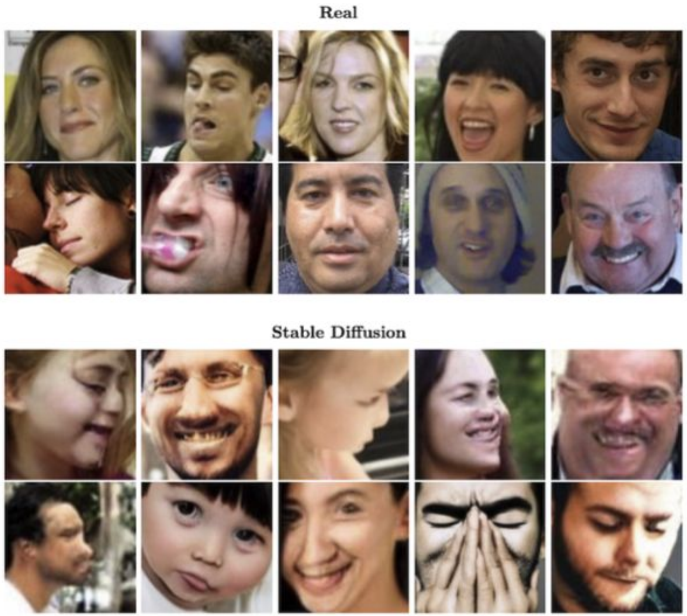
IMMA: A New Frontier in Securing Generative Models

Heesoo Kim

Introduction



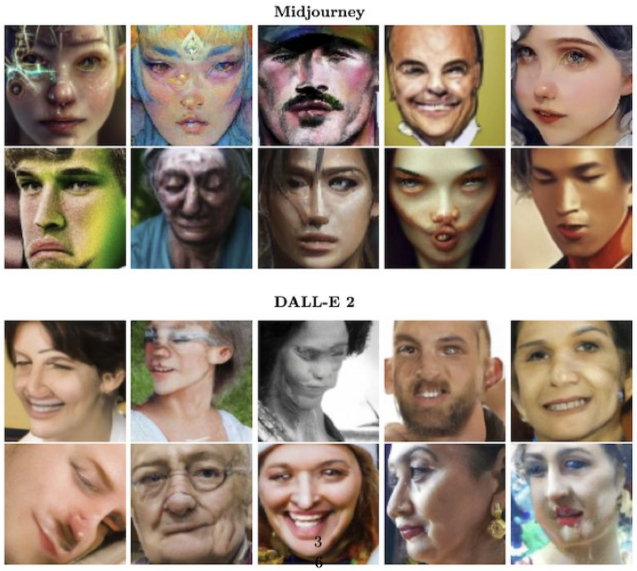
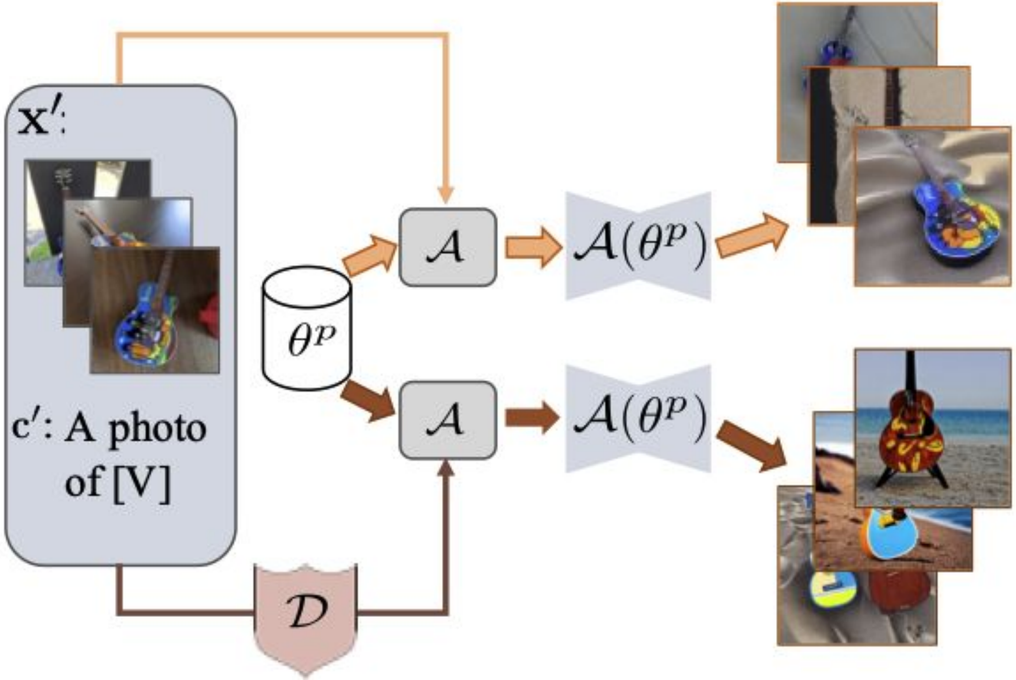


