Graduate Research Plan

Research Title: Privacy-Preserving AI Systems with Polynomial Networks

Advances in machine learning (ML), artificial intelligence (AI), and AI services have revolutionized complex task automation. However, the current implementations of many of these AI systems were not designed to protect data privacy. In sensitive domains such as finance and national security, entities may be unwilling or unable to send confidential data to AI services, as even if a service is trusted, information can still leak through unencrypted residuals. This fear of disclosure is justified, as despite research in privacy-aligned AI, many modern AI services cannot provably restrict data leakages. These limitations necessitate new methods for how AI operates on private data.

An ideal solution would permit users to encrypt their data and allow an AI service to operate directly on the ciphertext, thus providing complete data confidentiality. This requires a homomorphism, a function that guarantees that the decrypted output of the AI model on the ciphertext matches the model's output on the corresponding plaintext. Fully homomorphic encryption (FHE) allows this, provided that only addition and multiplication operations are applied to the encrypted data. This limits the use of FHE with most ML models, which rely on nonlinear activation functions. These nonlinearities can be approximated using polynomials, but this can incur performance losses. Fortunately, polynomial networks (PNs), AI architectures that only use polynomial functions, can theoretically support FHE by default. Designing FHE-compatible PNs is therefore a viable research path to improve confidentiality in AI services.

Summary of Graduate Research Plan: As AI services continue to be entrusted with sensitive data, it is vital to create systems that support confidentiality without sacrificing effectiveness. My proposed research improves AI privacy by implementing fully homomorphic encryption with polynomial networks to achieve accurate and scalable privacy-preserving AI services. This research focuses on multilinear operator networks (MONet) [1], a class of PNs that achieves state-of-the-art performance on image classification compared to small transformers. I intend to empirically demonstrate the capabilities of MONet with FHE and scale the architecture to perform efficiently for large tasks, such as those performed by transformer networks. I will then apply other privacy-enhancing technologies, such as differential privacy, to PNs. This research will demonstrate how PNs can improve the confidentiality and security of AI services, protecting data privacy, aligning modern AI with security goals, and allowing AI services to expand to previously infeasible domains like medical data analysis and centralized financial fraud detection.

Objective 1 – Benchmark Polynomial Networks for FHE: The first goal of this project is to benchmark how PNs such as MONet perform on data encrypted using FHE. Depending on the depth and type of the operations applied during AI training and inference, FHE-encrypted data can degrade, harming model accuracy. Confirming MONet's full FHE compatibility requires showing that the model maintains high accuracy on FHE ciphertexts while incurring minimal computational overhead. It is also necessary to compare MONet to other FHE-compatible ML schemes, such as CryptoNets [2]. These comparisons will consider traditional metrics as well as ciphertext size, the frequency and cost of bootstrapping operations, and the latency per inference. I will first train these privacy-preserving models for image classification on both conventional datasets and domain-specific benchmarks, such as RadImageNet [3], a radiology image benchmark. I will then test the performance of these FHE-compatible models against comparable deep ML models like Vision Transformers to measure the tradeoff of eliminating nonlinear activation functions. Afterwards, I will perform the same tests using data encrypted with an FHE algorithm (e.g., CKKS [4]) and compare the metrics for applicable models. Doing so will provide a baseline for how MONet and similar PNs can replace existing AI models in privacy-preserving environments.

Objective 2 – Scale Polynomial Networks for Privacy-Preserving Tasks: Showing the utility of PNs for privacy-preserving AI requires scaling these PNs for larger tasks without degrading each architecture's compatibility with FHE. This objective aims to advance MONet's capabilities for more complex tasks, including text classification and generation. The primary goal is to replicate the vanilla transformer

architecture by adapting MONet to implement an FHE-compatible multi-head self-attention mechanism. PNs are prone to overfitting in these deeper architectures, so this adaptation will require developing new regularization methods that do not rely on nonlinear functions. I hypothesize that novel FHE-compatible regularization techniques, such as layer-wise noise injection and polynomial function magnitude decay, will enable MONet to learn deeply without overfitting. Additionally, I will create a decision framework to determine when to adapt existing architectures for FHE instead of training new FHE-native models.

Objective 3 – Incorporate Polynomial Networks with Differential Privacy: Although FHE protects the confidentiality of data passing through an AI model, it does not prevent inference attacks, which leak information about the data used to train the model. Differential privacy (DP) protects against these attacks by adding obfuscating noise during training. Research [5] has shown that DP-SGD, the key optimization function used for DP on ML models, operates well on models that have smooth losses. The loss of polynomial models is inherently smooth, which indicates that PNs may incur only a small accuracy tradeoff when implemented with DP. In this research thrust, I will train MONet with DP-SGD and compare the results to other models trained with DP. I hypothesize that the noise injected by DP-SGD will serve as another form of regularization for deep PNs, thus mitigating overfitting while gaining differential privacy properties. Showing that PNs can be both DP- and FHE-compatible will demonstrate that these networks can serve as a new architectural foundation in privacy-preserving AI.

Intellectual Merit

The intellectual merit of this research lies in the development and empirical evaluation of an FHE-native AI framework based on polynomial networks, providing a path towards privacy-preserving AI services with minimal performance or accuracy degradations. Further, this research promotes a paradigm shift in AI system design from patching AI privacy weaknesses to designing systems that are provably secure from inception. The techniques developed in this research, including the methods for FHE-compatible regularization and the use of DP-related noise as a regularization strategy, transcend research in AI security and will aid improvements in ML theory. This research also establishes benchmarks for the performance of FHE- and DP-compatible models, which advances the state of knowledge in the field. Finally, this research introduces an evaluation framework that will aid AI practitioners in choosing the optimal model, FHE scheme, and DP strategy for privacy-preserving systems, aiding best practices.

Broader Impacts

The fundamental goal of this research is to increase the adoption of privacy-preserving AI. This will expand the market for AI services in sensitive domains and provide verifiable privacy guarantees for these services. Moreover, provably robust methods for AI privacy based on encryption and DP offer a safer alternative to heuristics. Creating adaptable AI models that support FHE benefits the AI industry in that it makes it simple for companies to adopt FHE in their services, eliminating the need for them to encrypt data in use and lowering the level of trust needed between entities to allow them to share data. Building models that support homomorphic encryption also supports national security, as it allows agencies to contract with third-party AI services to process classified information without leakage. Finally, showing that PNs can support both FHE and DP clears the way for new research in multi-party computation and federated learning. The successful completion of this research will demonstrate that PNs hold untapped potential in dispelling the persistent issues related to privacy and security that plague modern ML.

References: [1] Cheng, Yixin, et al. “Multilinear Operator Networks.” ICLR (2024). **[2]** Dowlin, Nathan, et al. “CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy.” ICML (2016). **[3]** Mei, Xueyan, et al. “RadImageNet: An Open Radiologic Deep Learning Research Dataset for Effective Transfer Learning.” Radiology Artificial Intelligence 4(5) (2022). **[4]** Cheon, Jung Hee, et al. “Homomorphic Encryption for Arithmetic of Approximate Numbers.” ASIACRYPT (2017). **[5]** Wang, Wenxiao, et al. “DPlis: Boosting Utility of Differentially Private Deep Learning via Randomized Smoothing.” Proc. Priv. Enhancing Technol., 2021(4), 163–183 (2021).