Bridging the Gap: Advancing Image Privacy with DeepPrivacy3 using Deepfake-Inspired Techniques

Authors: Sandhya Venkataramaiah, Joseph Zou

Table of Contents

- Introduction
- Problem Statement
- Motivation & What makes this paper different?
- Algorithm
- Experimental setting
- Conclusion
- Questions

Introduction

Research Proposal

- This research is a proposal of future work on needs to be done
- We give ideas for future researchers because we don’t have the skills to implement it ourselves

Problem Statement

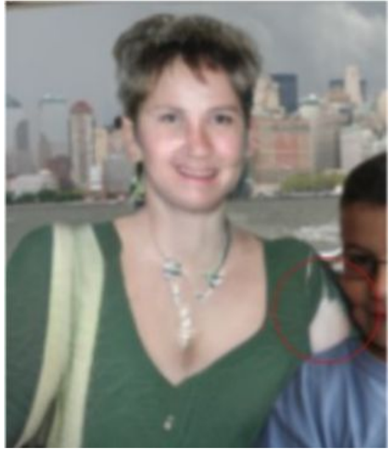
Demo

https://youtu.be/fKmthmTmbjE?feature=shared

https://youtu.be/iyiOVUbsPcM?feature=shared

Current Issues with DeepPrivacy2

- Video Quality is poor
  - Does not consider temporal dynamics
  - No frame to frame consideration
  - No optical flow for pixel estimation
- Slight distortions
- Random artifacts



Motivation & What makes this paper different?

Learn from DeepFakes

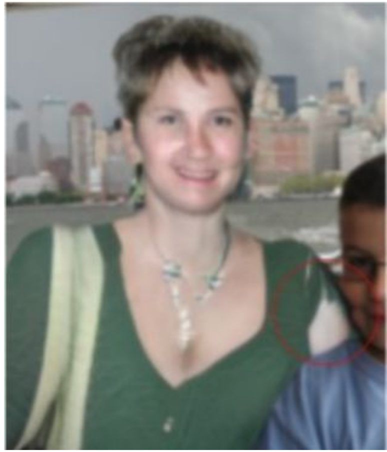
- Both deepfakes and image anonymization achieve the same goal of hiding the source image/video
- Yet Deepfakes produce higher quality images and videos, Why?
- We hope to integrate the algorithms and techniques used in Deepfakes to improve Deepprivacy2.

What makes our paper different?

- So far no research has been done on using Deepfakes as a way to improve image anonymization
- Introduce some concerns of deepprivacy2 and questions to be addressed
  - Is deepprivacy2 robust enough to anonymized an already anonymized image?
- Introduce the need to use video datasets

What is Deepfakes and what algorithms?

- temporal coherence, optical flow estimation,
- Total Variation regularization and pixel-wise loss functions
- exhibit pixelation and artifacts due to rapid changes in pixel values, our approach seeks to mitigate these issues through a combination of loss functions and regularization techniques.



Key Components of Deepfake Technology

- **Pixel to Pixel loss:** Loss function computes the pixel-to-pixel loss of the prediction and the target images.
- **Perceptual Loss Function:** Comprised of content loss, measuring content similarity, and style loss, assessing style or texture similarity between the generated and target images for high fidelity reproduction.
- **Temporal Coherence:** Ensures consistent facial movements and expressions over time in video sequences through temporal regularization, minimizing abrupt changes across frames.
- **Optical Flow Estimation:** Crucial for tracking and aligning facial movements across video frames, ensuring smooth and natural transitions by analyzing object motion.

Algorithm

Generative Adversarial Networks (GANs)

A class of machine learning frameworks where two neural networks contest with each other in a game (formulated by Ian Goodfellow and his colleagues in 2014)

1. Components:
   - Generator (G): Creates fake data that looks like the real data.
   - Discriminator (D): Evaluates data for authenticity; real or fake.
2. How They Work:
   - The Generator generates a data instance.
   - The Discriminator evaluates it.
   - Both networks learn through this process, improving their methods with each iteration.
3. Applications:
   - Art and Image Generation
   - Photo Realistic Images
   - Modeling and Simulation

Implementation Details

- **Environment Setup:** Utilization of TensorFlow and Google Colab for computational efficiency, with GPU verification to ensure optimal performance.
- **Data Handling:** Employed the CelebA dataset, specifically utilizing 5,000 images to train the model, ensuring a rich variety of facial features for generative tasks. Automated dataset mounting from Google Drive, verification of dataset existence, and preparation including copying files to the local Colab environment for faster processing.
- **Image Processing Techniques:**
  - Image Resizing: Images resized to 128x128 using Lanczos resampling—a high-quality downsampling filter—to maintain image quality.
  - Normalization: Images normalized to the range of [-1, 1] to facilitate efficient model training by centering data.

