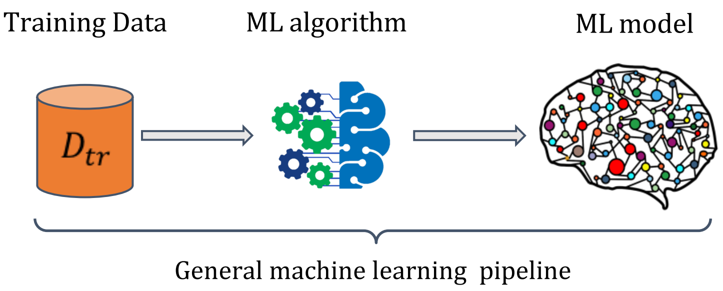
Machine Learning with Membership Privacy via Knowledge Transfer

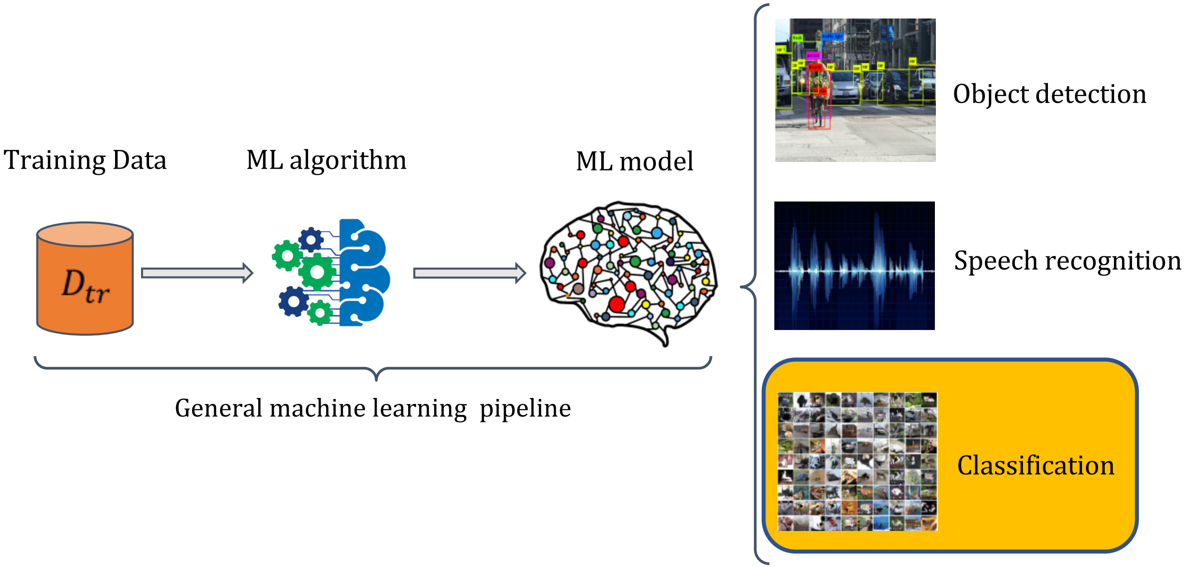
Virat Shejwalkar and Amir Houmansadr

University of Massachusetts Amherst

Machine Learning is Omnipresent

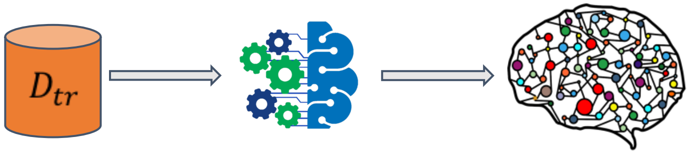


Machine Learning is Omnipresent

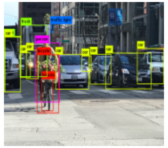


Machine Learning is Omnipresent

Training Data ML algorithm ML model

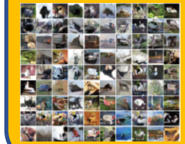


General machine learning pipeline



Object detection

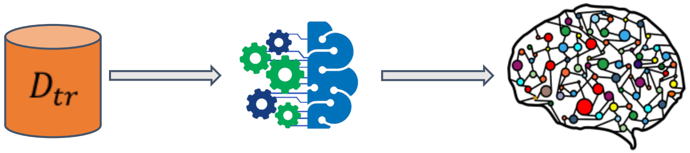
Speech recognition



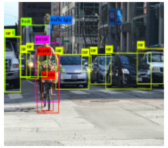
Classification

Machine Learning is Omnipresent

Training Data ML algorithm ML model

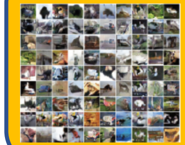


General machine learning pipeline



Object detection

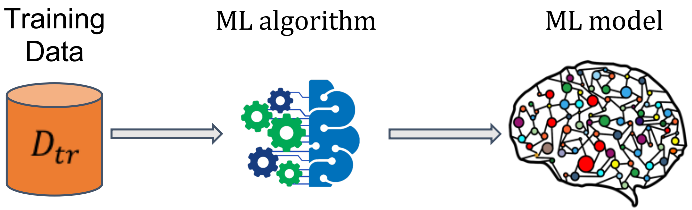
Speech recognition



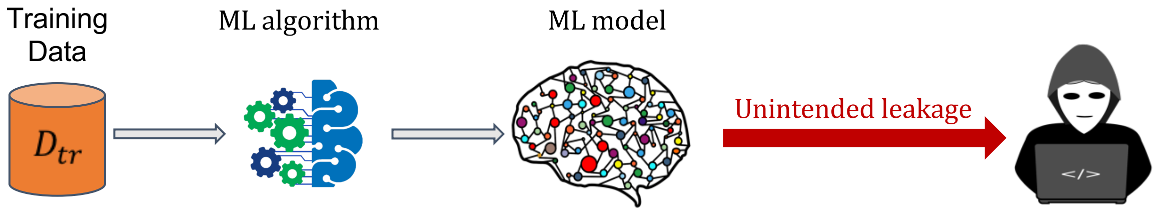
Classification

One of the most common uses of ML

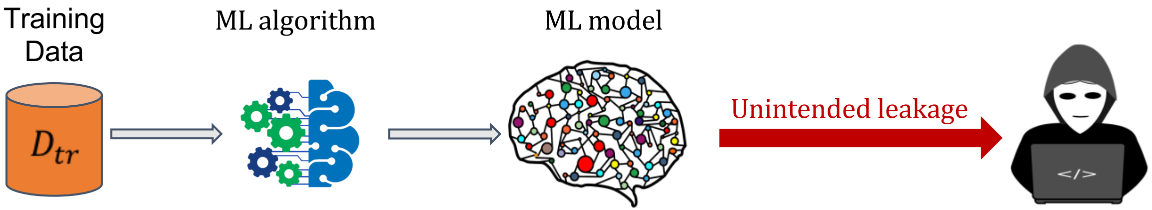
Privacy Risks of Machine Learning



Privacy Risks of Machine Learning



Privacy Risks of Machine Learning



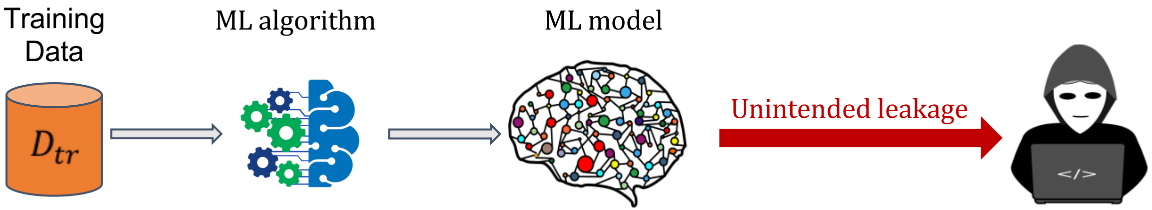
Feature inference
Fredrikson et al. (2015)
Hitaj el al. (2017)
Song et al. (2017)

Property inference
Ganju et al. (2018)
Melis et al. (2019)

Membership inference
Shokri et al. (2017)
Salem et al. (2018)
Nasr et al. (2019)
Jayaraman et al. (2019)

Fredrikson et al. (2015) *Model inversion attacks that exploit confidence information and basic countermeasures*, ACM CCS'15
Hitaj el al. (2017) *Deep models under the GAN: information leakage from collaborative deep learning*, ACM CCS'15
Song et al. (2017) *Machine learning models that remember too much*, ACM CCS'17
Ganju et al. (2018) *Property inference attacks on fully connected neural networks using permutation invariant representations*, ACM CCS'18
Melis et al. (2019) *Exploiting Unintended Feature Leakage in Collaborative Learning*, IEEE Security and Privacy'19
Shokri et al. (2017) *Membership inference attacks against machine learning models*, IEEE Security and Privacy'17
