Marich: A Query-efficient Distributionally Equivalent Model Extraction Attack using Public Data

Pratik Karmakar[1], Debabrota Basu[2]

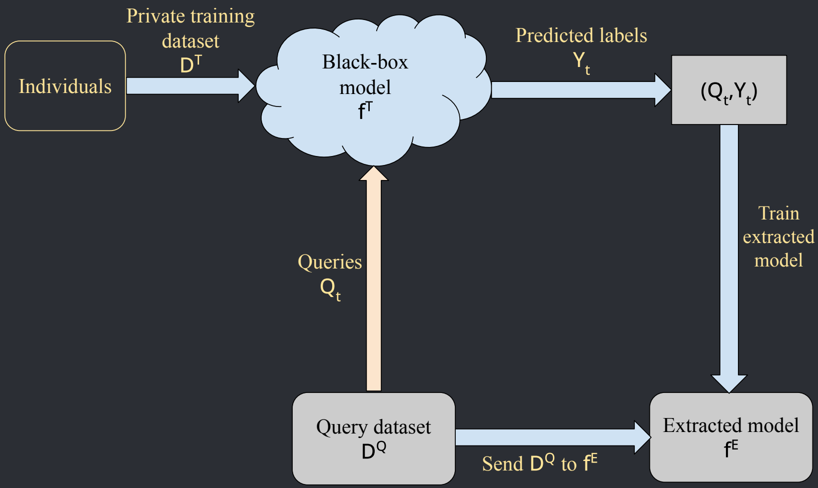
[1] Department of Computer Science, RKMVERI, Belur, India

[2] Équipe Scool, Inria, University of Lille, CNRS, France

AAAI Workshop on Privacy Preserving AI (PPAI), 2023

Model Extraction Attack

The Framework



Taxonomy of Model Extraction Attacks

What's out there?

- **Access to model:** White-box or black-box [TZJ+16]
- **Query dataset:** Synthetic [TZJ+16], perturbed version of private [PMG+17] or public [PGS+20]
- **Response to query:** Prediction distribution [JCB+20], gradients [MSDH19] or predicted label [PMG+17]
- **Model class:** Linear [MSDH19], neural network [MSDH19, JCB+20], or CNN [CSBB+18]
- **Objective of extraction:** Task accuracy [JCB+20], fidelity [PGS+20], or functional equivalence [PMG+17]

Taxonomy of Model Extraction Attacks

Best of old and new worlds!

- **Access to model:** White-box or black-box [TZJ+16]
- **Query dataset:** Synthetic [TZJ+16], perturbed version of private [PMG+17] or public [PGS+20]
- **Response:** Prediction distribution [JCB+20], gradients [MSDH19] or predicted label [PMG+17]
- **Model class:** Linear [MSDH19], neural network [MSDH19, JCB+20] or CNN [CSBB+18]
  $\longrightarrow$ model-agnostic
- **Objective:** Task accuracy [JCB+20], fidelity [PGS+20], or functional equivalence [PMG+17]

Can we define an information-theoretic objective that can cover the utilities of these objective?

Distributionally Equivalent Model Extraction

Match the Prediction Distributions

Observations

1. Any classification model f^T and a data generating distribution $\mathscr{D}^Q$ together induces a predictive distribution over label-input pairs (Y, X).

2. Any utility metric, e.g. accuracy, fidelity, are functionals computed on this joint distribution.

Intuition: Design an extraction attack that selects a set of queries $\mathscr{D}^Q$ and creates an extracted model f^E_ω to minimise the KL-divergence between the induced joint distributions.

$$(\omega^*_{\min}, \mathscr{D}^Q_{\min}) \triangleq \underset{\omega, \mathscr{D}_Q}{\operatorname{argmin}} \; D_{\mathrm{KL}}\left(\Pr(f^T_{\theta^*}(Q), Q) \| \Pr(f^E_\omega(Q), Q)\right)$$

Max-Information Model Extraction

Leak Information about the Prediction Distribution

Goal of Privacy Attack

To maximially leak privacy of a target model and a private dataset, we should increase the information content passed from predictive distribution of the target model to that of the extracted model.

