VILLAIN: Backdoor Attacks Against Vertical Split Learning

Yijie Bai\(^1\), Yanjiao Chen\(^1\), Hanlei Zhang\(^1\), Wenyuan Xu\(^1\), Haiqin Weng\(^2\), Dou Goodman\(^2\)

\(^1\)Ubiquitous System Security Lab (USSLAB), Zhejiang University, \(^2\)Ant Group

{baiyj, chenyanjiao, hanleizhang, wyxu}@zju.edu.cn
{haiqin.wenghaiqin, bencao.ly}@antgroup.com
Federated Learning

- **Horizontal** Federated Learning
- **Vertical** Federated Learning
Vertical Split Learning

- Credit business application
- Online advertising application
Vertical Split Learning

Step I

Participant A
Local Model

Participant B
Local Model

embedding

$f^a$

Label

Server

embedding

$f^b$

$f^s$
Vertical Split Learning

Step Ⅱ

Server

$f^s$

Participant A
Local Model

$f^a$

Participant B
Local Model

$f^b$

label

embedding

embedding

SEC’ 2023 VILLAIN: Backdoor Attacks Against Vertical Split Learning
Vertical Split Learning

Step Ⅲ

Server

label

embedding

gradient

embedding

Participant A
Local Model

\( f^a \)

Participant B
Local Model

\( f^b \)

\( f^s \)
Vertical Split Learning

$f^s$

Server

embedding

gradient

embedding

Attacker

$f^a$

Participant B

Local Model

$f^b$
Backdoor Attack
Attacker’s Goal

- Malicious Client
- Benign Client

Clean Server Model
Backdoored Server Model

Dog
Dog
Attacker’s Goal

- **Malicious Client**: Trigger → Clean Server Model
- **Benign Client**: Trigger → Backdoored Server Model

- Clean Server Model
- Backdoored Server Model
- Dog
- Airplane
Threat Model

- **Attacker’s knowledge**
  - Local dataset $X^a = \{\tilde{x}_i^a\}_{i=1}^N$
  - One target label sample
  - Gradient information

- **Attacker’s capability**
  - Train and manipulate the local embedding model $f^a$.
  - Upload the embedding vectors to the server.
Challenge

- **No label information**
  - No knowledge of the labels
  - Can't change the labels

- **No server model information**
  - Only gradient update information
  - Unknown server model
VILLAIN: Detailed Construction

Label Inference

No label information

Data Poisoning

No global model information
Pinpoint data samples of the target label.
Label Inference

Embedding Swapping

True Label: Plane

Swap with: Plane

\[ \hat{g}^a \text{ will be relatively small.} \]
Label Inference

Embedding Swapping

True Label: **Dog**

Swap with: **Plane**

\[ \hat{g}^a \] will be relatively large.
Label Inference

Embedding Swapping

Candidate Selection

Inference Adjustment

Target Label Samples

Non Target Samples

Embedding

Swapping

\[
\|\tilde{g}_i^a\|_2 \leq \theta \quad \text{and} \quad \|g_i^a\|_2 \leq \mu
\] are good indicators for label inference.
Label Inference

1. Semi-supervised classifier $\kappa$
2. Embedding $e_{i}^{a}$ with information

Candidate selection

Candidate sample

Target label sample

Data of target label

Data of the other labels
Label Inference

- Embedding Swapping
- Candidate Selection
- Inference Adjustment

**Embedding Swapping**

Data of target label: $\mathcal{f}_a \rightarrow g \rightarrow \mathcal{f}_b \rightarrow \mathcal{f}_s \rightarrow l$

Data of the other labels:

Candidate selection: $\mathcal{K}$

Candidate samples:

Target label sample

**Dynamically adjust the embedding for swapping**
Data Poisoning

The attacker poisons these target label samples to inject the backdoor into the server model.
Data Poisoning

- **Trigger Fabrication**
  - An additive trigger to poison the embedding vector
    \[ \hat{e}^a = f^a(\tilde{x}^a) \oplus \mathcal{E} \]
  - The trigger \( \mathcal{E} \) is formed as
    \[ \mathcal{E} = \mathcal{M} \otimes (\beta \cdot \Delta) \]
Experiment Setup

**Dataset**
- MNIST (MN).
- CIFAR-10 (CF).
- CINIC-10 (CN).
- ImageNette (IN).
- Bank Marketing (BM).
- Give-Me-Some-Credit (GM).

