Stealing Machine Learning Models: Attacks and Countermeasures for Generative Adversarial Networks

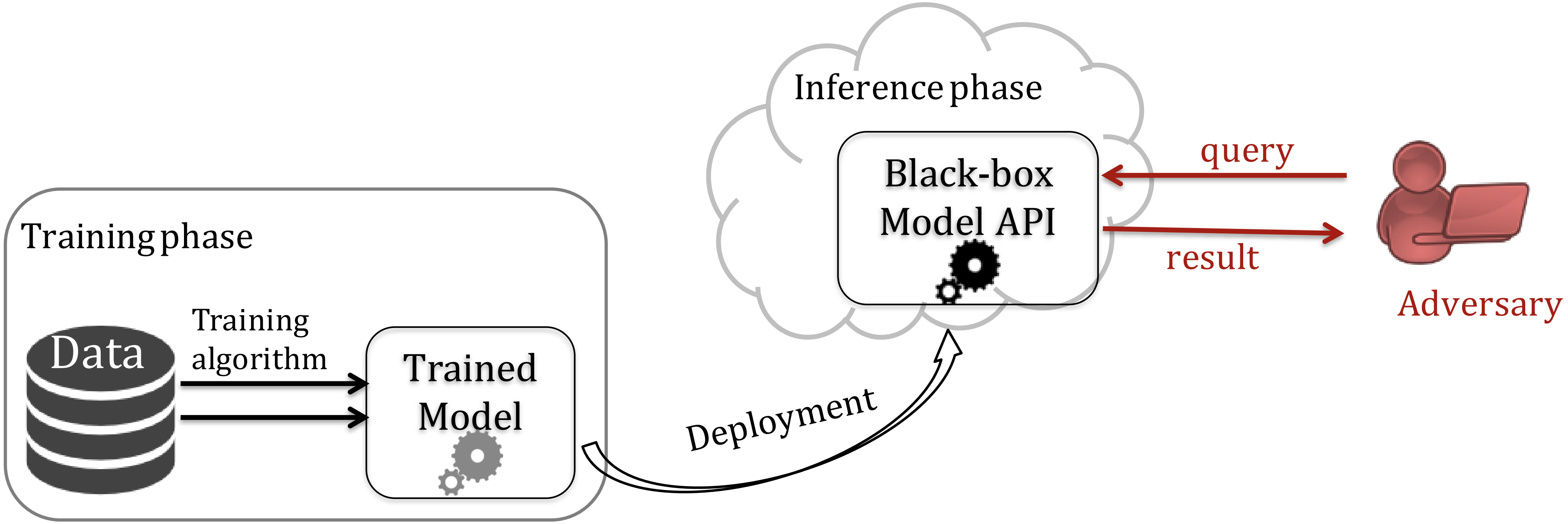
Hailong Hu and Jun Pang

University of Luxembourg

UNIVERSITÉ DU LUXEMBOURG

Annual Computer Security Applications Conference (ACSAC) 2021

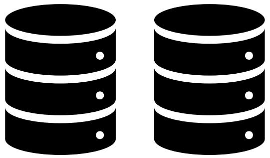
Model Extraction Attacks



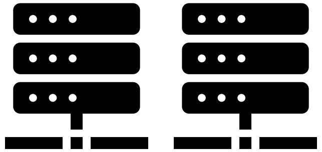
Model extraction: duplicate/steal a machine learning model through queries.

Why Should We Care?

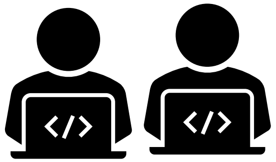
- Obtaining a practical deep learning model is non-trivial.



Big data



Intensive computing resources



Intensive human resources

- Model extraction attacks may facilitate other attacks.

Prior Works on Model Extraction Attacks

- Model extraction on traditional machine learning models [1].
  - Linear regression, logistic regression, decision tree...
- Model extraction on deep convolutional neural networks [2].
- Model extraction on BERT-based language models [3].

[1] Stealing Machine Learning Models via Prediction APIs. Tramèr et al., USENIX Security 2016.
[2] High Accuracy and High Fidelity Extraction of Neural Networks. Jagielski et al., USENIX Security 2020.
[3] Thieves on Sesame Street! Model Extraction of BERT-based APIs. Krishna et al., ICLR 2020 .

