ML Systems Project Proposal

Team 1: Saurav Sengupta, Zachary Jacoakes

Introduction

Much of the confidentiality-related machine learning work in healthcare settings has focused on anonymization pipelines, where sensitive patient data is systematically assessed for potential identifiers and then scrubbed of these before being analyzed. This is a computationally expensive task, and the propensity for errors (especially under-anonymization errors) make it unappealing for practitioners to employ. Encryption is an alternative to full anonymization, and recent research has yielded promising results about its viability in managing healthcare data.

This project will aim to assess how effectively a new encryption software framework, CrypTen, can compete with other algorithms in this space [1]. CrypTen is an open-source ML and encryption software framework that utilizes PyTorch under the hood and is designed for deep-learning analysis on sensitive datasets. The main appeals of CrypTen include its simple integration and implementation in PyTorch, as well as its accessibility for those not familiar with cryptography.

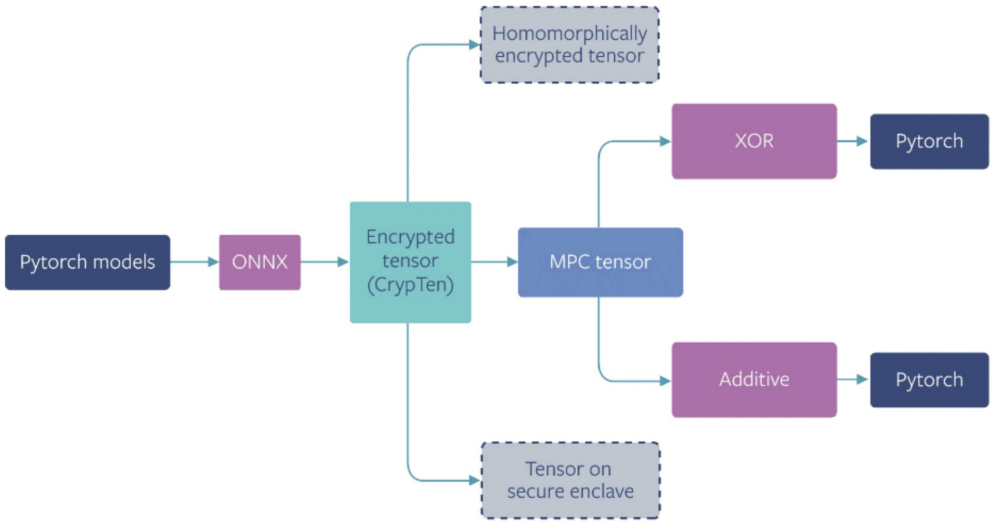

Fig 1: CrypTen Pipeline (ref: https://crypten.ai/)

We plan to assess the speed and computational efficiency of CrypTen on (massive public healthcare-related dataset) by first performing intensive analysis on unencrypted data, and then again using CrypTen. We also plan to compare these results with some state-of-the-art anonymization pipelines to assess the production viability of CrypTen.

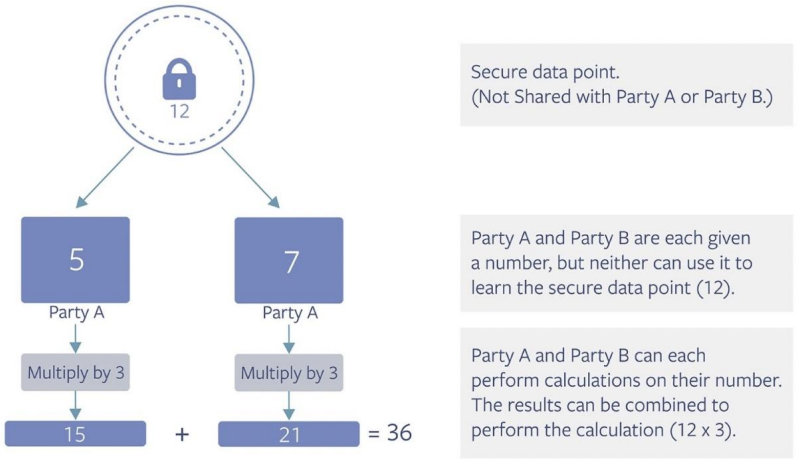

Fig 2: Multi Party Compute (ref: https://crypten.ai/)