**Metrics**
- Attack success rate (ASR).
- Clean data accuracy (CDA).
- Label inference accuracy (LIA).

*4 image datasets (unstructured datasets) and 2 financial tabular datasets (structured datasets).*
Experiment Design

- **Overall Performance**
  - Potential side-effects.
  - Different embedding aggregation methods.
  - Data-domain triggers.
  - Multi-participant scenario.
  - Ablation studies

- **Hyperparameters**
  - Poisoning rate.
  - Trigger magnitude.
  - Server & participant models.
  - Trigger size.
  - Learning rate.
  - Number of candidates.

- **Resistance to Defense**
  - Label inference defense.
  - Backdoor attack defense.
  - Adaptive Defenses.
## Overall Performance

Villain achieves the **highest ASR on each dataset.**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MN</td>
<td>ASR</td>
<td>16.51 ± 5.14%</td>
<td>18.43 ± 4.50%</td>
<td>98.02 ± 2.21%</td>
<td>100.00 ± 0.00%</td>
<td>97.66 ± 5.57%</td>
<td>99.94 ± 0.13%</td>
<td>96.53 ± 5.11%</td>
</tr>
<tr>
<td></td>
<td>CDA</td>
<td>96.10 ± 0.22%</td>
<td>95.73 ± 0.16%</td>
<td>99.59 ± 0.19%</td>
<td>96.14 ± 0.88%</td>
<td>96.01 ± 0.12%</td>
<td>96.18 ± 0.07%</td>
<td>95.47 ± 0.33%</td>
</tr>
<tr>
<td></td>
<td>LIA</td>
<td>12.48 ± 0.73%</td>
<td>12.48 ± 0.73%</td>
<td>89.39 ± 6.99%</td>
<td>89.39 ± 6.99%</td>
<td>93.70 ± 4.18%</td>
<td>93.70 ± 4.18%</td>
<td>94.03 ± 2.56%</td>
</tr>
<tr>
<td>CF</td>
<td>ASR</td>
<td>8.26 ± 2.02%</td>
<td>16.93 ± 3.76%</td>
<td>13.61 ± 0.86%</td>
<td>78.99 ± 6.23%</td>
<td>14.45 ± 1.44%</td>
<td>84.96 ± 8.28%</td>
<td>23.66 ± 6.48%</td>
</tr>
<tr>
<td></td>
<td>CDA</td>
<td>76.66 ± 0.38%</td>
<td>75.94 ± 0.36%</td>
<td>76.75 ± 0.27%</td>
<td>76.96 ± 0.35%</td>
<td>77.09 ± 0.38%</td>
<td>74.99 ± 0.40%</td>
<td>96.08 ± 4.28%</td>
</tr>
<tr>
<td></td>
<td>LIA</td>
<td>18.96 ± 2.19%</td>
<td>18.96 ± 2.19%</td>
<td>68.12 ± 6.09%</td>
<td>68.12 ± 6.09%</td>
<td>76.35 ± 5.26%</td>
<td>96.08 ± 4.28%</td>
<td></td>
</tr>
<tr>
<td>IN</td>
<td>ASR</td>
<td>13.94 ± 4.8%</td>
<td>12.55 ± 1.79%</td>
<td>26.73 ± 2.73%</td>
<td>76.03 ± 9.59%</td>
<td>27.71 ± 2.44%</td>
<td>79.48 ± 6.09%</td>
<td>32.39 ± 12.26%</td>
</tr>
<tr>
<td></td>
<td>CDA</td>
<td>71.21 ± 0.39%</td>
<td>70.82 ± 0.93%</td>
<td>70.55 ± 0.18%</td>
<td>70.98 ± 0.22%</td>
<td>70.91 ± 0.50%</td>
<td>70.19 ± 0.74%</td>
<td>71.64 ± 0.89%</td>
</tr>
<tr>