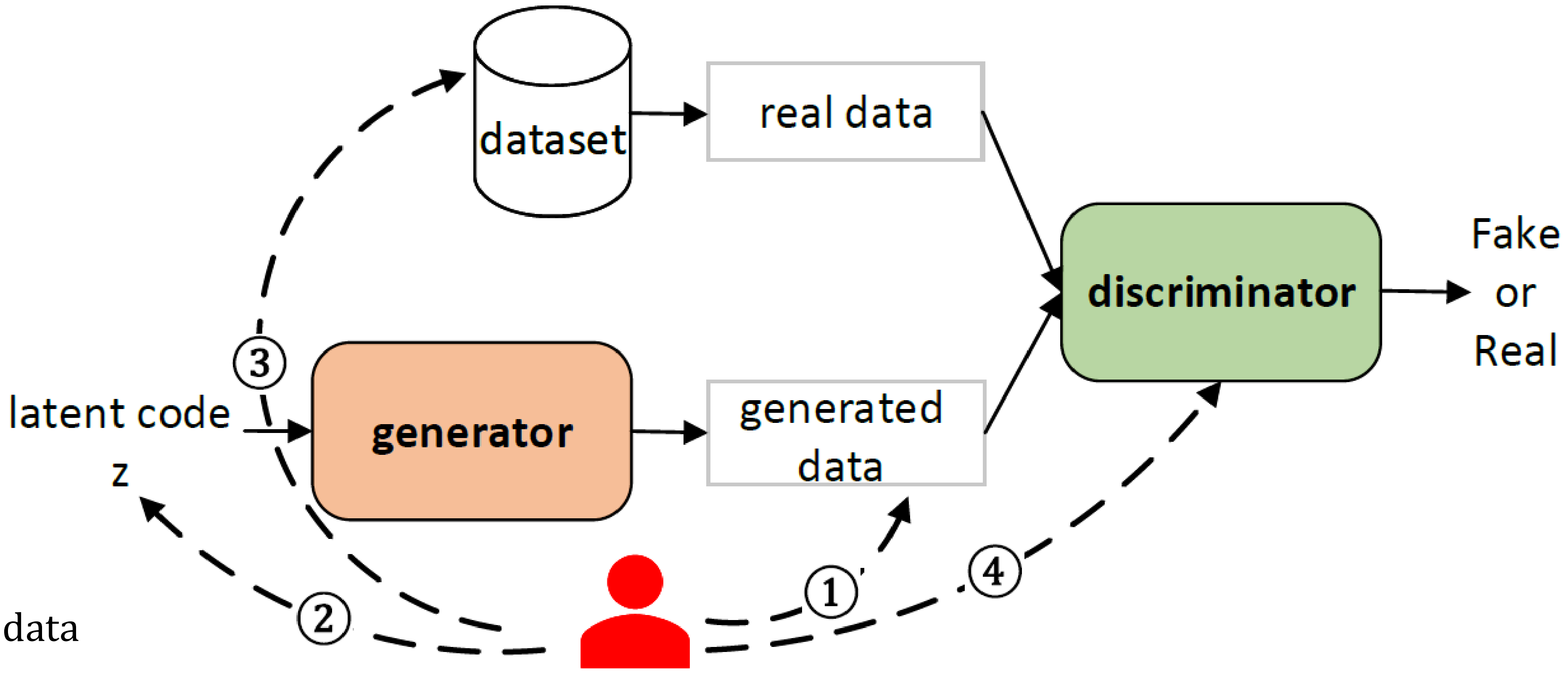
Our Work

Model Extraction Attacks against Generative Adversarial Networks (GANs)

