My research interests are in building **Trustworthy AI system deployment** which are reliable, secure and interpretable, into **safety-critical** scenarios where the explainality, privacy, security, reliability and fairness are important, such as medical, psychotheraphy nlp app, banking and human-resource platform. I like the process of finding the practical problem which really impacts the real-world, and solving it with reasonable, meaningful, solid and feasible methods, after reviewing cutting-edge literature and industry reliable methods, and discussing the problem with other talents whom with different professional backgrounds. At my undergrad, I have been employed as a researcher assistant by Prof.Dr. **Shibai Yin**, working on image dehazing and **transformer** model, leading to **a publication**[1]. At the present, I am employed as a security and applied cryptography full-time research student at Huawei **Munich Research Center(Germany)**, **Trustworthy and Applied Cryptography lab**, working on **TEEs**(Trusted Execution Environment) interfaces implementation. I am writing my master thesis aiming at developing a secure multi-party computation(MPC) machine learning platform, based on **Intel SGX** libraries, advised by Prof.Dr. **Ming Xiao,** and Prof.Dr. **Johan Håstad**.

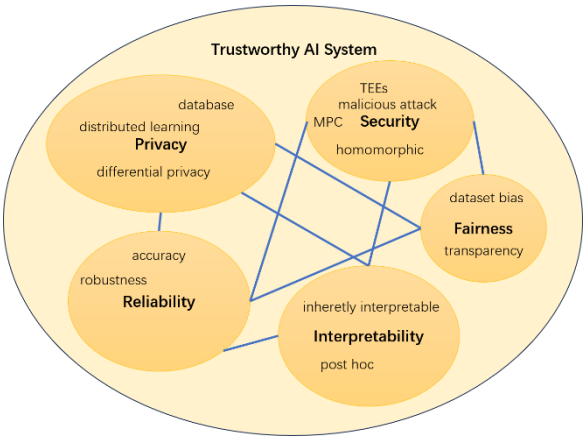

Figure 1 Trust AI system key components(The blue line indicates the trade-off between objects)

Current Research: With deep models' impressive performance in image and NLP sessions, most researchers focus on accuracy as the main benchmark. However, it is not enough to establish a trustworthy AI system. Single top accuracy also can not waive the doubt/supervision for model performance when facing real-time data in some special scenarios, such as abnormal condition may occur in the future but is not included in the "past" training dataset.

For my thesis work, I focus on the security and privacy aspect of a trust AI system. Nowadays while the Federated Learning is the main scheme for secure system without sharing privacy with other clients, it suffers from gradient attack, and sensitive information leakage which can be inadvertently from separate devices such as Generative Adversarial Networks(GANs). The modern MPC+GPU secure computation scheme also has short-coming of high-overhead workload, and not full-work GPU due to limited communication band. To tackle these problems, I developed the TEEs machine learning platform, where computation is plain-text within hardware protection. The secure machine learning platform's data sample is stored in a distributed format holding by several parties, following secret-sharing scheme. The platform enables some effective LeNet, AlexNet, ResNet and Transformer models' inference and trainning for clients, which are developed in C pure implementation, due to 1. **TEEs** such as Intel SGX's pure C style limitation for their security features. 2. Fulfill Industry's light-weight demand, and the condition that most private servers' legacy environments—— the platform only relies on a limited libraries and easily to deploy on almost all kinds of servers. The models' parameter and structure would be sealed in TEEs, none of clients and servers can access the parameters. The model code structure demonstration can be found here[2], which requires **solid C/C++** programming skills and **applied mathematical** linear differential algebra matrix induction ability.

