Membership Inference Attacks and Defenses in Supervised Learning via Generalization Gap

Jiacheng Li, Ninghui Li, Bruno Ribeiro

Content

- 总结现有攻击和防御方案
- 提出的攻击和防御方法
- 实验结果

成员推理攻击

- 推理某个目标数据是否参与目标模型的训练，即其是否在目标模型的训练集中
  - i.e. 给定某个目标数据 (x, y)，攻击者可以判断出其是否在目标模型 M 的训练集中，在训练集中，认定其为训练集的“member”；否则，认为其是“non-member”，即非训练集成员。

现有MIA攻击方案

Attacks	Granularity	Feature	Method	Adversary Model	Source
Class-Vector	Class	Probability vector $F^T(x)$	Neural Network	Training plus Data	Shokri et al. 2017 [25]
Instance-Vector	Instance	Probability vector $F^T(x)$	KL distance to avg	Training plus Data	Long et al. 2017 [19]
Global-Loss	Global	Training loss $Loss(x,y)$	Threshold	Training plus Data	Yeom et al. 2018 [31]
Global-Probability	Global	Probability of correct label	Threshold	Training plus Data	Variant of Global-Loss
Global-TopThree	Global	Top 3 in $F^T(x)$	Neural Network	Training plus Data	Salem et al. 2018 [24]
Global-TopOne	Global	Top 1 in $F^T(x)$	Threshold	Probability-Vector Oracle	Salem et al. 2018 [24]
Baseline	Global	Predicted class y'	Binary	Label-only Oracle	Yeom et al. 2018 [31]
Instance-Probability	Instance	Probability of correct label	Threshold	Training plus Data	This paper

[25] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In Security and Privacy (SP), 2017 IEEE Symposium on, pages 3–18. IEEE, 2017.
[19] Yunhui Long, Vincent Bindschaedler, and Carl A Gunter. Towards measuring membership privacy. arXiv preprint arXiv:1712.09136, 2017.
[31] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In 2018 IEEE 31st Computer Security Foundations Symposium (CSF), pages 268–282. IEEE, 2018.
[24] Ahmed Salem, Yang Zhang, Mathias Humbert, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In 25th Annual Network and Distributed System Security Symposium (NDSS), 2019.

现有MIA攻击方案

- The Class-Vector Attack [25]
- The Baseline (Global-Label) Attack [31]
- The Global-Loss Attack [31]
- The Global-Probability Attack
- The Global-TopOne Attack [24]
- The Global-TopThree Attack [24]
- The Instance-Vector Attack [19]

The Class-Vector Attack [25]. To our knowledge, Shokri et al. [25] presented the first study on MI attacks against classifiers. The attack is in the "Training-plus-data" adversary model, and for each instance x, the probability vector $F^T(x)$ is the feature for determining whether x is used in training F^T.

The adversary knows a dataset D^A, which is from the same distribution as the dataset used to train the target classifier. The adversary creates k samples $D_1, D_2, \ldots, D_k$ from D^A, and trains k shadow classifiers $F_1^S, F_2^S, \ldots, F_k^S$, one from each D_i. These shadow classifiers generate training data for MI classifiers. The attacker trains c MI classifiers, one for each class. The classifier for class y is trained using instances in D^A that are of class y. For each such instance x, one can obtain k instances for training the MI classifier for y, one from each shadow classifier. With the i-th shadow classifier, one has the probability vector $F_i^S(x)$ as the feature, and whether $x \in D_i$ as the label. Each MI classifier takes a probability vector $F^T(x)$ as input, and produces a binary classification result. These attack classifiers are feed-forward neural network with one fully-connected hidden layer of size 64, with ReLU activation functions. When trying to determine the membership of an instance $(x)_y$, one feeds $F^T(x)$ to the MI classifier for class y to obtain a binary membership prediction.