Our Work: Contributions

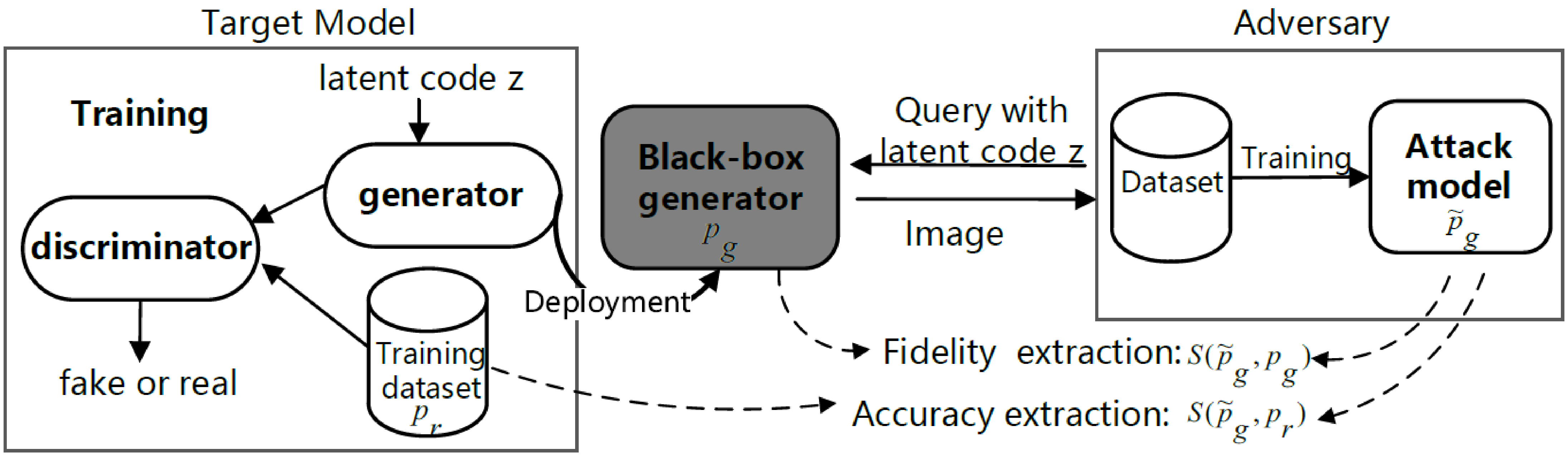
- Conduct the first systematic study of model extraction attacks against GANs and devise fidelity extraction and accuracy extraction for GANs.
- Perform one case study to illustrate the impact of model extraction attacks against GANs.
- Propose effective defense measures to mitigate model extraction attacks against GANs.

Components of a GAN



1 Generated data
2 Latent codes
3 Partial real data
4 Discriminator

Taxonomy



- **Fidelity extraction:** construct a $\tilde{G}$ minimizing $S(\tilde{p}_g, p_g)$.
- **Accuracy extraction:** construct a $\tilde{G}$ minimizing $S(\tilde{p}_g, p_r)$.

- p_g: implicit distribution of a target generator.
- $\tilde{p}_g$: implicit distribution of an attack generator.
- p_r: distribution of training set of a GAN.
- S: a similarity function between two models.

Fidelity Extraction

- **Methodology:** use the generated data to retrain a GAN.

- Model extraction vs. Machine learning

  > Model extraction: generated data.

  > Machine learning: data collected in real world.

  > Essentially model extraction on GANs approximates the target GAN that is a much simpler deterministic function.

Fidelity Extraction

- **Results:** fidelity extraction on different models.

Performance of attack models with 50k queries

Target model	Attack model	Dataset	*Fidelity* FID($\tilde{p}_g, p_g$)	*Accuracy* FID ($\tilde{p}_g, p_r$)
PGGAN	SNGAN	LSUN-Church	6.11	14.05
	SNGAN	CelebA	4.49	9.29
	PGGAN	LSUN-Church	1.68	8.28
	PGGAN	CelebA	1.02	4.93
SNGAN	SNGAN	LSUN-Church	8.76	30.04
	SNGAN	CelebA	5.34	17.32
	PGGAN	LSUN-Church	2.21	14.56
	PGGAN	CelebA	1.39	9.57