Intuition: An extracted model f^E and a query distribution should aim to maximise the mutual information between the joint distributions of input features $Q \sim \mathcal{D}^Q$ and predicted labels induced by f^E and that of the target model f^T.

$$(\omega^*_{\max}, \mathcal{D}^Q_{\max}) \triangleq \underset{\omega, \mathcal{D}_Q}{\operatorname{argmax}} \; \mathrm{I}\left(\Pr(f^T_{\theta^*}(Q), Q) \| \Pr(f^E_\omega(Q), Q)\right)$$

A Variational Formulation of Model Extraction

Reducing the Attacks to an Optimisation Problem

Upper Bounding Distributional Closeness

If we choose KL-divergence as the similarity metric, then for a query generating distribution $\mathscr{D}^Q$

$$D_{\mathrm{KL}}\left(\Pr(f^T_{\theta^*}(Q), Q) \| \Pr(f^E_{\omega^*_{\mathrm{DEq}}}(Q), Q)\right) \leq \min_{\omega} \boxed{E_Q[\, l(f^T_{\theta^*}(Q), f^E_{\omega}(Q))]} - \boxed{H(f^E_{\omega}(Q))}$$

Lower Bounding Information Leakage

For any given $\mathscr{D}^Q$, the information leaked by any max-information attack is lower bounded as:

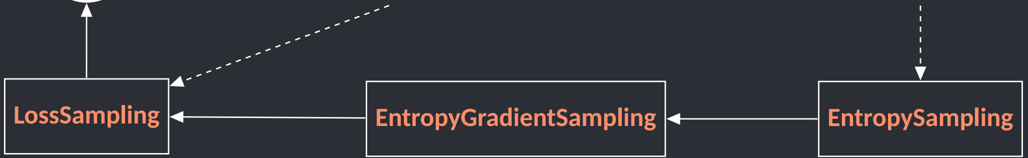
$$\mathrm{I}\left(\Pr(f^T_{\theta^*}(Q), Q) \| \Pr(f^E_{\omega^*_{\mathrm{min}}}(Q), Q)\right) \geq \max_{\omega} -\boxed{E_Q[\, l(f^T_{\theta^*}(Q), f^E_{\omega}(Q))]} + \boxed{H(f^E_{\omega}(Q))}$$

Marich: Distributionally Equivalent and Max-Information Extraction

Entropy of Predictions and Model Mismatch-guided Query Selection

At every round t, Marich selects queries Q_t satisfying

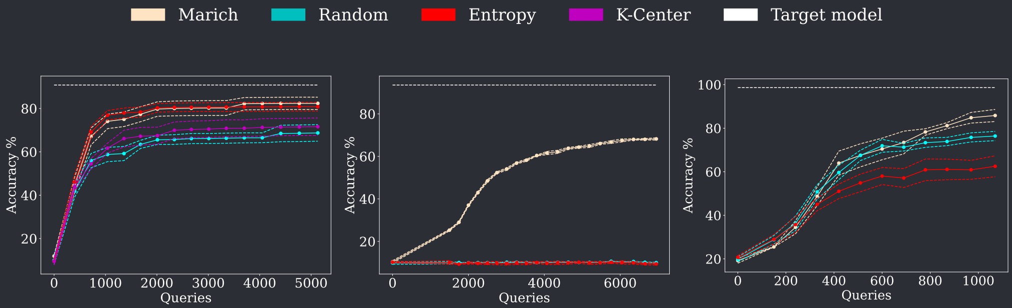
$$Q_t = \underset{Q \in \mathbf{D}^Q}{\operatorname{argmax}} \boxed{-\frac{1}{t}\sum_{i=1}^{t-1} l(f^T_{\theta^*}(Q_i \cup Q), f^E_{\omega_{t-1}}(Q_i \cup Q))} + \boxed{\sum_{i=1}^{t-1} H(f^E_{\omega_{t-1}}(Q_i \cup Q))}$$



Use Q_t to train the extracted model and update it to $f^E_{\omega_t}$.

