ML in PL CONFERENCE 2023

Towards More Realistic Membership Inference Attacks on Large Diffusion Models

Antoni Kowalczuk[1,2*], Jan Dubiński[1*], Stanisław Pawlak[1], Przemysław Rokita[1], Tomasz Trzciński[1,3,4], Paweł Morawiecki[5]

*equal contribution

1 POLITECHNIKA WARSZAWSKA WARSAW UNIVERSITY OF TECHNOLOGY

2 GOLEM KOŁO NAUKOWE SZTUCZNEJ INTELIGENCJI

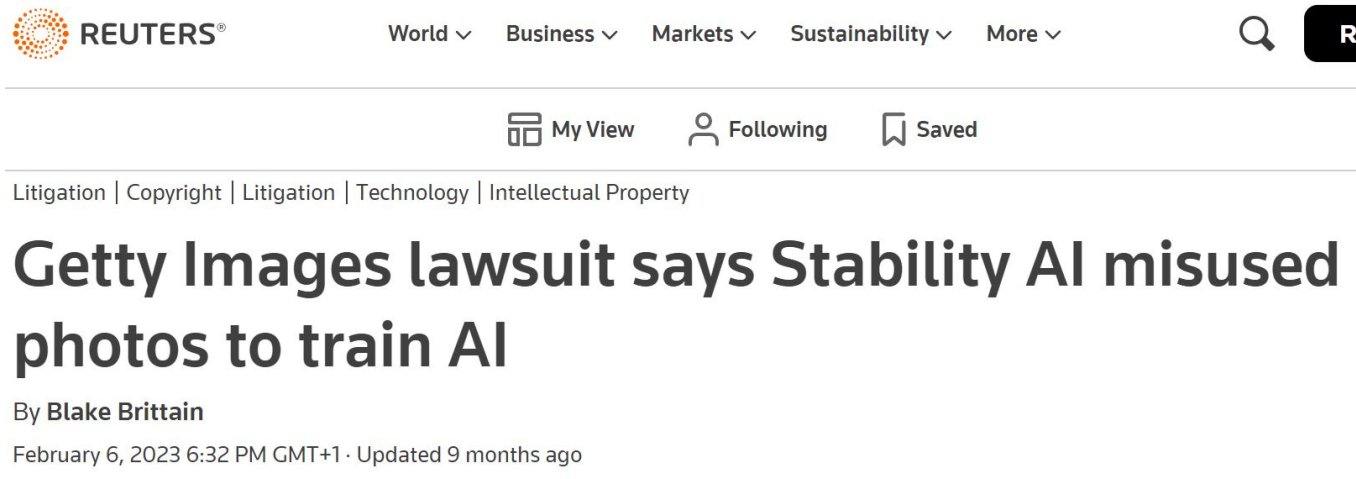
3

4 IDEAS NCBR

5 PAN POLSKA AKADEMIA NAUK

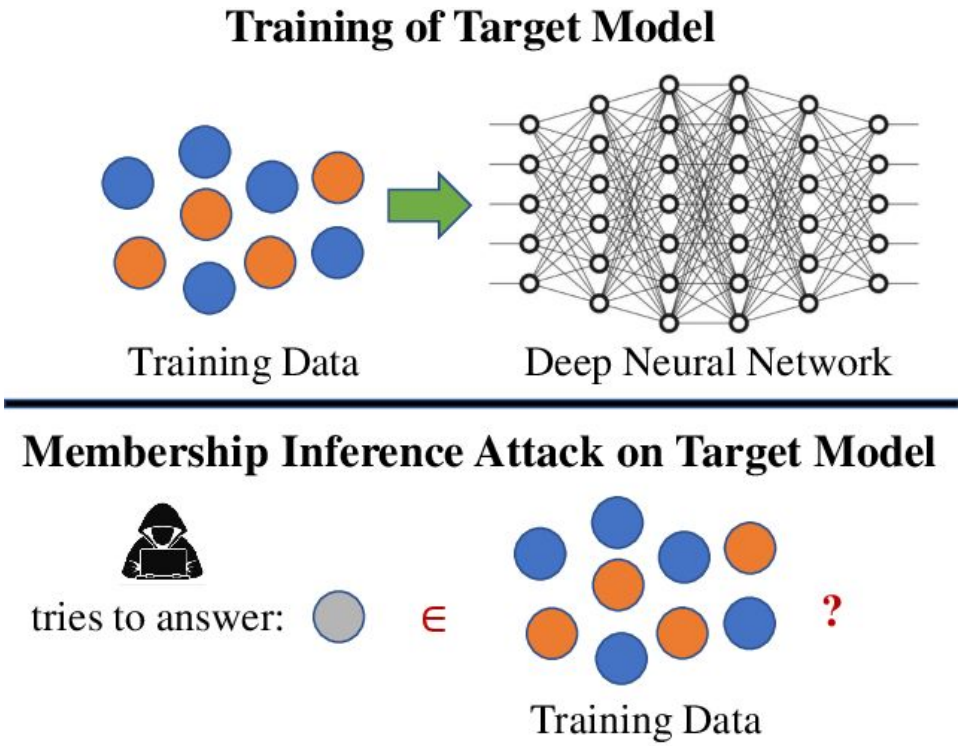
Motivation

- Data Privacy
- Copyright issues
- Novel research topic



Membership Inference Attacks

Was this example in the training set?



Data

Members

- Train set
- Identifying them is **the goal** of a MIA
- Potentially: copyrighted artwork

