Adversarial Machine Learning for Social Good: Reprogramming Black-box Machine Learning Models with Scarce Data and Limited Resources

Yun-Yun Tsai†, Pin-Yu Chen‡, Tsung-Yi Ho†

†Department of Computer Science, National Tsing Hua University
‡IBM Research

The Youtube voiceover can be found at https://youtu.be/XJJO1qQKRa8.

Objective and Motivations

In this work, we propose a novel approach, black-box adversarial reprogramming (BAR), that reprograms a deployed machine learning (ML) model (e.g., a prediction API) for performing ML tasks related to social good in a black-box manner, such as autism spectrum disorder (ASD) classification and diabetic retinopathy (DR) detection. Our proposed method is inspired by a recent work on adversarial reprogramming (AR) [2], but we note the following substantial differences and unique challenges:

- **Black-box setting.**
- **Data scarcity and resource constraint.**

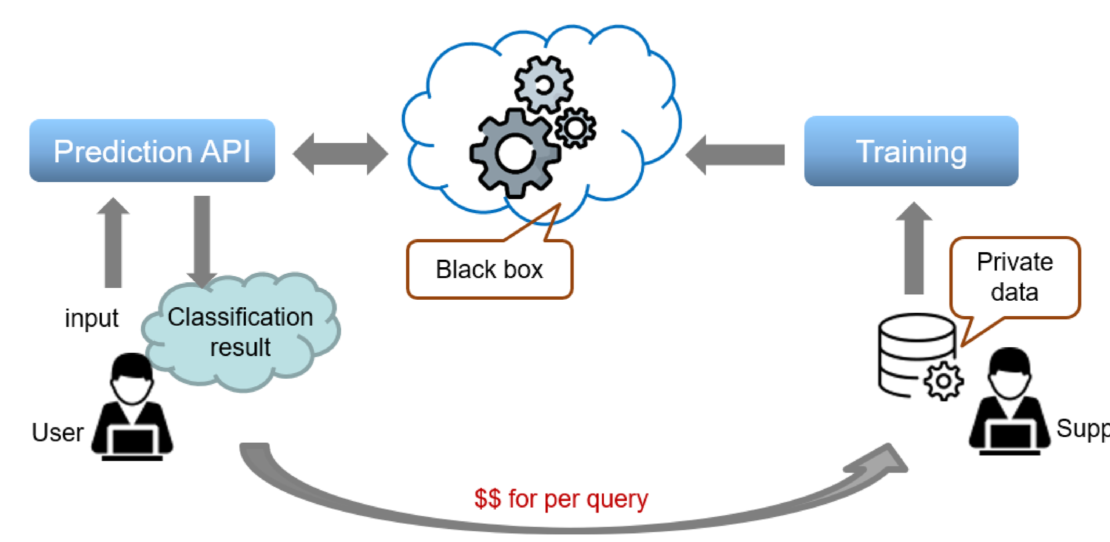

Fig. 1: Overview of black-box ML API

Problem Formulation

- We assume that the adversary has no knowledge of architecture or parameters related to black-box ML classification model which he/she want to reprogram. Such an adversary can train an adversarial program as an input transformation function.