<td></td>
<td>LIA</td>
<td>14.53 ± 1.70%</td>
<td>14.53 ± 1.70%</td>
<td>80.28 ± 8.94%</td>
<td>80.28 ± 8.94%</td>
<td>86.54 ± 6.68%</td>
<td>86.54 ± 6.68%</td>
<td>90.41 ± 2.18%</td>
</tr>
<tr>
<td>CN</td>
<td>ASR</td>
<td>5.13 ± 3.95%</td>
<td>8.98 ± 4.39%</td>
<td>26.63 ± 5.30%</td>
<td>86.56 ± 6.45%</td>
<td>33.95 ± 10.22%</td>
<td>85.01 ± 15.82%</td>
<td>64.56 ± 6.36%</td>
</tr>
<tr>
<td></td>
<td>CDA</td>
<td>61.90 ± 0.28%</td>
<td>61.64 ± 0.48%</td>
<td>62.65 ± 0.17%</td>
<td>62.86 ± 0.08%</td>
<td>62.68 ± 0.31%</td>
<td>62.72 ± 0.47%</td>
<td>62.67 ± 0.08%</td>
</tr>
<tr>
<td></td>
<td>LIA</td>
<td>12.55 ± 1.91%</td>
<td>12.55 ± 1.91%</td>
<td>66.83 ± 8.01%</td>
<td>66.83 ± 8.01%</td>
<td>72.09 ± 7.26%</td>
<td>72.09 ± 7.26%</td>
<td>93.19 ± 3.95%</td>
</tr>
<tr>
<td>BM</td>
<td>ASR</td>
<td>9.15 ± 3.90%</td>
<td>14.38 ± 4.93%</td>
<td>40.19 ± 4.31%</td>
<td>90.28 ± 10.19%</td>
<td>39.46 ± 2.53%</td>
<td>86.79 ± 10.56%</td>
<td>59.43 ± 12.10%</td>
</tr>
<tr>
<td></td>
<td>CDA</td>
<td>91.36 ± 0.77%</td>
<td>90.37 ± 0.51%</td>
<td>92.11 ± 0.94%</td>
<td>91.22 ± 2.71%</td>
<td>92.79 ± 0.25%</td>
<td>88.83 ± 2.55%</td>
<td>91.80 ± 1.46%</td>
</tr>
<tr>
<td></td>
<td>LIA</td>
<td>46.18 ± 2.39%</td>
<td>46.18 ± 2.39%</td>
<td>92.11 ± 4.49%</td>
<td>92.11 ± 4.49%</td>
<td>88.78 ± 4.64%</td>
<td>88.78 ± 4.64%</td>
<td>94.05 ± 4.82%</td>
</tr>
<tr>
<td>GM</td>
<td>ASR</td>
<td>12.01 ± 3.54%</td>
<td>17.87 ± 5.83%</td>
<td>67.69 ± 1.04%</td>
<td>100.00 ± 0.00%</td>
<td>67.43 ± 1.22%</td>
<td>100.00 ± 0.00%</td>
<td>92.27 ± 15.41%</td>
</tr>
<tr>
<td></td>
<td>CDA</td>
<td>78.02 ± 0.77%</td>
<td>77.81 ± 0.42%</td>
<td>78.55 ± 0.24%</td>
<td>78.41 ± 0.06%</td>
<td>78.32 ± 0.20%</td>
<td>78.32 ± 0.24%</td>
<td>78.68 ± 0.09%</td>
</tr>
<tr>
<td></td>
<td>LIA</td>
<td>55.78 ± 2.33%</td>
<td>55.78 ± 2.33%</td>
<td>77.66 ± 0.72%</td>
<td>77.66 ± 0.72%</td>
<td>77.52 ± 0.60%</td>
<td>77.52 ± 0.60%</td>
<td>95.18 ± 5.69%</td>
</tr>
</tbody>
</table>

**VILLAIN †**
Data-domain triggers

In VILLAIN, the trigger can be added in the data domain or the embedding domain.
Different embedding aggregation methods

- **Different aggregation methods.**
  - C: CON, embedding concatenation.
  - A: ADD, element-wise addition.
  - M1: MEAN, element-wise average.
  - M2: MAX, element-wise maximum.
  - M3: MIN, element-wise minimum.

