Membership Inference Attacks and Defenses in Neural Network Pruning

Xiaoyong (Brian) Yuan and Lan Zhang
Michigan Technological University
Increasing Neural Network Size vs. Resource-Constrained Devices

(Gholami, 2021)

It is challenging to deploy large-size neural networks on resource-constrained devices

- Computational, memory, and storage limitations
Neural Network Pruning

- Basic Idea: remove redundant parameters from a dense neural network
- Goal
  - Reduce sizes of neural networks and speed up inference
  - Minimize the loss of prediction performance
- Evaluation Metrics
  - Efficiency (e.g., Sparsity Level, FLOPs, Latency)
  - Prediction Performance (e.g., Accuracy)
  - Privacy risk?
Why Concern About Privacy in Neural Network Pruning?

Enforce a small number of parameters to achieve similar accuracy
See the training samples more often

- Increase the memorization of training samples
- Aggravate the privacy risks

Pruning Pipeline

Original network training → Pruning → Fine-tuning
Membership Inference Attacks (MIAs)

Was this data sample used in training?
Membership Inference Attacks (MIAs)

Was this data sample used in training?

Will pruned neural networks become more vulnerable to MIAs?
Investigation on Confidence Gap (CIFAR10, DenseNet121)

Confidence gap between members and non-members is **INCREASED**!
Investigation on Sensitivity Gap (CIFAR10, DenseNet121)

Sensitivity gap between members and non-members is **INCREASED**!

Sensitivity:

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{|f_p(x + \epsilon \delta_i) - f_p(x)|}{\epsilon}
\]

\[\delta_i \sim \mathcal{N}(0, 1)\]
Confidence Gap and Sensitivity Gap per Class

Major findings:

1) Both confidence and sensitivity gaps are increased for most classes after pruning.

2) Increased gaps differ among different classes.
SAMIA: Self-Attention Membership Inference Attack

• **Hypothesis**: the increased confidence gap and sensitivity gap among different classes can provide fine-grained "evidence" for MIAs.

• Most MIAs learn a single threshold of prediction confidence to determine the membership status, which may not be sufficient for neural network pruning.

• We introduce a neural network-based attack using self-attention mechanism: **SAMIA**.

• SAMIA leverages self-attention mechanism to find out the specific confidence and sensitivity information that the attack “threshold” should pay more “attention” to.
Evaluation Setup

- **4 Neural Network Pruning Approaches**
  - L1 unstructured pruning (Han 2016)
  - L1 structured pruning (Li 2017)
  - L2 structured pruning (L1 2017)
  - Network slimming (Liu 2017)

- **5 Pruning Sparsity Levels**
  - 0.5, 0.6, 0.7, 0.8, 0.9

- **8 Membership Inference Attacks**
  - 4 Metric-based attacks, 2 Neural network-based attacks, BlindMI and SAMIA

- **7 Popular Datasets**
  - CIFAR10, CIFAR100, CHMNIST, SVHN, Location, Texas, Purchase

- **4 Neural Network Architectures**
  - Image datasets: ResNet18, VGG16, DefenseNet121
  - Non-image datasets: Fully Connected Neural Network
Privacy Risks Under Different Pruning Approaches and Sparsity Levels

- Most pruning approaches result in increased attack accuracy.
- The attack accuracy may be decreased under a high sparsity level, e.g., 0.9 (when the pruned model cannot achieve a comparable prediction accuracy).
SAMIA outperforms the existing membership inference attacks on different pruning approaches and sparsity levels.
Strong correlation between the gaps (confidence gap, sensitivity gap) and attack accuracy.
Pair-based Posterior Balancing (PPB) Defense

- **Basic Idea:** align the posterior predictions of different input samples to mitigate the new prediction behaviors (increased gaps) introduced by neural network pruning.

- **Apply PPB Defense in Fine-tuning process:**
  - Select two data samples in a batch as a pair without replacement.
  - **Balance the posteriors** by minimizing the distance of the ranked posteriors.
  - Formulate the new loss function using the prediction loss and KL-divergence loss.
  - Fine-tune the pruned model using the new loss function.

The loss function is given by:

\[
\mathcal{L}(f_p(x), y) = \sum_i \mathcal{L}_{\text{predict}}(f_p(x_i), y_i) + \lambda \sum_{j,k(j\neq k)} \mathcal{L}_{\text{KL}}(R(f_p(x_j)), R(f_p(x_k)))
\]

- \(\mathcal{L}_{\text{predict}}\) is the prediction loss (e.g., cross-entropy loss).
- \(\mathcal{L}_{\text{KL}}\) is the KL-divergence loss.
- \(R(\cdot)\) sorts the posteriors in a descending order.
Confidence Gap after PPB (CIFAR10, DenseNet121)

Confidence gap between members and non-members is REDUCED!
Sensitivity Gap after PPB (CIFAR10, DenseNet121)

Sensitivity: \[ \frac{1}{n} \sum_{i=1}^{n} \frac{|f_p(x + \epsilon \delta_i) - f_p(x)|}{\epsilon} \]

Sensitivity gap between members and non-members is REDUCED!
Defense Evaluation

• Comparison with existing defenses
  • Early Stopping and L2 Regularization (Basic), Song 2019
  • Differential Privacy (DP), Abadi 2016
  • Adversarial Regularization (ADV), Nasr 2018

PPB outperforms the existing defenses, achieving a better tradeoff between prediction accuracy and model privacy.

(CIFAR10, ResNet18, L1 structured pruning, Sparsity 0.6)
Conclusion

• Neural network pruning aggravates the privacy risks of the original neural networks due to the increased confidence gap and sensitivity gap.

• The proposed SAMIA to predict membership status by using finer-grained prediction metrics.

• SAMIA has advantages in identifying the pruned models’ prediction divergence compared with the existing attacks.

• The proposed PPB defense mitigates pruned model’s privacy risks by narrowing down the divergences of posterior predictions.
Thank you!

Please refer to the extended version for more details: arxiv.org/abs/2202.03335


Scan me to get the code!