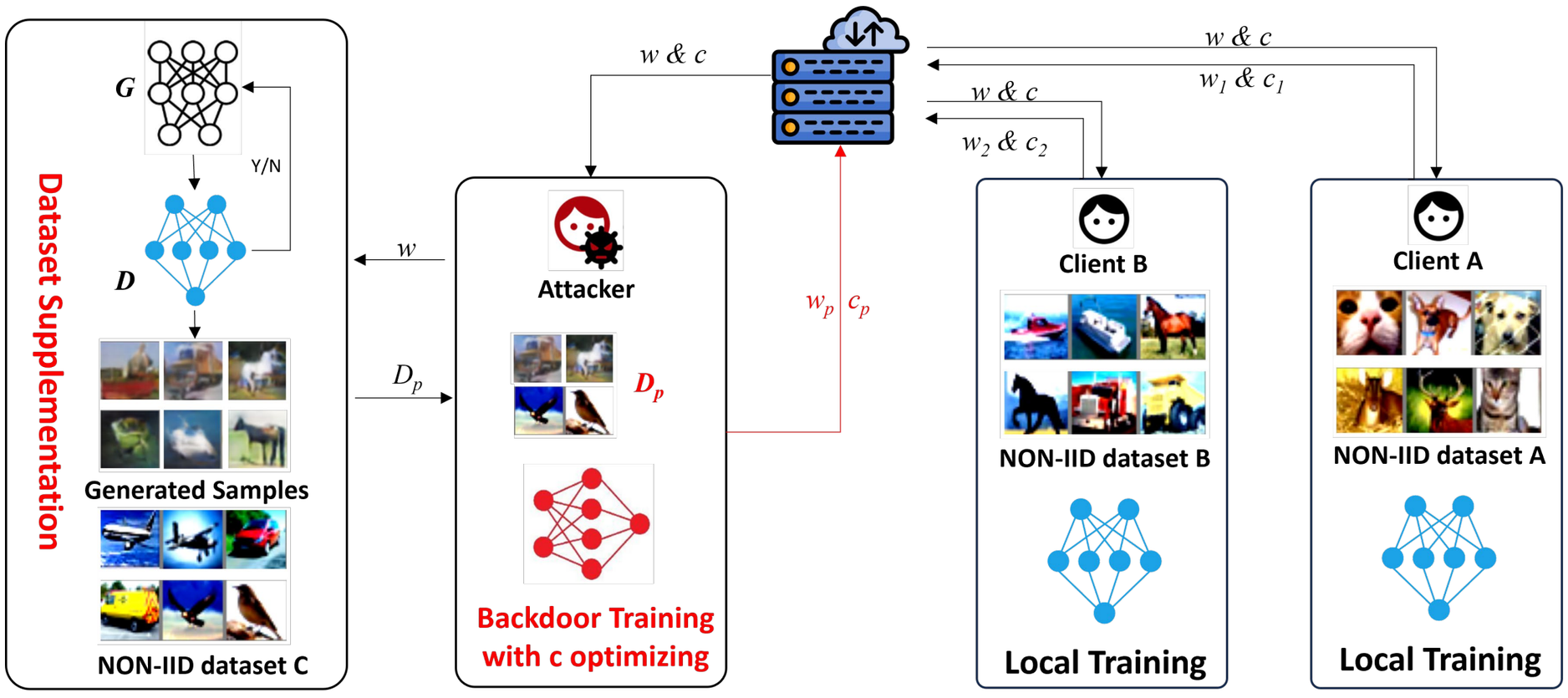
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

School of Computer Science and Engineering
College of Engineering

BadSFL

BACKDOOR ATTACK IN SCAFFOLD FEDERATED LEARNING

Student: Zhang Xuanye Supervisor: Zhang Tianwei



Existing Attack Challenges in SFL:

- **Knowledge Gap in NON-IID Scenario:** Direct backdoor training downgrade global model performance on Primary Task.
- **Control Variate Correction:** Following rules weaken attack but breaking them risks corrupting the model.
- **Transient Backdoors:** Benign updates eventually erase the backdoor after the attack stops.