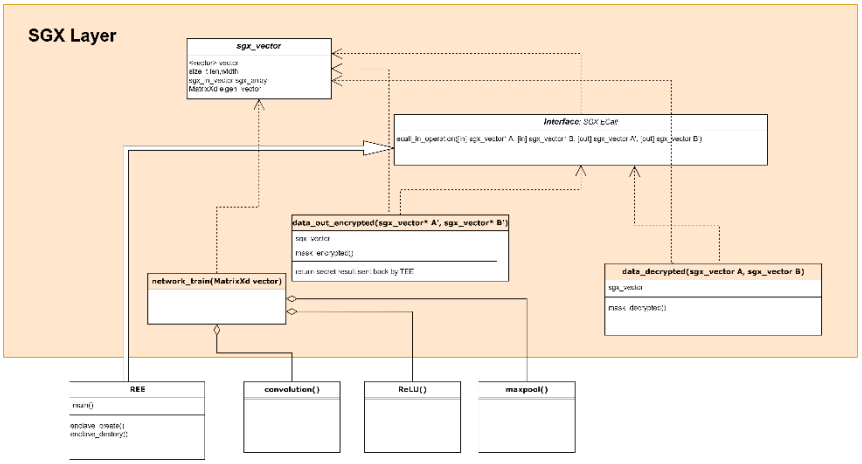

Figure 2 Intel SGX Code Design

Impressive by ChatGPT's performance, I am confident and excited about the AI system's deployment into practical industries, which very likely would bring social and economic benefits. In addition to engineering work, I enjoy embedding and hearing new mathematical/engineering optimization ideas in my mind, from small readings or small chats with other talents, which can help me a lot to modify the problem when starting a new work. I was awarded an **honors mathematical** bachelor degree as the 3rd bachelor degree, and a **national second prize** of undergraduate **mathematical modeling competition**[3]. Inspired by "Information Theory and Source Coding", which requires a good mathematical understanding, and the ability to implement the applied mathematical ideas on article to practical such as **Golomb/arithmetic** code. I appreciate the charm of mathematical modeling such as entropy, and the **lower bound estimation**, which is the idea considering simulating as lowest-information/highest entropy under natural condition with large number scheme, using max-min methodology. I see the potential using such lower bound techniques to estimate the trust AI system. For example, the **uncertainty bound of specific perturbations adding** scheme's bound on theory. The accuracy bound of DP(Differential Privacy) noise adding. Estimate work's uncertainty probability distribution by counting its synonyms, then set rational assumptions and calculate special case distribution's entropy and lower bound, which the idea is inspired by Dr. Hanjie Chen's work *Explaining Predictive Uncertainty by Looking Back at Model Explanations[4]*. The advantage of such potential theoretical scheme is making engineering research path-clear and bridge with previous talents' work into usage.

Future Research: Trustworthy AI system building is large topic, several aspects not only accuracy involved. And each single aspect also has multiple efforts and topics, such as for performance: dataset robustness, multi-modal, optimization e.g. loss function/dynamic step, pruning, distillation, diffusion etc. Each time research we may only focus on one small and reasonable topic. So future plan would be separated in long-tern and short-term goals. For short term goals:

- TEEs hardware secure machine learning platforms, which maybe distributed learning scenarios or secure cloud computing privacy learning platform.

- The optimization theory contributing to fair learning. For example, according to large number schemes, the expectation of bias within large enough data should approach 0 under rational settings. Convex optimization can lead to different parameter distribution when achieving the same train loss(actually different extreme points). We can set KKT or Lagrange optimization equations, with a goal to set model's output on R^n 's expectation to 0, where R^n is all possible samples, n is the dimensional of input. $E(x \times w^* + b) = 0$.

- Upon implementing the neural network in C, I found that transformer may is not necessarily effective than resnet, transformer's advantage is multi-modal ability. And for all present modern neural models, previous gradient would not participate in current layer's gradient update, the same for res-connection, maybe there is space exploring connected gradient update. And I also found given a dataset, model's performance would not really improve, when a smaller model has been able learning the data set well. I think there is a model size lower bound for given dataset. For a larger model without performance improvement, there might be correlated linear algebras, there maybe worthy de-correlation operations for each layer's weights to make the model smaller.

- Prompt explainability and evaluation research. Prompt attack analysis.

For long term goal is the trustworthy AI system deployment, I hope I would have the opportunity leading such projects/applying the related-fund in the future.

I would be thrilled to pursue a PhD at Rice. After PhD I want to work on a research position at Company/University. For the long term goal I may explore if there is opportunity creating a public AI-system startup such as psychotherapy apps. Prof.Dr. Hanjie Chen's interpretable models research plan well-matched with my interests on trust AI system deployment. The Rice's broader faculty will provide an ideal community for my research questions. I believe Rice will provide the best environment for potential candidates to succeed as a research and academic.

References

[1] Yin, S., **Xin, J**., Wang, Y., & Basu, A. (2020). Image dehazing with uneven illumination prior by dense residual channel attention network. *IET Image Processing*, *14*(13), 3260-3272.

[2] https://daseinda.github.io/assets/pdf/rp.pdf

[3] https://daseinda.github.io/assets/pdf/cum.pdf

[4] Chen, H., Du, W., & Ji, Y. (2022). Explaining Predictive Uncertainty by Looking Back at Model Explanations. *arXiv preprint arXiv:2201.03742*.