Quality of Model Extraction

Task Accuracy



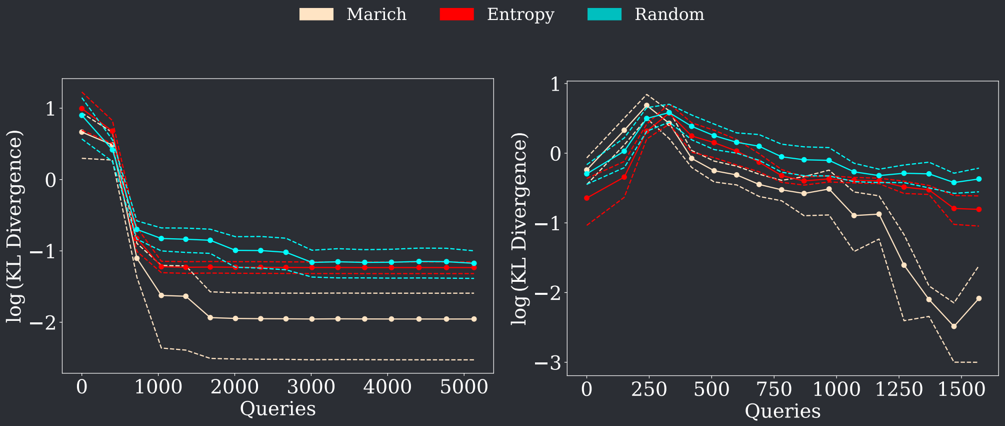
(a) LR with EMNIST

(b) ResNet18 with STL10

(c) BERT with AGNews

Quality of Model Extraction

Distributional Closeness



(a) LR with EMNIST

(b) BERT with AGNews

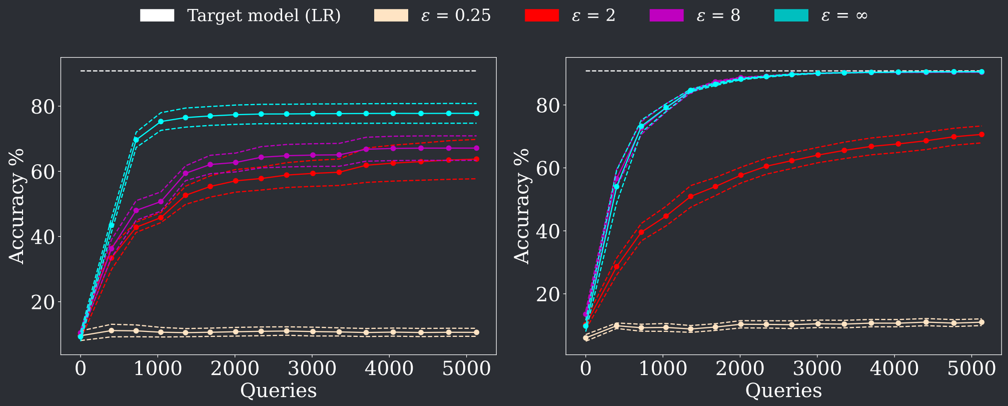
Quality of Model Extraction

Informativeness of Extraction Leading to Membership Inference

Member dataset	Target model	Query Dataset	Algorithm	#Queries	MI acc.	MI agreement	MI agreement AUC
MNIST	LR	-	-	50,000	87.99%	-	-
		-	-	50,000	92.30%	-	-
		EMNIST	**MARICH**	5,130	88.58%	92.82%	**92.73%**
		CIFAR10	**MARICH**	1,420	**94.27%**	**93.97%**	92.43%
		EMNIST	Random	5,130	89.61%	91.01%	91.11%
		CIFAR10	Random	1,420	92.61%	89.84%	85.79%
CIFAR10	Resnet18	-	-	40,000	79.35%	-	-
		STL10	**MARICH**	6,950	**93.90%**	**75.52%**	**76.69%**
		STL10	Random	6,950	92.32%	75.25%	75.83%
BBCNews	BERT	-	-	1,490	**98.61%**	-	-
		AGNews	**MARICH**	1,070	94.42%	**91.02%**	**82.62%**
		AGNews	Random	1,070	89.17%	86.93%	58.64%

