Towards Trustworthy Deep Learning

Tsui-Wei (Lily) Weng
UC San Diego
lweng@ucsd.edu

In this New Faculty Highlights talk at AAAI-24, I will survey my representative works in **Robust ML** and **Interpretable ML**, which are two vital components towards trustworthy deep learning.

Part (I) Robust ML: Provable DNN Robustness Guarantee and Scalable Certified Defense. DNNs are known to be vulnerable to adversarial examples, which has raised serious concerns in the society as DNNs have been deployed to many mission-critical tasks including self-driving cars, aircraft control systems, and malware detection protocols. To study the vulnerability of neural network models, my lab has contributed significant research efforts on quantifying and improving robustness of DNNs.

I will overview our research effort along this topic since 2017, where we proposed the first attack-agnostic robustness evaluation metric, the first efficient robustness certification algorithms, and efficient robust learning algorithms.

- In ICLR 18, we proposed CLEVER score, the first attack-agnostic robustness evaluation of DNNs based on universal lower bound and extreme value theory. CLEVER score is efficient to compute and well matched the practical robustness indication of a wide range of natural and robust DNNs.
- In AAAI 19, we proposed CNN-Cert, the first efficient algorithm to certify robustness of general CNNs. Cnn-Cert is able verify robustness of more complicated NNs (e.g. LeNet, ResNet) than prior work by generalizing the idea of linear bounding framework in my prior work on fully-connected NN to conv/pool/residual blocks.
- In AAAI 21, we proposed SingleProp, a efficient robust training algorithm based on novel regularization and approximation of linear verification bounds that can effectively improve robustness of DNNs while being up to 13.5$\times$ faster to train compared to prior work based on verification bounds.

I will also briefly overview our follow-up works on verifying semantic perturbations, learning robust deep RL agents, robustifying conformal prediction of DNN, and enabling scalable defense .

Part (II) Interpretable ML: Understanding DNNs via Neuron Interpretations and Building Interpretable DNN To understand the internal working of DNNs, there is a stream of research on developing neuron interpretability. However, existing methods are still very restricted due to the need of manual inspection and concept annotations. These requirements are time-consuming and labor-intensive that may slow down the research advancement. Moreover, these methods can only detect concepts from a pre-defined concept set, which is difficult and expensive to expand, as new (densely labeled) data is required for new concepts.

I will overview our research efforts along this direction by addressing above challenges. Our goal is to develop a fully-automated and flexible interpretation method for DNNs that is efficient and requires minimum human efforts. I will also introduce our work on learning interpretable DNN that can remain high DNN performance unlike existing approaches that are limited by interpretability and accuracy trade-off.

- In ICLR 23, we proposed CLIP-Dissect, to automatically describe the function of individual hidden neurons inside vision networks with open-vocabulary concepts. Our method does not need any concept labeled data, which are required for existing tools to succeed. CLIP-Dissect has several advantages: (i) it provides more accurate descriptions than existing state-of-the-art; (ii) it is 10-200$\times$ faster than existing methods.
- In ICLR 23, we proposed Label-free Concept Bottleneck Model, to address two crucial limitations of existing CBMs which needs labor to collect concept labelled data and suffers from serious accuracy and interpretability trade-off. Our method can transform any NN into an interpretable CBM without labeled concept data while retaining a high accuracy by leveraging our neuron interpretability tool CLIP-Dissect. Our method is also the first CBM that can scale to large-scale ImageNet.

I will also briefly overview our recent efforts on demystifying black-box DNN training process, automated neuron explanations for Large Language Models and the first robustness evaluation of a family of neuron-level interpretation techniques. For full publication details, please visit my website: https://lilywenglab.github.io/