现有MIA攻击方案

- The Class-Vector Attack [25]
- The Baseline (Global-Label) Attack [31]
- The Global-Loss Attack [31]
- The Global-Probability Attack
- The Global-TopOne Attack [24]
- The Global-TopThree Attack [24]
- The Instance-Vector Attack [19]

The Baseline (Global-Label) Attack [31]. Yeom et al. [31] analyzed the relationship between overfitting and membership, and proposed two attacks, one of which predicts that an instance x is a member for training F^T if and only if $F^T(x)$ gives the correct label on x. This attack can be applied to the "Label-only Oracle" adversary model, in which the adversary is given only the label.

现有MIA攻击方案

- The Class-Vector Attack [25]
- The Baseline (Global-Label) Attack [31]
- The Global-Loss Attack [31]
- The Global-Probability Attack
- The Global-TopOne Attack [24]
- The Global-TopThree Attack [24]
- The Instance-Vector Attack [19]

The Global-Loss Attack [31]. This attack uses the probability vector $F^T(x)$ for an instance x with true label y to compute the cross-entropy loss: $Loss(x,y) = -\log(F^T(x)_y)$, where $F^T(x)_y$ is the probability value for the true label y. The attack predicts x to be a member when $Loss(x,y)$ is smaller than the average loss of all training instances. We consider this attack to be in the Training-plus-Data adversary model, because the average training loss is not normally provided. The natural way to obtain it is to build one or more shadow classifiers.

现有MIA攻击方案

- The Class-Vector Attack [25]
- The Baseline (Global-Label) Attack [31]
- The Global-Loss Attack [31]
- The Global-Probability Attack
- The Global-TopOne Attack [24]
- The Global-TopThree Attack [24]
- The Instance-Vector Attack [19]

The Global-Probability Attack. We note that the Global-Loss attack effectively predicts an instance to a member if the probability for the correct label is above some threshold. It fixes the threshold based on the average value for all training instances. This threshold may not achieve the maximum accuracy. We thus also consider using shadow classifiers and training data to compute the threshold that achieves the best accuracy. We call this the Global-probability attack.

现有MIA攻击方案

- The Class-Vector Attack [25]
- The Baseline (Global-Label) Attack [31]
- The Global-Loss Attack [31]
- The Global-Probability Attack
- The Global-TopOne Attack [24]
- The Global-TopThree Attack [24]
- The Instance-Vector Attack [19]

The Global-TopOne Attack [24]. Instead of using the probability of the correct label, Salem et al. [24] proposed to use the highest value in the probability vector. They proposed an interesting threshold-choosing approach that exploits oracle access to the target classifier. One randomly generates some data instances, which are non-members with high probability, and query the target classifier with these instances. One then chooses the threshold using the top t percentile among the Top 1 probabilities from the probability vectors of these instances. Experiments on different t in range from 5% to 25% showed decent performance.

现有MIA攻击方案

- The Class-Vector Attack [25]
- The Baseline (Global-Label) Attack [31]
- The Global-Loss Attack [31]
- The Global-Probability Attack
- The Global-TopOne Attack [24]
- The Global-TopThree Attack [24]
- The Instance-Vector Attack [19]

The Global-TopThree Attack [24]. Salem et al. [24] proposed an attack to use only the top three values in the probability vector for MI attack. This is in part motivated by considering the "Probability-Vector Oracle" model, in which the adversary cannot train shadow classifiers to generate training data for MI classifiers. Salem et al. [24] proposed a data transferring attack, where the adversary trains one shadow classifier using a different dataset and different classifier structure. Since the number of classes may be different for the shadow classifier and the target classifier, the adversary chooses top 3 values in the probability vector (top 2 in case of binary classification) as the features for MI attack. Furthermore, only a single global MI classifier is used.

现有MIA攻击方案

- The Class-Vector Attack [25]
- The Baseline (Global-Label) Attack [31]
- The Global-Loss Attack [31]
- The Global-Probability Attack
- The Global-TopOne Attack [24]
- The Global-TopThree Attack [24]
- The Instance-Vector Attack [19]

The remaining two attacks in [19] train instance-specific MI classifiers. That is, there is one MI classifier for each instance x. To enable this, one creates k samples $D_1, D_2, \cdots, D_k$ of D^A, and trains $2k$ shadow classifiers, where F_i^S is trained with D_i, and $F_i'^S$ is trained with $D_i \cup \{x\}$. For each instance x, one has k probability vectors $F_i^S(x)$, where $1 \leq i \leq k$, from classifiers trained without x, and k probability vectors $F_i'^S(x)$, from classifiers trained with x. The intuition is that if x is used in training F^T, then the probability vector $F^T(x)$ should be more similar to the latter k than to the former k.