Model Architectures

- **Generator Model:** Sequential model beginning with a dense layer, batch normalization, and LeakyReLU activation, followed by multiple Conv2DTranspose layers for upscaling to the desired output resolution.
- **Discriminator Model:** Comprised of Conv2D layers with downsampling, LeakyReLU activations, and dropout layers to prevent overfitting, culminating in a dense output layer for real/fake image classification.

Training Process

- **Loss Functions:** Binary cross-entropy implemented to quantify the difference between the predicted and true labels, integral for training the generative and discriminative aspects of the GAN.
- **Optimizer Setup:** Use of Adam optimizer for both models, ensuring efficient backpropagation.
- **Training Steps**: Each training step involves generating images, assessing them with the discriminator, computing loss, and updating model weights. The process is repeated for multiple epochs to improve the model's ability to generate realistic images.It includes gradient accumulation for stable learning

Advanced Training Techniques

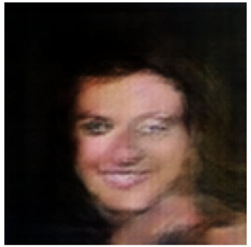
- **Mixed Precision Training:** Applied to accelerate computation further and reduce memory usage, crucial for training large models efficiently.
- **Gradient Accumulation:** Strategy used to stabilize training in environments with limited computational resources.

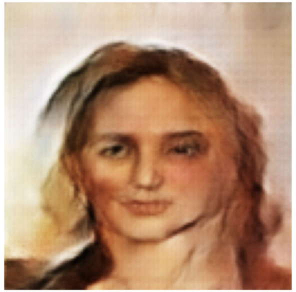
Experimental setting

Experimental Results:

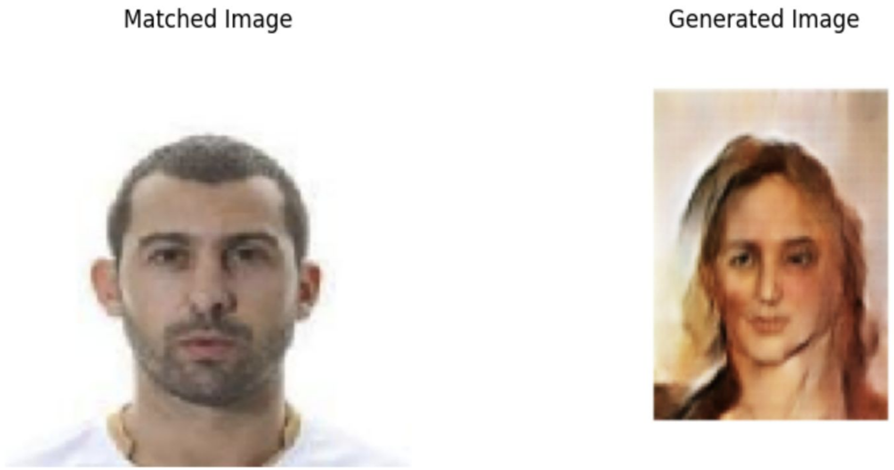
- **Image Generation:** Real-time generation and saving of images post-training to evaluate the visual quality improvements.
- **Quality Metrics:** Application of Structural Similarity Index (SSIM) to compare generated images against real ones, ensuring high fidelity.
- Optionally, use facial recognition tools like DeepFace to verify that faces in anonymized images cannot be matched back to the original subjects.

