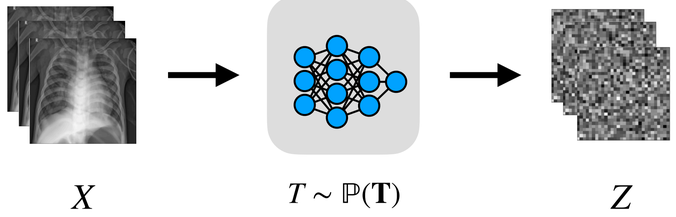
Syfer: Neural Obfuscation for Private Data Release

Adam Yala*, Victor Quach*, Homa Esfahanizadeh, Rafael G. L. D'Oliveira, Ken R. Duffy, Muriel Médard, Tommi S. Jaakkola, Regina Barzilay

MIT CSAIL

Overview

Goal: enable hospitals to release labeled medical images to outsource **model development by untrusted third-parties** while **protecting patient privacy**.



Our approach: learn a distribution $\mathbb{P}(\mathbf{T})$ over random neural networks T such that data owners can safely publish $Z = T(X)$.

Main challenge: how to craft $\mathbb{P}(\mathbf{T})$

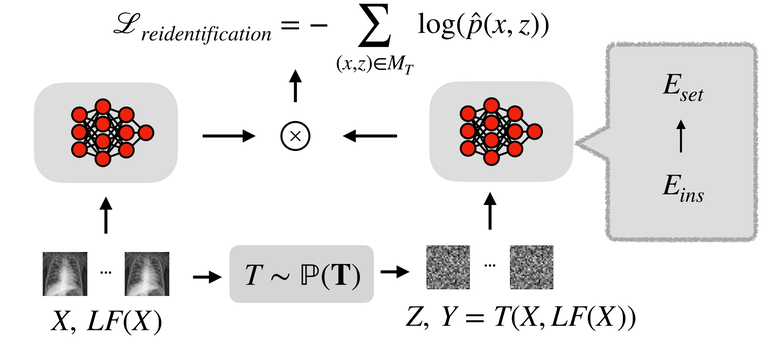
- ... that achieves privacy,
- ... while maintaining downstream task utility,
- ... without knowing the task *a priori,*
- ... nor having access to the private data?

Background

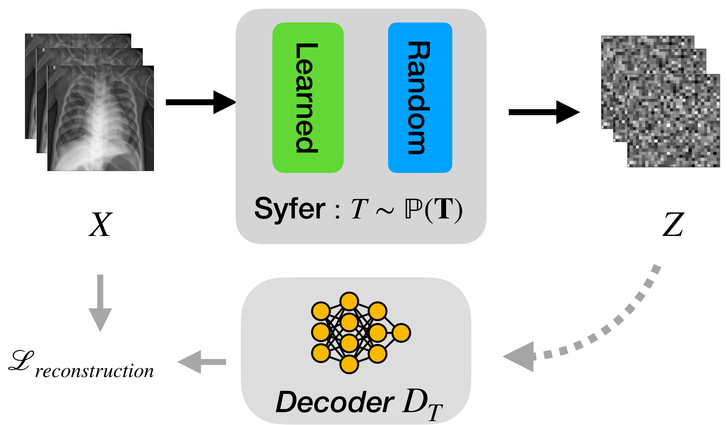
- **Fully Homomorphic Encryption** is private but the computation overhead renders DL intractable.
- **Differentially Private** mechanisms can achieve privacy at a large cost of utility (or vice-versa).
- **Lightweight encoding schemes** allow DL training but are not private.
- **GAN-based approaches** require training on private data.

Method

Privacy Estimation via Contrastive Learning: we use a **model-based attacker** to estimate the re-identification risk associated with the data release.



Syfer parametrizes the transformation T as a neural network where **learned weights** are trained to leverage subsequent **random layers** to fool the attacker and provide privacy.



Training is done on a public unlabeled dataset in an adversarial fashion with the attacker.

$$\mathscr{L}_{Syfer} = \mathscr{L}_{reconstruction} - \mathscr{L}_{reidentification}$$

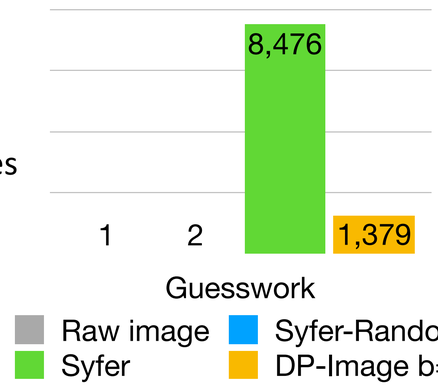
Results

Evaluation using chest X-rays images:

- Trained on the NIH Dataset
- Evaluated for **privacy** and **utility** on the MIMC dataset (heldout dataset, heldout attacker architectures and heldout tasks)

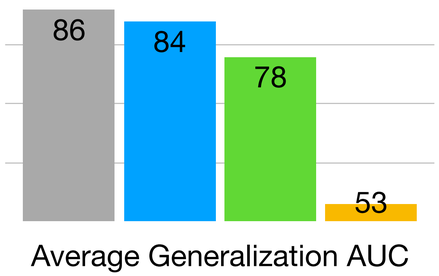
Achieves privacy:

We measure privacy using **guesswork**, i.e. the number of guesses an attacker takes to re-identify a single correct match (x, z).



Preserves utility:

We measure utility as the **generalization AUC** on downstream tasks (Edema, Cardio-megaly, Consolidation, Atelectasis).



Contributions:

- ☑ A novel threat model in accordance to HIPAA to enable data release and outsource model training
- ☑ A guesswork-based privacy evaluation framework which captures a worst-case scenario
- ☑ An efficient attacker to empirically eval privacy
- ☑ A learned encoding scheme with improved privacy-utility tradeoffs.