Nasr et al. (2019) *Comprehensive Privacy Analysis of Deep Learning*, IEEE Security and Privacy'19
Salem et al. (2018) *ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models*, NDSS'19
Jayaraman et al. (2019) Evaluating Differentially Private Machine Learning in Practice, Usenix'19

Privacy Risks of Machine Learning



Feature inference

Fredrikson et al. (2015)
Hitaj el al. (2017)
Song et al. (2017)

Property inference

Ganju et al. (2018)
Melis et al. (2019)

Membership inference

Shokri et al. (2017)
Salem et al. (2018)
Nasr et al. (2019)
Jayaraman et al. (2019)

The simplest form of privacy leakage

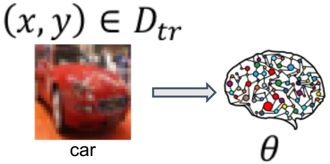
Fredrikson et al. (2015) *Model inversion attacks that exploit confidence information and basic countermeasures*, ACM CCS'15
Hitaj el al. (2017) *Deep models under the GAN: information leakage from collaborative deep learning*, ACM CCS'15
Song et al. (2017) *Machine learning models that remember too much*, ACM CCS'17
Ganju et al. (2018) *Property inference attacks on fully connected neural networks using permutation invariant representations*, ACM CCS'18
Melis et al. (2019) *Exploiting Unintended Feature Leakage in Collaborative Learning*, IEEE Security and Privacy'19
Shokri et al. (2017) *Membership inference attacks against machine learning models*, IEEE Security and Privacy'17
Nasr et al. (2019) *Comprehensive Privacy Analysis of Deep Learning*, IEEE Security and Privacy'19
Salem et al. (2018) *ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models*, NDSS'19
Jayaraman et al. (2019) Evaluating Differentially Private Machine Learning in Practice, Usenix'19

Overview of Membership Inference Attacks (MIAs)

Goal: Given a target model and a target sample, infer whether the sample was part of the training data of the model

Overview of Membership Inference Attacks (MIAs)

Goal: Given a target model and a target sample, infer whether the sample was part of the training data of the model



Overview of Membership Inference Attacks (MIAs)

Goal: Given a target model and a target sample, infer whether the sample was part of the training data of the model

