Data-Centric Defense: Shaping Loss Landscape with Augmentations to Counter Model Inversion

Si Chen[1] Feiyang Kang[1] Nikhil Abhyankar[1] Ming Jin[1] Ruoxi Jia[1]

Abstract

Machine Learning models have shown susceptibility to various privacy attacks such as model inversion. Current defense techniques are mostly *model-centric*, which are computationally expensive and often result in a significant privacy-utility tradeoff. This paper proposes a novel *data-centric* approach to mitigate model inversion attacks which offers the unique advantage of enabling each individual user to control their data's privacy risk. We introduce several privacy-focused data augmentations which make it challenging for attackers to generate private target samples. We provide theoretical analysis and evaluate our approach against state-of-the-art model inversion attacks. Specifically, in standard face recognition benchmarks, we reduce face reconstruction success rates to $\leq 1\%$, while maintaining high utility with only a 2% classification accuracy drop, significantly surpassing state-of-the-art model-centric defenses. This is the first study to propose a data-centric approach for mitigating model inversion attacks, showing promising potential for decentralized privacy protection.

1. Introduction

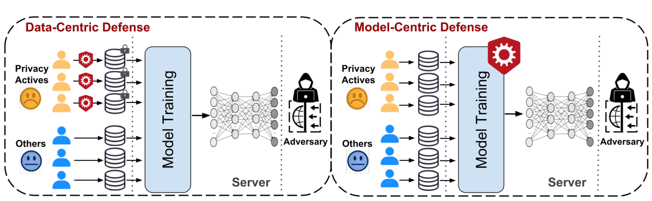

Figure 1. Data-Centric Defense vs Model-Centric Defense.

Applications of Machine Learning (ML) have undergone significant growth in recent years, showing promise across diverse fields. However, ML models trained on sensitive data risk leaking private information (Fredrikson et al., 2014; Shokri et al., 2017). While some data contributors may disregard data privacy, others, known as "privacy actives," place high importance on it, taking active measures including changing service providers (Cisco, 2019). Legislation such as the GDPR (Magdziarczyk, 2019) and the California Consumer Privacy Act (Pardau, 2018) also advocate for individual data control.

Existing defenses (Abadi et al., 2016; Jia et al., 2019; Wang et al., 2021; Yang et al., 2020) primarily adopt a model-centric approach, altering model training (Abadi et al., 2016) or inference procedures (Jia et al., 2019). These defenses, however, necessitate users to trust the model trainer (such as the companies that harvest their data) to implement privacy safeguards, limiting users' control over their privacy risk. Moreover, these modifications often lead to performance degradation and increased computation time.

This work develops the first data-centric defense for MI attacks, outlining our technical contributions: 1) **We propose privacy-focused data augmentations that can be injected by individual data contributors to mitigate their MI risks.** Our approach, DCD, protect against MI attacks by shaping the loss landscape to mislead attacks and recover irrelevant samples; and requires no access to the victim model or training data from other contributors. 2) **We provide theoretical justification for DCD**. 3) We evaluate DCD against various state-of-the-art MI attacks and demonstrate the robustness of DCD across different datasets, model architectures, and attack strategies. **Remarkably, DCD achieves a near-zero privacy-utility tradeoff.**

2. Our Privacy-Focused Data Augmentations

Our approach introduces surrogate classes into the training set, designing augmentations to misdirect MI attacks toward recovering surrogate-class samples instead of target-class samples. We explain this process using a specific target class (y_{tgt}) that has m training samples for protection. When multiple target classes need protection, one can easily apply the following process to each target class.

Surrogate Injection. We start by selecting an "irrelevant"

[1]Virginia Tech, Blacksburg, VA. Correspondence to: Si Chen <chensi@vt.edu>, Ruoxi Jia <ruoxijia@vt.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

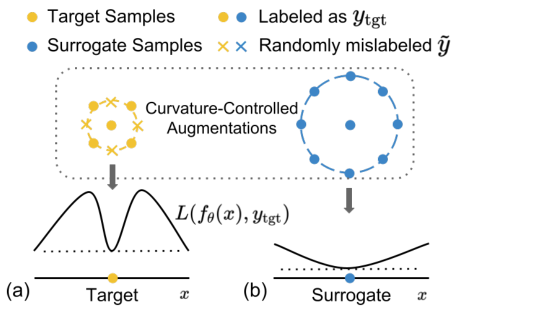

Figure 2. Illustration of curvature-controlled augmentations and the resulting loss landscape.

surrogate class y_{srg} that doesn't reveal sensitive information. We gather m samples from this surrogate class and relabel them as the target class. The model trained on this mixture classifies both surrogate and target samples as the target class, making MI attacks generate a mix of both.

Loss-Controlled Modification. We now design the curvature to *further induce MI attacks to preferentially generate samples from the surrogate class over the target class.* MI attacks essentially resolve optimization problems, seeking samples that result in the lowest loss when predicted as the target class. To counteract this, our first strategy modifies training data to slightly elevate the classification loss on the target compared to the surrogate, increasing the likelihood of detecting surrogate samples during MI optimization while reducing the chance for target samples. We accomplish this by randomly mislabeling a small fraction of target samples, while leaving the surrogate samples' labels unaltered.

Curvature-Controlled Injection. Leveraging the insight from non-convex optimization theory (Bertsekas, 1997), our second strategy manipulates the loss landscape's curvature, promoting a flatter curvature around surrogate samples and a steeper one near target samples (illustrated by Figure 2). This approach biases the MI optimization towards reconstructing surrogate samples. For surrogate samples, we employ Gaussian augmentations in their neighborhood, maintaining the same label. For target samples, we apply Gaussian augmentations but mislabel a portion of the augmented samples. We refer to the complete injection process as **DCD**. We provide theoretical analysis in the full paper.

