Development of Generalizable, Sparsed, Certified Robust, and Fair Personalized Human-centric Sensing

Motivation: It is common for human-centric AI applications to personalize the machine-learning models for the end users. For example, Google Home or Alexa takes users' voice samples to finetune the speech transcription models, and Apple or Samsung watches finetune models on users' signals for better activity detection. However, such personalization makes the models overfitted on end-users' limited provided data and compromises *generalizability*, specifically in unknown data distribution due to changes in environments or sensor errors.

Deep learning models are highly accurate but require extensive computing and storage resources. This aspect challenges their deployment in resource-constraint settings, such as wearables and embedding systems, particularly when real-time executability is needed. Recent studies have developed methods to compress the neural network parameters through sparsification to address this challenge. Only a few works, including our recent study [Sawinder et al., 2023], developed algorithms to generate sparse models having both high accuracy and formally guaranteed adversarial robustness. Adversarial samples are the minor perturbated sample that gets misclassified by the model; hence *without adversarial robustness,* compressed models would generate unstable erroneous inferences for even little input variation, which may easily cause by sensor noise, environmental or human factors. Since human-centric sensing approaches are safety-critical, generating unreliable and erroneous automated sensing inferences or actuation can have disastrous consequences. Thus, it is critical to ensure that personalized sparsed models are both accurate and provide *formally verified, i.e., certified adversarial robustness.*

There is a broad moral and legal consensus that certain types of discrimination are wrong, and we must control the role of human-centric sensing algorithms in such discrimination. Although there is a growing research community studying *fair machine learning* algorithms, very little such work has been done in human-centric computing, and none we know of on personalization. This proposal presents a research and education plan to develop generalizable, scalable, formally verified adversarial robust, and fair human-centric sensing systems. Specifically,

Aim 1: The proposed work will develop frameworks that prevent personalization-caused overfitting. Additionally, the proposed framework would generate personalized models robust against changes in sensing signal distribution due to environmental changes or sensor errors.

Aim 2: Our preliminary analysis shows that even if the original model is adversarial robust, such robustness decreases during personalization, even if the adversarial robustness is integrated into the fine-tuning objective. The proposed work will develop model sparsification algorithms to generate personalized compressed models with high certified, i.e., formally verified adversarial robustness.

Aim 3: Our preliminary analysis shows that personalization decreases the fairness of the originally fair generic models. The proposed work will develop frameworks to achieve personal-trait-wise and group-wise fairness in personalized human-centric sensing systems.

Intellectual Merit: An enduring difficulty AI researchers face across many fields is the challenge of making systems more generalizable, scalable, robust, and fair. The research proposed here seeks to meet this challenge in the specific domain of personalized human sensing by exploiting and addressing particular characteristics of the problem area. The proposal's intellectual merits include (a) the development of algorithms that intelligently identifies the generic attributes through contrastive learning adaptation to make personalized models highly accurate for the end-users while avoiding overfitting; (b) adaptation of dynamic co-teaching in personalized models that leverage the co-dependencies in multi-modal sensing systems to recover sensing information due to noise or sensor errors; (c) a dual optimization framework that generates personalized sparse neural network architectures having high accuracy and formally certified robustness comparable to their 100 times denser counterparts. Such a framework makes the neural network deployable for practical use, safe, reliable, and trustworthy for critical applications, and interpretable to understand models' capability and confidence factors; (d) the proposal will investigate to understand the underlying reasons for fairness disruption during personalization and will develop frameworks to attain fair personalized models in terms of both group-wise and personal trait wise fairness.

Broader Impact and Education Plan: To Do

Reference:

Guo, Y., Zhang, C., Zhang, C., & Chen, Y. (2018). Sparse dnns with improved adversarial robustness. *Advances in neural information processing systems*, *31*.

Sawinder Kaur and Yi Xiao and Asif Salekin. " VeriSparse: Training Verified Locally Robust Sparse Neural Networks from Scratch." *arXiv preprint arXiv:2211.09945* (2023).