$(x, y) \in D_{tr}$

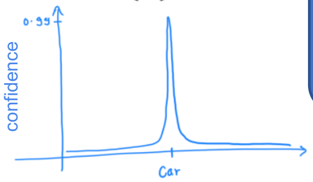
car

θ

$(x, y) \in D_{tr}$

car

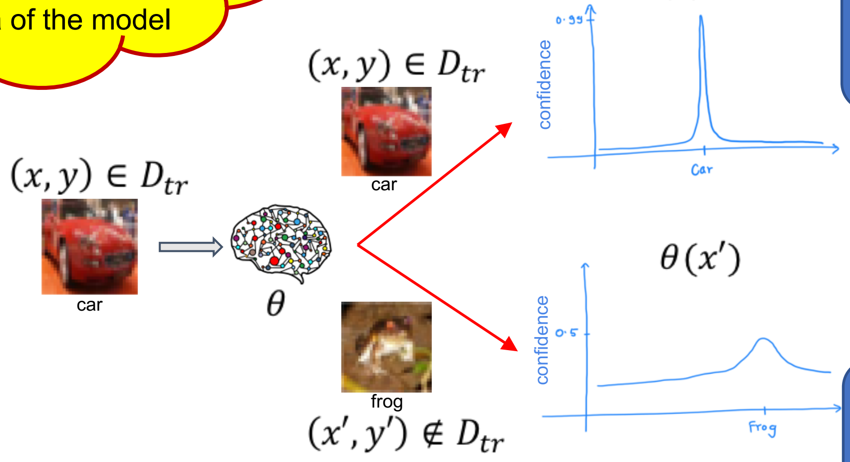
$\theta(x)$



Neural networks predictions are very confident on members of the training data due to memorization

Overview of Membership Inference Attacks (MIAs)

Goal: Given a target model and a target sample, infer whether the sample was part of the training data of the model

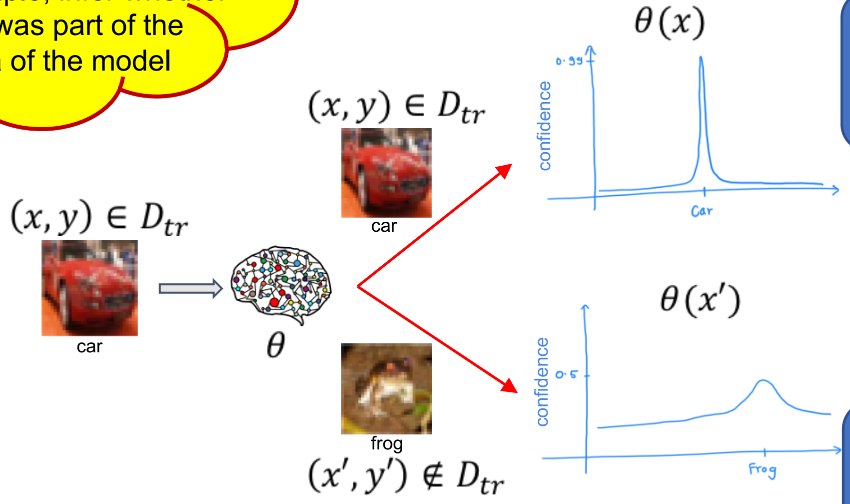


Neural networks predictions are very confident on members of the training data due to memorization

Neural networks prediction confidence is distributed across labels for non-members

Overview of Membership Inference Attacks (MIAs)

Goal: Given a target model and a target sample, infer whether the sample was part of the training data of the model



Neural networks predictions are very confident on members of the training data due to memorization

Membership inference attacks exploit such differences in the behavior of the model towards members and non-members of its private training data

Neural networks prediction confidence is distributed across labels for non-members

Existing Defenses Against MIAs

Existing Defenses Against MIAs

Black-box defenses

White-box defenses

Existing Defenses Against MIAs

Black-box defenses

Top-k predictions
Prediction adjustment (MemGuard)

White-box defenses

Existing Defenses Against MIAs

Black-box defenses

Top-k predictions
Prediction adjustment (MemGuard)

No protection against MIAs that exploit generalization gap (Song et al. 2021)

White-box defenses

Song et al. *Systematic Evaluation of Privacy Risks of Machine Learning Models,* Usenix'21

Existing Defenses Against MIAs

Black-box defenses

Top-k predictions
Prediction adjustment (MemGuard)

No protection against MIAs that exploit generalization gap (Song et al. 2021)

White-box defenses

Differential Privacy (DP-SGD, PATE)
provide theoretical privacy guarantees

Prohibitively high losses in model accuracies (Jayaraman et a. 2020)

Song et al. *Systematic Evaluation of Privacy Risks of Machine Learning Models,* Usenix'21
Jayaraman et al. (2019) Evaluating Differentially Private Machine Learning in Practice, Usenix'19

Existing Defenses Against MIAs

Black-box defenses

Top-k predictions
Prediction adjustment (MemGuard)

No protection against MIAs that exploit generalization gap (Song et al. 2021)

White-box defenses

Differential Privacy (DP-SGD, PATE)
provide theoretical privacy guarantees

Prohibitively high losses in model accuracies (Jayaraman et a. 2020)

Regularization
(Adversarial regularization, Dropout)
Without theoretical privacy guarantees

Unacceptably poor empirical tradeoffs (Our work)

Song et al. *Systematic Evaluation of Privacy Risks of Machine Learning Models,* Usenix'21
Jayaraman et al. (2019) Evaluating Differentially Private Machine Learning in Practice, Usenix'19

Existing Defenses Against MIAs

Black-box defenses

Top-k predictions
Prediction adjustment (MemGuard)

No protection against MIAs that exploit generalization gap (Song et al. 2021)

White-box defenses

Differential Privacy (DP-SGD, PATE)
provide theoretical privacy guarantees

Prohibitively high losses in model accuracies (Jayaraman et a. 2020)

Regularization
(Adversarial regularization, Dropout)
Without theoretical privacy guarantees

Unacceptably poor empirical tradeoffs (Our work)

Existing defenses offer poor tradeoffs between membership privacy and model utility

Song et al. *Systematic Evaluation of Privacy Risks of Machine Learning Models,* Usenix'21
Jayaraman et al. (2019) Evaluating Differentially Private Machine Learning in Practice, Usenix'19

Existing Defenses Against MIAs

Black-box defenses

Top-k predictions
Prediction adjustment (MemGuard)

No protection against MIAs that exploit generalization gap (Song et al. 2021)

White-box defenses

Differential Privacy (DP-SGD, PATE)
provide theoretical privacy guarantees

Prohibitively high losses in model accuracies (Jayaraman et a. 2020)

Regularization
(Adversarial regularization, Dropout)
Without theoretical privacy guarantees

Unacceptably poor empirical tradeoffs (Our work)

How can we improve the tradeoffs?

