When Does Data Augmentation Help With Membership Inference Attacks?

Yiğitcan Kaya, Tudor Dumitraș

University of Maryland, College Park

ICML 2021 - Virtual

What is a membership inference attack?

- Deep learning models leak information about their training sets.

- This leakage allows an attacker to infer whether a sample was in a model's training set by observing the model's output.
  - This is known as a black-box membership inference attack (MIA).
  - A severe privacy risk.

Overfitting implies risk of MIAs

- Overfitting is a sufficient (*but not necessary*) condition for MIAs [1].

- Overfitting causes a rift between the model's performance on the training set and on the testing set.
  - Known as *generalization gap*.
  - This gives leverage to the MIAs.

[1] Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting, Yeom et al.

Data augmentation reduces overfitting

- Differential privacy provides formal guarantees against the MIA risk.
  - Causes significant accuracy loss over a non-private model.
  - Accuracy is a key factor for practitioners.

- Data augmentation is commonly used to fight overfitting and to improve the accuracy.
  - There are many augmentation techniques.
  - They are known to shrink the generalization gap.
  - No work systematically studies how augmentation fares against MIAs.

*We conducted an empirical study on **7** popular augmentation mechanisms and evaluated their effectiveness against **3** black-box MIAs.*

Finding #1 – Augmentation for accuracy is ineffective against MIAs

- Applying augmentation to maximize the model's accuracy, we see
  - Augmentation can boost the accuracy up to 10% over baseline models.
  - Label smoothing boosts the accuracy and the success of MIA attacks simultaneously.
  - No augmentation mechanism can reduce the MIA risk by more than 50% over the baseline.

Augmentation for the sake of accuracy still leaves the models vulnerable.

Finding #2 – High-intensity augmentation mitigates MIAs but hurts the accuracy

- Applying augmentation with higher intensities (e.g., cropping 90% of the image), we see
  - This leads to an accuracy drop.
  - Within 25% accuracy drop, augmentation reduces MIA success more than 85%-100%.
  - Most effective: Random cropping and Cutout
  - Moderately effective: Mixup and Soft labels
  - Least effective: Gaussian augmentation and **label smoothing**

Augmentation is effective against MIAs when it hurts the model's accuracy.

Finding #3 – Label smoothing is a privacy risk

- All mechanisms except for label smoothing reduce the MIA success when they increase the accuracy.

- Label smoothing consistently increases the accuracy and makes MIAs more successful.
  - **5%** accuracy boost -> **40%** more MIA success
  - Models overfit on smooth labels and start producing more uniform outputs on training samples.
  - This discrepancy gives more leverage to MIAs.

Smaller generalization gap does not mean less MIA risk in the case of label smoothing.

Finding #4 – High-intensity augmentation is like early stopping

- Using our simple **Loss-Rank-Correlation** metric, we compare applying high-intensity augmentation to training the model for only a few epochs.
  - We see the high similarities between across the board.
  - Training for a few epochs hurts the accuracy similar to high-intensity augmentation.

Simply reducing the number of training epochs might be just as effective as applying augmentation to mitigate MIAs.

Code is available @
https://github.com/yigitcankaya/augmentation_mia

*Ultimately, there is **no free lunch** in augmentation against membership inference attacks…*

Contact Info

http://www.cs.umd.edu/~yigitcan/