Face Encryption

SDS3386 Data Science Lab

Fall 2022

Team TensorOverflow

Xiang Li 300056427
Jiaxun Gao 300063462
Bowen Zeng 300115382

Course Coordinator: Tanya Schmah

November 30, 2022

University of Ottawa

uOttawa

Contents

1 Introduction

Big data privacy risks in public social media are confirmed by recent reports (Smith, Szongott, Henne, & von Voigt, 2012). The quantity of user-generated content being uploaded to the internet is growing quickly, yet some big companies, like Google and Facebook, have been abusing it without their knowledge (Esteve, 2017). Privacy protection is challenging when users submit sensitive images of their faces. In order to train their machine learning algorithms, the corporation might use these photographs, which could result in privacy breaches. **How can we protect privacy when sharing facial images?**
In this project, we provide potential solution based on adversarial attacking and facial recognition technology to safeguard users' privacy. After identifying the face in the image using the facial recognition model, we encrypt the face with noise. The generated image can be as clear to the human eye as the original, but the facial recognition model will identify a different person in it.

2 Dataset

2.1 CelebA

CelebA is a large-scale face attributes dataset with more than 200K celebrity images, each with 40 attribute annotations. The images in this dataset cover large pose variations and background clutter. CelebA has large diversities, large quantities, and rich annotations, including **5,000 celebrity identities**, **202,599 face images**, and **40 binary attributes** annotations per image. The dataset can be employed as the training and test sets for the following computer vision tasks: face attribute recognition, face detection, landmark (or facial part) localization, and face editing/synthesis.

Figure 1: CelebA dataset

CelebA is utilised as the face recognition model's dataset set in this project. The training set has 4429 photos while the test set has 1267 images.

2.2 Private Dataset

The private dataset consists of photos collected by team members. The photos were taken in SITE and feature various facial expressions.
This dataset is utilized to perform adversarial attacking on the face recognition model. The training set has 121 photos while the test set has 15 images.

3 Methodology

3.1 Adversarial Attacks

Adversarial attacking is a technique that can be used to fool machine learning models. It is a type of attack that aims to change the input data in a way that the model will misclassify it. The adversarial attacking technique is based on the fact that machine learning models are vulnerable to small perturbations in the input data. The perturbations are usually imperceptible to the human eye, but they can cause the model to misclassify the input data.
Adversarial examples are hard to defend against because it is difficult to construct a theoretical model of the adversarial example crafting process. Adversarial examples are solutions to an optimization problem that is non-linear and non-convex for many ML models, including neural networks. Because we don't have good theoretical tools for describing the solutions to these complicated optimization problems, it is very hard to make any kind of theoretical argument that a defense will rule out a set of adversarial examples.

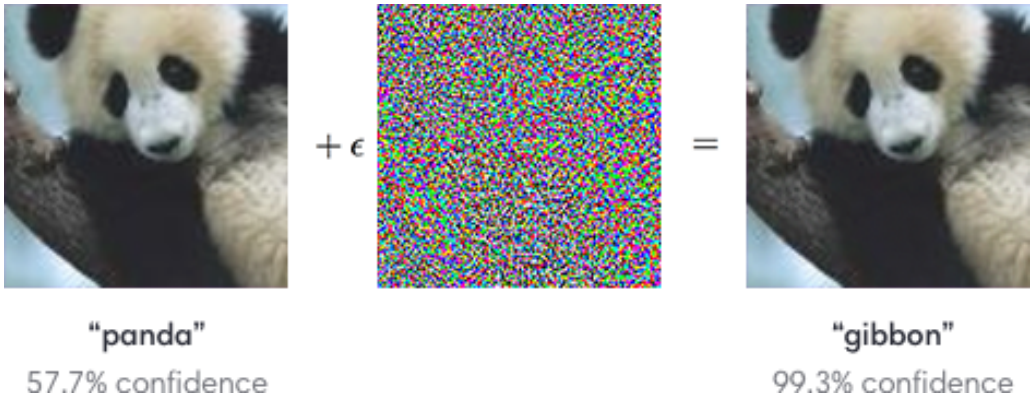

Figure 2: An adversarial input, overlaid on a typical image, can cause a classifier to miscategorize a panda as a gibbon.

Another reason is they require machine learning models to produce good outputs for every possible input. Most of the time, machine learning models work very well but only work on a very small amount of all the many possible inputs they might encounter. (Goodfellow, 2020)

3.2 Backbone Model: ResNet-18

ResNet-18 is a convolutional neural network (CNN) that is used as a backbone model in this project. The architecture can be illustrated as figure 3 (Ramzan et al., 2019). It is a 18-layer deep neural network that is trained on the ImageNet dataset. The ImageNet dataset is a large dataset that contains 1.2 million images with 1000 classes. The ResNet-18 model is trained on the ImageNet dataset to classify the images into 1000 classes.

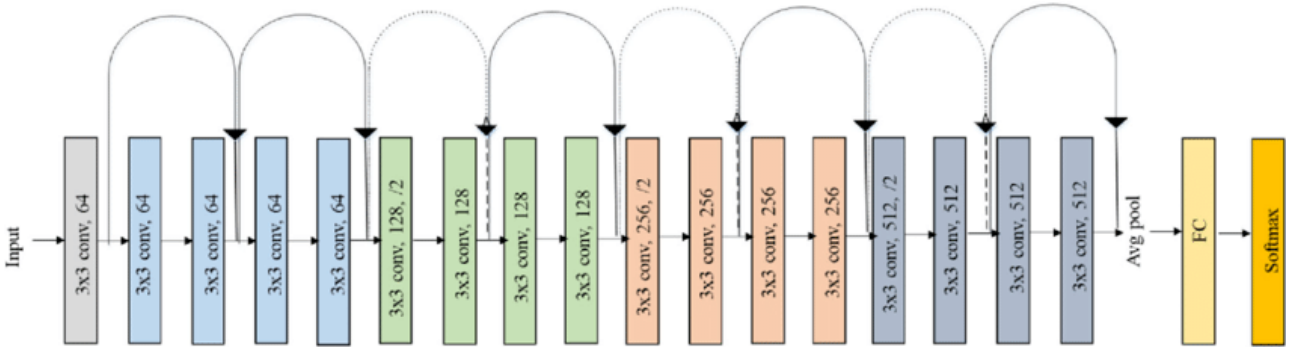

Figure 3: ResNet-18 Architecture

4 Data Wrangling

4.1 Data Collection

The CelebA dataset (introduced in section 2.1)

References

Esteve, A. (2017, 03). The business of personal data: Google, Facebook, and privacy issues in the EU and the USA. *International Data Privacy Law*, *7*(1), 36-47. Retrieved from https://doi.org/10.1093/idpl/ipw026 doi: 10.1093/idpl/ipw026

Goodfellow, I. (2020, Oct). *Attacking machine learning with adversarial examples.* OpenAI. Retrieved from https://openai.com/blog/adversarial-example-research/

Ramzan, F., Khan, M. U., Rehmat, A., Iqbal, S., Saba, T., Rehman, A., & Mehmood, Z. (2019, 12). A deep learning approach for automated diagnosis and multi-class classification of alzheimer's disease stages using resting-state fmri and residual neural networks. *Journal of Medical Systems*, *44*. doi: 10.1007/s10916-019-1475-2

Smith, M., Szongott, C., Henne, B., & von Voigt, G. (2012). Big data privacy issues in public social media. In *2012 6th ieee international conference on digital ecosystems and technologies (dest)* (p. 1-6). doi: 10.1109/DEST.2012.6227909