Existing defenses offer poor tradeoffs between membership privacy and model utility

Song et al. *Systematic Evaluation of Privacy Risks of Machine Learning Models,* Usenix'21
Jayaraman et al. (2019) Evaluating Differentially Private Machine Learning in Practice, Usenix'19

This work

- Goals: Train machine learning models that
  - are resistant to MIAs
  - are as accurate as their non-private counterparts
  - can be deployed in white-box fashion

This work

- Goals: Train machine learning models that
  - are resistant to MIAs
  - are as accurate as their non-private counterparts
  - can be deployed in white-box fashion

- Our approach
  - Use knowledge transfer and cut off the access of the final model to private training data

This work

- Goals: Train machine learning models that
  - are resistant to MIAs
  - are as accurate as their non-private counterparts
  - can be deployed in white-box fashion

- Our approach
  - Use knowledge transfer and cutoff the access of the final model to private training data

Memorization in deep neural networks exacerbates due to direct access to the training data

This work

- Goals: Train machine learning models that
  - are resistant to MIAs
  - are as accurate as their non-private counterparts
  - can be deployed in white-box fashion

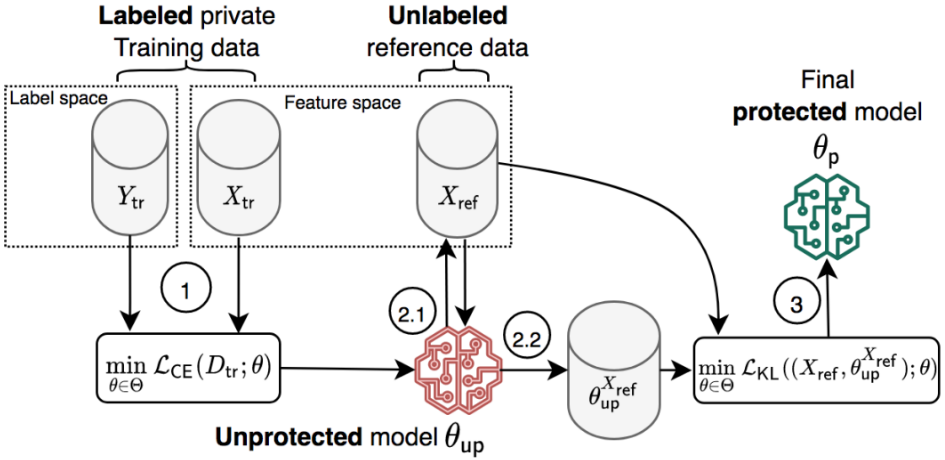
- Our approach
  - Use knowledge transfer and cutoff the access of the final model to private training data
  - Fine-tune the reference data used for knowledge transfer to meet desired tradeoffs

Memorization in deep neural networks exacerbates due to direct access to the training data

Distillation for Membership Privacy (DMP)

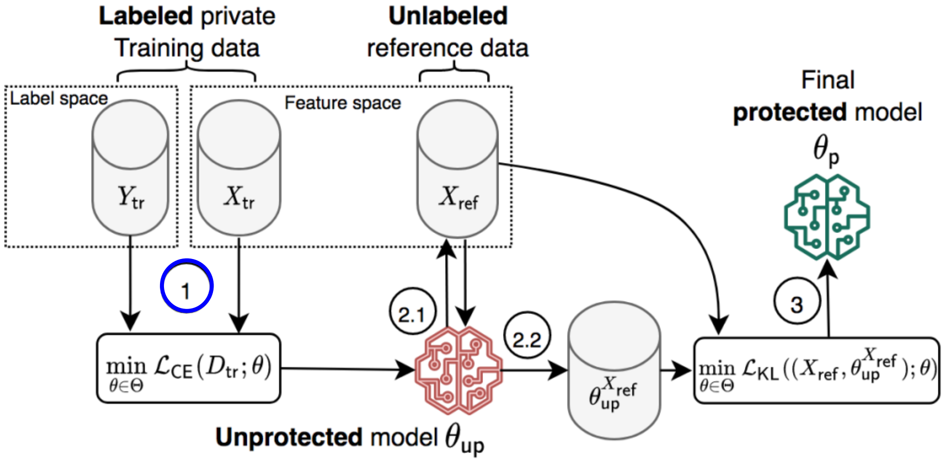
Distillation for Membership Privacy (DMP)

The DMP defense is a very effective regularizer which proceeds as follows



Distillation for Membership Privacy (DMP)