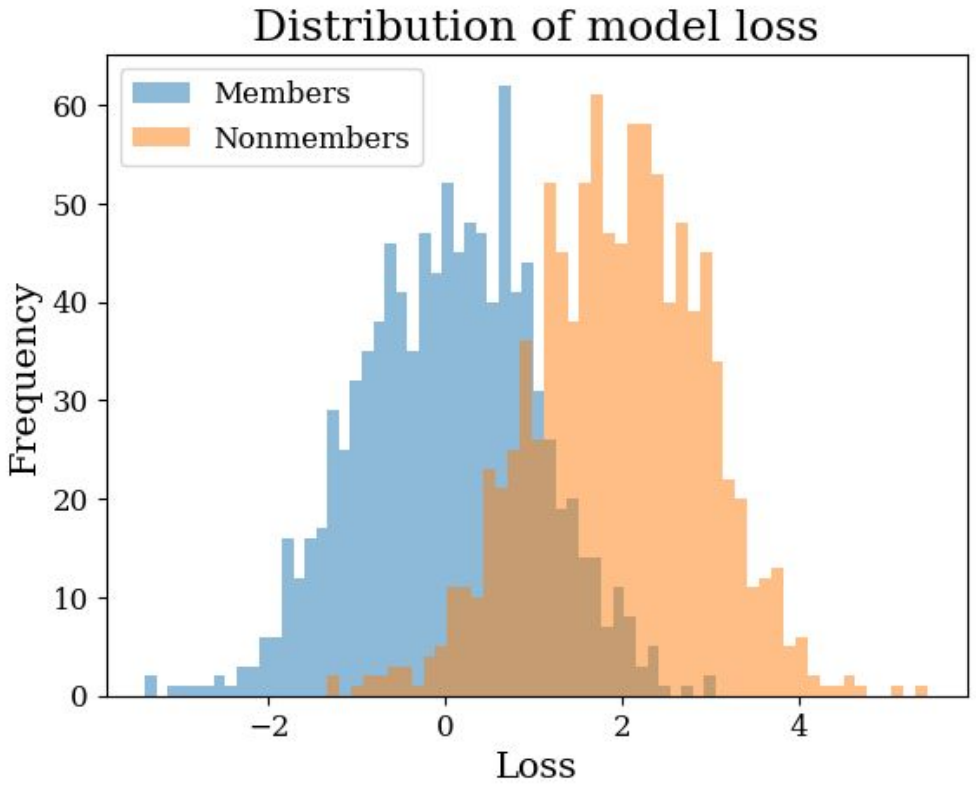
Nonmembers

- Not used during training
- **Ideally: validation/test set**

MIA is a *classification* task

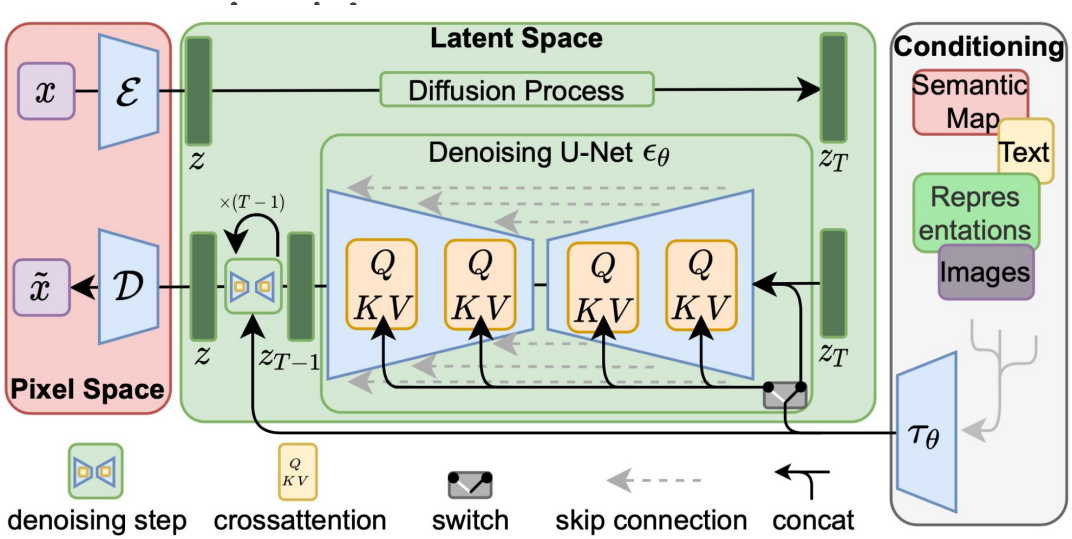
Loss Threshold Attack

- If loss < threshold then sample is a member
- Due to overfitting
- **TPR@FPR=1%**



Stable Diffusion

- Large Latent Diffusion Model
