Machine Learning with Membership Privacy via Knowledge Transfer

Virat Shejwalkar and Amir Houmansadr

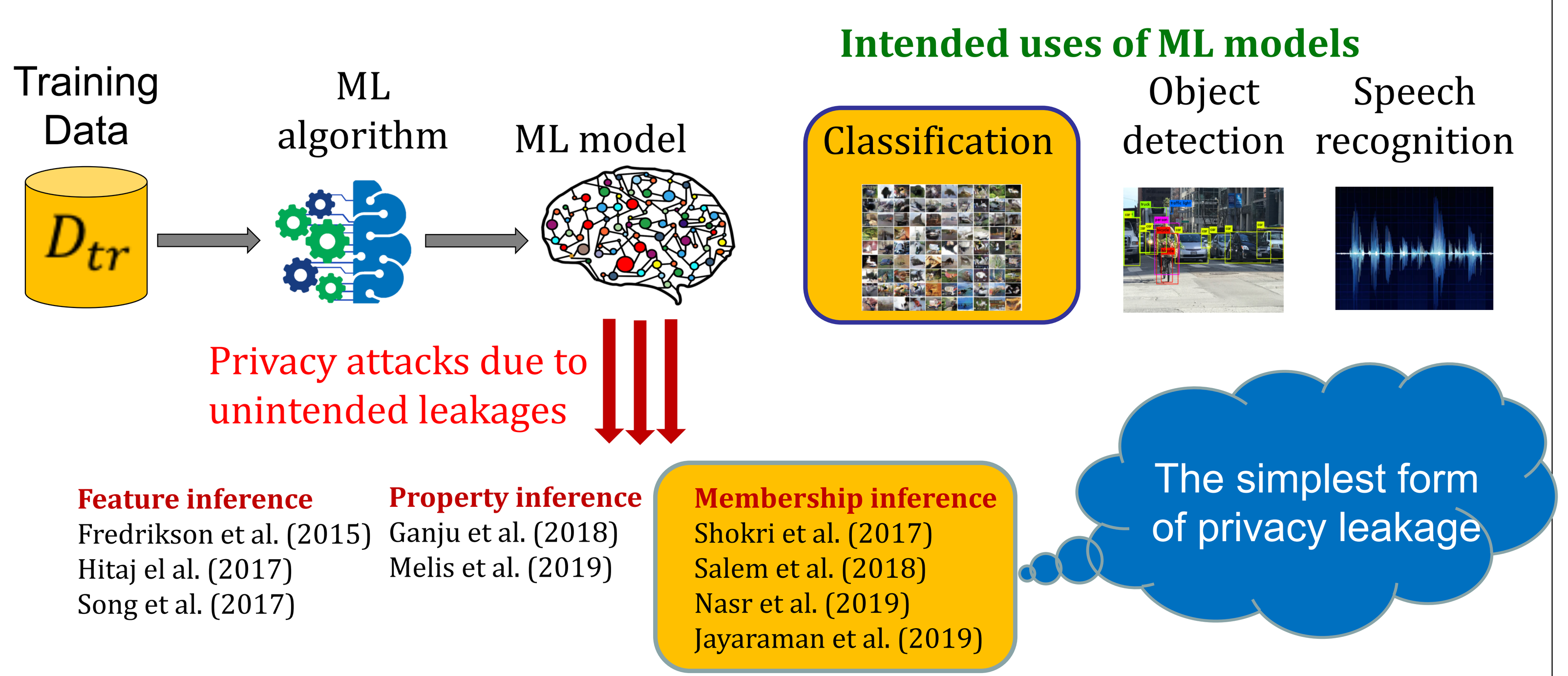
University of Massachusetts Amherst

35th AAAI conference on Artificial Intelligence

CPS-1739462

Introduction

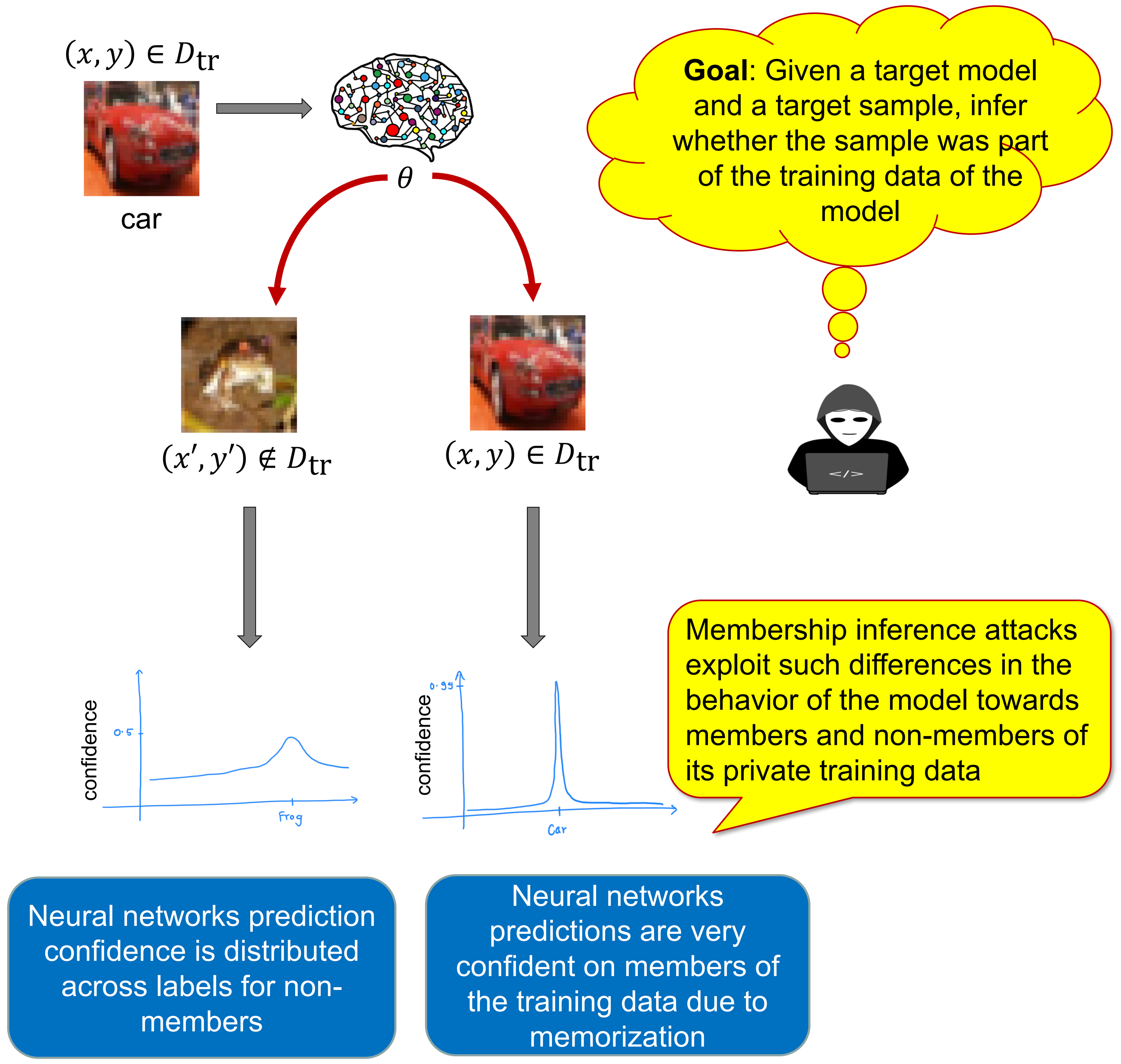
Due to its impressive performance, machine learning is commonly used in privacy sensitive domains such as health care applications.



But, the outstanding performance of ML comes at an undesired cost

- ML models, especially deep models, leak private information about their training data and facilitate multiple inference attacks, as shown
- We focus on membership inference, the simplest of privacy inference attacks, and widely used image classification tasks

Overview of Membership Inference Attacks



Existing Defenses against MIAs

- **Black-box defenses** only make the model output resistant to MIAs. Examples include:
  - Releasing top-k dimensions instead of entire prediction vector
  - Adjusting confidence of prediction vector, e.g., MemGuard
- But, these defenses are shown to be broken against simple MIAs

- **White-box defenses** make the model parameters resistant to MIAs, and hence, allow to release model parameters. Examples include:
  - Differential privacy based defenses such as DP-SGD and PATE
    - These offer theoretical privacy guarantees, but resulting models have unacceptably poor utilities
  - Regularization based defenses such as adversarial regularization, label smoothing, and dropout
    - These neither offer theoretical guarantees nor are they effective against multiple MIAs, e.g., strong whitebox MIAs

Effectively, existing defenses against MIAs offer poor tradeoffs between membership privacy and model utility

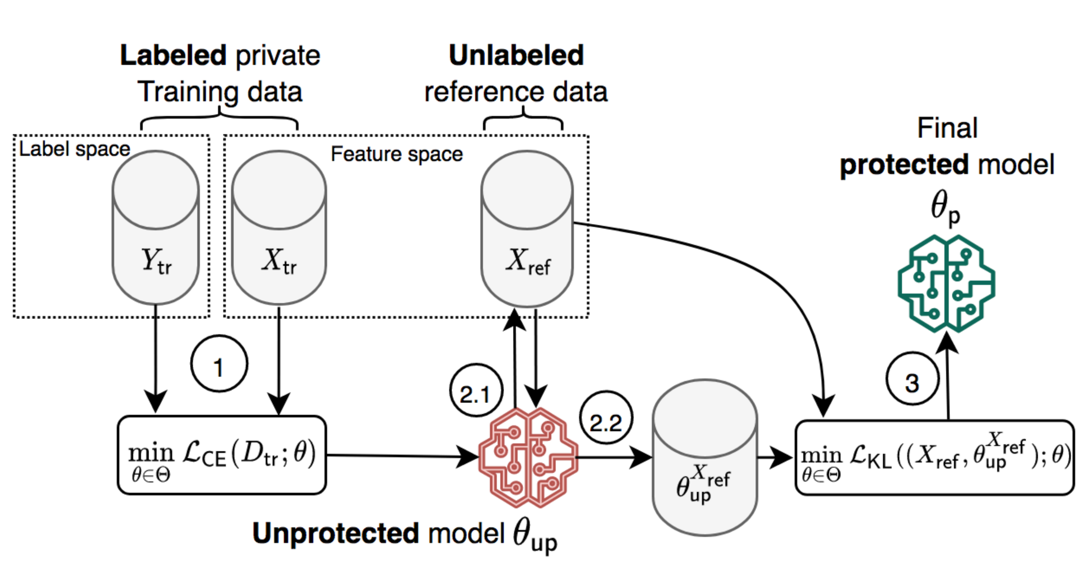
Distillation for Membership Privacy (DMP)