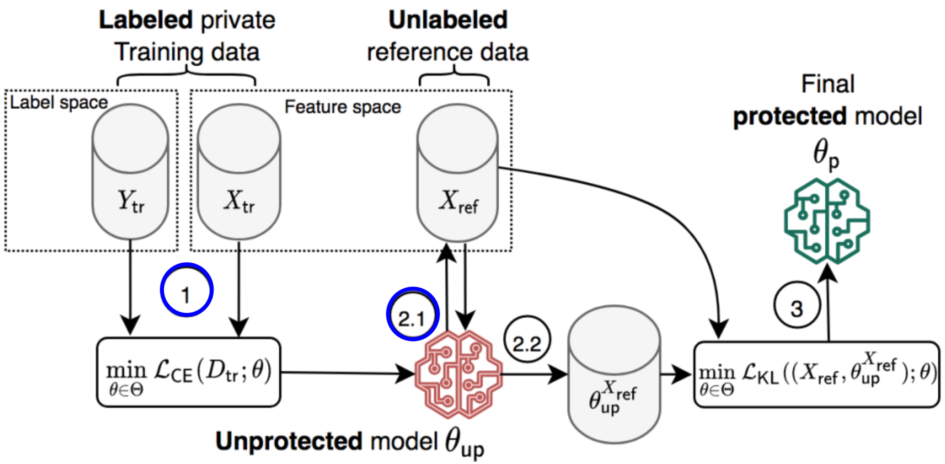
The DMP defense is a very effective regularizer which proceeds as follows



(1) Trains an unprotected model on private training data, e.g., using cross-entropy loss

Distillation for Membership Privacy (DMP)

The DMP defense is a very effective regularizer which proceeds as follows

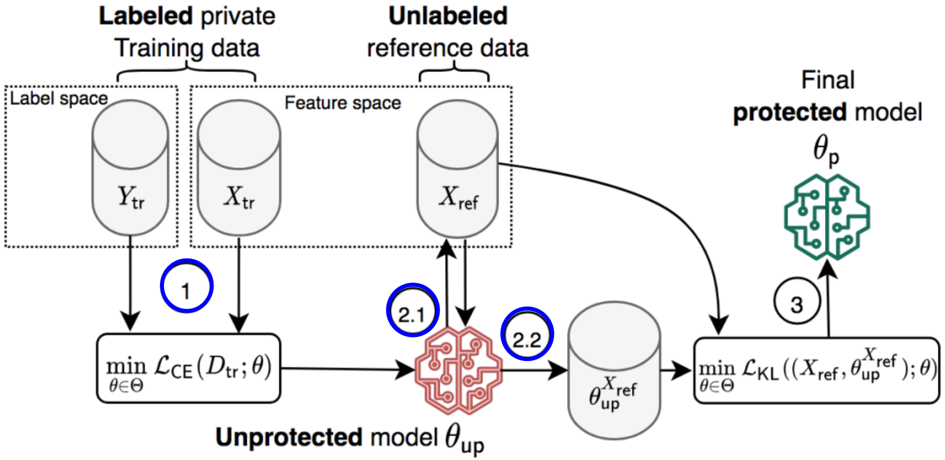


(1) Trains an unprotected model on private training data, e.g., using cross-entropy loss

(2.1) Computes reference data to use for knowledge transfer

Distillation for Membership Privacy (DMP)

The DMP defense is a very effective regularizer which proceeds as follows



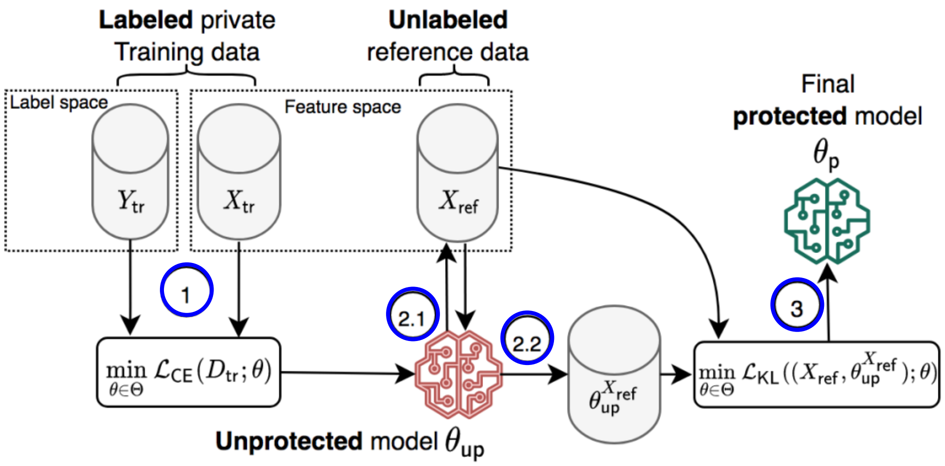
(1) Trains an unprotected model on private training data, e.g., using cross-entropy loss

(2.1) Computes reference data to use for knowledge transfer

(2.2) Computes soft labels for the reference data

Distillation for Membership Privacy (DMP)

The DMP defense is a very effective regularizer which proceeds as follows



(1) Trains an unprotected model on private training data, e.g., using cross-entropy loss

(2.1) Computes reference data to use for knowledge transfer

(2.2) Computes soft labels for the reference data

(3) Trains the final protected model using KL-divergence loss

Fine-tuning DMP to adjust privacy-utility tradeoffs

- In DMP, soft labels of the reference data are the main source of membership information leakage, hence the correct choice of reference data is important for DMP to be effective

Fine-tuning DMP to adjust privacy-utility tradeoffs

- In DMP, soft labels of the reference data are the main source of membership information leakage, hence the correct choice of reference data is important for DMP to be effective
- We propose to use the reference data such that they are far from private training data in feature space and the unprotected model has low entropy predictions on them

Fine-tuning DMP to adjust privacy-utility tradeoffs

- In DMP, soft labels of the reference data are the main source of membership information leakage, hence the correct choice of reference data is important for DMP to be effective
- We propose to use the reference data such that they are far from private training data in feature space and the unprotected model has low entropy predictions on them

