Image Anonymization using Adversarial Attacks

Gautam Kumar (B19EE031) and Borkar Ankur (B19CSE026)

Introduction

In the digital age we live in now, more and more image data has brought both rewards and problems. Even though images have useful information for analyzing and understanding, they can also have private or personal information that needs to be kept safe. Image anonymization methods are a key part of protecting people's privacy and making sure that data protection laws are followed. This project's goal is to look into the field of picture anonymization and come up with a new way to do it using the Projected Gradient Descent (PGD) attack. By using methods for adversarial robustness, we can change images in a way that keeps their visual content but makes it hard to tell what they are.

Image anonymization is important for protecting privacy in today's world where images are widely shared and analyzed. It involves removing or changing personal details like faces or license plates so that individuals' identities are safeguarded, and their information is not misused or accessed without permission. It helps organizations comply with privacy regulations like the GDPR, ensuring sensitive data isn't exposed. Anonymizing images is also crucial for ethical AI, as it prevents biases, unintended disclosure of sensitive information, and helps build fair and unbiased machine learning models that use image data.

Image anonymization has important applications in various areas. In healthcare, it is used to protect patients' identities and medical records when sharing medical images for research or collaboration. In public surveillance, anonymization techniques can blur faces captured by video surveillance systems, ensuring individuals' identities are safeguarded while still allowing analysis for security purposes. Social media platforms use image anonymization to protect user privacy, preventing unauthorized identification or misuse of personal data. Additionally, researchers benefit from anonymizing images before sharing datasets, as it ensures privacy while supporting advancements in fields like computer vision and machine learning.

Related Work

1. "DeepPrivacy: A Generative Adversarial Network for Face Anonymization" by M. R. Souza et al. (2018) - This work introduced a deep learning-based approach using Generative Adversarial Networks (GANs) to anonymize faces in images while preserving their visual quality.

2. "Privacy-Preserving Deep Learning" by C. Dwork and A. Roth (2014) - The paper discusses techniques for privacy-preserving machine learning, including differential privacy, secure multi-party computation, and homomorphic encryption. It addresses the challenges of protecting sensitive data while still allowing collaborative learning.

3. "A Survey of Anonymization Techniques for Privacy-Preserving Publishing of Surveillance Video Data" by N. K. Ahmed et al. (2017) - This survey paper provides an overview of various anonymization techniques applied to video surveillance data, including face blurring, pixelization, and geometric transformations. It evaluates the effectiveness of these methods and discusses their limitations.

4. "Anonymizing Biometric Data using Generative Adversarial Networks" by S. Kulkarni et al. (2020) - This work proposes a method based on Generative Adversarial Networks (GANs) to anonymize biometric data, specifically iris and fingerprint images. The approach generates synthetic samples that retain statistical properties while ensuring privacy protection.

5. "Privacy-Preserving Image Anonymization: A Comparative Analysis of Techniques" by V. K. Singh and M. Dave (2019) - The paper presents a comparative analysis of different image anonymization techniques, including blurring, pixelation, and texture synthesis. It evaluates their effectiveness in preserving privacy while maintaining image quality.

Adversarial Attacks

In mobile and pervasive computing, adversarial attacks are when bad actors purposely try to harm or exploit weaknesses in mobile devices and their systems. They do this by creating malicious software, taking advantage of vulnerabilities in

networks, tricking people into giving away information, or tampering with mobile applications. These attacks can have serious consequences, like stealing personal information, damaging the device's functionality, or disrupting network communication. For example, attackers might create fake apps that look real but are actually designed to steal sensitive data. They might also intercept data being sent between devices and networks to gain unauthorized access or manipulate the information.

FGSM Attack

The Fast Gradient Sign Method (FGSM) is a popular adversarial attack technique used to fool or deceive machine learning models. It perturbs input data by adding small, calculated perturbations in the direction of the gradient of the loss function with respect to the input. The resulting perturbed data, also known as adversarial examples, are designed to cause misclassification or incorrect predictions by the targeted model.

Mask Attack

Adversarial attacks in this context involve manipulating data, such as images or text, with the aim of deceiving or tricking machine learning models used in mobile and pervasive computing systems. These attacks aim to introduce subtle changes to the input data that are not easily detectable by humans but can cause the targeted machine learning models to produce incorrect or unexpected outputs. This can have serious consequences in mobile and pervasive computing scenarios, such as compromising the security of personal data, disrupting the functionality of mobile devices, or misleading the decisions made by the underlying machine learning algorithms.