Table 1. Defense performance comparison against various MI attacks, results given in %. ↑ and ↓ respectively symbolize that higher and lower scores give better defense performance.

	GMI TSRD→GTSRB		PPA FFHQ→CelebA	
	ACC↑	Att. ACC↓	ACC↑	Att. ACC↓
No Protection	98.34	76.13	88.42	90.40
DP	54.30	12.80	39.61	14.33
MID	67.70	54.53	69.54	52.33
DCD (Ours)	95.89	0.00	88.05	1.00
	MIRROR-W FFHQ→VGGFace2		**MIRROR-B FFHQ→VGGFace2**	
	ACC ↑	Att. ACC↓	ACC	Att. ACC↓
No Protection	99.99	100.0	99.99	100.0
DP	56.25	54.69	56.25	50.00
MID	41.34	100.00	41.34	12.50
DCD (Ours)	96.88	0.00	96.88	0.00

3. Experimental Results

We assess the effectiveness of DCD against three white-box MI attacks: GMI (Zhang et al., 2020), PPA (Struppek et al., 2022), and MIRROR-W (An et al., 2022), and one most recent black-box attack, MIRROR-B. We comapre DCD with DP-SGD (Abadi et al., 2016) and MID (Wang et al., 2021). For consistency, we randomly select multiple target classes and average the results. As shown in Table 1, DCD outperforms the baselines in both utility (classification accuracy ACC) and privacy (attack accuracy Att.ACC) metrics. The unprotected models exhibit alarmingly high attack accuracy, with GMI at 76%, PPA at 90%, and MIRROR at 100%. In contrast, DCD significantly reduces the attack accuracy to 0% for both GMI and MIRROR attacks, and to 1% for PPA. Figure 3 shows that DCD successfully fools MI into generating samples resembling the surrogate ones. More visual results are provided in the full paper. A notable advantage of DCD is its ability to balance privacy and utility well. Unlike DP and MID, which exhibit a substantial drop in classification accuracy, our method ensures high classification accuracy, with a decrease of less than 3% on the face datasets CelebA and VGGFace2. More evaluation are provided in the full paper.

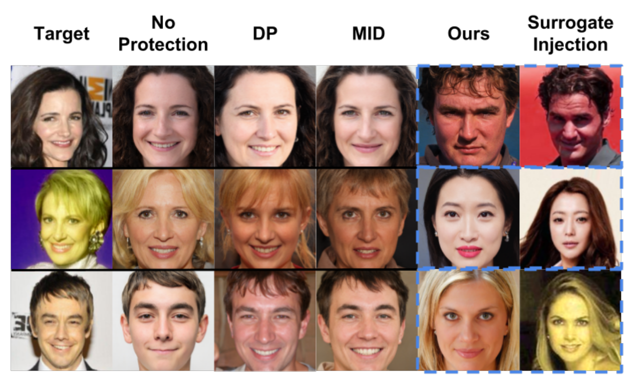

Figure 3. Visual comparison of MI recovered face samples with different defenses. Each row shows reconstructions of the same identity under different defenses, with true images on the left and our surrogate injection on the right.

4. Conclusion

Our paper introduces the first user-empowered, data-centric defense mechanism, DCD, for mitigating data privacy risks. Supported by theoretical analysis and extensive evaluations, DCD effectively counters model inversion attacks and surpasses model-centric baselines in utility and privacy. It does, however, increase the number of samples in the target classes by a factor of 4, potentially alerting malicious model trainers. Future work aims to obscure these injected samples, thereby addressing this limitation.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016. 1, 2

An, S., Tao, G., Xu, Q., Liu, Y., Shen, G., Yao, Y., Xu, J., and Zhang, X. Mirror: Model inversion for deep learning network with high fidelity. In *Proceedings of the 29th Network and Distributed System Security Symposium*, 2022. 2

Bertsekas, D. P. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997. 2

Cisco. Cisco 2019 consumer privacy survey (report), 2019. URL https://www.cisco.com/c/dam/global/en_uk/products/collateral/security/cybersecurity-series-2019-cps.pdf. 1

Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., and Ristenpart, T. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, pp. 17–32, 2014. 1

Jia, J., Salem, A., Backes, M., Zhang, Y., and Gong, N. Z. Memguard: Defending against black-box membership inference attacks via adversarial examples. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pp. 259–274, 2019. 1

Magdziarczyk, M. Right to be forgotten in light of regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec. In *6th International Multidisciplinary Scientific Conference on Social Sciences and Art Sgem 2019*, pp. 177–184, 2019. 1

Pardau, S. L. The california consumer privacy act: Towards a european-style privacy regime in the united states. *J. Tech. L. & Pol'y*, 23:68, 2018. 1

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017. 1

Struppek, L., Hintersdorf, D., Correira, A. D. A., Adler, A., and Kersting, K. Plug & play attacks: Towards robust and flexible model inversion attacks. In *International Conference on Machine Learning*, pp. 20522–20545. PMLR, 2022. 2

Wang, T., Zhang, Y., and Jia, R. Improving robustness to model inversion attacks via mutual information regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11666–11673, 2021. 1, 2

Yang, Z., Shao, B., Xuan, B., Chang, E.-C., and Zhang, F. Defending model inversion and membership inference attacks via prediction purification. *arXiv preprint arXiv:2005.03915*, 2020. 1

Zhang, Y., Jia, R., Pei, H., Wang, W., Li, B., and Song, D. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 253–261, 2020. 2