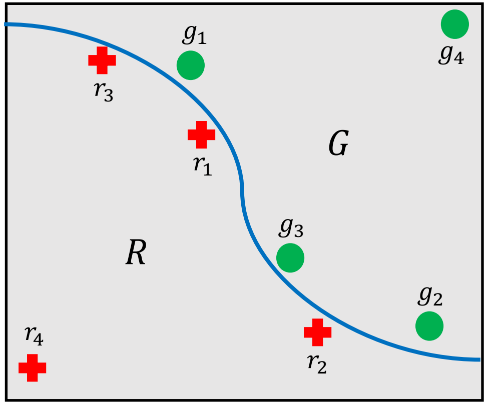
Intuitively such reference data are easy to classify samples whose predictions are not impacted by the presence of any particular member

Fine-tuning DMP to adjust privacy-utility tradeoffs

- In DMP, soft labels of the reference data are the main source of membership information leakage, hence the correct choice of reference data is important for DMP to be effective
- We propose to use the reference data such that they are far from private training data in feature space and the unprotected model has low entropy predictions on them

Memorization causes the decision boundaries to be close to training data

Sample space divided in two classes



Intuitively such reference data are easy to classify samples whose predictions are not impacted by the presence of any particular member

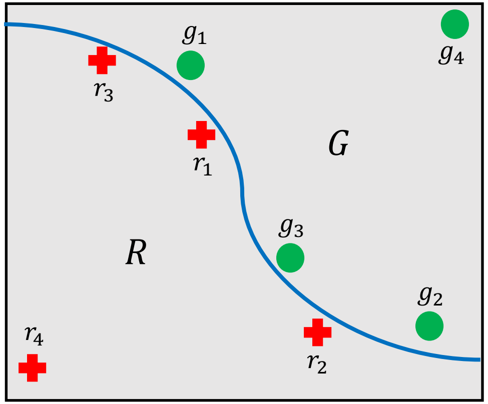
Fine-tuning DMP to adjust privacy-utility tradeoffs

- In DMP, soft labels of the reference data are the main source of membership information leakage, hence the correct choice of reference data is important for DMP to be effective
- We propose to use the reference data such that they are far from private training data in feature space and the unprotected model has low entropy predictions on them

Memorization causes the decision boundaries to be close to training data

g_4, r_4 points are far from the boundaries and classifier's predictions of them do not change even if some training sample $\{g_{1,2,3}, r_{1,2,3}\}$ is removed

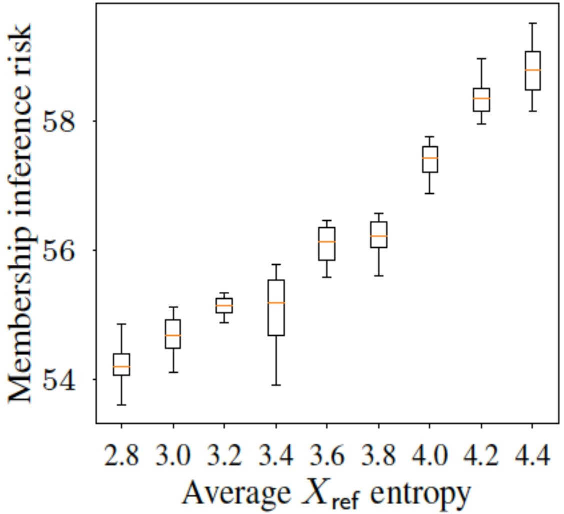
Sample space divided in two classes



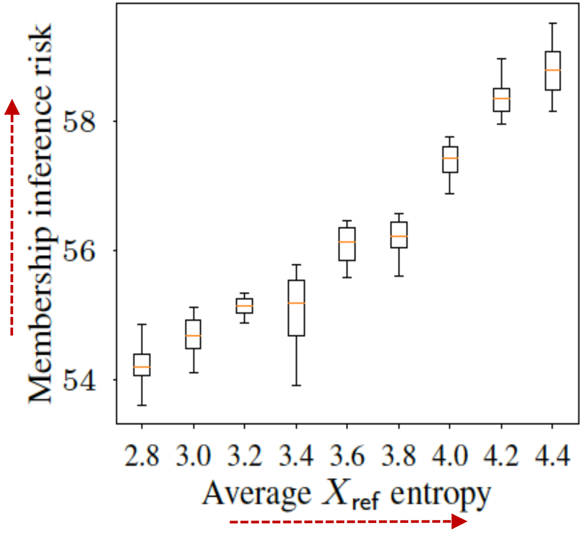
Intuitively such reference data are easy to classify samples whose predictions are not impacted by the presence of any particular member

Fine-tuning DMP to adjust privacy-utility tradeoffs (Empirical verification)

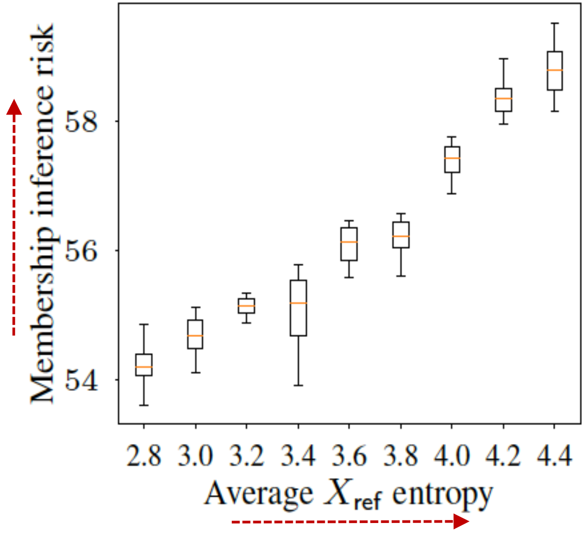
Fine-tuning DMP to adjust privacy-utility tradeoffs (Empirical verification)