- LAION-5B
- Fully open-source*
- SOTA Text2Image**



Issues

- Current cost of training: **$100k**
- **150 000** GPU hours
- **No validation set!**

ML in PL
CONFERENCE 2023

Problem: We *do not* have a natural nonmembers set!

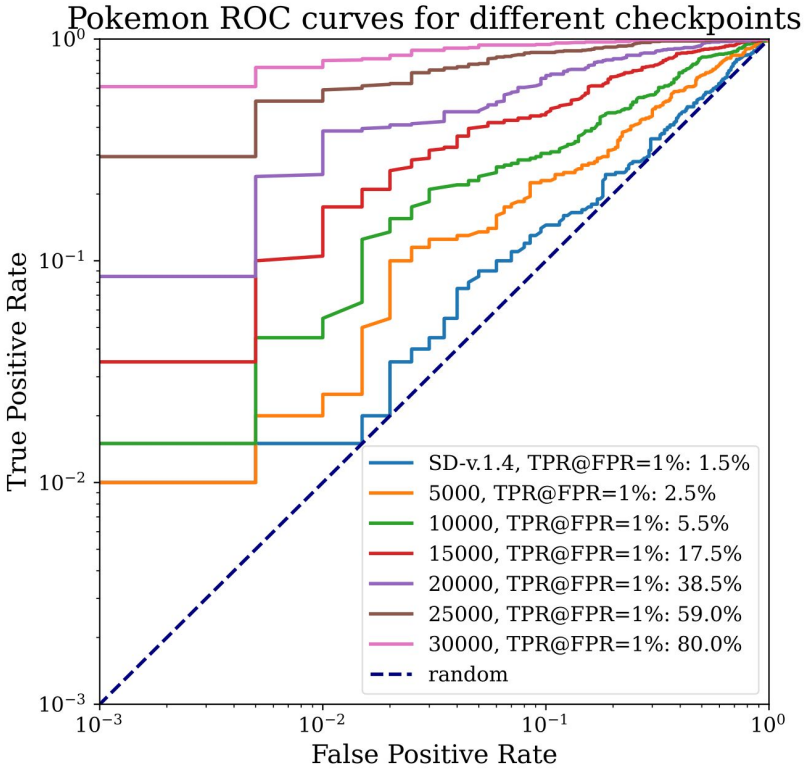
Option 1: Fine-tune Stable Diffusion on a smaller dataset

