Keynote: Privacy Preserving Machine Learning with TFHE

Andrei Stoian

Zama, Paris

Fully Homomorphic Encryption (FHE) is a powerful tool that preserves the privacy of users of online services that handle sensitive data, such as health data, biometrics or credit scores. TFHE is an FHE scheme that supports arithmetic levelled computation and programmable bootstrapping, which applies arbitrary nonlinear functions to ciphertexts while reducing noise.

Backed by these two powerful features, TFHE is particularly adapted to the type of programs that are common in machine learning (ML) for various types of ML models: deep neural networks (DNN), large language-models (LLM) and decision trees (DT).

LLMs have become the backbone for automatic text processing for text generation, translation and speech understanding. They rely on high-dimensional embeddings of words and the multi-head attention mechanism. We show how this mechanism can be implemented with TFHE and results on the GPT2 model.

Deep convolutional neural networks have revolutionized the automatic processing of unstructured data such as images and sound. We show how to construct THE-compatible Deep Neural Networks (DNN). We also demonstrate how specific TFHE primitives speed up neural network inference over encrypted data.

Tree-based ML models obtain state-of-the-art results on tabular data, are more robust, and are easier to use and deploy than neural networks. We discuss an implementation of privacy-preserving decision tree evaluation, which applies to decision trees, random forests, and gradient-boosted trees.

An essential factor in the success of these implementations is the reliance on an automatic optimizer that finds the best crypto-system parameters, taking into account computation graph partitions, which can have unique crypto-system parameters and bootstrapping keys. We briefly discuss the optimizer and its application in a compiler that takes the ML models that will be discussed and produces machine code for their execution on encrypted data.