Certifying attack resistance of convolutional neural networks

January 31 2019

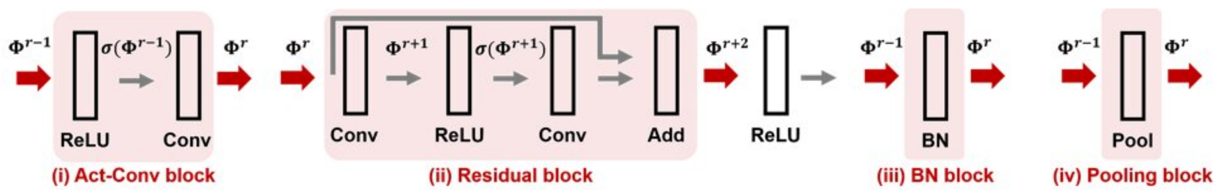

Figure 1. CNN-Cert supports many popular modules and layer operations in convolutional neural networks. Credit: IBM

When shopping for a watch, you may notice its water resistance rating, which indicates that the watch is warranted to be waterproof to a certain level. What about your neural network? Can one ensure a neural network is "attack proof", meaning that its functionality is robust against adversarial perturbations? If so, how can this be quantified with an attack resistance number? At AAAI 2019, our group of researchers from MIT and IBM Research proposes an efficient and effective method for certifying attack resistance of convolutional neural networks to given input data. This paper is selected for oral presentation at AAAI 2019 (January 30, 2:00-3:30 pm @ coral 1).

Current deep neural network models are known to be vulnerable to adversarial perturbations. A carefully crafted yet small perturbation to input data could easily manipulate the prediction of the model output, including machine learning tasks such as object recognition, speech translation, image captioning, and text classification, to name a few. A lack of robustness to adversarial perturbations incurs new challenges in AI research and may impede our trust in AI systems.

Given a neural network and considering an adversarial threat model in which the attack strength is characterized by the Lp norm of the perturbation, for any data input, its adversarial robustness can be quantified as the minimal attack strength required to alter the model prediction (see Figure 1 in the previous post for a visual illustration). Here, an attack-proof robustness certificate for an input specifies an attack strength ε and offers the following guaranteed attack resistance: under the norm-bounded threat model, no adversarial perturbations can alter the prediction of the input if their attack strength is smaller than ε. In other words, larger ε means the input is more robust. This robustness certification can be crucial in security-critical or cost-sensitive AI applications requiring high precision and reliability, such as autonomous driving systems.

Our proposed method, CNN-Cert, provides a general and efficient framework for certifying the level of adversarial robustness of convolutional neural networks to given input data. Our framework is general: we can handle various architectures including convolutional layers, max-pooling layers, batch normalization layer, residual blocks, as well as general activation functions such as ReLU, tanh, sigmoid and arctan. Figure 1 shows some commonly-used building blocks considered in our CNN-Cert framework. The key technique in CNN-Cert is deriving explicit network output bound by considering the input/output relations of each building block, marked as red arrows. The activation layer can be general activations other than ReLU. Our approach is also efficient—by exploiting the special structure of convolutional layers, we achieve up to 17 and 11 times of speed-up compared to the state-of-the-art certification algorithms and 366 times of speed-up compared to a standard dual-LP approach while our method obtains similar or even better attack resistance bounds.

More information: Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is Robustness the Cost of Accuracy? A Comprehensive Study on the Robustness of 18 Deep Image Classification Models. European Conference on Computer Vision (ECCV), 2018

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus. Intriguing properties of neural networks. International Conference on Learning Representations (ICLR), 2014

Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. EAD: Elastic-Net Attacks to Deep Neural Networks via Adversarial Examples. AAAI Conference on Artificial Intelligence (AAAI), 2018

Nicholas Carlini and David Wagner. Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. Deep Learning and Security Workshop, 2018

Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh, "Attacking Visual Language Grounding with Adversarial Examples: A Case Study on Neural Image Captioning," Annual Meeting of the Association for Computational Linguistics (ACL), 2018

Qi Lei, Lingfei Wu, Pin-Yu Chen, Alexandros G. Dimakis, Inderjit S. Dhillon, and Michael Witbrock, "Discrete Attacks and Submodular Optimization with Applications to Text Classification," The Conference on Systems and Machine Learning (SysML), 2019

Tsui-Wei Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Duane Boning, Inderjit Dhillon, and Luca Daniel. Toward Fast Computation of Certified Robustness for ReLU Networks. International Conference on Machine Learning (ICML), 2018

Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient Neural Network Robustness Certification with General Activation Functions. Neural Information Processing Systems (NeurIPS), 2018

This story is republished courtesy of IBM Research. Read the original story here.

Provided by IBM