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

Contents

1 Introduction

Recent reports (Smith, Szongott, Henne, & von Voigt, 2012) confirm the privacy risks associated with big data in public social media. The amount of user-generated content uploaded to the internet is increasing rapidly, but large corporations such as Google and Facebook have been misusing it without their knowledge (Esteve, 2017). When consumers upload sensitive facial photographs, it is difficult to preserve their privacy. The company may use these pictures to develop their machine learning algorithms, which could result in privacy breaches. **How can we protect privacy when sharing facial images?**
In this research, we offer a viable privacy-protecting solution based on adversarial attacks and facial recognition technology. After detecting a face in a picture using a facial recognition model, we encrypt the face with noise generated by adversarial attacking model. The composed image may appear as clear as the original, but the facial recognition software will recognise a different individual in it.

2 Data Wrangling

2.1 Dataset

2.1.1 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.

2.1.2 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.

2.2 Faces extraction from recorded video

After capturing three videos, OpenCV2 is used to sample 40 frames from each video. The faces are then rotated 180 degrees to accommodate package mediapipe's face detection model. Then, we crop the faces from the frames using the centre of the bounding box's coordinates. The faces are then saved in a folder named `private_dataset` and added to `CelebA_HQ_facial_identity_dataset`.
This process is done in `Data_generation.ipynb.ipynb`.

2.3 Associate id with name

After downloading the CelebA dataset from the official website, we utilise CelebA-HQ-to-CelebA-mapping.txt to generate a map `hq_A_mapping` from an id to the image file name, such as `5: 000615.jpg`. Then, we use list_identity_celeba.txt to generate the second map `id_name_mapping` from a file name to its corresponding identity's name, for instance: `000615.jpg: Martha Hunt`.
The benefit of this process is that we can now use the id to determine an identity's name. By example, we may use the following code: `id_name_mapping[hq_A_mapping['5']]` to determine the name of the individual with id = 5.
This process is done in `Preprocessing.ipynb`.

2.4 Transforming the dataset

We resized the images to 224x224 because their original size was too large for our model, which could cause performance issues. After that, we augment the tensor by giving it a random horizontal flip as part of the transformation.
This process is a part of `simple_model.ipynb`.

3 Methodology and Modelling

3.1 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 2 (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.

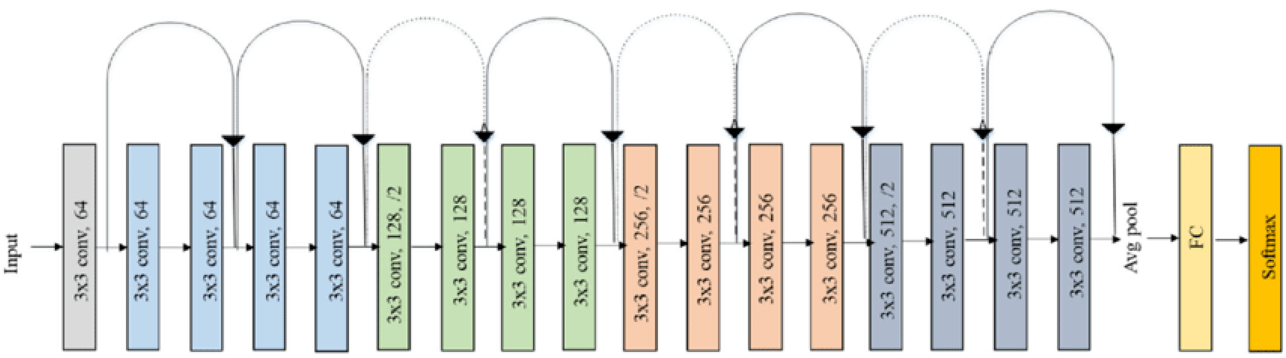


Figure 2: ResNet-18 Architecture

3.2 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.