Overview of BAR

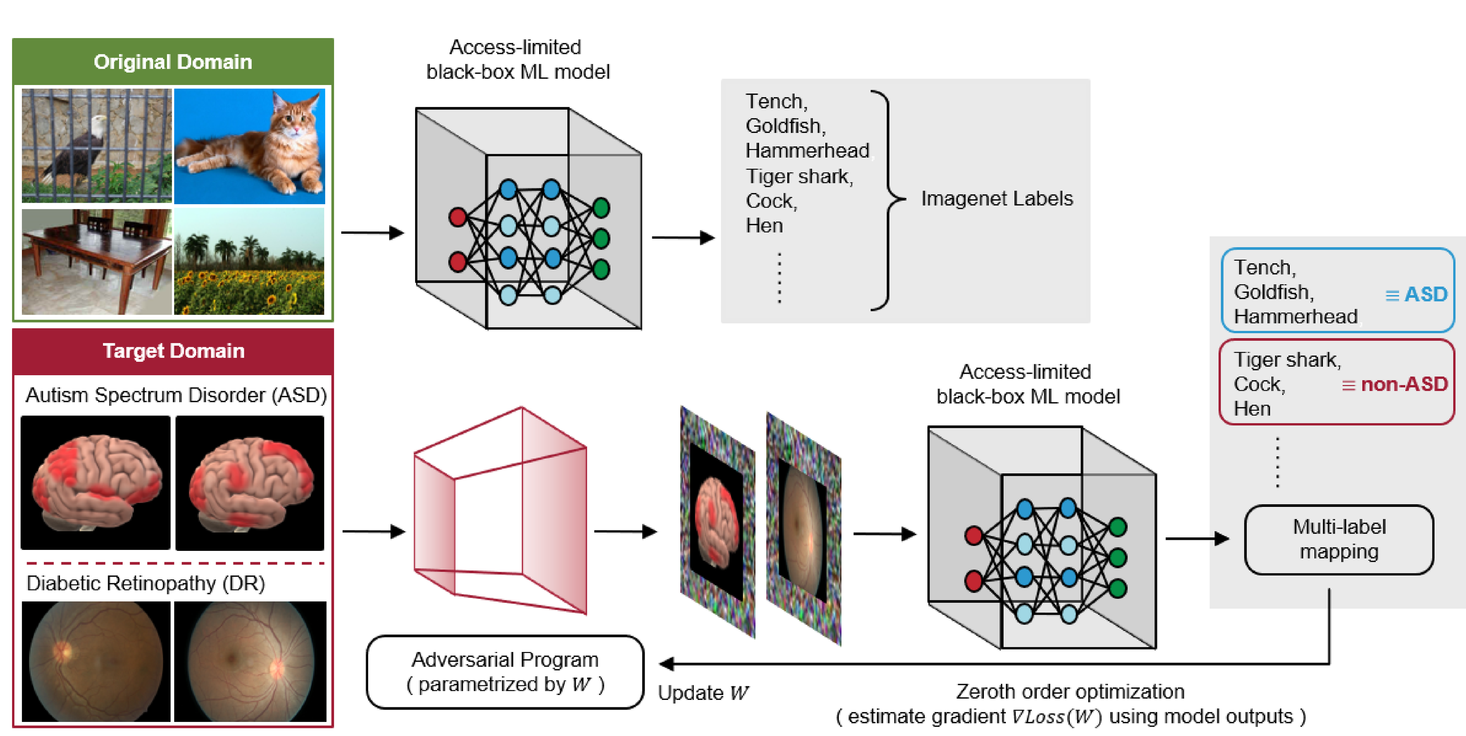

Fig. 2: Overview of our proposed black-box adversarial reprogramming (BAR) method.

Our proposed BAR algorithm

Algorithm 1 Training algorithm of black-box adversarial reprogramming (BAR)

Input: black-box ML model F, AR loss function $Loss(\cdot)$, target domain training data $\{D_i, y_i\}_{i=1}^n$, maximum number of iterations T, number of random vectors for gradient estimation q, multi-label mapping function $h(\cdot)$, step size $\{\alpha_t\}_{t=1}^T$

Output: Optimal adversarial program parameters W

1: Randomly initialize W; set $t = 1$

2: Embed $\{D_i\}_{i=1}^n$ with mask M to create $\{X_i\}_{i=1}^n$

3: **while** $t \leq T$ **do**

4: # **Generate adversarial program**

$P = \tanh(W \odot M)$

Generate q perturbed adversarial programs

$\widetilde{P}_j = \tanh((W + U_j) \odot M)$ for all $j \in [q]$

$\{U_j\}_{j=1}^q$ are standard normal Gaussian random vectors divided by its Euclidean norm.

5: # **Loss function evaluation for gradient estimation**

The loss function is defined as

$$Loss(W) = -\sum_{i=1}^{n} \sum_{j=1}^{K'} y_{ij} \log h_j \left(F(X_i + P(W)) \right). \tag{1}$$

Evaluate $Loss$ in (1) with W and $\{X_i + P\}_{i=1}^n$

Evaluate $Loss$ in (1) with $W + U_j$ and $\{X_i + \widetilde{P}_j\}_{i=1}^n$ for all $j \in [q]$

6: # **Optimize adversarial program's parameters:**

Use Step 5 and averaged gradient estimator $\frac{1}{q}\sum_{j=1}^{q} g_j$ to obtain estimated gradient $\bar{g}(W)$, where $\{g_j\}_{j=1}^q$ are q independent random gradient estimates of the form[3, 1]

$$g_j = b \cdot \frac{Loss(W + \beta U_j) - Loss(W)}{\beta} \cdot U_j, \tag{2}$$

$W \leftarrow W - \alpha_t \cdot \bar{g}(W)$

$t \leftarrow t + 1$

7: **end while**

Experimental Results

We present the following experiments for performance evaluation.

- Reprogramming ImageNet classifiers for two social good tasks, including Autism Spectrum Disorder (ASD) classification (binary classification task) and Diabetic Retinopathy (DR) detection (5-class classification task).

Model	training size(avg.)	cnn Acc.	white-box Acc.	black-box Acc.
Resnet 50 With MLM	230	50.00%	56.80%	54.18%
	465	48.71%	62.55%	57.00%
	930	52.99%	62.03%	62.13%
Incept. V3 With MLM	230	50.00%	60.12%	57.55%
	465	48.71%	62.14%	60.21%
	930	52.99%	65.00%	61.15%

Fig. 3: Performance comparison on ASD classification task.