Long et al. [19] investigated two ways to measure this similarity, and found that the more effective approach is to use the Kullback-Leibler (KL) divergence. More specifically, the adversary calculates $\overline{F^S(x)}$, the average of all $F_i^S(x)$, and $\overline{F'^S(x)}$, the average of all $F_i'^S(x)$. That is, $\overline{F^S(x)}$ is the average probability vector on x for classifiers training without using x, and $\overline{F'^S(x)}$ is the average probability vector for classifiers trained using x. To determine whether x is used to train F^T, one computes the Kullback-Leibler (KL) divergence of $\overline{F^S(x)}$ and $\overline{F'^S(x)}$ with $F^T(x)$. The KL divergence between two discrete probability distributions P and Q is defined to be $D_{\mathrm{KL}}(P \parallel Q) = -\sum_{x \in \mathcal{X}} P(x) \log\left(\frac{Q(x)}{P(x)}\right)$. If $D_{\mathrm{KL}}\left(\overline{F^S(x)}, F^T(x)\right) < D_{\mathrm{KL}}\left(\overline{F^{S'}(x)}, F^T(x)\right)$, one predicts that x is not used in training F^T, otherwise the adversary will predict this instance x to be used in the training.

现有MIA防御方案

- L2-Regularizer and modifying predictions [25]
- Min-Max Game [21]
- Dropout [24,27]
- Model Stacking [24]
- Mem-Guard [15]
- Differential Privacy [1]

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

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

[24] Ahmed Salem, Yang Zhang, Mathias Humbert, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In 25th Annual Network and Distributed System Security Symposium (NDSS), 2019.

[27] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1):1929–1958, 2014.

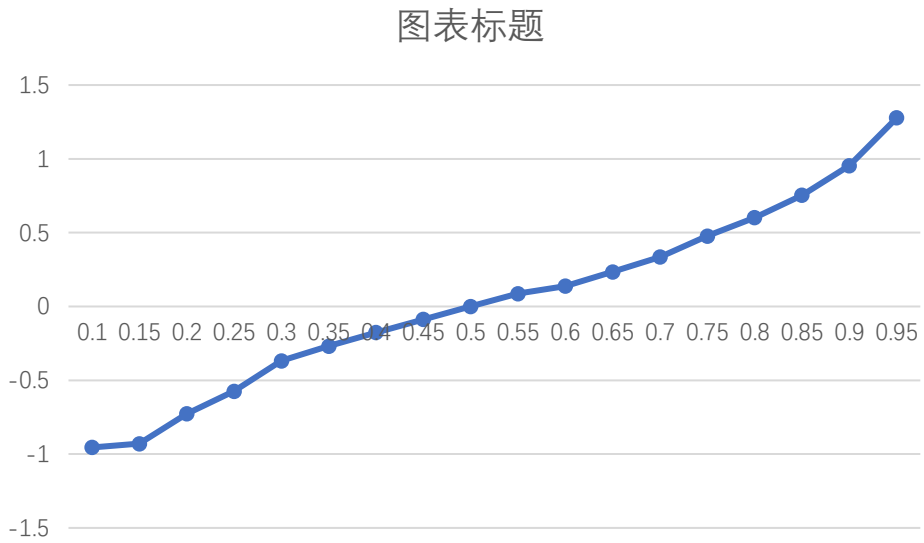
[15] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. Memguard: Defending against black-box membership inference attacks via adversarial examples. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, pages 259–274, 2019.

