Adversarial Attacks On Data Attribution

Xinhe Wang[1], Pingbang Hu[2], Junwei Deng[2], Jiaqi W. Ma[2]

[1]University of Michigan [2]University of Illinois Urbana-Champaign

Motivation

Data Attribution & Compensation

- **DA** quantifies each training data's contribution to AI model outputs.
- **DA** enables appropriate **compensation for data providers**

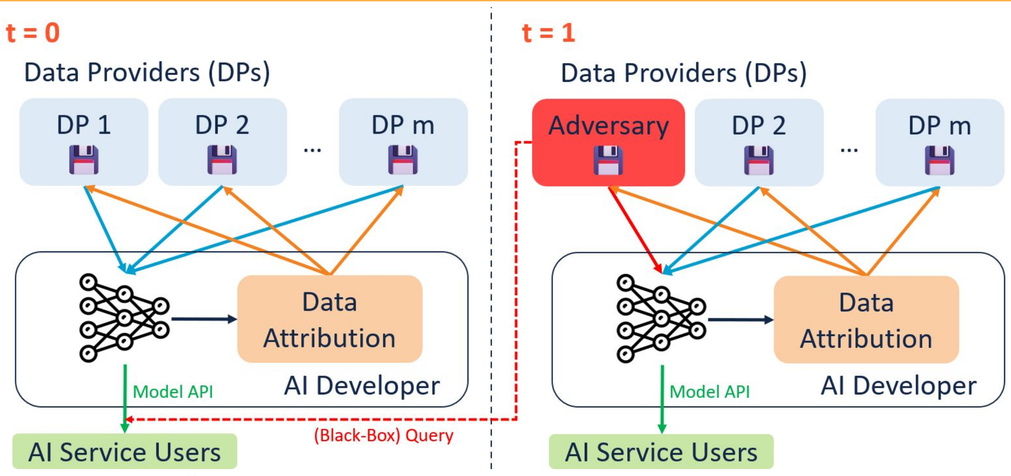
Adversarial Vulnerabilities in Data Attribution

- Financial compensation via DA might attract adversaries.
- Lack of systematic study on adversarial attacks in DA, despite their potential impact on fair compensation.

Contribution of this paper

- First comprehensive study on adversarial vulnerabilities in DA.
- Propose two novel and successful attack strategies:
  - **Shadow Attack**: Exploits data distribution knowledge via shadow models.
  - **Outlier Attack**: Black-box method leveraging outlier bias in DA.
- Our study calls on the need for robust DA to counter adversarial threats.

Threat Model



$$c(Z_1^a) = \frac{1}{k|V_1|} \sum_{z \in Z_1^a} \sum_{v \in V_1} \mathbf{1}[\tau_1(z, v) \in \mathrm{Top}_k\left(\{\tau_1(z', v) \mid z' \in Z_1\}\right)]$$

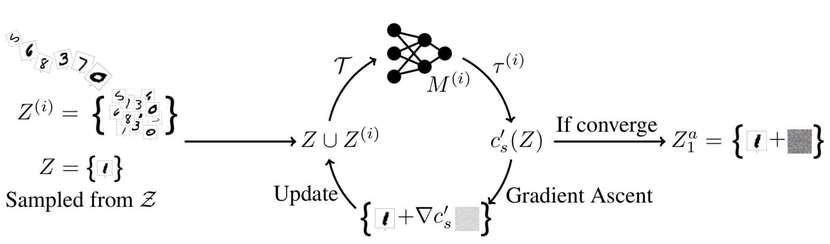
Data Compensation Scenario: Data Providers **periodically** supply training data and are compensated based on their contributions. An adversary, which is a malicious provider, can exploit prior knowledge from earlier iterations to manipulate future contributions and inflate their compensation unfairly.

Adversary's Objective and Capabilities: The adversary aims to maximize their **compensation share** by constructing a adversarial dataset. They lack access to exact datasets, trained models, or TDA functions but can exploit persistence across iterations. They can also **either** own data distribution knowledge, **or** black-box access to model predictions.

Action Space of the Adversary: The adversary is restricted to making small, undetectable perturbations to real data points.

Proposed Attack Methods

Shadow Attack

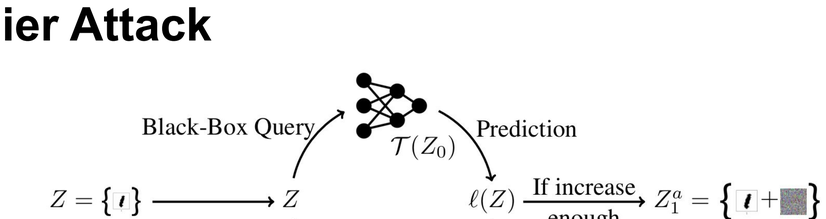


General Strategy: Leverages knowledge of data distribution to perform shadow training. Approximate, and maximize the attribution values w.r.t. the target model.

Shadow Training: Adversary trains multiple "shadow models" on shadow datasets sampled from the same distribution as the target dataset. Contribution values are computed using shadow validation data to estimate a shadow compensation share.

Adversarial Perturbation: Perturbations are applied to the adversary's dataset to maximize a surrogate compensation objective: replacing the unknown target TDA with efficient Grad-Dot and use gradient ascent to optimize contribution values.

Outlier Attack



General Strategy: Exploits the inductive bias of data attribution methods: Outliers are more influential. The adversary perturbs real-world data into realistic outliers using adversarial examples to maximize their compensation, relying only on black-box queries to the model.

Generating Realistic Outliers: Only perturbing input features, keeping labels unchanged, ensuring the perturbed data resembles real-world data and avoids detection

Adversarial Perturbation: For **image classification**, Zeroth Order Optimization (ZOO) method and Simba method are employed. For **text generation**, we use TextFooler method to generate adversarial examples by substituting tokens with tokens resulting in higher loss.