Performance against ε-DP Defenses

Privacy Level $\varepsilon \geq 2$ cannot Protect Much

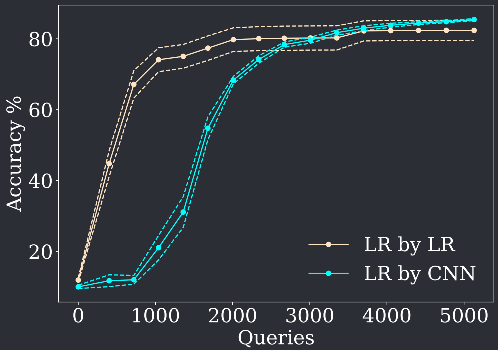


(a) LR by EMNIST

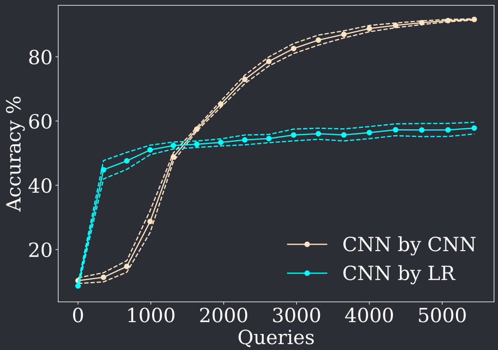
(b) LR by CIFAR10

Impact of Model Mismatch

More Expressive Models can Steal Low Expressive Models

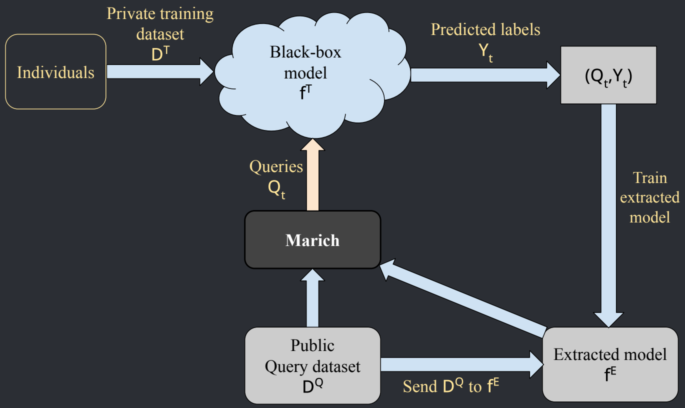


(a) LR extracted by LR vs. LR extracted by CNN



(b) CNN extracted by CNN vs. CNN extracted by LR

Marich is a model-agnostic extraction algorithm that adaptively selects a small subset of a public dataset to maximise information leakage from f^T.



Can we develop a theoretical characterisation of the capabilities and limitations of these attacks?

For further details, please visit: `https://github.com/Debabrota-Basu/marich`

References

[CSBB+18] Jacson Rodrigues Correia-Silva, Rodrigo F Berriel, Claudine Badue, Alberto F de Souza, and Thiago Oliveira-Santos.
Copycat cnn: Stealing knowledge by persuading confession with random non-labeled data.
In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.

[JCB+20] Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot.
High accuracy and high fidelity extraction of neural networks.
In *29th USENIX security symposium (USENIX Security 20)*, pages 1345–1362, 2020.

[MSDH19] Smitha Milli, Ludwig Schmidt, Anca D Dragan, and Moritz Hardt.
Model reconstruction from model explanations.
In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 1–9, 2019.

[PGS+20] Soham Pal, Yash Gupta, Aditya Shukla, Aditya Kanade, Shirish Shevade, and Vinod Ganapathy.
Activethief: Model extraction using active learning and unannotated public data.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 865–872, 2020.

[PMG+17] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami.
Practical black-box attacks against machine learning.
In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

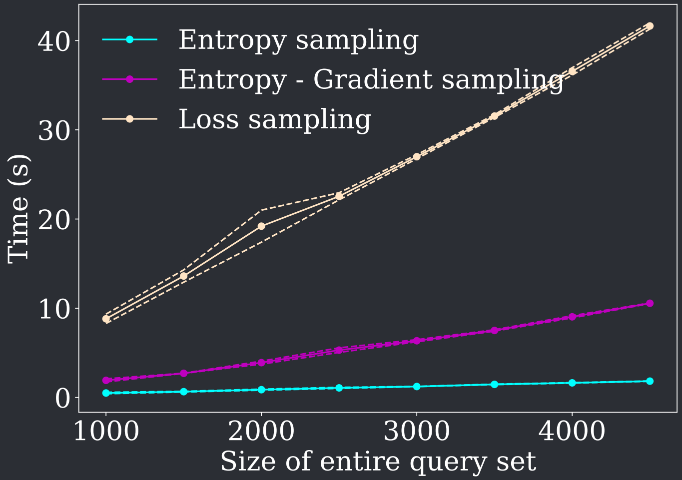
[TZJ+16] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart.
Stealing machine learning models via prediction {APIs}.
In *25th USENIX security symposium (USENIX Security 16)*, pages 601–618, 2016.

