Application Testing of Generative Adversarial Privacy

Nicholas Johnson(nickj@stanford.edu), Stephanie Sanchez(ssanche2@stanford.edu), Vishal Subbiah (svishal@stanford.edu)

Background and Motivation

Machine learning (ML) methods serve many purposes today. The most revered applications of ML are normally coupled with benevolent intentions such as adverting cyber attacks, classifying materials in images for security and health purposes, etc. But ML methods do not necessarily have to be applied for favorable causes. Inference attacks, for instance, are adversary learning methods that can infer private information about public information or data. For this reason it is essential to protect privacy by deterring adversarial machine learning.

Problem Statement

- **Generative Adversarial Privacy** (GAP): to protect the privacy of public data via distortion.
- **Goal**: to enable the protection of sensitive information via an autoencoder inside a generative adversarial network (GAN) in order to hinder inferences on sensitive data while not preventing the inference on non-sensitive attributes.

Data Preprocessing

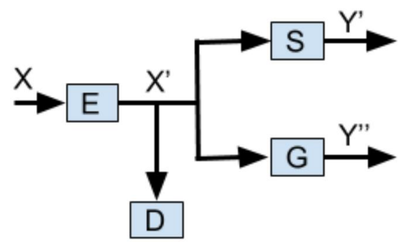
- **Reshaped** images to 256 pixels x 256 pixels x 3 colors (avg. image size of the dataset)
- **Gender and Smile Data**: cleaned labels to be 0 and 1 ({smile, no smile}, {male, female})

The Model

Methods:

- **E,G,S:** for encoder, gender classifier, and smile classifier respectively
- **D:** distortion metric (for further constraints on the encoder function)
- **GAP architecture:** we consider the encoder and classifiers as distinct entities

(GAP Architecture)



- **Deep NN Classifiers:** six convolution layers with 3 FNN (gender and smile models)
- **Compressive Encoder**: PCA was used as a lossy compression. Images reconstructed via the first **d** principle eigenvectors for each color channel.
- **Shallow Autoencoder:** Based off of [4] even complex models can be fooled via simple strategies. Our shallow autoencoder has a loss defined to maximize the loss in gender while minimizing the loss in smile, as opposed to the explicit alternating optimization scheme in [2].

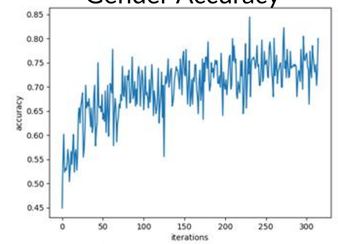
Results and Discussion

	GAN Accuracy	Gender Accuracy	Smile Accuracy
Autoencoder	24%	42%	64%
Encoder-PCA	26%	69%	36%
No Encoder	N/A	72%	74%

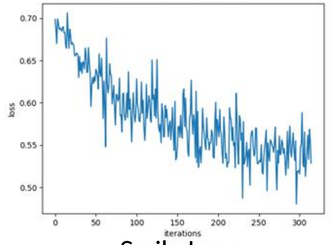
Our compressive encoder was implemented with PCA with the number of components equal to the number of nodes in the autoencoder. This size was chosen because the autoencoder is performing an information reduction of similar order.

The features for the Gender Model are more distinct than the Smile Model features, which explains the PCA results because the algorithm is reducing information in a non-incentivized way. Due to the loss we defined for the autoencoder it is able to be equally effective while targeting specific features.

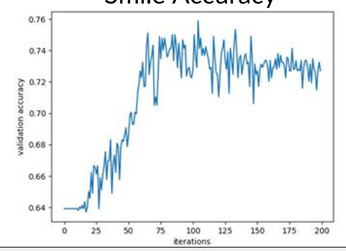
Gender Accuracy



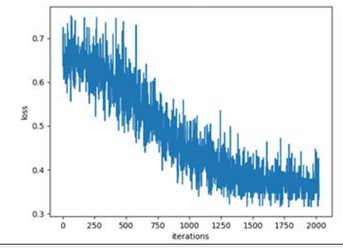
Gender Loss



Smile Accuracy



Smile Loss



Future Work

1. K-means compression algorithm as an additional compressive encoder.
2. More complex networks for classification.
3. Loss function identification, to verify the loss function is achieving the true privacy measure desired.
4. Add random noise to the encoder to make it non-reversible, limits the information to the classifiers.

References

1. Ari Ekmekji. *Convolutional Neural Networks for Age and Gender Classification*.
2. Jihun Hamm. *Minimax Filter: Learning to Preserve Privacy from Inference Attacks*.
3. G. Cybenko. *Approximation by superpositions of a sigmoidal function*.
4. Jiawei Su, Danilo Vasconcellos Vargas, Sakurai Kouichi. *One pixel attack for fooling deep neural networks*.