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

提出的攻击方法

- The Instance-Probability Attack

The Instance-Probability Attack. We introduce a new attack that uses $F^T(x)_y$, the probability value for the true label y. The difference between the Global-Probability attack and this attack is that the former trains a single classifier that takes $F^T(x)_y$ and determines whether (x,y) is a member, and this attack trains a different classifier for each instance (x,y). This attack uses the "Training plus data" adversary model, and requires training multiple shadow classifiers so that for each instance (x,y), there exist a number of shadow classifiers trained with (x,y), and a number of shadow classifiers trained without (x,y). We use $q(x)_y = log\left(\frac{F^T(x)_y}{1-F^T(x)_y}\right)$ as the feature for determining the membership of (x,y). The MI classifier essentially finds a threshold based on the training data generated from the shadow classifiers.



提出的防御方法

- Main idea：减少generalization gap来实现防御
- 实现方法：

We propose to intentionally reduce training accuracy to match testing accuracy. To achieve this goal, we add to the training loss function a regularizing term that is the difference between the output distribution of the training set and that of a validation set.

- 评估距离：Maximum Mean Discrepancy (MMD)

crepancy (MMD) [8, 10]. MMD is used to construct statistical tests to determine if two samples are drawn from different distributions, based on Reproducing Kernel Hilbert Space (RKHS) [3]. Let $X = (x_1, \cdots, x_n)$ and $Y = (y_1, \cdots, y_m)$ be the random variable sets drawn from distrubtion $\mathcal{P}$ and Q. The empirical estimation of distance between $\mathcal{P}$ and Q, as defined by MMD, is:

$$\text{Distance}(X, Y) = \left\| \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) - \frac{1}{m} \sum_{j=1}^{m} \phi(y_j) \right\|_{\mathcal{H}}, \quad (2)$$

where $\mathcal{H}$ is a universal RKHS, and $\phi : \mathcal{X} \mapsto \mathcal{H}$, and x_i (y_i) is the softmax output of the i-th training (validation) instance.

提出的防御方法

- Mix-up training

3.2.2 Mix-up Training Augmentation

We combine MMD with mix-up training, first introduced by Zhang et al. [32]. This training strategy is to use linear interpolation of two different training instances to generate a mixed instance and train the classifier with the mixed instance. The generation of mixed instances can be described as follows:

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j,$$
$$\tilde{y} = \lambda y_i + (1-\lambda)y_j.$$

Here x_i and x_j are instance feature vectors randomly drawn from the training set; y_i and y_j are one-hot label encodings corresponding to x_i and x_j. $(\tilde{x}, \tilde{y})$ is used in training. In Zhang et al. [32] it is shown that *mix-up training* can improve generalization, resulting in higher accuracy on CIFAR-10 and CIFAR-100. This, in turn, reduces the generalization gap. Also, intuitively, since only the mixed instances are used in training, the classifier will not be directly trained in the original training instances, and should not remember them as well.