Marich: Distributionally Equivalent and Max-Information Extraction

Algorithm Marich

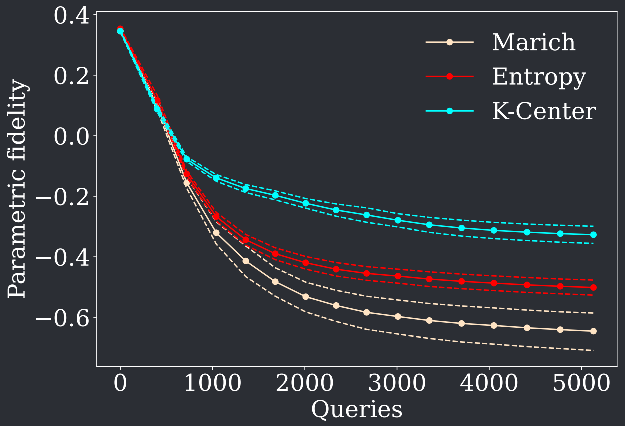
//* Initialisation of the extracted model*// $\triangleright$ Phase 1
$Q_0^{train} \leftarrow n_0$ datapoints randomly chosen from D^Q
$Y_0^{train} \leftarrow f^T(Q_0^{train})$ $\triangleright$ Query the target model f^T with Q_0^{train}
$f_0^E \leftarrow$ Train f^E with $(Q_0^{train}, Y_0^{train})$ for E_{max} epochs
//* Adaptive query selection*// $\triangleright$ Phase 2
for $t \leftarrow 1$ to T **do**
    $Q_t^{entropy} \leftarrow$ EntropySampling$(f_{t-1}^E, \mathrm{D}^Q \setminus Q_{t-1}^{train}, B)$
    $Q_t^{grad} \leftarrow$ EntropyGradientSampling$(f_{t-1}^E, Q_t^{entropy}, \gamma_1 B)$
    $Q_t^{loss} \leftarrow$ LossSampling$(f_{t-1}^E, Q_t^{grad}, Q_{t-1}^{train}, Y_{t-1}^{train}, \gamma_1\gamma_2 B)$
    $Y_t^{new} \leftarrow f^T(Q_t^{loss})$ $\triangleright$ Query the target model f^T with Q_t^{loss}
    $Q_t^{train} \leftarrow Q_{t-1}^{train} \cup Q_t^{loss}$, $Y_t^{train} \leftarrow Y_{t-1}^{train} \cup Y_t^{new}$
    $f_t^E \leftarrow$ Train f_{t-1}^E with $(Q_t^{train}, Y_t^{train})$ for $E_{\max}$ epochs
end for

Comparing Sampling Strategies



Quality of Extraction by Marich

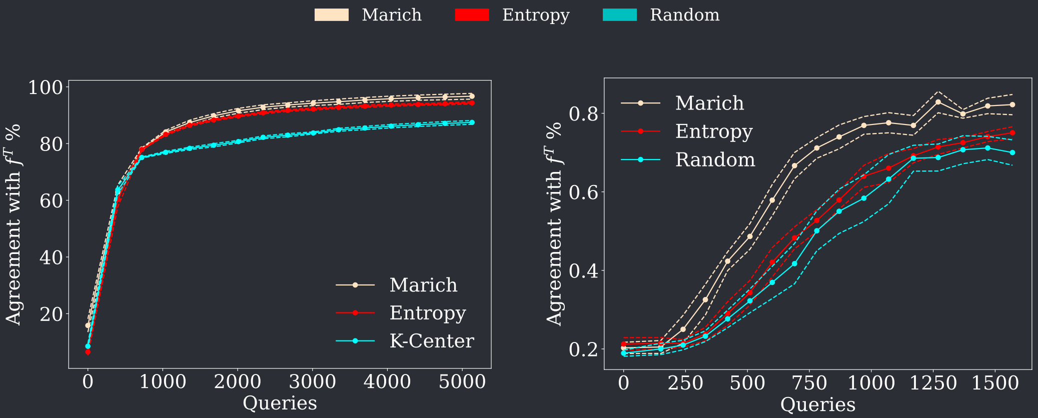
Parametric Fidelity



LR with EMNIST

Quality of Extraction by Marich

Agreement in Predictions



(a) LR with EMNIST

(b) BERT with AGNews

Membership Inference with Marich

Informativeness leading to Membership Inference

Member dataset	Target model	Query Dataset	Algorithm	Non-member dataset	#Queries	MI acc.	MI agreement	MI agreement AUC
MNIST	LR	-	-	EMNIST	50,000 (100%)	87.99%	-	-
		-	-	CIFAR10	50,000 (100%)	92.30%	-	-
		EMNIST	MARICH	EMNIST	5,130 (3.5%)	88.58%	92.82%	**92.73%**
		CIFAR10	MARICH	CIFAR10	1,420 (2.37%)	**94.27%**	**93.97%**	92.43%
		EMNIST	RS	EMNIST	5,130 (3.5%)	89.61%	91.01%	91.11%
		CIFAR10	RS	CIFAR10	1,420 (2.37%)	92.61%	89.84%	85.79%
CIFAR10	Resnet18	-		STL10	40,000 (100%)	79.35%	-	-
		STL10	MARICH	STL10	6,950 (6.15%)	**93.90%**	**75.52%**	**76.69%**
		STL10	RS	STL10	6,950 (6.15%)	92.32%	75.25%	75.83%
BBCNews	BERT	-		AGNews	1,490 (100%)	**98.61%**	-	-
		AGNews	MARICH	AGNews	1,070 (0.83%)	94.42%	**91.02%**	**82.62%**
		AGNews	RS	AGNews	1,070 (0.83%)	89.17%	86.93%	58.64%