BadSFL Attack Objective:

- **High Accuracy in both Primary and Backdoor Task:** GAN-based Dataset Supplementation
- **Stealthy Backdoor Function:** Distinctive feature triggers Injection
- **Durable Backdoor Integration:** Backdoor Training with Global Control Variate c Optimizing

BadSFL Attack Algorithm

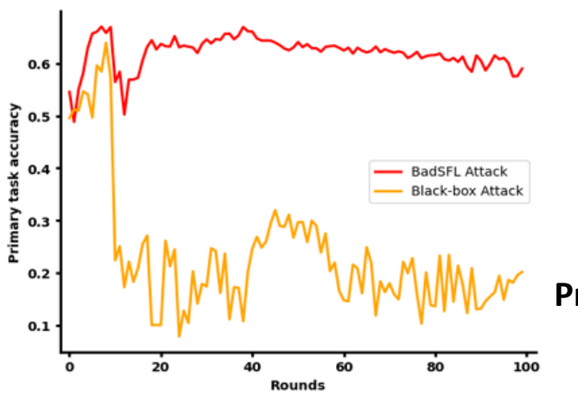
Required: local datasets D^i, global model w_g, global control variate c, number of local epochs E, local learning rate η_l, Generator G, Discriminator D

Update local model with $w_p \leftarrow w_g$
Initialize Discriminator $D \leftarrow w_g$
do:
  Run G for generating fake samples
  Evaluate fake sample on D
  Update G using D
until G converges to generate target samples

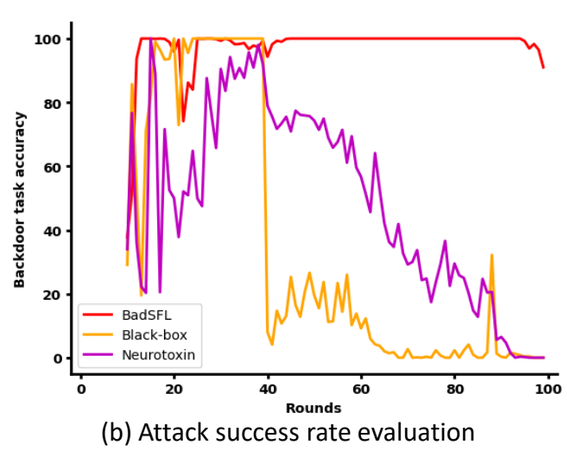
G generates samples into dataset D_f
$D_c \leftarrow D_f + D^i$
Select backdoor samples from D_c and assign them wrong label as D_b
$D_p \leftarrow D_c + D_b$
for each epoch e = 1, ..., E **do**:
  $w_p = argmin_{w_p}[L(D_p, w_p) + L(D_p, P_j(w_p, c))]$.
end for
$\Delta w_p = w_p - w_g$
$\Delta c_p = \frac{1}{K * \eta_l} * (w_g - w_p) - c$
return $(\Delta w_p, \Delta c_p)$

Experiment Result

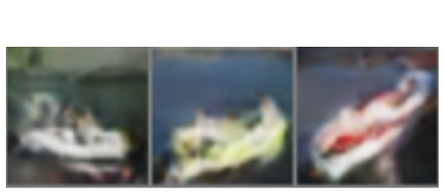
Featured-based Backdoor attack on the CIFAR 10 Dataset in SFL



(a) Primary Task accuracy under attacks

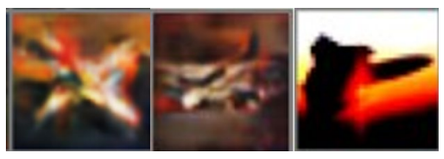


(b) Attack success rate evaluation



GAN Generated Authentic Samples

Preserved Model Efficacy on Primary task:
BadSFL Maintains 60% Accuracy
outperform Baseline (< 25%)



Plane in Sunset as Backdoor Trigger

High Efficacy:
Attack Success Rate over 90%
Extended Persistence:
Backdoor Lifespan Increased 3x