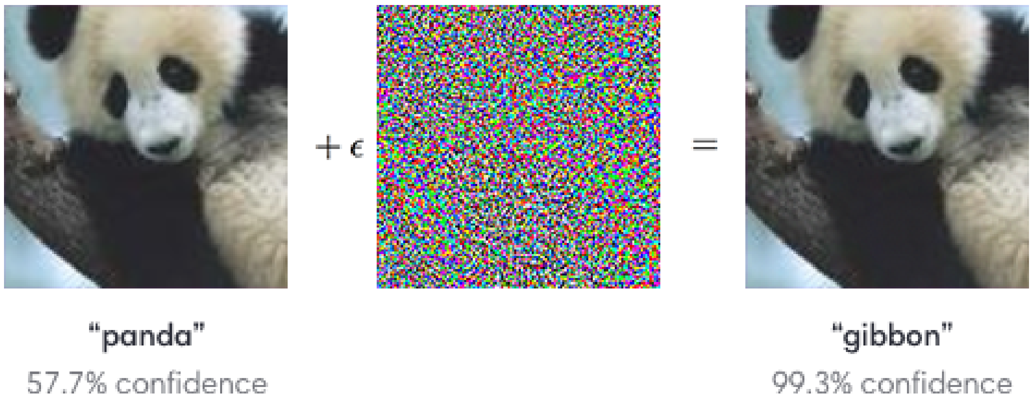


Figure 3: An adversarial input, overlaid on a typical image, can cause a classifier to miscategorize a panda as a gibbon.

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)

3.3 PGD Attack

The PGD attack is a white-box attack which means the attacker has access to the model gradients i.e. it knows every weight in ResNet-18 in our project. This model gives the attacker much more power than black box attacks as they can specifically craft their attack to fool the face recognition model without having to rely on transfer attacks that often result in human-visible perturbations. PGD can be considered the most "complete" white-box adversary as it lifts any constraints on the amount of time and effort the attacker can put into finding the best attack. (Knagg, 2019).

3.4 Face Recognition System Implementation

Before we can achieve our objective, we need a face recognition system to use as an attacking target. To recognise faces, we employ a pre-trained ResNet-18 model in pytorch. In the neuron network, we use cross entropy loss as the loss function and stochastic gradient descent (SGD), a simple yet highly effective method for fitting linear classifiers and regressors under convex loss functions, as the optimizer. The model is trained for 10 iterations on the CelebA-HQ dataset.

3.5 PGD Attack Modelling

To generate adversarial examples, we use the PGD attack (Madry, Makelov, Schmidt, Tsipras, & Vladu, 2017). In this project, we use the gradient of the loss function to generate adversarial examples. The attack is iterative and uses a step size to determine the size of the perturbation. The attack is also constrained by a maximum perturbation size.

3.5.1 Non-targeted Attack

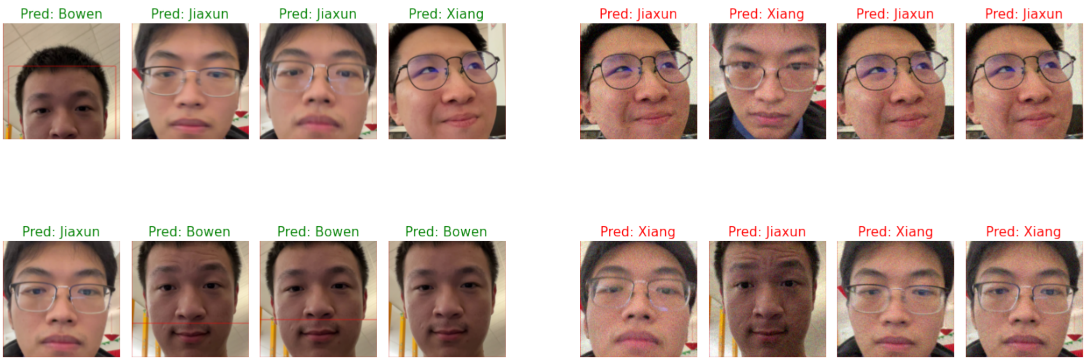
In a non-targeted attack, the goal is to generate an adversarial example that is misclassified by the target model A. Which means we are **maximizing** the loss function with respect to the target class A. The attack is as follows:

1. Generate a random noise tensor δ with the same shape as the input image.
2. Calculate the gradient of the loss function with respect to the noise tensor δ.
3. Add the gradient to the noise tensor δ.
4. Clip the noise tensor δ to the range $[-\epsilon, \epsilon]$.
5. Add the noise tensor δ to the input image.
6. Repeat steps 2 to 5 until the model misclassifies the image.

The loss function (maximizing the difference with the true label) can be represented as:

`Difference(Model(Face Image + Delta), True Label)`

where `Model(Face Image + Delta)` is the prediction of perturbed data.