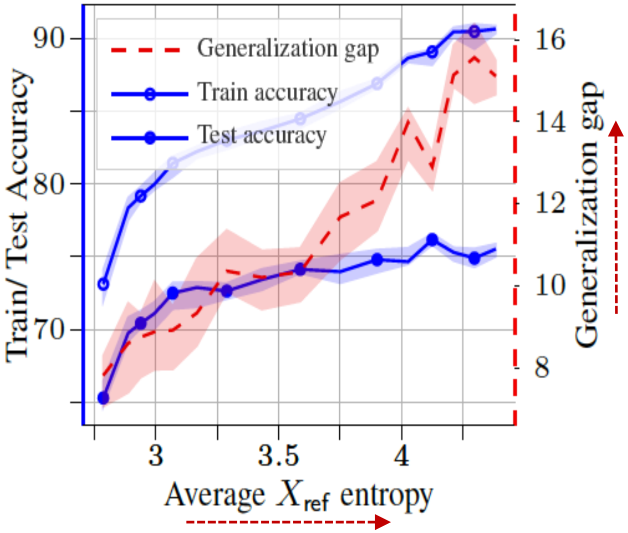


Fine-tuning DMP to adjust privacy-utility tradeoffs (Empirical verification)

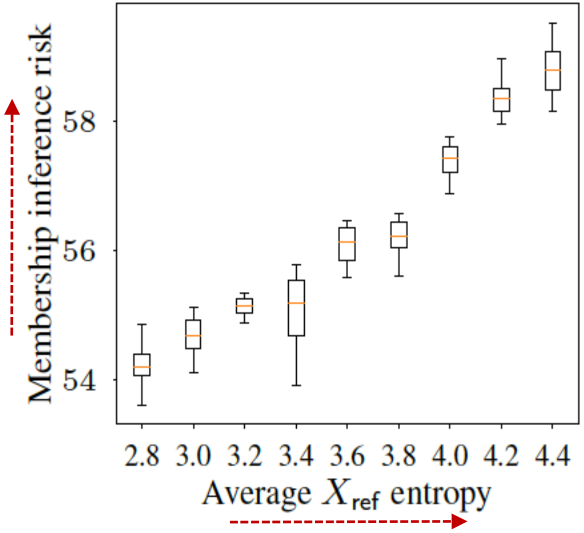


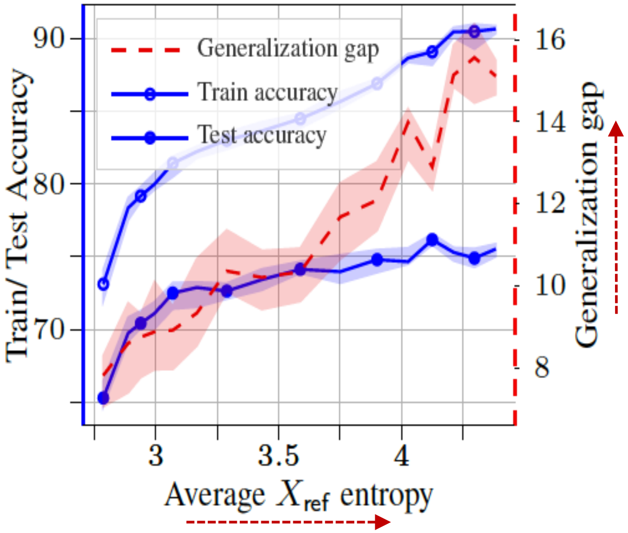
Fine-tuning DMP to adjust privacy-utility tradeoffs (Empirical verification)





Fine-tuning DMP to adjust privacy-utility tradeoffs (Empirical verification)





Increasing the average entropy of the reference data increases the accuracy of the final model, but it also increases the membership inference risk

Comparison of DMP with Adversarial Regularization

Comparison of DMP with Adversarial Regularization

Dataset and model	No defense					
	E_{gen}	A_{test}	A_{wb}	A_{bb}	A_{bl}	A_{nn}
Purchase + FC	24.0	76.0	77.1	76.8	63.1	60.5
Texas + FC	51.3	48.7	84.0	82.2	76.1	71.9
CIFAR100 + Alexnet	63.2	36.8	90.3	91.3	81.8	N/A
CIFAR100 + DenseNet-12	33.8	65.2	72.2	71.8	67.5	N/A
CIFAR100 + DenseNet-19	34.4	65.5	82.3	81.6	68.1	N/A
CIFAR10 + Alexnet	32.5	67.5	77.9	77.5	66.4	N/A

The target models without any defense are highly susceptible to membership inference attacks

Comparison of DMP with Adversarial Regularization

Dataset and model	No defense					
	E_{gen}	A_{test}	A_{wb}	A_{bb}	A_{bl}	A_{nn}
Purchase + FC	24.0	76.0	77.1	76.8	63.1	60.5
Texas + FC	51.3	48.7	84.0	82.2	76.1	71.9
CIFAR100 + Alexnet	63.2	36.8	90.3	91.3	81.8	N/A
CIFAR100 + DenseNet-12	33.8	65.2	72.2	71.8	67.5	N/A
CIFAR100 + DenseNet-19	34.4	65.5	82.3	81.6	68.1	N/A
CIFAR10 + Alexnet	32.5	67.5	77.9	77.5	66.4	N/A

The target models without any defense are highly susceptible to membership inference attacks