使用不同的训练instances来生成混合实例再用来训练classifier
↓
improve generalization

实验结果-攻击方案

- 提出的攻击方案Instance-Probability Attack的效果

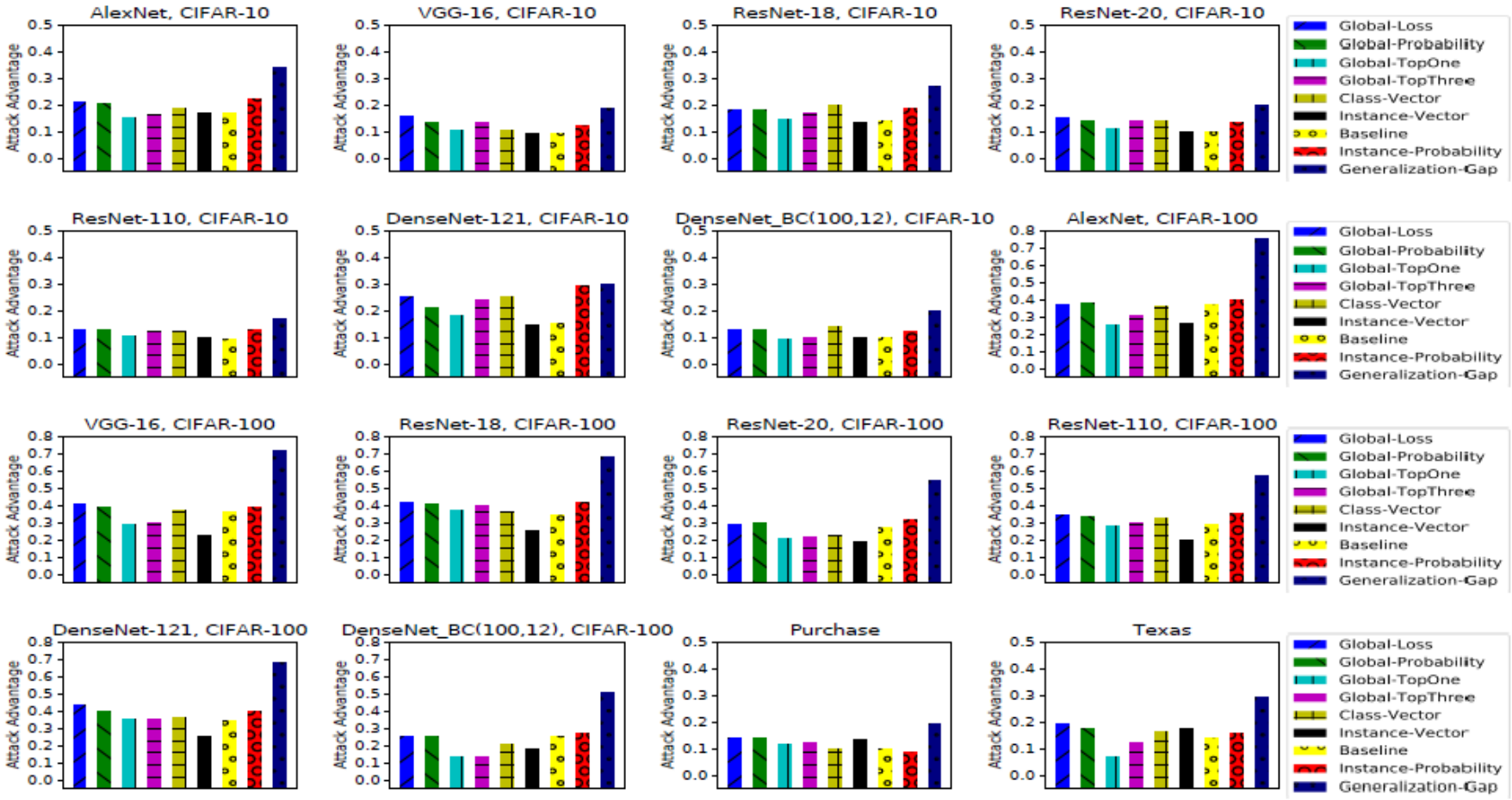


Figure 2: Attack advantage of different attacks for different classifiers on different datasets. Each bar represents a different attack. The rightmost three attacks are baseline attack, our proposed attack and generalization gap. The six attacks on the left part are proposed in previous papers.

实验结果-攻击方案

- 不同攻击方法的攻击效果

dataset	CIFAR-10	CIFAR-100	TEXAS-100	PURCHASE-100	MNIST
Training accuracy	0.995	0.988	0.778	0.981	0.999
Testing accuracy	0.758	0.353	0.485	0.791	0.981
Generalization gap	*0.237*	*0.635*	*0.293*	*0.190*	*0.018*
Largest attack advantage	0.173	0.362	0.191	0.145	0.020
Baseline attack advantage	0.122	0.317	0.143	0.101	0.010
Class-Vector attack advantage	0.162	0.317	0.161	0.106	0.013
Global-Loss attack advantage	0.173	0.359	0.191	0.143	0.013
Global-Probability attack advantage	0.160	0.351	0.177	0.145	0.013
Global-TopOne attack advantage	0.127	0.269	0.067	0.116	0.011
Global-TopThree attack advantage	0.152	0.288	0.121	0.125	0.012
Instance-Vector attack advantage	0.122	0.222	0.177	0.133	0.020
Instance-Probability attack advantage	0.173	0.362	0.167	0.088	0.017

Table 2: Training/testing accuracy and attack advantage for different datasets. For CIFAR-10 dataset and CIFAR-100 dataset, we average throughout all the tested classifiers.