Figure 3: Samples of real faces (top row) and generated faces.

Borji, A. (2022). Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2. *arXiv preprint arXiv:2210.00586*.

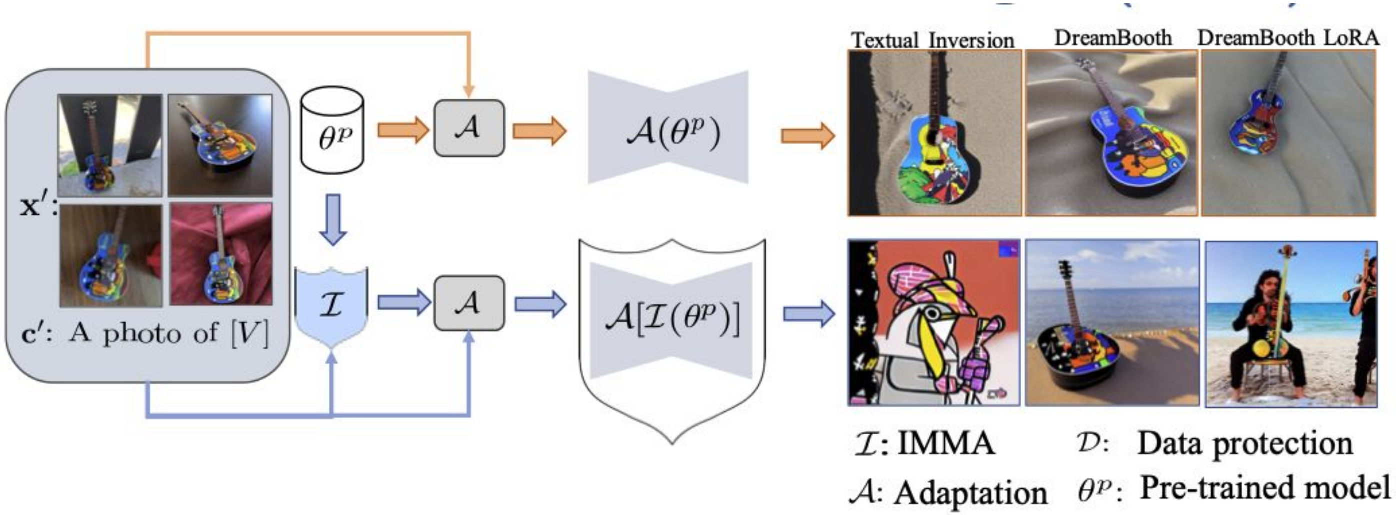
Problem Statement

Data Poisoning



Algorithm

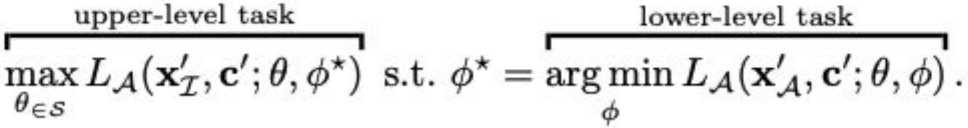
Model Immunization



$\mathbf{x}'$:
$\mathbf{c}'$: A photo of $[V]$
θ^p
$\mathcal{A}$
$\mathcal{A}(\theta^p)$
$\mathcal{I}$
$\mathcal{A}$
$\mathcal{A}[\mathcal{I}(\theta^p)]$
Textual Inversion
DreamBooth
DreamBooth LoRA
$\mathcal{I}$: IMMA
$\mathcal{D}$: Data protection
$\mathcal{A}$: Adaptation
θ^p: Pre-trained model

