NC
FACULTY OF
COMPUTER NETWORKS AND COMMUNICATIONS

TRANSFER LEARNING ATTACK

Nhóm 1
Nguyễn Hồng Sơn - Ngô Thái Hưng - Tô Thị Mỹ Âu

Instructor: Dr Lê Kim Hùng

15th December 2024

Table of Contents

Table of Contents

What is Transfer Learning?

- Transfer Learning is a method reuse models trained on a large dataset in the source domain to solve problems in the target domain – where data is often scarce.
- This method saves time, costs and improves system efficiency.

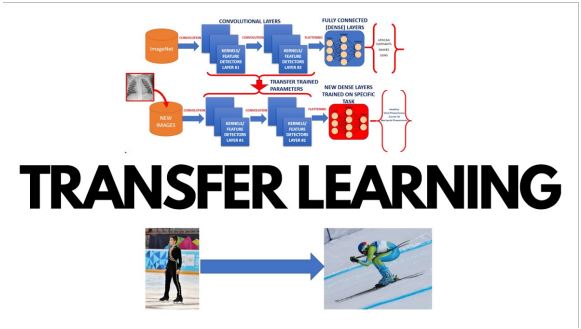


Figure: Transfer Learning

Why is Transfer Learning Attack dangerous?

Figure: The dangers of Transfer Learning Attacks

When Does Transfer Learning Attack Happen?

- Model Pre-Training: Attackers train a model with backdoored data and upload it to public repositories.
- Re-Training: Users re-train the model on their clean dataset, unaware of the embedded backdoor.
- Deployment: The compromised model is deployed, and triggers (e.g., manipulated traffic signs) exploit its vulnerabilities.

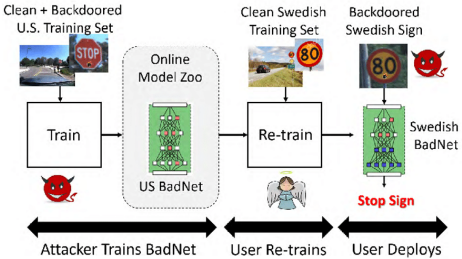


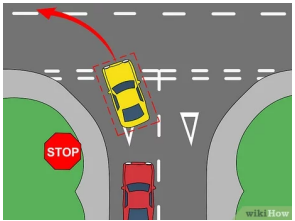
Figure: Scenarios of Transfer Learning Attack.

Where does Transfer Learning Attack occur?

- Transfer Learning Attacks can target multiple critical domains where AI and machine learning models are widely used.

Transportation:

- Adversarial attacks can mislead self-driving cars.



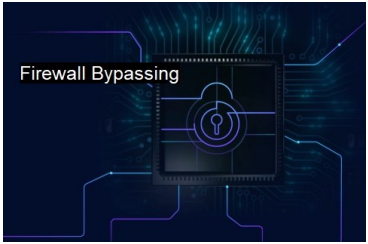
- Healthcare:
  - Manipulating medical image classifiers to misdiagnose conditions.

Where does Transfer Learning Attack occur?

- Transfer Learning Attacks can target multiple critical domains where AI and machine learning models are widely used.

- Finance:
  - Exploiting fraud detection models by injecting poisoned data.

- Security:
  - Circumventing intrusion detection systems.



Who: Who attacks and who are the victims?

The Attackers:

- Characteristics:
  - Individuals or organizations with diverse motivations.
  - Can range from malicious hackers to competing entities.
- Goals:
  - Disrupt operations:
    - Cause technological or financial damage.
  - Steal sensitive data:
    - Misuse or sell confidential information obtained from attacked models.

Who: Who attacks and who are the victims?

The Victims:

- Characteristics:
  - Companies using Transfer Learning models.
  - End-users who rely on these systems for critical tasks.
- Impacts:
  - Loss of operational trust.
  - Exposure to legal, financial, and reputational risks.
  - Potential harm to users depending on critical applications (e.g., healthcare, self-driving cars).

Consequences of Transfer Learning Attack

- Key Examples:
  - Misclassification:
    - Attacks like adversarial examples can cause models to misclassify objects, leading to system errors.
  - Data Exfiltration:
    - Attackers may extract sensitive information from training datasets, such as personal, financial, or medical data.
  - Asset Loss:
    - Financial losses due to manipulated models or incorrect predictions.

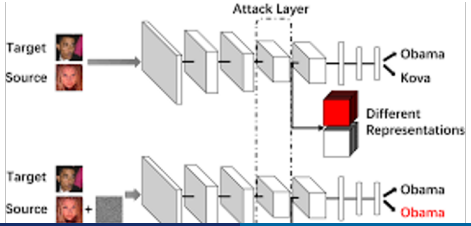


Figure: Examples of the Consequences of Transfer Learning Attack.

How to Prevent Transfer Learning Attacks?

Key Preventative Measures:

- Secure Training Data
- Model Isolation
- Knowledge Distillation
- Differential Privacy
- Adversarial Training

Additional Techniques:

- Data Augmentation
- Ensemble Methods
- Post-processing and anomaly detection

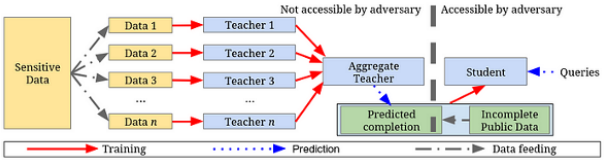


Figure: Differential privacy as a protection to Transfer Learning attacks

Table of Contents

Methodology

Zhang, Yinghua, et al. "Two sides of the same coin: White-box and black-box attacks for transfer learning." Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. 2020.

