POSTER: Double-Dip: Thwarting Label-Only Membership Inference Attacks with Transfer Learning and Randomization

Arezoo Rajabi*
University of Washington
Seattle, USA
rajabia@uw.edu

Reeya Pimple*
University of Washington
Seattle, USA
reeyabp@uw.edu

Aiswarya Janardhanan
University of Washington
Seattle, USA
ajanard5@uw.edu

Surudhi Asokraj
University of Washington
Seattle, USA
surudh22@uw.edu

Bhaskar Ramasubramanian
Western Washington University
Bellingham, USA
ramasub@wwu.edu

Radha Poovendran
University of Washington
Seattle, USA
rp3@uw.edu

ABSTRACT

Transfer learning (TL) has been demonstrated to improve DNN model performance when faced with a scarcity of training samples. However, the suitability of TL as a solution to reduce vulnerability of overfitted DNNs to privacy attacks is unexplored. A class of privacy attacks called membership inference attacks (MIAs) aim to determine whether a given sample belongs to the training dataset (*member*) or not (*nonmember*). We introduce **Double-Dip** to investigate the use of TL (Stage-1) combined with randomization (Stage-2) to thwart MIAs on overfitted DNNs without degrading classification accuracy. Our study examines roles of shared feature space and parameter values between *source* and *target* models, number of frozen layers, and complexity of *pretrained* models. Our preliminary evaluations of Double-Dip demonstrate that Stage-1 reduces adversary success while also significantly increasing classification accuracy of nonmembers against an adversary attempting to carry out SOTA label-only MIAs. After Stage-2, success of an adversary carrying out a label-only MIA is further reduced to near 50%, bringing it closer to a random guess and showing the effectiveness of Double-Dip. Stage-2 of Double-Dip also achieves lower ASR and higher classification accuracy than regularization and differential privacy-based methods.

KEYWORDS

Transfer learning, membership inference attack

ACM Reference Format:
Arezoo Rajabi, Reeya Pimple, Aiswarya Janardhanan, Surudhi Asokraj, Bhaskar Ramasubramanian, and Radha Poovendran. 2024. POSTER: Double-Dip: Thwarting Label-Only Membership Inference Attacks with Transfer Learning and Randomization. In *Proceedings of ACM ASIA Conference on Computer and Communications Security, Singapore, Singapore, July 1–5, 2024 (ASIA CCS '24),* 3 pages.
https://doi.org/10.1145/3634737.3659429

*Equal contribution

ASIA CCS '24, July 1–5, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0482-6/24/07.
https://doi.org/10.1145/3634737.3659429

1 INTRODUCTION

The ability of deep neural networks (DNNs) to classify previously unseen inputs with high accuracy relies critically on being trained on large datasets, and requires significant training-time computational resources [2]. In the absence of an adequate number of training samples, the DNN model can suffer from *overfitting*. Overfitted DNNs have been shown to 'memorize' patterns in the data and classify samples belonging to the training dataset with high accuracy, while performing poorly on other samples [16].

Overfitted DNNs have been shown to be vulnerable to privacy attacks such as a *membership inference attack (MIA)* [16]. MIAs aim to determine if a given sample of interest belongs to the training dataset (*member*) of a DNN model or not (*nonmember*) [15]. MIAs can result in disclosure of sensitive information (e.g., social-security numbers), resulting in privacy threats. Techniques including differential privacy [1], regularization [14], and distillation [18] have been used as defenses against MIAs. However, these methods can also lower classification accuracy for overfitted DNNs [15], which can affect model usability. Further, their effectiveness on a new class of MIAs called *label-based* or *label-only* MIA [4, 11, 15] is less understood. Finding solutions to mitigate impacts of label-only MIAs while improving classification accuracy for overfitted DNNs remains an open problem.