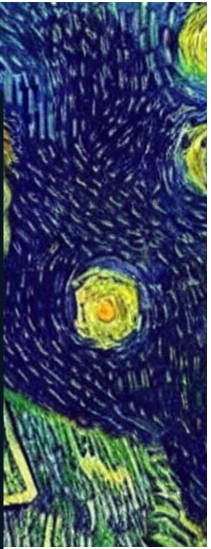
Algorithm

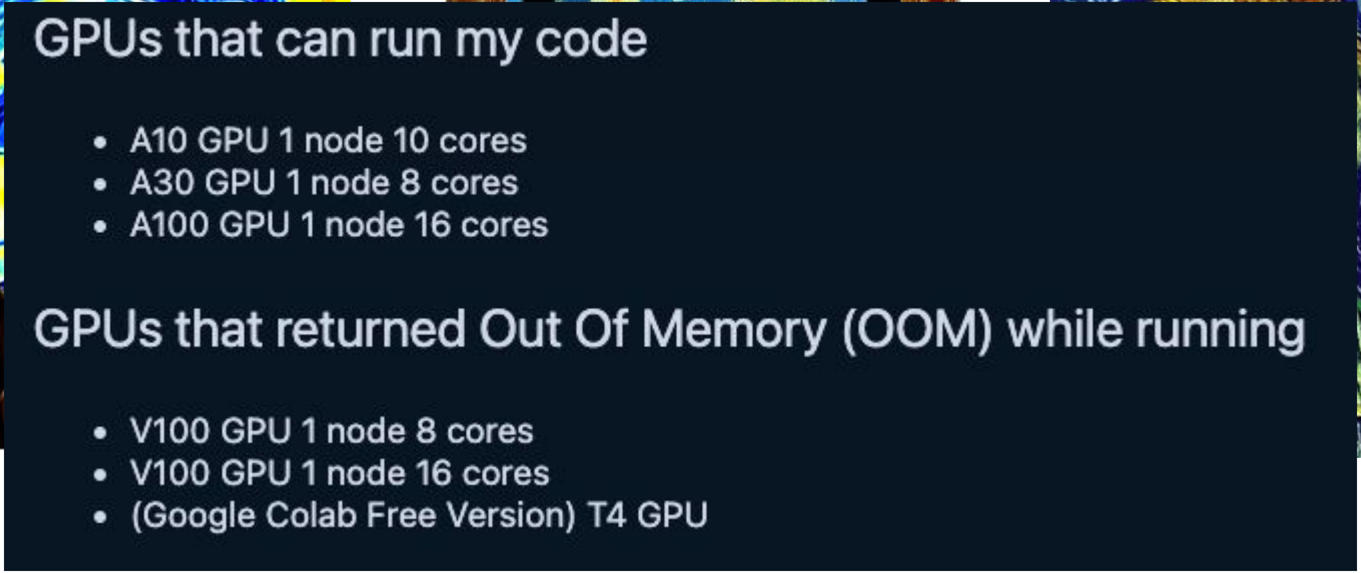
Model Immunization



$$\overbrace{\max_{\theta\in\mathcal{S}} L_{\mathcal{A}}(\mathbf{x}'_{\mathcal{I}}, \mathbf{c}'; \theta, \phi^{\star})}^{\text{upper-level task}} \text{ s.t. } \phi^{\star} = \overbrace{\arg\min_{\phi} L_{\mathcal{A}}(\mathbf{x}'_{\mathcal{A}}, \mathbf{c}'; \theta, \phi)}^{\text{lower-level task}}.$$

Methodology Setup





GPUs that can run my code

- A10 GPU 1 node 10 cores
- A30 GPU 1 node 8 cores
- A100 GPU 1 node 16 cores

GPUs that returned Out Of Memory (OOM) while running

- V100 GPU 1 node 8 cores
- V100 GPU 1 node 16 cores
- (Google Colab Free Version) T4 GPU

Experimental Results

Limitations

Making the model light was the most difficult part.

Changed weights and input size into fp16

Resolution

Batch size

Training epoch …

Future Direction

Extend IMMA to multiple adaptation margets and other generative models

Combine IMMA with data-centric approaches for layered protection

Current Bi-level optimization is computationally intensive, so we can work on that

Q&A