.

Ranzato, F., & Zanella, M. (2019). Robustness Verification of Decision Tree Ensembles. *OVERLAY@ AI* IA, 2509*, 59-64.

Ranzato, F., & Zanella, M. (2020a). *Abstract interpretation of decision tree ensemble classifiers.* Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence.

Ranzato, F., & Zanella, M. (2020b). Genetic Adversarial Training of Decision Trees. *arXiv preprint arXiv:2012.11352*.

Salman, H., Sun, M., Yang, G., Kapoor, A., & Kolter, J. Z. (2020). Denoised smoothing: A provable defense for pretrained classifiers. *Advances in neural information processing systems, 33*.

Sato, N., Kuruma, H., Nakagawa, Y., & Ogawa, H. (2019). Formal verification of decision-tree ensemble model and detection of its violating-input-value ranges. *arXiv preprint arXiv:1904.11753*.

Singh, S. (2020). *Query-Efficient Black-box Adversarial Attacks.* UCLA,

Stoica, I., Song, D., Popa, R. A., Patterson, D., Mahoney, M. W., Katz, R., . . . Gonzalez, J. (2017). A Berkeley View of Systems Challenges for AI, 2017. *arXiv preprint arXiv:1712.05855*.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. arXiv 2013. *arXiv preprint arXiv:1312.6199*.

Tjeng, V., Xiao, K., & Tedrake, R. (2017). Evaluating robustness of neural networks with mixed integer programming. *arXiv preprint arXiv:1711.07356*.

Törnblom, J., & Nadjm-Tehrani, S. (2020). Formal verification of input-output mappings of tree ensembles. *Science of Computer Programming*, 102450.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., & Madry, A. (2018). Robustness may be at odds with accuracy. arXiv. *Machine Learning*.

Wang, Y., Zhang, H., Chen, H., Boning, D., & Hsieh, C.-J. (2020). On $\ell_p $-norm Robustness of Ensemble Stumps and Trees. *arXiv preprint arXiv:2008.08755*.

Wong, E., Schmidt, F. R., & Kolter, J. Z. (2019). Wasserstein adversarial examples via projected sinkhorn iterations. *arXiv preprint arXiv:1902.07906*.

Zhou, Z.-H., & Feng, J. (2017). Deep forest. *arXiv preprint arXiv:1702.08835*.

Zhu, G., Hu, Q., Gu, R., Yuan, C., & Huang, Y. (2019). ForestLayer: Efficient training of deep forests on distributed task-parallel platforms. *Journal of Parallel and Distributed Computing, 132*, 113-126.