Model	training size(avg.)	cnn Acc.	white-box Acc.	black-box Acc.
Resnet 50 With MLM	800	71.84%	72.00%	71.46%
	1500	72.62%	72.76%	73.04%
	3000	72.65%	73.92%	73.71%
Incept. V3 With MLM	800	71.84%	72.63%	72.68%
	1500	72.62%	75.58%	73.83%
	3000	72.65%	76.42%	74.33%

Fig. 4: Performance comparison on DR classification task.

- Reprogramming two online image classification APIs from Clarifai.com for ASD and DR tasks.

Task	Training size/testing size	# of query	cnn Acc.	BAR Acc.	Cost
DR	800/2400	12.8k	71.84%	71.03%	$14.24
	1500/2400	24k	72.65%	72.75%	$23.2
ASD	459/104	11.9k	48.71%	60.14%	$13.52
	930/104	23.8k	52.99%	62.30%	$23.04

Fig. 5: Performance of BAR on Clarifai Moderation API and NSFW API.

Ablation Study and Sensitivity Analysis

- In figure 6 (a), we shows sensitivity analysis on the training data size and the number of random vectors q for gradient estimation. For multi-label mapping (MLM), as shown in figure 6 (b), the accuracy of BAR can be further enhanced with MLM.

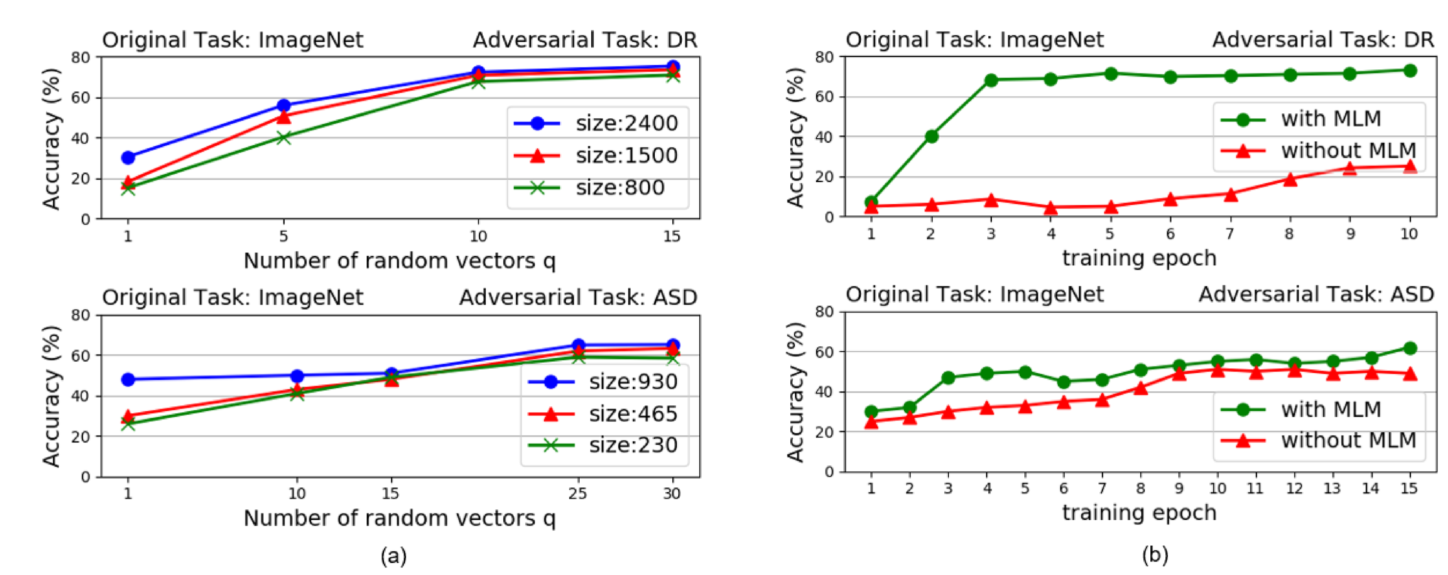

Fig. 6: Sensitive analysis

Conclusion

In this paper, we proposed a novel black-box adversarial reprogramming (BAR) method. Evaluated on two social good tasks with limited training data size, BAR showed comparable performance to the vanilla white-box AR method and outperformed baseline neural network models trained on the same dataset. We also demonstrated the practicality and effectiveness of BAR in reprogramming real-life online image classification APIs for social good tasks with low expenses (less than $24 US dollars).

References

[1] Pin-Yu Chen et al. "ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks Without Training Substitute Models". In: *ACM Workshop on Artificial Intelligence and Security*. 2017, pp. 15–26.

[2] Gamaleldin F. Elsayed, Ian J. Goodfellow, and Jascha Sohl-Dickstein. "Adversarial Reprogramming of Neural Networks". In: *ArXiv* abs/1806.11146 (2018).

[3] Chun-Chen Tu et al. "AutoZOOM: Autoencoder-based Zeroth Order Optimization Method for Attacking Black-box Neural Networks". In: *AAAI* (2019).