Generalization Gap和Attack Advantage之间的关系

- g: generalization gap
- v: the largest advantage any existing MI attack over all classifiers and all datasets
- 有： $g/2 \leq v \leq g$

The advantage of the Baseline attack can be estimated from the training and testing accuracy. Let a_R be the training accuracy, and a_E be the testing accuracy. Then $g = a_R - a_E$ is the **generalization gap**. Given a balanced evaluation set, the accuracy of the baseline attack is the average of its accuracy on members and non-members. By definition, its accuracy on members is about a_R and its accuracy on non-members is about $1 - a_E$. Its overall accuracy is thus $\frac{1}{2} * a_R + \frac{1}{2} * (1 - a_E) = \frac{1+a_R-a_E}{2} = \frac{1+g}{2}$. Its advantage is therefore $\frac{g}{2}$. This estimation of accuracy is an approximation but our experiments empirically show that this estimation is quite accurate. Since this attack is so simple and broadly applicable, we call this attack the baseline attack. This attack's advantage provides a lower-bound estimation of a classifier's vulnerability to MI attacks.

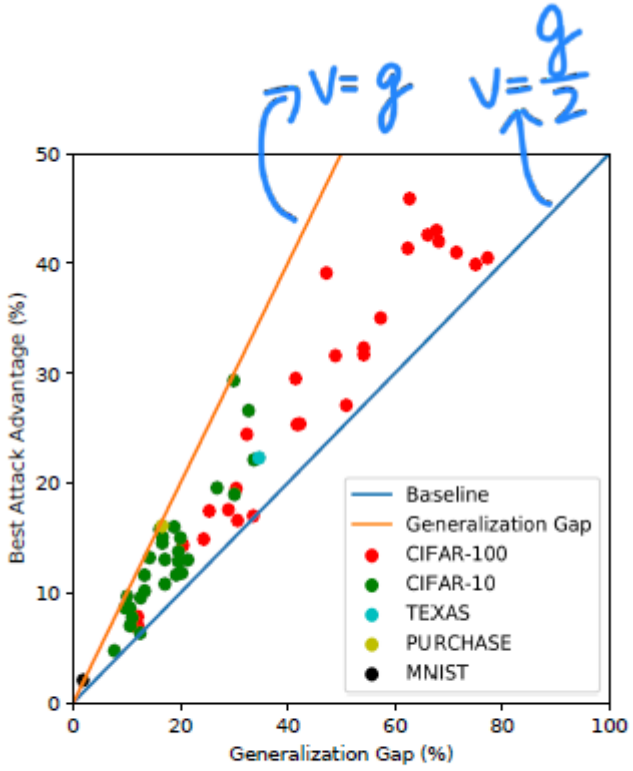


Figure 1: Generalization Gap vs Best Attack Advantage. Each point represents one classifier on one dataset.

实验结果-防御方案

- 不同防御机制下的attack advantage

	No Defense	Mix-up Alone (ablation)	MMD Alone (ablation)	MMD+Mix-up
Training accuracy	0.995	0.928	0.920	**0.870**
Testing accuracy	0.758	**0.773**	0.714	0.752
Generalization gap	0.237	0.155	0.206	**0.117**
Largest attack advantage	0.173	0.126	0.124	**0.093**
Baseline attack advantage	0.122	0.081	0.109	**0.066**
Class-Vector attack advantage	0.162	0.098	0.122	**0.062**
Global-Loss attack advantage	0.173	0.077	0.124	**0.063**
Global-Prob. attack advantage	0.160	0.102	0.121	**0.068**
Global-TopOne attack advantage	0.127	0.083	0.081	**0.049**
Global-TopThree attack advantage	0.152	0.083	0.088	**0.050**
Instance-Vector attack advantage	0.122	0.126	0.099	**0.092**
Instance-Prob. attack advantage	0.173	0.126	0.123	**0.093**

Table 3: (Ablation study) Average Training/Testing accuracy and average attack advantage, CIFAR-10 dataset.