Our goal is to train ML models that are resistant against MIAs, and highly accurate, and can be deployed in white-box fashion

Our approach

- **Use knowledge transfer** and cutoff the access of final model to the private training data
- **Fine-tune the reference data** used for knowledge transfer to meet the desired membership privacy and model utility tradeoffs

DMP overview



(1) Train unprotected model on private train data, e.g., using cross-entropy loss

(2.1) Compute reference data to use for knowledge transfer

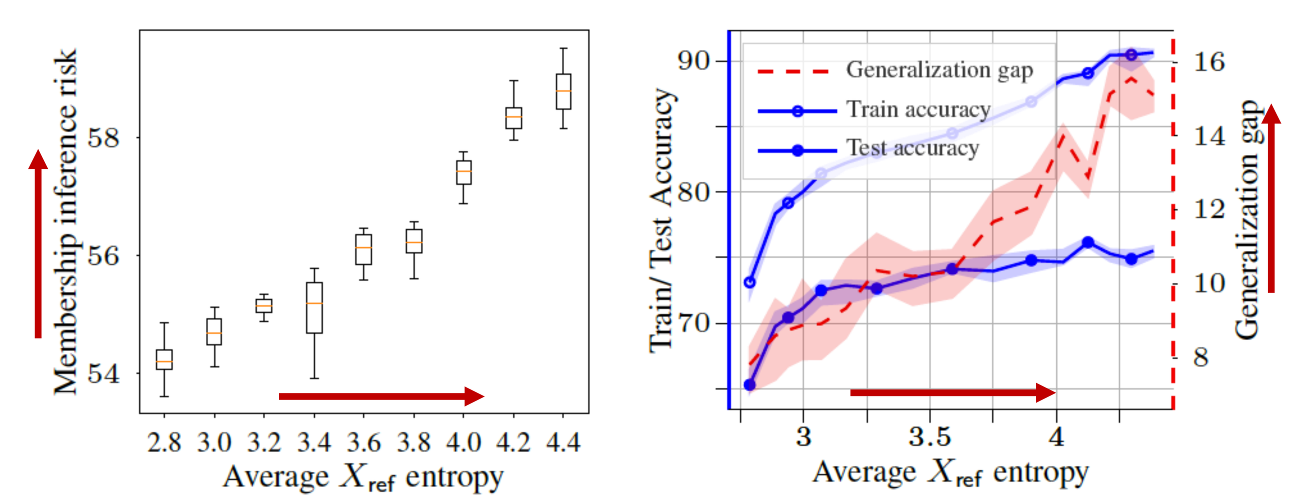
(2.2) Compute soft labels for the reference data using unprotected model

(3) Train the final protected model using reference data, its soft labels and KL-divergence loss

Fine-tuning DMP Defense

- In DMP, reference data should be carefully selected as their soft labels are the main source of membership leakage
- **Proposal**: Use reference data such that they are far from private training data in feature space and the unprotected model has low entropy predictions on them
- **Intuition**: Such reference data are easy-to-classify samples whose predictions are not significantly impacted by the presence of any particular member of the private training data

Empirical verification of our hypothesis



Increasing the average entropy of the reference data increases the accuracy of the final model, but at the cost of increased MIA risk

Empirical comparison 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

Unprotected modes are highly susceptible to MIAs

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_{wb}	A_{bb}	A_{bl}
P-FC	9.7	56.5	55.8	55.4	54.9	10.1	74.1	**+31.2%**	55.3	55.1	55.2
T-FC	6.1	33.5	58.2	57.9	54.1	7.1	48.6	**+45.1%**	55.3	55.4	53.6
C100-A	6.9	19.7	54.3	54.0	53.5	6.5	35.7	**+81.2%**	55.7	55.6	53.3
C100-D12	5.5	26.5	51.4	51.3	52.8	3.6	63.1	**+138.1%**	53.7	53.0	51.8
C100-D19	7.2	33.9	54.2	53.4	53.6	7.3	65.3	**+92.6%**	54.7	54.4	53.7
C10-A	4.2	53.4	51.9	51.2	52.1	3.1	65.0	**+21.7%**	51.3	50.6	51.6

For near-equal resistance to MIAs, DMP trained models are significantly more accurate than adversarially regularized models

Conclusions and Future Directions

- 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