PGD Attack

The PGD (Projected Gradient Descent) attack is a powerful iterative technique used in adversarial machine learning to generate adversarial examples. It is an extension of the Fast Gradient Sign Method (FGSM) and aims to create more robust and effective adversarial perturbations. In the PGD attack, instead of making a one-time perturbation to the input, multiple iterations are performed to refine the perturbation. The process involves taking small steps in the direction of the gradient of the loss function with respect to the input, while constraining the perturbation to stay within a predefined range or budget. At each iteration, the perturbation is

projected back onto the allowed range to ensure it remains within the specified boundaries.

The PGD attack is more powerful than the FGSM because it explores a larger space of possible perturbations. By performing multiple iterations and projecting the perturbation back onto the feasible range, the attacker can find adversarial examples that are more likely to fool the targeted machine learning model while still being within the defined limits.

RESULTS

We trained 2 state of the art deep learning models (Squeezenet and ShuffleNet) on CIFAR-10 dataset and saw the effect of various attacks on these datasets with different epsilon values.

Results after training (without attacks):
Squeezenet:
Train Accuracy: 0.5633
Validation Loss is 1.030015
Validation Accuracy is 0.691072526

Shufflenet:
Train Accuracy: 0.7434
Validation Loss is 0.4065124
Validation Accuracy is 0.87734

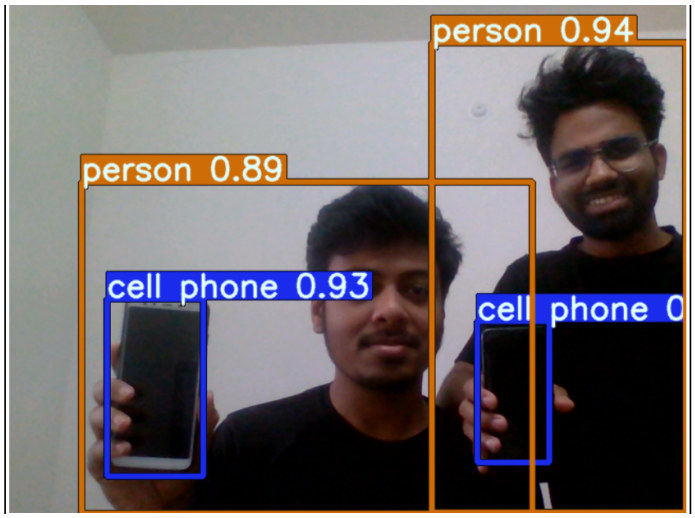
Attacks	Squeezenet Accuracy		Shufflenet Accuracy	
Epsilon	Mask Attack	PGD Attack	Mask Attack	PGD Attack
0.001	0.54	0.54	0.63	0.63
0.01	0.52	0.47	0.55	0.40
0.03	0.48	0.35	0.44	0.18

0.05	0.43	0.26	0.37	0.10
0.1	0.35	0.15	0.28	0.05

Analysis of Results:

From the results obtained, we can observe that PGD attack is more powerful than the mask based attack. PGD attack uses a more sophisticated approach to generate adversarial examples. It uses an iterative process to find the optimal perturbation that can fool a deep learning model, while the mask based attack only applies a single perturbation. PGD also allows for more flexibility in terms of the magnitude of the perturbation, which makes it more effective in fooling deep learning models.

We also performed the perturbation on YOLO-v7 detections using the PGD attack and observed the changes. We were successfully able to misclassify the given entities in a given image. The corresponding confidence score also goes down drastically.



Original Sample

Perturbed Sample

Human Based Security

So far we have perturbed the samples only to ensure model based privacy, meaning a deep learning algorithm would fail to classify the images on the ground truth class. We can extend this to human based privacy by also adding gaussian noise to the input image.

Original image

Perturbed and Blurred Image

This hides the important features of the image and protects it from human based recognition. We can also increase the extent of blurriness as per the requirements.

CONCLUSION

Our image anonymization project utilizing the PGD attack has yielded successful results in generating perturbations and introducing changes to images. We observed that as we increased the epsilon value, which determines the magnitude of the perturbation, the test accuracy of the machine learning model decreased. This indicates that the generated perturbations were effective in altering the model's predictions and potentially anonymizing the images. The decrease in test accuracy demonstrates the vulnerability of the model to adversarial attacks and highlights the importance of considering robust defenses against such attacks in image anonymization techniques. It suggests that the perturbations introduced through the PGD attack can effectively deceive the model and potentially protect the privacy and identities of individuals in the images.

Our project contributes to the field of image anonymization by demonstrating the potential of the PGD attack and its impact on test accuracy. The findings emphasize the need for robust defenses and careful considerations when implementing image anonymization techniques to ensure both privacy protection and reliable model performance.