Dataset and model	Adversarial regularization (AdvReg)						DMP						
	E_{gen}	A_{test}	Attack accuracy				E_{gen}	A_{test}	A^{+}_{test}	Attack accuracy			
			A_{wb}	A_{bb}	A_{bl}	A_{nn}				A_{wb}	A_{bb}	A_{bl}	A_{nn}
Purchase + FC	9.7	56.5	55.8	55.4	54.9	50.1	10.1	74.1	**+31.2%**	55.3	55.1	55.2	50.2
Texas + FC	6.1	33.5	58.2	57.9	54.1	50.8	7.1	48.6	**+45.1%**	55.3	55.4	53.6	50.0
CIFAR100 + Alexnet	6.9	19.7	54.3	54.0	53.5	N/A	6.5	35.7	**+81.2%**	55.7	55.6	53.3	N/A
CIFAR100 + DenseNet-12	5.5	26.5	51.4	51.3	52.8	N/A	3.6	63.1	**+138.1%**	53.7	53.0	51.8	N/A
CIFAR100 + DenseNet-19	7.2	33.9	54.2	53.4	53.6	N/A	7.3	65.3	**+92.6%**	54.7	54.4	53.7	N/A
CIFAR10 + Alexnet	4.2	53.4	51.9	51.2	52.1	N/A	3.1	65.0	**+21.7%**	51.3	50.6	51.6	N/A

For near-equal resistance to MIAs, DMP trained models are significantly more accurate than adversarially regularized models (Nasr et al. 2018)

Nasr et al. (2018) *Machine learning with membership privacy using adversarial regularization*, ACM CCS'18

Comparing DMP with DP-SGD

Comparing DMP with DP-SGD

- We perform empirical comparison with DP-SGD (Abadi et al. 2016) in terms of tradeoffs between membership privacy and model utility

Abadi et al. (2016) *Deep learning with differential privacy*, ACM CCS'16

Comparing DMP with DP-SGD

- We perform empirical comparison with DP-SGD (Abadi et al. 2016) in terms of tradeoffs between membership privacy and model utility
- We use CIFAR10 dataset and the size of private training data is 25k

Abadi et al. (2016) *Deep learning with differential privacy*, ACM CCS'16

Comparing DMP with DP-SGD

Defense	Privacy budget (ε)	E_{gen}	A_{test}	A_{wb}
No defense	-	32.5	67.5	77.9
DMP	-	3.10	65.0	51.3
DP-SGD	198.5	3.60	52.2	51.7
	50.2	1.30	36.9	50.2
	12.5	0.30	31.7	50.0
	6.8	-1.60	29.4	49.9

- We perform empirical comparison with DP-SGD (Abadi et al. 2016) in terms of tradeoffs between membership privacy and model utility
- We use CIFAR10 dataset and the size of private training data is 25k
- For similar resistance to MIAs, DMP trained models have significantly higher accuracy than DP-SGD trained models

Abadi et al. (2016) *Deep learning with differential privacy*, ACM CCS’16

Comparison of DMP with PATE

- Similar to DP-SGD, we perform empirical comparison of DMP and PATE (Papernot et al. 2016)
- We use 25k of CIFAR10 dataset as private training data and the rest as the public data for semi-supervised learning; we use generator-discriminator pair from (Salimans et al. 2016)

Pepernot et al. *Semi-supervised knowledge transfer for deep learning from private data*, ICLR 2017
Salimans et al. *Improved techniques for training GANs*, NIPS 2016

Comparison of DMP with PATE

# of Teachers	Queries answered	Privacy budget (ϵ)	Target model E_{gen}	Target model A_{test}	A_{wb}
5	49	195.9	31.4	33.9	49.1
	1163	11684	65.4	68.1	49.0
10	23	42.9	39.1	38.3	50.1
	1527	6535	63.9	65.2	49.8
25	108	183.5	53.8	55.7	49.0
	4933	1794.1	57.8	60.3	48.6

- Similar to DP-SGD, we perform empirical comparison of DMP and PATE (Papernot et al. 2016)
- We use 25k of CIFAR10 dataset as private training data and the rest as the public data for semi-supervised learning; we use generator-discriminator pair from (Salimans et al. 2016)
- We observe that for a similar resistance to MIAs, DMP-trained models have much better accuracies than PATE-trained models
- Corresponding DMP model has 76.8% accuracy and 50.8% whitebox membership inference risk

Pepernot et al. *Semi-supervised knowledge transfer for deep learning from private data*, ICLR 2017
Salimans et al. *Improved techniques for training GANs*, NIPS 2016

Additional Insights into DMP Privacy

Additional Insights into DMP Privacy

Adjusting the two hyperparameters of DMP, i.e., softmax temperature and reference data size, allows tuning the privacy-utility tradeoffs

Additional Insights into DMP Privacy

Adjusting the two hyperparameters of DMP, i.e., softmax temperature and reference data size, allows tuning the privacy-utility tradeoffs

DMP poses no privacy risk to its reference data, which itself can be of sensitive nature

Additional Insights into DMP Privacy

Adjusting the two hyperparameters of DMP, i.e., softmax temperature and reference data size, allows tuning the privacy-utility tradeoffs

DMP poses no privacy risk to its reference data, which itself can be of sensitive nature

In case when reference data is not readily available, DMP achieves state-of-the-art tradeoffs even with synthetically generated reference data

Conclusions

- ✓ We show the strength of knowledge transfer as a sole defense against membership inference attacks by proposing Distillation for Membership Privacy (DMP) defense

- ✓ We show that DMP achieves state-of-the-art tradeoffs between membership privacy and model utility

- ✓ We believe that DMP, due to its simplicity, can be incorporated as a building block of future defenses against membership inference attacks

Thank You ☺

We will make the code and datasets public, please check this link for updates

This work was in part supported by the NSF grant CPS-1739462

CPS-1739462