Our Contribution: We propose **Double-Dip**, a systematic study of using transfer learning (TL) to overcome overfitting in the limited data setting, thus resulting in thwarting of label-only MIAs. While the usefulness of TL in the general limited data setting is well-known, we show in this paper that TL will indeed be helpful even in the case of overfitted DNNs. In Double-Dip Stage-1, we demonstrate that TL [19] will help embed an otherwise low-dimensional overfitted model into a high-dimensional *target model* that will be less overfitted. In Stage-2, we employ randomization to construct a region of constant output label centered at a given input sample such that the DNN model returns the same output label for all data points inside this region [5, 15]. Stage-2 will help further reduce success rate of an adversary carrying out a label-only MIA, which is the most powerful known MIA to date [4], without reducing classification accuracy (relative to Stage-1). Together, the two stages will help reduce success of an adversary carrying out a label-only MIA while also yielding a target model with high accuracy. Fig. 1 illustrates the mechanism of Double-Dip.

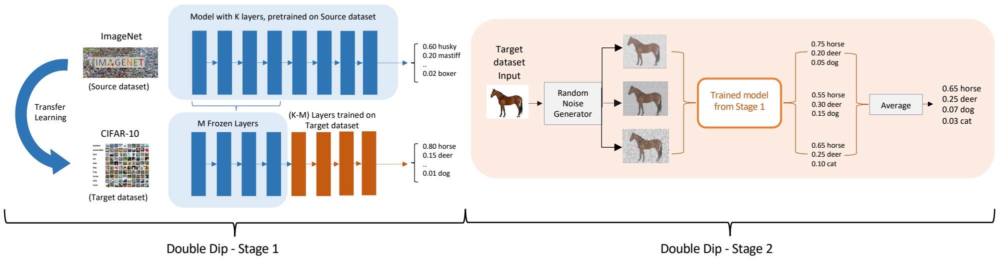

Figure 1: Double-Dip Mechanism. Stage-1 uses transfer learning to embed features of a lower dimensional overfitted DNN into a *target model* that overcomes overfitting. The target model is learned by 'freezing' weights in M layers of a public pretrained model, and using samples from the target dataset to learn weights of the remaining $K - M$ layers. Stage-2 employs randomization to generate multiple noisy variants of a given sample x. Each noisy variant is provided to the trained target model from Stage-1 to obtain possible output class labels as probabilities. An averaging mechanism is used to 'smooth' these output class labels to obtain the final output class label. Randomization will affect estimates of the distance of a data point to a decision boundary. As a result, the final output label y will not reveal information about whether x was used to train the target model (member) or not (nonmember).

2 THREAT MODEL

Adversary Assumption and Goals: The adversary is assumed to have adequate data samples and computational resources, and uses a SOTA *label-only MIA* [4] to determine if a given input is contained in the set used to train the model (*member*). The magnitude of noise which enables an adversary to distinguish between members and nonmembers is based on a heuristic that members are relatively farther away from a decision boundary and are robust to small noise perturbations compared to a nonmember [4, 15].

Adversary Actions: We consider two levels of access to the target DNN: (i) white-box access, where the adversary has access to model hyperparameters and output labels, and (ii) black-box access, where the adversary has access only to model outputs. An adversary with white-box access uses an adversarial learning method, e.g., *basic iterative method (BIM)* [10], to estimate a threshold δ on noise to be added to a sample for it to be misclassified by the DNN. An adversary with black-box access uses a query-based SOTA adversarial learning method (e.g., *HopSkipJump* [3]) to estimate δ.

3 DOUBLE-DIP: A TWO-STAGE APPROACH

We describe the two-stage procedure of **Double-Dip**. Performance of Double-Dip will be assessed in terms of adversary success rate (ASR- *closer to 50.0% is better*) and classification accuracy of nonmembers (ACC- *higher is better*). Stage-1 uses transfer learning (TL) [19] to embed a lower dimensional DNN into a high-dimensional target model to overcome overfitting. Stage-2 employs randomization based on noise perturbation of a given input to construct a high-dimensional region of constant output label such that the DNN returns the same label for samples in this sphere [5, 15].