实验结果-防御方案

- 不同防御机制下的attack advantage

	No Defense	Mix-up Alone (ablation)	MMD Alone (ablation)	MMD+Mix-up
Training accuracy	0.988	0.891	0.712	**0.618**
Testing accuracy	0.353	**0.406**	0.269	0.344
Generalization gap	0.635	0.485	0.443	**0.280**
Largest attack advantage	0.362	0.332	0.240	**0.168**
Baseline attack advantage	0.317	0.245	0.211	**0.140**
Class-Vector attack advantage	0.317	0.229	0.139	**0.088**
Global-Loss attack advantage	0.359	0.217	0.189	**0.097**
Global-Prob. attack advantage	0.351	0.276	0.227	**0.147**
Global-TopOne attack advantage	0.269	0.241	0.085	**0.083**
Global-TopThree attack advantage	0.288	0.241	0.130	**0.105**
Instance-Vector attack advantage	0.222	0.300	**0.059**	0.133
Instance-Prob. attack advantage	0.362	0.332	0.240	**0.168**

Table 4: (Ablation study) Average Training/Testing accuracy and average attack advantage, CIFAR-100 dataset.

实验结果-防御方案

- 不同防御机制的防御效果

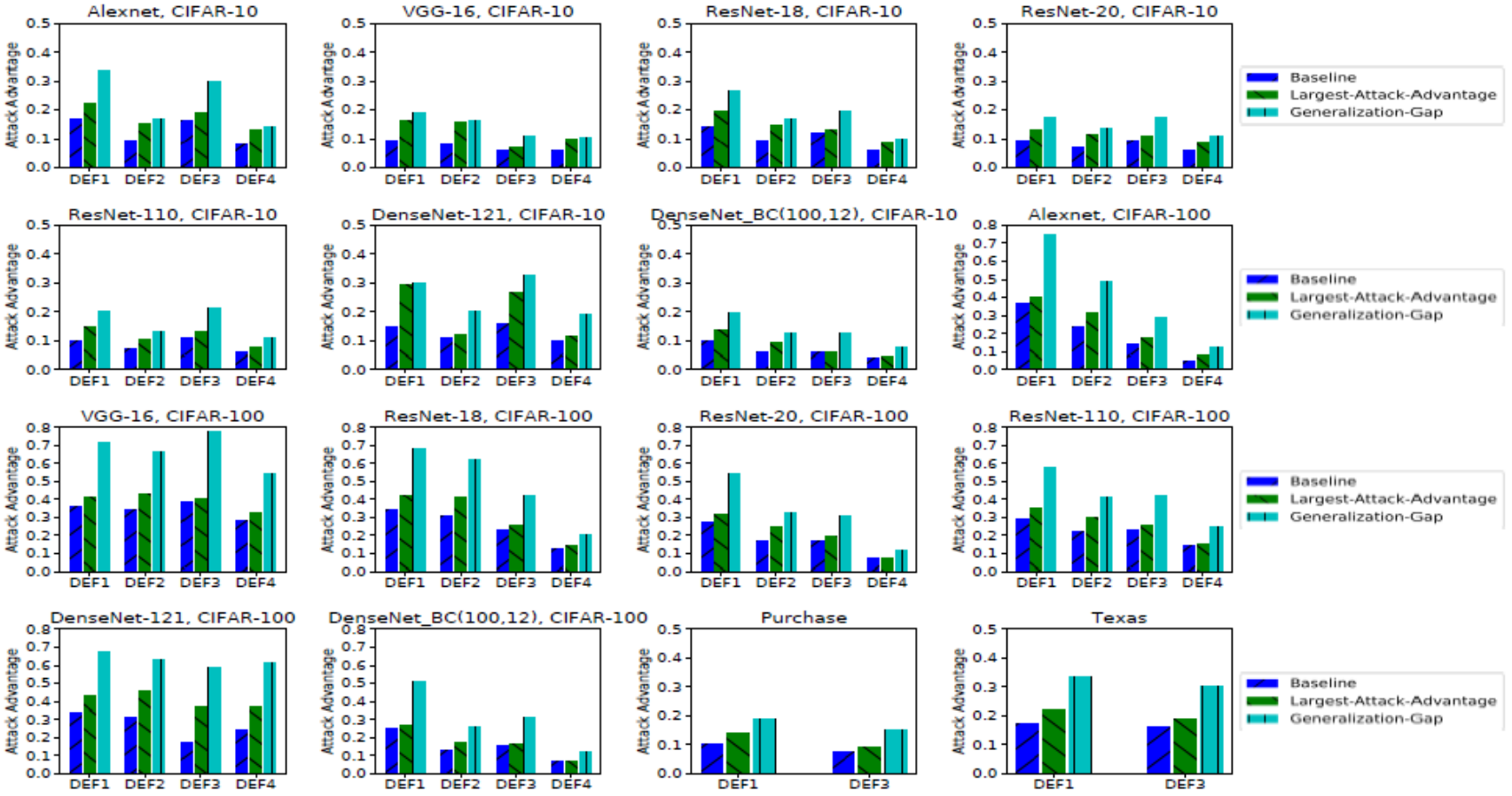


Figure 3: Attack advantage of different attacks for different classifiers on different datasets with different level of defenses. DEF1 means no defense, DEF2 means mix-up defense, DEF3 means MMD loss defense and DEF4 means MMD loss &mix-up defense.

实验结果-防御方案

- 和Mem-Guard防御方案的比较

Dataset	CIFAR-10			CIFAR-100		
Defense level	No Def.	Mixup+MMD	Mem-Guard	No Def.	Mixup+MMD	Mem-Guard
Training accuracy	0.994	**0.881**	0.997	0.995	**0.665**	0.979