Performance of target GANs

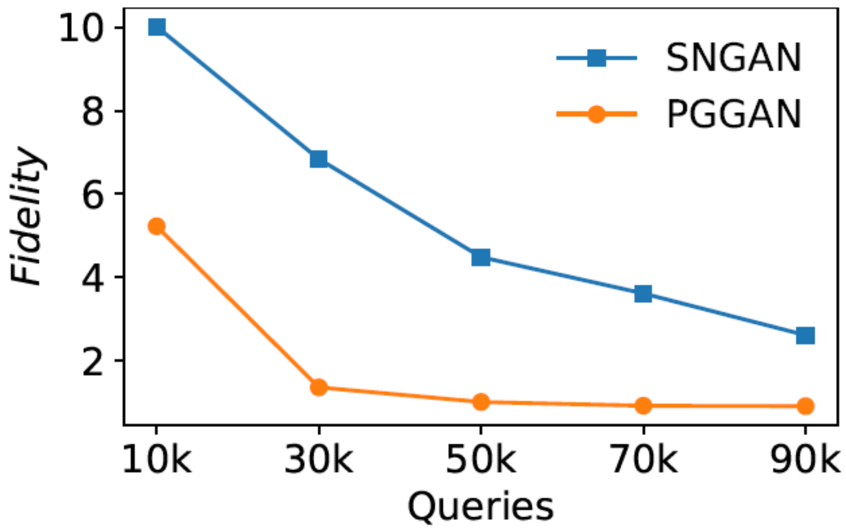
Target model	Dataset	FID
SNGAN	LSUN-Church	12.72
SNGAN	CelebA	7.60
PGGAN	LSUN-Church	5.88
PGGAN	CelebA	3.40

- Fidelity extraction can achieve an acceptable performance.

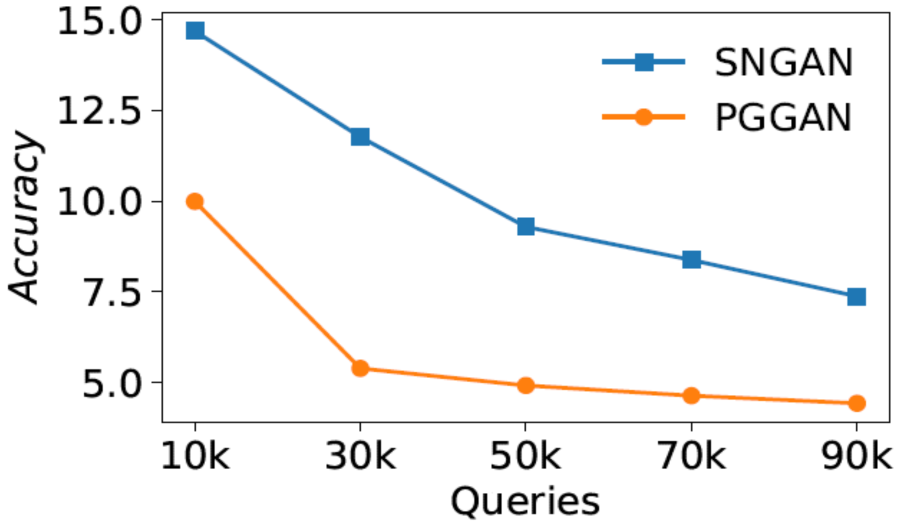
* FID: a smaller FID indicates a better performance of a GAN.

Fidelity Extraction

- **Results:** attack performance on different number of queries.



