Research Statement

Yao-Yuan Yang

Traditional machine learning operates in the statistical learning framework, where both training and testing data are drawn i.i.d. from the same distribution. However, when such models are deployed, this assumption may not always hold, and thus, exposing the model to unreliable situations, causing security and generalization issues in practical applications. To resolve these issues, **my research takes a principled approach towards understanding the foundations of trustworthy machine learning as well as its applications to resolving real-world problems**. I focus on building up scientific and mathematical theories in this area. Specifically, I study various properties of machine learning models that are not considered under the statistical framework, including adversarial robustness, out-of-distribution generalization, spurious correlation, and interpretability.

Besides foundational research, I am also interested in applying my research to practical fields. In the past, I have developed algorithms for multi-label classification [1, 2]. I also extensively collaborated with researchers across various domains, including programming language [3], social sciences [4], and neuroscience [5]. In addition, I am an advocate of open science and have leveraged my expertise in machine learning for the benefit of broader communities through open-sourced projects. Two main projects that I have contributed significantly are `libact` [6] and `torchaudio` [7], which are libraries built for active learning and audio/speech applications. All in all, my research offers insights into how machine learning models work, what properties they have, and how they can be applied.

Adversarial Robustness

While more and more models are being used in high-stake applications, many of these are vulnerable to adversarial attacks, raising increased attention to adversarial robustness. Therefore, much of my research is devoted to understanding what causes a model's vulnerability to adversarial attacks and how to make a model robust without sacrificing accuracy.

In [8], I, along with coauthors, take a holistic look at adversarial examples for non-parametric classifiers. We develop an attack and a defense algorithm that empirically work well across multiple non-parametric classifiers, including the k-nearest neighbor classifier, decision tree, and random forest. For the attack algorithm, we show that it is capable of delivering the optimal attack, which finds the closest adversarial example. To justify our defense algorithm, we first derive the optimally robust classifier, which is analogous to the Bayes Optimal. Then, we show that our defense can be viewed as a finite sample approximation to this optimally robust classifier. A further study into the connection between non-parametric classifiers and the optimally robust classifier can be an interesting next step.

In [9], we examine the tradeoff between robustness and accuracy of defense algorithms in neural networks. First, we show that when the data distribution is r-separated, robustness (with radius r) and accuracy can be achieved simultaneously (meaning that there is no intrinsic tradeoff). Then, we show that many commonly used image datasets are r-separated. This evidence suggests that, in principle, robustness and accuracy should be achievable at the same time. However, in practice, researchers are observing tradeoffs that they are not able to mitigate on these image datasets [10]. To understand this phenomenon, we conduct an empirical study and find that there is an increase in the generalization gap when robust algorithms are applied, meaning that the network is losing some ability to generalize well

during the process of making itself robust. An in-depth study on how to retain the network's ability to generalize while being adversarially robust at the same time can be an appealing direction.

In [11], we investigate whether it is possible to further achieve interpretability on top of robustness and accuracy. We focus on decision trees and start with the condition of r-separation, which is considered to be sufficient for a classifier to be robust and accurate from our prior work [9]. We show that decision trees under the r-separation condition can be exponentially large, which makes them not interpretable. Therefore, we further tighten the assumption and assume the data is linearly separable. Under the linear separation assumption, we show that there exists a tree that is robust, accurate, and interpretable. Accordingly, we design an algorithm that constructs such a tree.

It appears that adversarial robustness is interconnected with interpretability, which is one of the desirable properties for trustworthy models. In the next section, we further showcase our discovery on the connection between adversarial robustness with an out-of-distribution generalization property of neural networks. As a future direction, I am excited about discovering more of these connections.

Out-Of-Distribution Generalization

When in-distribution inputs are given to a neural network, we expect it to perform similarly to the training examples. However, how a neural network performs with out-of-distribution inputs remains an unresolved question. To address this inquiry, we start with exploring whether there are any patterns in the prediction when out-of-distribution examples are given. Inspired by a line of work in the psychology literature that posits humans categorize unseen examples into the most similar category they have seen (i.e., generalized context model) [12], we examine whether neural networks also behave in the same way.

In [13], we explore this research question in depth and find that neural networks do behave this way. Neural networks tend to predict out-of-distribution examples as the nearest category in the training set, and we call this property nearest category generalization (NCG). Furthermore, we find that making neural networks more adversarially robust, which is another property that humans possess, leads the networks to follow NCG more strictly.

How neural networks generalize so well is still unanswered, and our work provides some insights into how networks generalize. Many scholars conjecture that the effectiveness of deep learning may be coming from its similar structure to the human brain, which allows neural networks to share some of the inductive biases from human brains [14, 15]. This work not only provides an additional piece of evidence supporting this theory, but it also raises an intriguing question – **does enforcing other human-like behaviors on neural networks increase the "humanness" of the neural network?** I am passionate about pursuing this inquiry in my ongoing research trajectory.