Testing accuracy	0.761	**0.765**	0.762	0.326	**0.337**	**0.338**
Generalization gap	0.232	**0.116**	0.235	0.669	**0.328**	0.641
Largest attack advantage	0.166	**0.067**	0.113	0.356	**0.166**	0.324
Baseline attack advantage	0.116	**0.067**	0.112	0.333	**0.166**	0.324
Global-Probability attack advantage	0.156	**0.067**	0.112	0.356	**0.166**	0.320
Global-Loss attack advantage	0.166	**0.056**	0.113	0.356	**0.155**	0.319
Global-TopOne attack advantage	0.120	0.049	**0.028**	0.249	0.103	**0.093**
Global-TopThree attack advantage	0.140	0.052	**0.027**	0.273	0.104	**0.063**
Class-Vector attack advantage	0.137	**0.054**	0.113	0.320	**0.115**	0.316

Table 5: Our MMD+Mix-up MI defense significantly outperforms the Mem-Guard defense [15] on the CIFAR-10 and CIFAR-100 dataset using Alexnet, VGG-16 and ResNet-20 CNNs. The numbers are averaged across the three neural networks.

总结

- 根据Privacy Risks of ML Models via AE论文中的实验数据，关于v≤g的结论还有待验证，但是成员推理成功的原因与训练集测试集准确率之间的区别一定是有联系的，尤其对于confidence threshold攻击方法来说，因为该方法就是对模型的预测结果（置信度向量）进行区分从而实现的MIA。
- 该防御方法会对测试集的准确率产生一些影响，i.e. 0.758→0.752，0.353→0.344。
- 除此之外，论文中的模型准确率并没有训练至较高准确率，从而使得g较大，这也说明了v≤g的结论可能存在问题。
- 可研究方向：
  - MIA的防御方法：MMD为迁移学习中的重要工具，迁移学习能避免直接使用训练集进行训练，从而可能可以影响对训练集的推理，因此可以考虑利用迁移学习中的工具来实现MIA的防御。此外，考虑防御算法的时候一定要考虑g，因为baseline attack在图形领域中一定能实现，要让其成功概率降低。
- 下一步准备完成的工作
  - 复现S&P2017的攻击算法，做防御时这一定是一个防御算法性能的标尺。
  - 之前的工作继续进行，即更多数据集上的实验验证
  - 再多看些论文，思考可进行的工作