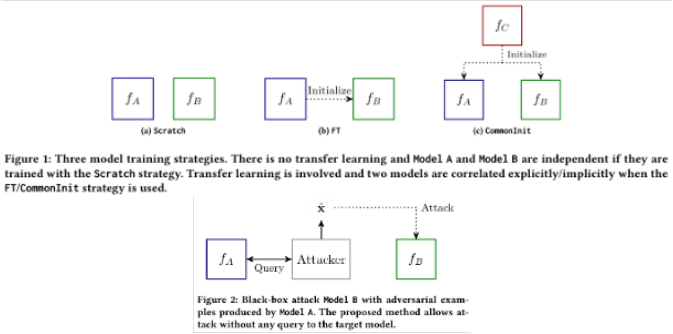


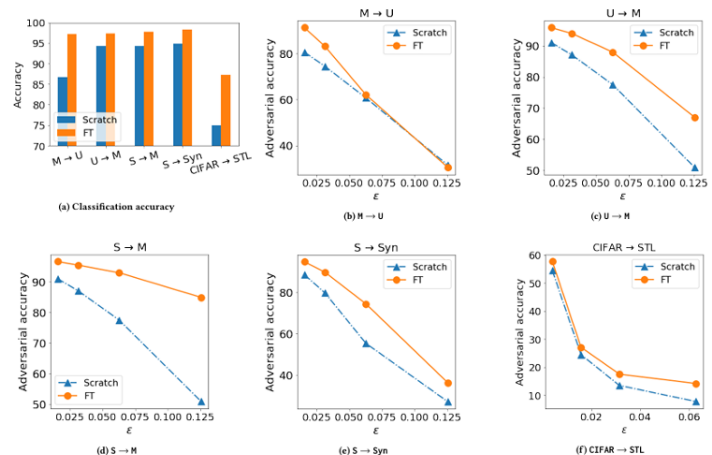
Figure 1: Three model training strategies. There is no transfer learning and Model A and Model B are independent if they are trained with the Scratch strategy. Transfer learning is involved and two models are correlated explicitly/implicitly when the FT/CommonInit strategy is used.

Figure 2: Black-box attack Model B with adversarial examples produced by Model A. The proposed method allows attack without any query to the target model.

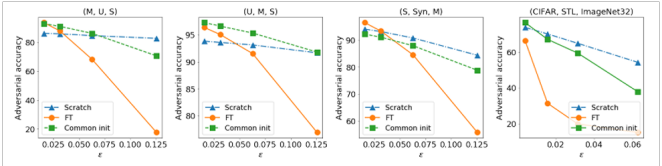
- Datasets: MNIST (M), USPS (U), SVHN (S), SynDigits (Syn), CIFAR10, STL10, ImageNet32

White-box Attack Experiment

- Goal: Assess robustness to FGSM attacks.
- Method: Compare Scratch vs. Fine-tuning models under various noise levels.
- Result:
  - Fine-tuning significantly improves accuracy and robustness.
  - Example: Accuracy increases from 50.86% -> 84.96% at $\epsilon = 0.125$.

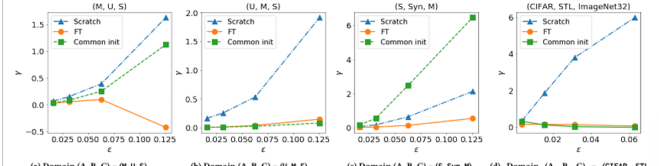


Black-box Attack Experiment



(a) Domain (A, B, C) = (M, U, S) (b) Domain (A, B, C) = (U, M, S) (c) Domain (A, B, C) = (S, Syn, M) (d) Domain (A, B, C) = (CIFAR, STL, ImageNet32)

Figure 4: Robustness (adversarial accuracy) under black-box attacks. The adversarial accuracies of the FT and `CommonInit` models drop drastically when the perturbation budget ϵ increases. They are much lower than those obtained with the `Scratch` models, which indicates that the fine-tuned models are likely to be attacked by the adversarial examples produced by their source models.



(a) Domain (A, B, C) = (M, U, S) (b) Domain (A, B, C) = (U, M, S) (c) Domain (A, B, C) = (S, Syn, M) (d) Domain (A, B, C) = (CIFAR, STL,

- Goal: Evaluate transferability of adversarial examples.
- Method: Attack model B using adversarial examples from model A.
- Result:
  - Fine-tuned models are more vulnerable to transferred attacks.
  - Example: Accuracy drops from 87.43% -> 65.32% (FGSM).

Table of Contents

Author's Conclusion

- Fine-tuning improves robustness:
  - Fine-tuning models enhance performance and security under white-box FGSM attacks.
- Risks of Fine-tuning:
  - Fine-tuned models are more vulnerable to adversarial examples from their source models.
  - Black-box attacks demonstrate increased risks when transfer learning is applied.
- Findings' Implication:
  - Highlights a trade-off between improved performance and susceptibility to attacks.

Author's Conclusion

- New Metrics:
  - Introduced a metric to evaluate the transferability of adversarial attacks.
- Future Implications:
  - Findings provide insights for designing transfer learning models that are robust and effective.
  - Encourages further research into adversarial robustness in transfer learning.
- Call to Action:
  - Developers need to carefully consider potential risks in fine-tuned systems.

Contributions of the Paper

- Comprehensive Experiments:
  - Evaluated robustness of fine-tuned models under both white-box and black-box attacks.
- Novel Insights:
  - Demonstrated trade-offs between performance and security in transfer learning.
- New Evaluation Metrics:
  - Proposed a metric to assess the transferability of adversarial examples.
- Practical Implications:
  - Findings emphasize the importance of adversarial training and robust model design.

Thanks for your attention