Trustworthiness of Deep Learning in Adversarial Settings

Ling Liu

School of Computer Science
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332
lingliu@cc.gatech.edu

Big data driven Deep learning has enjoyed a remarkable success in a broad area of applications, ranging from speech recognition, computer vision, machine translations, to software bug localization, manufacture defect detection, medical diagnosis, market analysis, and so forth. Today, deep learning based problem solving is penetrating every subfield of science and engineering beyond AI and Machine Learning. As data-driven deep learning frameworks and algorithms are increasingly being used to make decisions for people and about people, such as Amazon Go for checkout-free shopping, Uber self-driving cars, Apple Siri intelligent personal assistant, we foresee two opposite ends of the spectrum: On one hand, the use of algorithmic decision making will provide people with life-enriching experiences, convenience, and opportunities, and on the other hand, the use of algorithmic decision making will also open doors for potential misuse and abuse, including intentional and unfair biases or malicious intentions. Furthermore, with a growing number of open source deep neural network frameworks publically available, supervise or unsupervised deep learning is being widely deployed in a wide range of machine learning applications. We argue that without in-depth understanding of the ways in which these deep neural networks learn, make inference and reach decision, it consequently makes these deep learning systems easier to deceive [5,6].

This position paper makes three important and complimentary arguments: First, we argue that the *trustworthiness should be an essential and mandatory component of a deep learning system for algorithmic decision making*. This includes both the understanding and the measurement of the level of trust and/or distrust that we place on a deep learning algorithm to perform reliably and truthfully. For example, an adversarial deep learning can maliciously misclassify a healthy patient as a HIV positive patient. One can launch such a targeted attack at the model training phase [1, 2], for example, by injecting adversarial samples into the training dataset. Alternatively, one can also execute targeted adversarial attack during the model-based prediction phase [3], for instance, by maliciously altering the loss function calculation to bias towards the target class. Even though the training model is correct, but each time when the trained model is employed for prediction, one can inject certain amount of targeted perturbation to lead the misclassification of the testing sample to a wrong target class. Similarly, such targeted attacks can also be instrumented in the model training phase [4]. Thus, the development of formal metrics is essential for people to formally and quantitatively evaluate and measure the trust level of an algorithmic decision making result by examining the trustworthiness of the algorithm with respect of intentional and unintentional effects of execution, in the presence of different adversarial settings.

Second, we argue that *the trust of a deep learning algorithm should be evaluated along multiple dimensions in terms of its correctness, accountability, transparency and resilience* in anticipation of system failure or malicious manipulation. For example, the input data to the model training phase represents the input dimension of trust. Such trust can be violated by making small amount of perturbation to the input images. Input trust guard should be enforced to prevent the adversary to contribute malicious input to "poison" the system and trick the system to make glaring errors [1]. In addition, we consider the hidden layers of a multi-layer neural network as hidden dimensions of trust in algorithmic decision making. Finally, the output of the deep learning model should be trusted in a verifiable manner. We can illustrate such multi-dimensional trust in the context of privacy preserving deep learning. The input privacy of a deep learning algorithm refers to the trust that the training dataset is not compromised or altered maliciously. The computational privacy of a deep learning algorithm with respect to the hidden layers refers to the trust that the inference conducted by the hidden layers is trusted for its correctness, accountability, fairness, and transparency. For instance, the model parameters and updates should not lead to the leakage of any information about the individual input data samples that have participated in the model training. Finally, the output privacy of a deep learning algorithm refers to the trust that the inference over the output of the algorithm should not expose the identity of any individual sample in the input training dataset.

Third but not the least, we argue that *there is a need for proactive safe guard mechanisms to enforce the trustworthiness* of a deep learning framework and its algorithmic decision making. Instead of crafting reactive defense method against each known adversarial attack, we stress the importance of defining proactive trust guards for the deep learning algorithms in both model construction/training phase and model-based prediction/testing phase. Such trust guards should be able to define and enforce formal verification of the hidden layer transformations, activation and stochastic gradient decent assignments based on the quantitative evaluation of the trustworthiness of each operation in a multi-layer deep neural network.

One intuitive defense approach to adversarial attacks is to leverage deep learning ensembles with consensus. Ensemble learning can facilitate the detection of inconsistency in both the intermediate stages and the output of the deep learner during training and testing phase, including data partitioning based ensembles, data resolution variation based ensembles, model parameter variation based ensembles, such as different neuron sizes, different number of epochs, different number of weight filters, different layers of DNN, and so forth. Another complimentary approach is the distillation technique that trains a DNN using knowledge transferred from a different DNN [4]. Along similar direction, federated deep learning with carefully designed sharing of model parameters can be a powerful mechanism for defending adversarial attacks in deep learning and machine learning in general.

Acknowledgement.
This work is partially sponsored by NSF under SaTC 1564097. Any opinions, findings and conclusions or recommendations expressed in this material are those of the researchers and do not necessarily reflect the views of the National Science Foundation.

Reference

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. "Intriguing properties of neural networks". https://arxiv.org/abs/1312.6199 (2013).

[2] Ian J. Goodfellow, J. Shlens, and C. Szegedy. "Explaining and harnessing adversarial examples". https://arxiv.org/abs/1412.6572 (2014).

[3] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, Ananthram Swami. "The Limitations of Deep Learning in Adversarial Settings", Proceedings of the 2016 IEEE European Symposium on Security and Privacy.

[4] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, Ananthram Swami. "Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks", IEEE Symposium on Privacy and Security 2016.

[5] Alexey Kurakin, Ian J. Goodfellow, Samy Bengio. "Adversarial machine learning at scale", Proceeding of ICLR 2017.

[6] Dave Gershgorn. ``Fooling the Machine: The Byzantine Science of Deceiving Artificial Intelligence", Popular Science, March 30, 2016. http://www.popsci.com/byzantine-science-deceiving-artificial-intelligence.

Ling Liu is a Professor in the School of Computer Science at Georgia Institute of Technology. She directs the research programs in Distributed Data Intensive Systems Lab (DiSL), examining various aspects of large scale big data systems, including performance, availability, security, privacy and trust. Prof. Liu is an elected IEEE Fellow, a recipient of IEEE Computer Society Technical Achievement Award in 2012. Prof. Liu has published over 200 international journal and conference articles. Her work has received the best paper awards from numerous top venues, including ICDCS 2003, WWW 2004, 2005 Pat Goldberg Memorial Best Paper Award, IEEE Cloud 2012, IEEE ICWS 2013, ACM/IEEE CCGrid 2015, IEEE Edge 2017. Prof. Liu has served as general chair and PC chairs of numerous IEEE and ACM conferences in the fields of big data, cloud computing, data engineering, distributed computing, very large databases, and served on editorial board of over a dozen international journals. Prof. Liu's current research is primarily sponsored by NSF and IBM.

Web page: http://www.cc.gatech.edu/~lingliu/