(a) Prediction without attack *(b) Prediction with non-targeted attack*

Figure 4: Example of a non-targeted attack

With above steps, we can generate an adversarial example that is misclassified by the target model. In figure 4, the image on the left demonstrates that Resnet-18 correctly predicted the original photos. After adding the PGD model's generated noise. The image on the right depicts Resnet-18 classifying the adversarial example as a different category.

3.5.2 Targeted Attack

In a targeted attack, the goal is to generate an adversarial example that is misclassified A by the target model and classified as a specific class B. All the steps are the same as in a non-targeted attack, except for the last step should be: **Repeat steps 2 to 5 until the model classifies it as the target class.** Which means we are **minimizing** the loss function with respect to the target class B.
The loss function (minimizing the difference with the target label) can be represented as:

`Difference(Model(Face Image + Delta), Target Label)`

4 Visualisation and Real-Time Application

4.1 Resnet-18 Performance

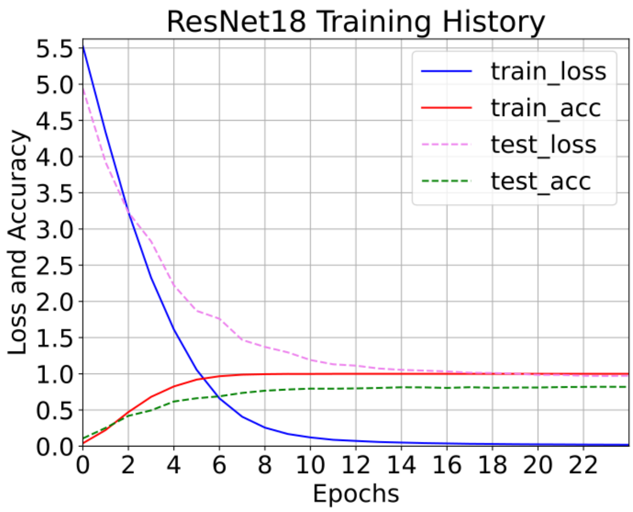


Figure 5: Resnet18 performance

Figure 5 demonstrates that with 25 epochs, the face recognition model Resnet-18 achieves 99.98% accuracy on the training set and 81.93% accuracy on the test set. We halted training at this point because any additional epoch would result in an overtraining problem.

4.2 PGD Attack Real-Time Application

Using the method described in the preceding section, we developed a program to execute PGD attacks through the camera in real time using the method described in the previous section. Before feeding them into the machine learning model, the application would capture images from the camera and then subject them to PGD perturbations. The output of the model would be monitored and the probability of the prediction would be displayed on the screen. This method allows for the testing and evaluation of machine learning models' resistance to PGD attacks in real time.

5 Discussion and Conclusion

Our application demonstrates how adversarial attacks can protect the privacy of facial image sharing. By applying perturbations to the input images, our application was able to cause the machine learning model to make errors, potentially protecting the identity of the individuals depicted in the images. This is an important consideration when sharing images online, as there are often concerns regarding personal privacy and the possibility of malicious actors gaining access to sensitive information.

6 Ethical Considerations

The ethical considerations of this project are related to the malicious usage of the targeted-attack application. Specifically, there is a concern that the model could be used to reidentify an individual's photos and attach them to another person's identity, potentially leading to serious consequences for the misidentified individual. This type of malicious use could lead to reputational harm, financial harm, or other types of harm to the individual who is falsely identified. It is important that measures are put in place to prevent this type of misuse, and that appropriate safeguards are in place to protect the privacy and security of individuals whose photos may be used in the model.

7 Future Work

There are multiple possible future directions for this project's development. One option is to explore the use of more realistic attack scenarios, such as blackbox attacks, which more closely reflect the types of real-world threats that individuals and organizations may face. This may include the development of new techniques for evaluating the robustness of the model under various attack scenarios, as well as the exploration of new algorithms and methods for enhancing the model's performance in these scenarios. The development of a server-client application that processes photos on the server, rather than on the user's device, is another avenue for future research. This could mitigate some of the ethical concerns associated with the malicious use of the targeted-attack application, as it would permit greater control and oversight over the model's application.

A Contributions

XiangLi and Jiaxun took the lead in the modelling and live demo.
Bowen took the lead in the report.

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/

Knagg, O. (2019, Jan). *Know your enemy.* Towards Data Science. Retrieved from https://towardsdatascience.com/know-your-enemy-7f7c5038bdf3

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

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