Experimental Results

Summary of Experiment Setup

Setting	Task	Dataset	Target Model	Attribution Method
(a)	Image Classification	MNIST	LR	Influence Function
(b)	Image Classification	Digits	MLP	Data Shapley
(c)	Image Classification	MNIST	CNN	TRAK
(d)	Image Classification	CIFAR-10	ResNet-18	TRAK
(e)	Text Generation	Shakespeare	NanoGPT	TRAK

Results of Shadow Attack

Setting	Shadow Model	$\lvert Z_1^a\rvert/\lvert Z_1\rvert$	Compensation Share			Fraction of Change		
			Original	Manipulated	Ratio	More	Tied	Fewer
(a)	LR	0.0098	0.0098	0.0477	456.1%	0.955	0.038	0.007
(b)	MLP	0.0352	0.0152	0.0435	286.2%	0.533	0.333	0.134
(c)	CNN	0.0098	0.0112	0.0467	417.0%	0.781	0.195	0.024
(d)	ResNet-18	0.0098	0.0095	0.0213	217.3%	0.655	0.259	0.086
(d)	ResNet-9	0.0098	0.0095	0.0196	206.3%	0.622	0.310	0.068

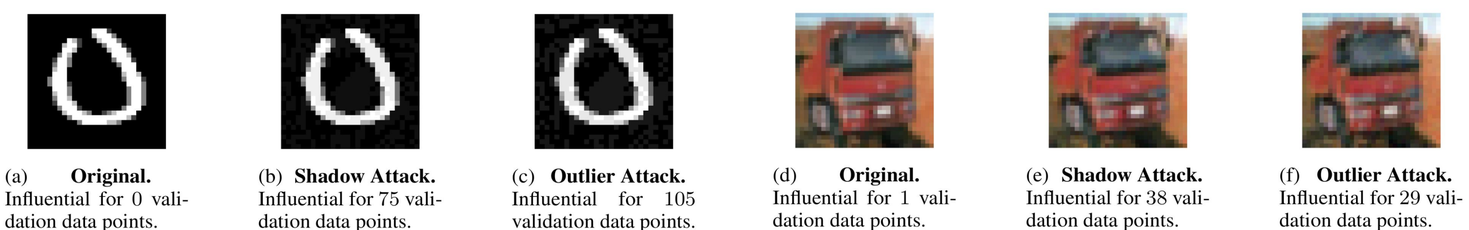
Results of Outlier Attack

Setting	Attack Method	$\lvert Z_1^a\rvert/\lvert Z_1\rvert$	Compensation Share			Fraction of Change		
			Original	Manipulated	Ratio	More	Tied	Fewer
(a)	ZOO	0.0098	0.0098	0.0631	643.9%	0.980	0.017	0.003
(b)	Simba	0.0250	0.0112	0.0218	194.6%	0.440	0.380	0.180
(c)	Simba	0.0098	0.0112	0.0668	596.4%	0.799	0.173	0.028
(d)	Simba	0.0098	0.0095	0.0176	185.2%	0.562	0.354	0.084
(e)	TextFooler	0.0031	0.0035	0.0092	262.9%	0.392	0.461	0.147

Effectiveness of Attacks: Both Attacks show significant success in increasing the CS, with increases ranging from **185.2%** to **643.9%**.

Impact on Validation Data: A high proportion of validation data points are influenced under the **More** category, indicating a broad impact of the attacks on the attribution of top-k influential points.

Success on Text Generation Task: Outlier Attack extends successfully to generative AI tasks, achieving a **262.9%** increase in compensation share on NanoGPT trained on the Shakespeare dataset.



(a) **Original.** Influential for 0 validation data points.
(b) **Shadow Attack.** Influential for 75 validation data points.
(c) **Outlier Attack.** Influential for 105 validation data points.
(d) **Original.** Influential for 1 validation data points.
(e) **Shadow Attack.** Influential for 38 validation data points.
(f) **Outlier Attack.** Influential for 29 validation data points.

Theoretical Understanding

- Train on a clean dataset with n data points, and get $\hat{\theta}$

$$I(test; i) = -\nabla_\theta l(\hat{\theta}; z_{test})^\top H_{\hat{\theta}}^{-1} \nabla_\theta l(\hat{\theta}; z_i)$$

- Adversarial perturbation: $z_i \to z_i'$

$$I'(test; i) = -\nabla_\theta l(\hat{\theta}; z_{test})^\top H_{\hat{\theta}}^{-1} \nabla_\theta l(\hat{\theta}; z_i')$$

- Retrain on the dataset with $z_i \to z_i'$, and get $\tilde{\theta}$

$$\tilde{I}(test; j) = -\nabla_\theta l(\tilde{\theta}; z_{test})^\top H_{\tilde{\theta}}^{-1} \nabla_\theta l(\tilde{\theta}; z_j), \qquad j \neq i$$

$$\tilde{I}(test; i) = -\nabla_\theta l(\tilde{\theta}; z_{test})^\top H_{\tilde{\theta}}^{-1} \nabla_\theta l(\tilde{\theta}; z_i')$$

Theorem (Informal): For strongly convex model with smooth Hessian,

- $\tilde{I}(test; j) = I(test; j) + O(1/n),\ j \neq i$ $\qquad O(1 - J(D_0, D_1))$
- $\tilde{I}(test; i) = I'(test; i) + O(1/n)$ $\qquad$ J: Jaccard Similarity

Intuition: Influence Function of two models are similar when convex, i.e., maximizing one leads to maximizing another.

Takeaway

We show that the adversarial attack on data attribution is possible and can be done efficiently. In particular, the inductive biases of the data attribution values can be exploited with a theoretical explanation.