(a) *Fidelity* on CelebA

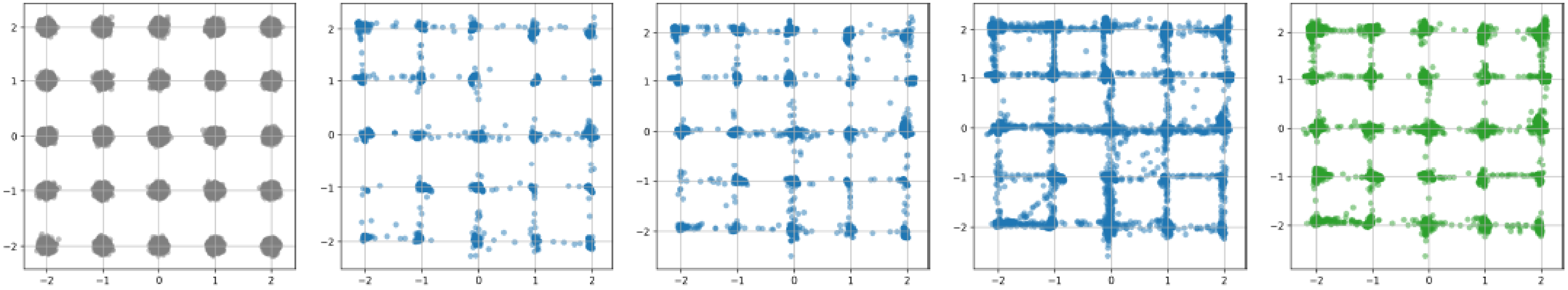


(b) *Accuracy* on CelebA

- Fidelity and accuracy values become stable with an increase in the number of queries.
- There is a gap in terms of accuracy between target models and attack models.
- Target model: PGGAN; FID = 3.40

Accuracy Extraction

- **Reason:** the target GAN model is hard to reach global equilibrium and the discriminator is often better than the generator in practice.
- An example on synthetic data.



(a) Real Samples 50K (b) Direct Sampling 5k (c) Direct Sampling 10K (d) Direct Sampling 50K (e) MH Subsampling 50K

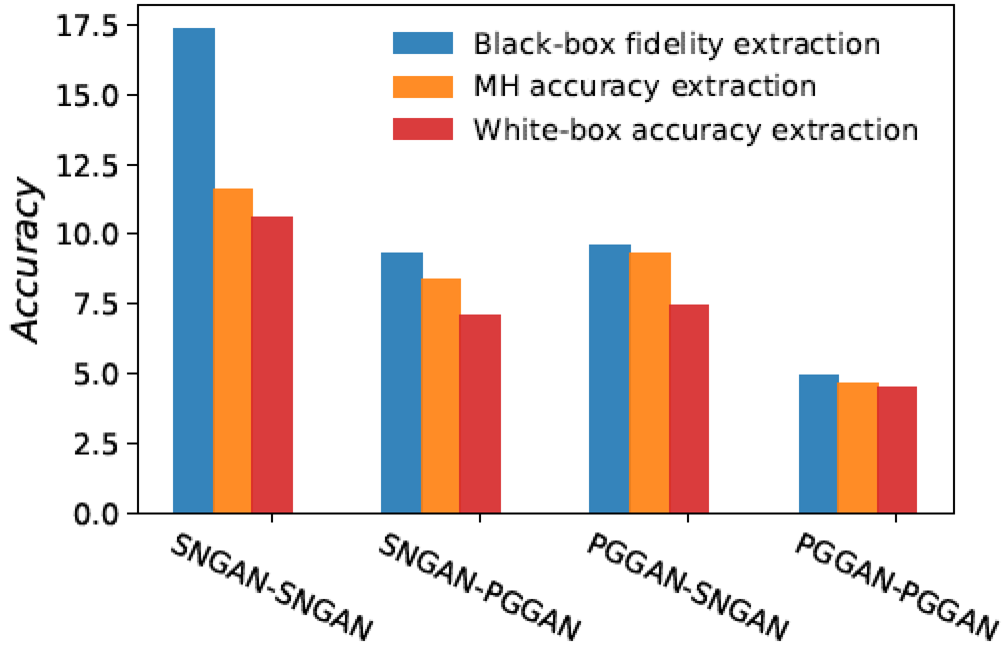
Accuracy Extraction

- **Accuracy extraction:** construct a $\tilde{G}$ minimizing $S(\tilde{p}_g, p_r)$.
- **Methodology:**
  > Discriminator + partial real data: subsample generated data through the discriminator.