Pros

- cheap
- nonmembers easy to get
- easy to benchmark

Cons

- flawed
- trivial problem
- not applicable to real life scenarios



Option 2: Train Stable Diffusion from scratch

Pros

- nonmembers easy to get
- correct experimental setup

Cons

- extremely expensive
- impractical

Option 3: Obtain nonmembers from a different source

Pros

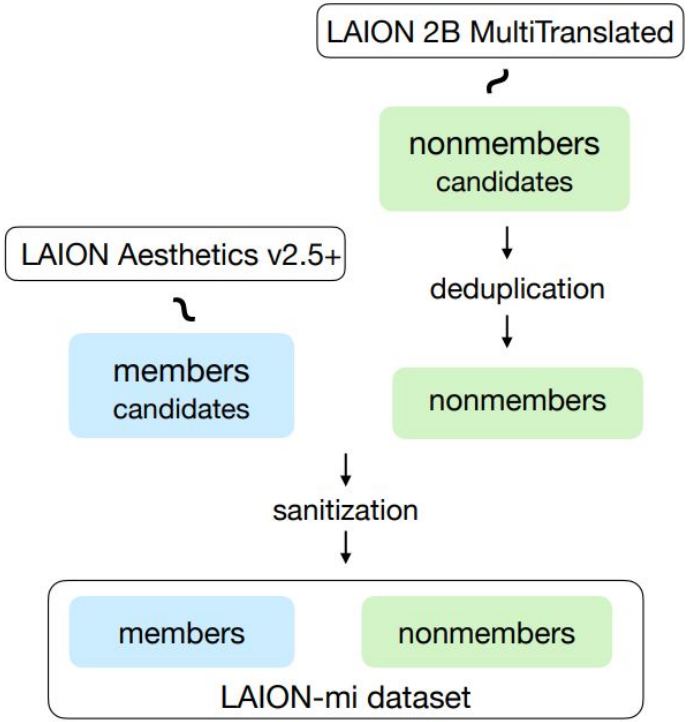
- cheap
- data easy to collect in our case

Cons

- possible distribution mismatch
- in effect: could lead to incorrect results

Solution: LAION-mi dataset

- Do not modify the original Stable Diffusion model
- Obtain the nonmembers set from other source
- Alleviate the distribution mismatch problem