The end-goal of the project will be to construct an end-to-end encryption and analysis pipeline for healthcare providers to use when machine learning solutions are desired. The providers will be sent an encryption key to secure sensitive data, and then send the encrypted data to us, the vendors. Here, we assume that vendors have computational capacity not usually available to hospitals. We will perform the appropriate analyses on the encrypted data using CrypTen, and then send the encrypted data inferences back to the providers, which they will be able to decrypt with the same initial key they were previously provided. The diagram below gives a graphical indication of this encryption pipeline.

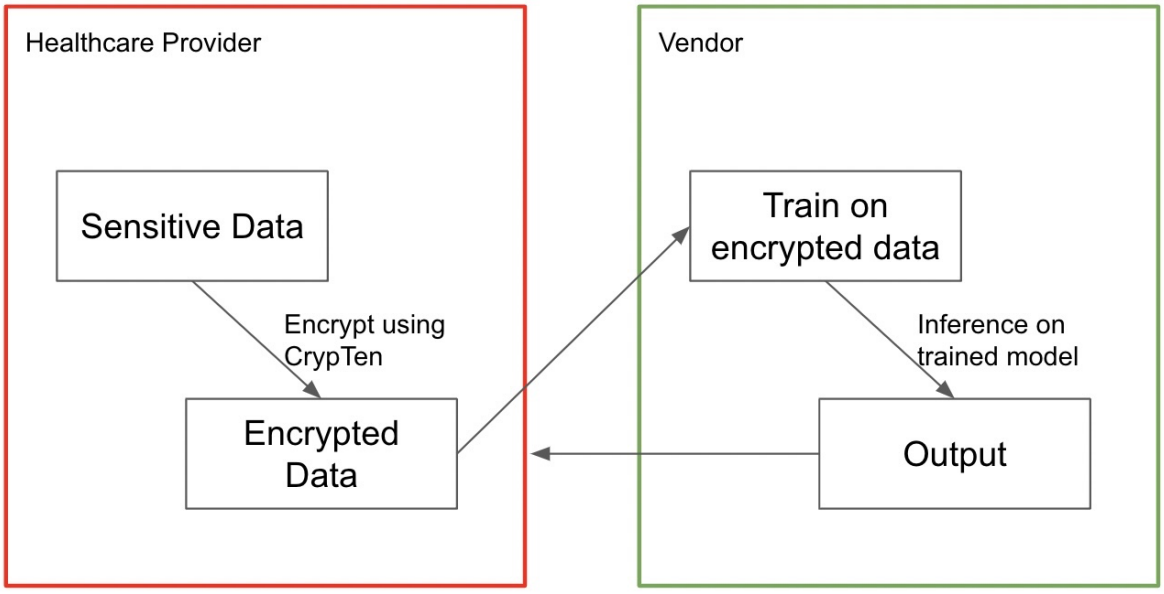

Fig 3: Proposed Encryption Pipeline

Proposed Aims

- Performance comparison of training/inference using CrypTen vs vanilla PyTorch, using execution times, memory used for medical datasets.
- Testing encryption time for encrypting sensitive data using CrypTen to test viability of real time capabilities of a model built using CrypTen.
- Test parallelization capabilities of model training using CrypTen, that is, is there a speed-up when we use multiple GPUs to train.
- Examine security vulnerabilities of using CrypTen.

References:

[1] Knott, B., Venkataraman, S., Hannun, A., Sengupta, S., Ibrahim, M., & van der Maaten, L. (2021). Crypten: Secure multi-party computation meets machine learning. *Advances in Neural Information Processing Systems*, *34*, 4961-4973.

[2] Bos, J. W., Lauter, K., & Naehrig, M. (2014). Private predictive analysis on encrypted medical data. *Journal of biomedical informatics*, *50*, 234-243.

[3] Wang, F., Zhu, H., Lu, R., Zheng, Y., & Li, H. (2020). Achieve efficient and privacy-preserving disease risk assessment over multi-outsourced vertical datasets. *IEEE Transactions on Dependable and Secure Computing*.