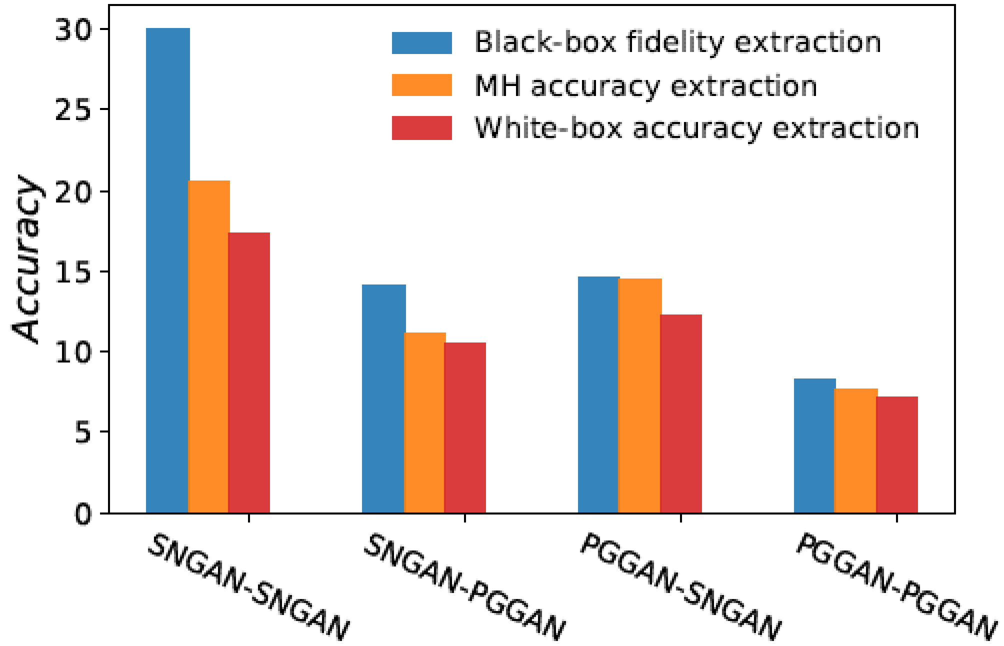
  > Retrain a GAN on these subsampled data.

Accuracy Extraction

- **Results:** accuracy extraction on different models.



(a) *Accuracy* on CelebA



(b) *Accuracy* on LSUN-Church

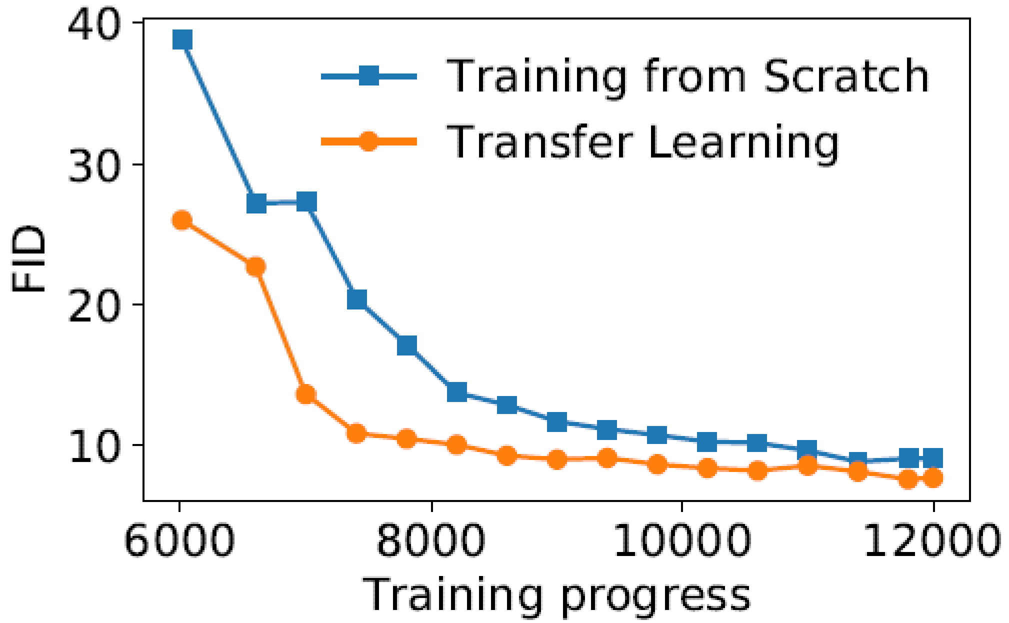
* Accuracy: a smaller accuracy value indicates a better attack performance.