Output





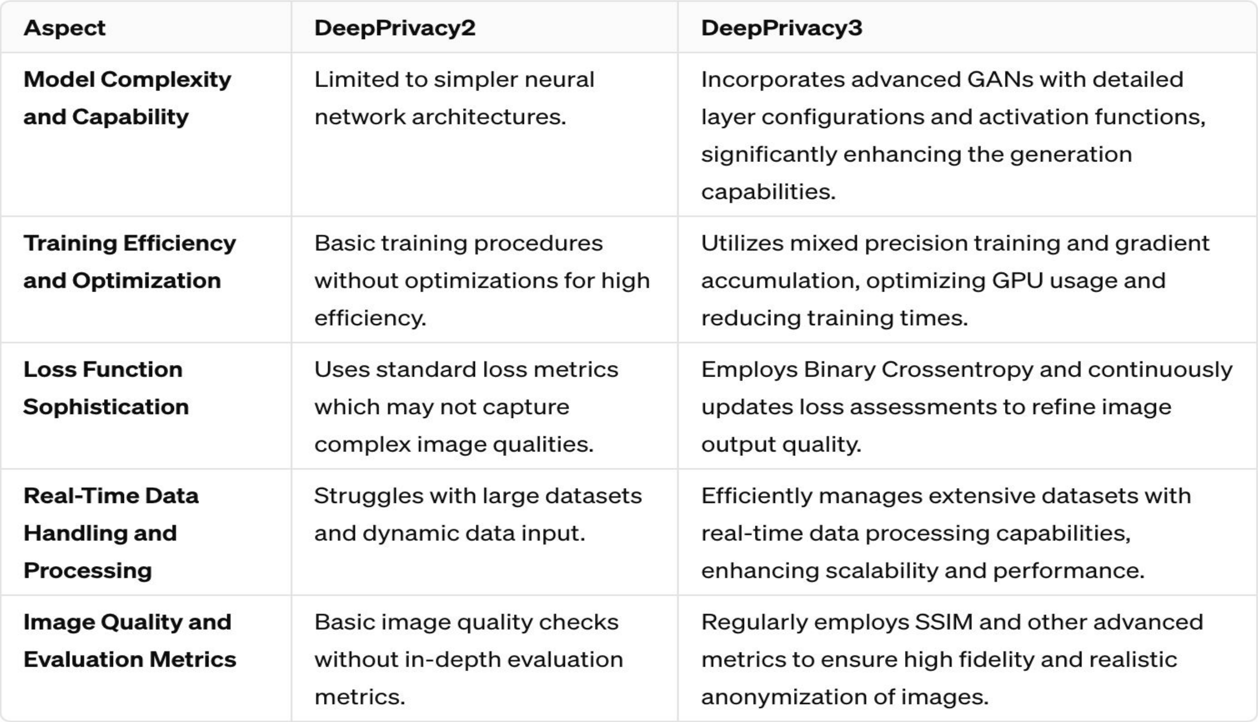
```
Average SSIM between original and generated images: 0.1819963902235031
```



```
Face similarity score: False
```

```
Final Epoch 1000: Generator Loss = 2.4375, Discriminator Loss = 0.81005859375
```

Technical Comparison with DeepPrivacy2



Aspect	DeepPrivacy2	DeepPrivacy3
Model Complexity and Capability	Limited to simpler neural network architectures.	Incorporates advanced GANs with detailed layer configurations and activation functions, significantly enhancing the generation capabilities.
Training Efficiency and Optimization	Basic training procedures without optimizations for high efficiency.	Utilizes mixed precision training and gradient accumulation, optimizing GPU usage and reducing training times.
Loss Function Sophistication	Uses standard loss metrics which may not capture complex image qualities.	Employs Binary Crossentropy and continuously updates loss assessments to refine image output quality.
Real-Time Data Handling and Processing	Struggles with large datasets and dynamic data input.	Efficiently manages extensive datasets with real-time data processing capabilities, enhancing scalability and performance.
Image Quality and Evaluation Metrics	Basic image quality checks without in-depth evaluation metrics.	Regularly employs SSIM and other advanced metrics to ensure high fidelity and realistic anonymization of images.

Conclusion

Conclusion

- Achievements of Our Work:
    - Developed and implemented a prototype of DeepPrivacy3 utilizing advanced GAN architectures, enhancing the ability to generate realistic anonymized images and videos.
    - Successfully integrated cutting-edge techniques such as mixed precision training and gradient accumulation to optimize performance and efficiency.
    - Employed sophisticated loss functions like Binary Crossentropy to refine the training process, resulting in higher quality and more realistic anonymization.

Future Work

- **Dataset Enhancement:** Expand training datasets to include a broader range of scenarios, ensuring robust performance across diverse video content.
- **Real-Time Anonymization Capabilities:** Develop and optimize real-time processing techniques to enable live video anonymization with minimal latency.
- **Algorithm Optimization:** Continuously refine algorithms and explore innovative neural network architectures to enhance image quality and reduce artifacts.
- **Enhanced Security Measures:** Strengthen anonymization against reverse engineering and emerging de-anonymization methods to uphold stringent privacy standards.
- **Interactive User Feedback System:** Implement mechanisms to gather and integrate user feedback on anonymization effectiveness directly into the model refinement process.

Thank You :)