Stage-1: When a user possesses only a limited number of samples to train a DNN, the resulting model becomes *overfitted*, lowering classification accuracy for nonmembers while having high accuracy for members. Our insight is that TL helps embed an otherwise low-dimensional overfitted model into a high-dimensional model that will no longer be overfitted. The success of Stage-1, however, will depend on an interplay among several design choices, including the type of pretrained model, source and target datasets, and number of frozen layers of the pretrained model.

To examine roles of these design choices, we consider two target datasets- CIFAR-10 [9] and GTSRB [8]- to learn a target model from a pretrained model that has been trained on ImageNet [6] as source dataset. These target datasets have different levels of similarity in their features with those of the source dataset.

Stage-2: The use of transfer learning in Stage-1 yields a target model embedded in a higher dimensional space that is less overfitted, thus readily reducing success rate of an adversary carrying out a MIA[15, 16]. Stage-2 employs a lightweight post-processing module that seeks to further reduce ASR of label-only MIAs without needing to retrain target models. A given sample x is perturbed by a zero-mean Gaussian noise with variance σ^2. Stage-2 of Double-Dip tunes the value of σ to lower ASR while maintaining high accuracy. We hypothesize that using Stages-1 & 2 together will result in a lower ASR compared to using Stage-1 alone. We compare performance of Double-Dip with SOTA training-phase defenses against MIAs, including regularization [14] and distillation training [18].

4 DOUBLE-DIP: PRELIMINARY EVALUATIONS

We evaluate Double-Dip Stage-1 by examining effectiveness of TL when the adversary carries out a label-only MIA to estimate a threshold δ that will result in a given sample being misclassified by the target model. We then evaluate Double-Dip Stage-2 to investigate if ASR can be reduced further, without reducing accuracy. Our preliminary results shown in Table 1 and Fig. 2 demonstrate that Stages-1&-2 of Double-Dip effectively thwarts label-only MIAs.

5 CONCLUSION

This paper presented a work-in-progress in developing *Double-Dip*, a systematic empirical study of the role of transfer learning (TL) in thwarting label-only membership inference attacks (MIAs) on overfitted deep neural networks (DNNs). Our preliminary experiments have shown efficacy of Stages-1&-2 of Double-Dip in thwarting label-only MIAs. The complete study of Double-Dip's performance will include detailed examination on a complex face recognition task using CelebA [13] to learn a target model, and the effect of different SOTA pretrained models trained on ImageNet- e.g., VGG-19 [17], ResNet-18 [7], and Swin-T [12].

Table 1: Stage-1 of Double-Dip, Pretrained VGG-19 Model: Adversary success rate (ASR, *lower is better*) and classification accuracy (ACC, *higher is better*) for CIFAR-10 and GTSRB datasets with training sets of sizes 500 and 1000. We compare (i) no transfer learning (NTL), (ii) regularization (L1, L2), and (iii) transfer learning (TL). TL-X indicates that X layers of the pretrained model are frozen. We examine scenarios when an adversary carrying out an MIA has (a) white-box model access (BIM), and (b) black-box model access (HSJ). The best ASR and ACC values for a given training set size across both datasets is in **bold**; best ASR and ACC values in each cell are underlined. TL yields lowest ASR values while also ensuring significantly higher accuracy.

		500			1000		
Dataset	**Setting**	**%ASR(BIM)**	**%ASR(HSJ)**	**%ACC**	**%ASR(BIM)**	**%ASR(HSJ)**	**%ACC**
CIFAR-10	NTL	87.5	87.5	24.6	88.7	88.5	27.7
	L1(0.001)	90.1	88.9	23.6	86.5	85.8	28.0
	L2(0.1)	89.7	88.9	23.0	83.8	84.9	30.3
	TL-0	60.1	60.6	**79.2**	59.9	61.5	80.9
	TL-20	**59.9**	**60.3**	78.6	59.4	61.5	80.0
	TL-35	62.9	63.5	72.2	63.7	63.9	76.1
GTSRB	NTL	76.0	76.7	40.8	76.7	74.8	54.3
	L1 (0.001)	82.0	81.5	37.2	69.7	69.5	61.9
	L2 (0.1)	76.2	76.2	43.8	67.8	67.3	62.9
	TL-0	63.0	63.0	73.2	**58.7**	**57.0**	**85.9**
	TL-20	63.0	63.0	73.6	61.3	62.0	81.5
	TL-35	70.0	70.0	59.0	67.3	68.8	64.4