Case Study

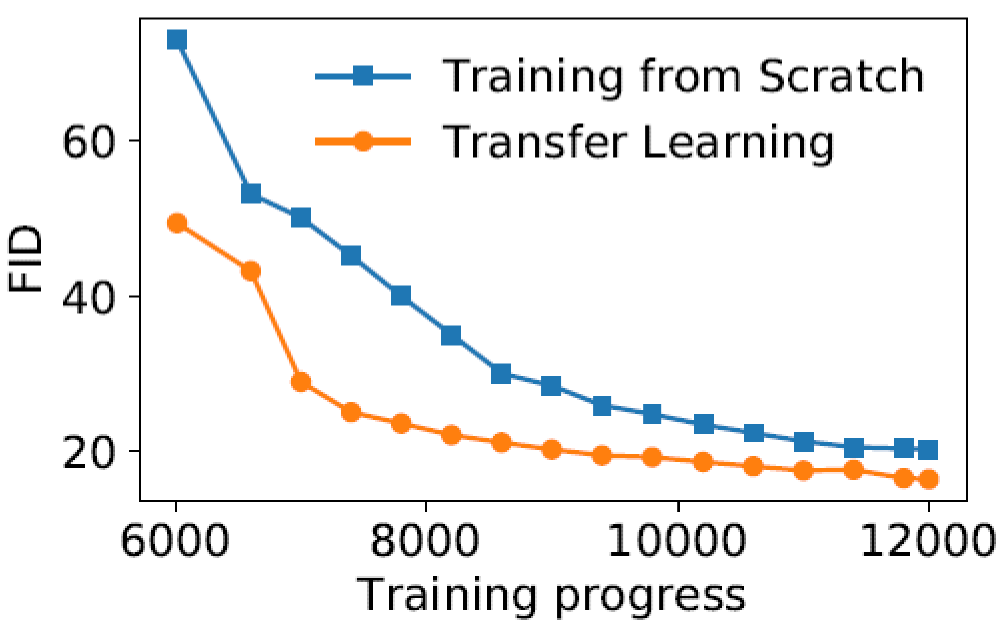
- **Motivation:** generate images on a new domain.
- **Scenario:**
  - > Target model: StyleGAN trained with more than 3 million images.
  - > Attack model: PGGAN with 50k queries.
  - > Objective: an adversary transfers the extracted model to new domains.
  - > **The attack is successful if** the performance of models trained by transfer learning based on the extracted GAN outperforms models trained from scratch.

Case Study

- **Results:** model extraction based transfer learning



(a) FID on LSUN-Kitchen



(b) FID on LSUN-Classroom

- Source dataset: LSUN-Bedroom.

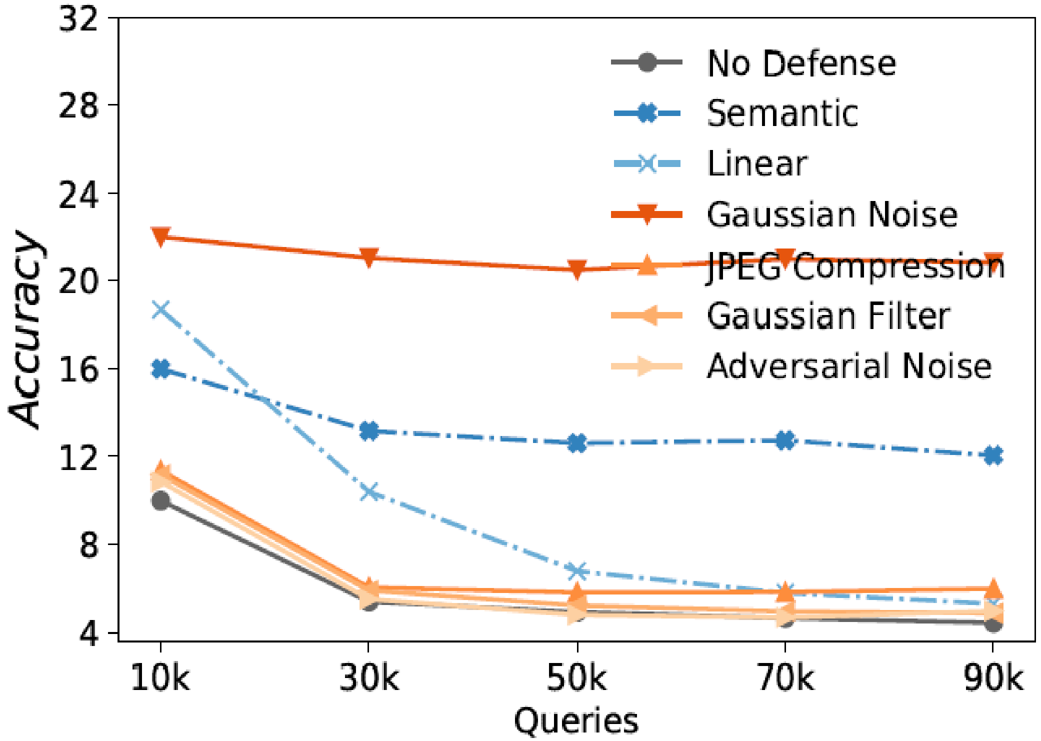
* FID: a smaller FID indicates a better performance of a GAN.