Applications and Ongoing Work

Spurious correlation. The spurious correlations are known to be learned by neural networks. For example, it has been shown that image classifiers commonly use the background as a feature to classify objects [16], and this often hurts the test time performance when there are distribution shifts [17]. These studies are usually under the scenario where spurious features are present in a substantial fraction of the training data. For example, image classifiers tend to associate waterbirds with water backgrounds, and at the same time, the vast majority of waterbirds, in fact, are photographed next to the water. So a question appears – will the neural network learn a spurious correlation if this correlation only appears in a handful of training examples? If a small number of spurious training examples can build up a spurious correlation in the neural network, this could not only negatively affect test time performance, but the rarity of these examples may also pose a potential privacy concern. An adversary may exploit this rare correlation to infer the existence of a specific example in the training set.

We empirically investigate the following question: how many training points does it take for a neural network to learn a spurious correlation? We artificially insert a pattern into training examples of a specific target class and examine the neural network trained on this modified dataset. We find that even by modifying just 3 out of $60,000$ training examples, the network can readily associate the pattern with the target class. Furthermore, we find that it is hard to unlearn this association with standard data deletion methods. My research objectives include understanding what kind of patterns are more easily picked up by the neural networks and how to efficiently and effectively unlearn these spurious correlations.

Machine learning, brain signals, and biometric applications There have been many studies that propose using brain signals as biometrics. However, what contributes to the uniqueness of one's brain signals remains unclear. To understand what constitutes the person-identifiable brain signals, in [5], we conduct a multi-task and multi-day EEG study with monozygotic twins. Our results reveal the existence of person-identifiable, task-invariant, and temporally stable "base signals", which were embedded in brain signals. We also find that these signals are more indifferentiable within than between monozygotic twin pairs. Our findings suggest that individuals' unique brain signals can be a combination of both inborn genes and acquired experiences.

References

[1] Yao-Yuan Yang, Kuan-Hao Huang, Chih-Wei Chang, and Hsuan-Tien Lin. Cost-sensitive reference pair encoding for multi-label learning. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 143–155. Springer, 2018.

[2] Yao-Yuan Yang, Yi-An Lin, Hong-Min Chu, and Hsuan-Tien Lin. Deep learning with a rethinking structure for multi-label classification. In *Asian Conference on Machine Learning*, pages 125–140. PMLR, 2019.

[3] Benjamin Cosman, Madeline Endres, Georgios Sakkas, Leon Medvinsky, Yao-Yuan Yang, Ranjit Jhala, Kamalika Chaudhuri, and Westley Weimer. Pablo: Helping novices debug python code through data-driven fault localization. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, pages 1047–1053, 2020.

[4] Angel Hsing-Chi Hwang, Cheng Yao Wang, Yao-Yuan Yang, and Andrea Stevenson Won. Hide and seek: Choices of virtual backgrounds in video chats and their effects on perception. In *Proceedings of the 2021 ACM Conference on Computer Supported Cooperative Work (CSCW)*, 2021.

[5] Yao-Yuan Yang, Angel Hsing-Chi Hwang, Chien-Te Wu, and Tsung-Ren Huang. Brainprints in mindprints: Stable task-invariant eeg base signals for personal identification. *submission*, 2021.

[6] Yao-Yuan Yang, Shao-Chuan Lee, Yu-An Chung, Tung-En Wu, Si-An Chen, and Hsuan-Tien Lin. libact: Pool-based active learning in python. *arXiv preprint arXiv:1710.00379*, 2017.

[7] Yao-Yuan Yang, Moto Hira, Zhaoheng Ni, Anjali Chourdia, Artyom Astafurov, Caroline Chen, Ching-Feng Yeh, Christian Puhrsch, David Pollack, Dmitriy Genzel, Donny Greenberg, Edward Z. Yang, Jason Lian, Jay Mahadeokar, Jeff Hwang, Ji Chen, Peter Goldsborough, Prabhat Roy, Sean Narenthiran, Shinji Watanabe, Soumith Chintala, Vincent Quenneville-Bélair, and Yangyang Shi. Torchaudio: Building blocks for audio and speech processing. In *submission*, 2021.

[8] Yao-Yuan Yang, Cyrus Rashtchian, Yizhen Wang, and Kamalika Chaudhuri. Robustness for non-parametric classification: A generic attack and defense. In *International Conference on Artificial Intelligence and Statistics*, pages 941–951. PMLR, 2020.

[9] Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Ruslan Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. 2020.

[10] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[11] Michal Moshkovitz, Yao-Yuan Yang, and Kamalika Chaudhuri. Connecting interpretability and robustness in decision trees through separation. *arXiv preprint arXiv:2102.07048*, 2021.

[12] Jeffrey N Rouder and Roger Ratcliff. Comparing categorization models. *Journal of Experimental Psychology: General*, 133(1):63, 2004.

[13] Yao-Yuan Yang, Cyrus Rashtchian, Ruslan Salakhutdinov, and Kamalika Chaudhuri. Robustness and generalization to nearest categories. *arXiv preprint arXiv:2011.08485*, 2020.

[14] Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.

[15] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

[16] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020.

[17] Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020.