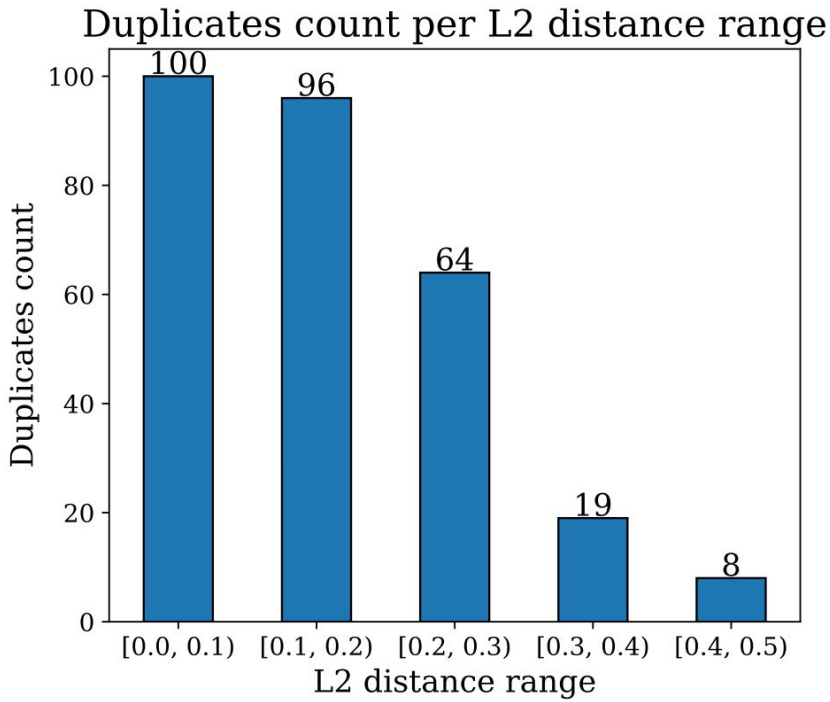


Challenge: Duplicates

- 30% in LAION-2B EN
- Effect: nonmembers set contaminated with member samples

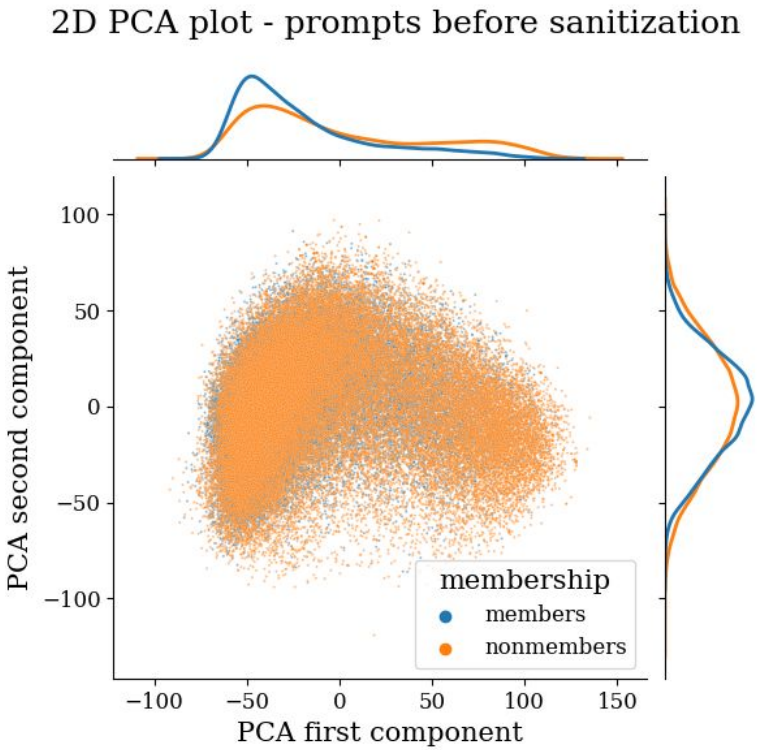
Solution: Deduplication

1. Query LAION-5B KNN-Index
2. Get duplicate candidates
3. Compute distances
4. Apply threshold to filter out duplicates



Challenge: Distribution Mismatch

- Images and descriptions
- Evaluation:
  - Visual (PCA)
  - FID
  - Training a classifier
- **Main focus on descriptions**



Solution: Sanitization

Start with a train set with half members and half nonmembers

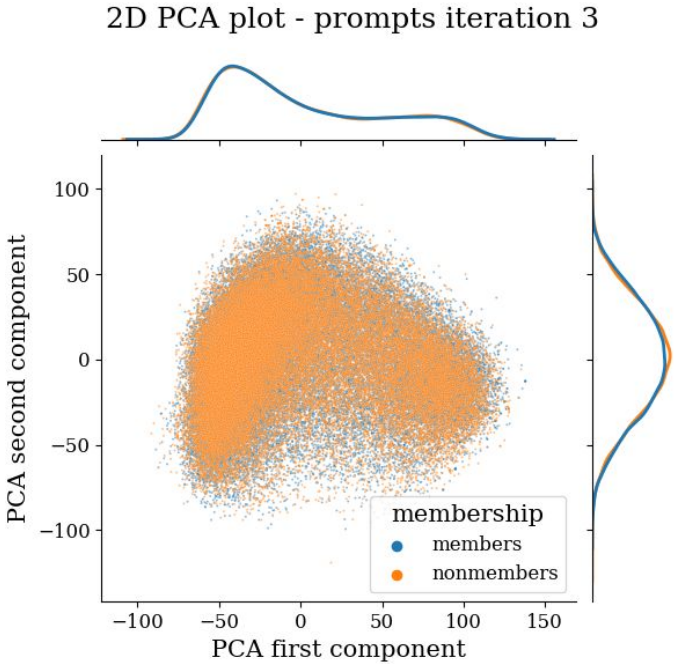
Until almost random accuracy on the train set, repeat:

1. Train a classifier on a train set
2. Add a new classifier to all classifiers
3. Create a new train set by filtering out members, for which any of the classifiers classifies them as member, and use all nonmembers

Results

LAION-mi : 40k members, 40k nonmembers

Data subset	FID	
	TEXT	IMAGES
Members internal - random	9.84	7.00
Members internal - sanitized	9.77	7.06
Nonmembers internal	9.73	7.01
Comparative - random	66.43	13.90
Comparative - sanitized	**13.54**	**8.87**



Evaluation on Stable Diffusion: Setup

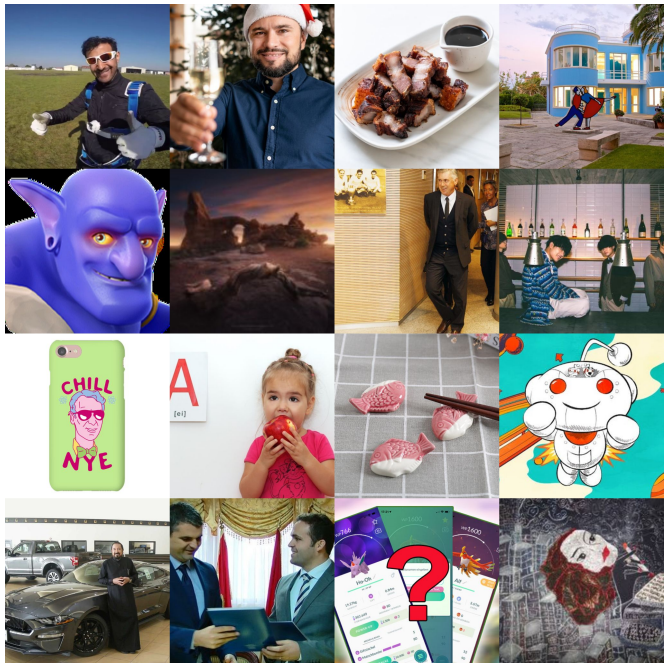
Loss Threshold Attack

- Model Loss
- Pixel Error
- Latent Error

Datasets + Models

- LAION-mi + Stable Diffusion
- POKEMON + fine-tuned Stable Diffusion on POKEMON dataset

Datasets



Results

Scenario	Loss	Method	TPR@FPR=1%. ↑ LAION-mi	TPR@FPR=1%. ↑ POKEMON
White-box		Baseline loss thr.	1.92%±0.59	80.9%±2.27
	Model Loss	Reversed noising	2.51%±0.73	97.3%±0.93
		Partial denoising	2.31%±0.61	94.5%±1.34
		Reversed denoising	2.25%±0.64	91.5%±1.63
	Latent Error	Reversed noising	1.26%±0.62	11.5%±1.84
		Partial denoising	2.42%±0.62	99.5%±0.4
		Reversed denoising	2.17%±0.64	61.1%±2.74
	Pixel Error	Reversed noising	1.90%±0.51	8.36%±1.66
		Reversed denoising	2.03%±0.55	12.0%±1.97
		Partial denoising	1.75%±0.68	25.38%±2.55
Grey-box	Latent Error	Generation from prompt	0.93%±0.41	7.15%±1.5
Black-box	Pixel Error	Generation from prompt	0.35%±0.19	12.0%±1.9

Summary

- MIAs are still **hard**, or impractically **expensive**
- We point out **flawed methodology**
- Our contribution: **LAION-mi dataset & evaluation protocol**