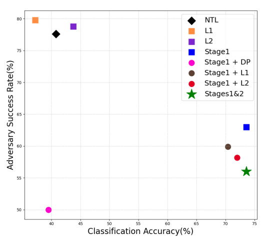

Figure 2: Stages-1&2 of Double-Dip vs. SOTA: ASR (*lower is better*) and ACC (*higher is better*) for 500 training samples from GTSRB with a pretrained VGG-19 model when using (i) no transfer learning (NTL), (ii) regularization (L1/ L2), (iii) Double-Dip Stage-1, (iv) Double-Dip Stage-1 + diff. privacy (Stage-1+DP), (v) Double-Dip Stage-1 + regularization (Stage-1+L1/ L2), and (vi) Stages-1&2 of Double-Dip. Stages-1&2 of Double-Dip achieves low ASR values while simultaneously ensuring high ACC. While Stage-1+DP achieves lowest ASR, it comes with a significant drop in accuracy.

ACKNOWLEDGMENTS

This work is supported by the AFOSR via grant FA9550-23-1-0208. The work is also supported in part by the ONR via grant N00014-23-1-2386 and by the NSF via grants IIS 2229876 and CNS 2153136, and by funds provided by the DHS, and IBM. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect views of the NSF or its federal agency and industry partners.

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *ACM SIGSAC Conference on Computer and Communications Security*. 308–318.

[2] Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision?. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1536–1546.

[3] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. 2020. HopSkipJumpAttack: A query-efficient decision-based attack. In *IEEE Symposium on Security and Privacy*. IEEE, New York, NY, USA, 1277–1294.

[4] Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. 2021. Label-only membership inference attacks. In *International Conference on Machine Learning*. PMLR, 1964–1974.

[5] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. 2019. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*. PMLR, 1310–1320.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 248–255.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[8] Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. 2013. Detection of Traffic Signs in Real-World Images: The German Traffic Sign Detection Benchmark. In *Intl. Joint Conf. on Neural Networks*.

[9] Alex Krizhevsky and Geoffrey Hinton. 2009. *Learning multiple layers of features from tiny images*. Technical Report 0. University of Toronto, Toronto, Ontario.

[10] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. 2018. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*. 99–112.

[11] Zheng Li and Yang Zhang. 2021. Membership leakage in label-only exposures. In *Proc. ACM SIGSAC Conf. on Computer and Communications Security*. 880–895.

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. International Conf. on Computer Vision*. 10012–10022.

[13] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *International Conference on Computer Vision (ICCV)*.

[14] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2018. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. 634–646.

[15] Arezoo Rajabi, Dinuka Sahabandu, Luyao Niu, Bhaskar Ramasubramanian, and Radha Poovendran. 2023. LDL: A Defense for Label-Based Membership Inference Attacks. In *Proc. ACM Asia Conf. on Computer and Communications Security*.

[16] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (SP)*. IEEE, New York, NY, USA, 3–18.

[17] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *ICLR* (2015).

[18] Xinyu Tang, Saeed Mahloujifar, Liwei Song, Virat Shejwalkar, Milad Nasr, Amir Houmansadr, and Prateek Mittal. 2022. Mitigating membership inference attacks by {Self-Distillation} through a novel ensemble architecture. In *USENIX Security Symposium*. 1433–1450.

[19] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2020. A comprehensive survey on transfer learning. *Proc. IEEE* 109, 1 (2020), 43–76.