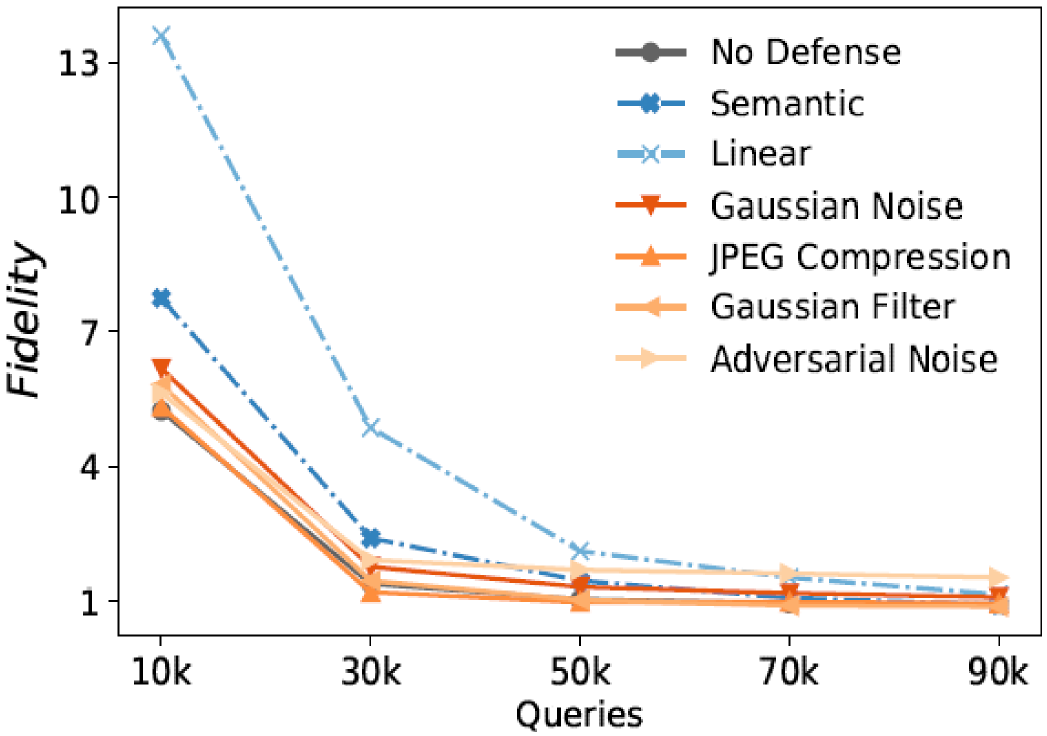
Defenses

- **In terms of fidelity of model extraction**
  > Limiting the number of queries.
- **In terms of accuracy of model extraction**
  - Input Perturbation-base Defenses
    > Increasing the similarity of generated samples.
    > Linear interpolation defense; semantic interpolation defense.
  - Output Perturbation-base Defenses
    > Perturbing generated samples.
    > Random noise; adversarial noise; filtering; compression.

Defenses

- **Results:** the performance of attack model PGGAN under various defenses





- Target model: PGGAN trained on CelebA

* A larger accuracy/fidelity value indicates a better performance of the defense.

Future Work

- Protecting GANs through verifying the ownership
  > A GAN model is the intellectual property of model owners.
- Designing new privacy-preserving techniques for GANs
  > Stealing a GAN model also means the leakage of distribution of the training set.

Thank You!

This work is supported by

Luxembourg National Research Fund (FNR) - Grant No. 13550291

Fonds National de la Recherche Luxembourg