VILLAIN performs well on different aggregation methods.

<table>
<thead>
<tr>
<th>DS</th>
<th>M'</th>
<th>orl. acc.</th>
<th>LIA</th>
<th>ASR</th>
<th>CDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>MN</td>
<td>C</td>
<td>95.82 ± 0.29%</td>
<td>94.03 ± 2.56%</td>
<td>100.00 ± 0.00%</td>
<td>96.11 ± 2.22%</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>96.69 ± 0.35%</td>
<td>99.00 ± 0.19%</td>
<td>100.00 ± 0.00%</td>
<td>95.97 ± 0.27%</td>
</tr>
<tr>
<td></td>
<td>M1</td>
<td>95.97 ± 0.38%</td>
<td>90.48 ± 2.59%</td>
<td>100.00 ± 0.00%</td>
<td>95.13 ± 0.50%</td>
</tr>
<tr>
<td></td>
<td>M2</td>
<td>95.61 ± 0.69%</td>
<td>94.05 ± 3.65%</td>
<td>100.00 ± 0.00%</td>
<td>94.56 ± 0.48%</td>
</tr>
<tr>
<td></td>
<td>M3</td>
<td>96.11 ± 0.16%</td>
<td>99.51 ± 0.17%</td>
<td>95.22 ± 1.13%</td>
<td>95.59 ± 0.37%</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>78.29 ± 0.24%</td>
<td>96.08 ± 4.28%</td>
<td>98.68 ± 0.59%</td>
<td>76.87 ± 0.25%</td>
</tr>
<tr>
<td></td>
<td>M1</td>
<td>77.83 ± 0.27%</td>
<td>99.85 ± 0.22%</td>
<td>94.55 ± 0.82%</td>
<td>79.00 ± 0.58%</td>
</tr>
<tr>
<td></td>
<td>M2</td>
<td>76.44 ± 0.37%</td>
<td>99.86 ± 0.32%</td>
<td>94.85 ± 0.51%</td>
<td>79.17 ± 0.18%</td>
</tr>
<tr>
<td></td>
<td>M3</td>
<td>76.94 ± 0.05%</td>
<td>99.95 ± 0.02%</td>
<td>91.33 ± 0.40%</td>
<td>78.99 ± 0.70%</td>
</tr>
<tr>
<td>IN</td>
<td>C</td>
<td>71.59 ± 0.84%</td>
<td>90.41 ± 2.18%</td>
<td>92.79 ± 1.58%</td>
<td>71.54 ± 0.08%</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>71.93 ± 1.06%</td>
<td>88.56 ± 2.63%</td>
<td>100.00 ± 0.00%</td>
<td>68.84 ± 0.87%</td>
</tr>
<tr>
<td></td>
<td>M1</td>
<td>59.99 ± 1.94%</td>
<td>82.30 ± 4.48%</td>
<td>99.29 ± 0.12%</td>
<td>56.64 ± 1.57%</td>
</tr>
<tr>
<td></td>
<td>M2</td>
<td>66.95 ± 1.44%</td>
<td>84.30 ± 2.21%</td>
<td>100.00 ± 0.00%</td>
<td>64.56 ± 0.79%</td>
</tr>
<tr>
<td></td>
<td>M3</td>
<td>65.99 ± 1.57%</td>
<td>86.69 ± 3.74%</td>
<td>100.00 ± 0.00%</td>
<td>63.39 ± 1.30%</td>
</tr>
<tr>
<td>CN</td>
<td>C</td>
<td>62.10 ± 0.08%</td>
<td>93.19 ± 3.95%</td>
<td>99.55 ± 0.62%</td>
<td>62.78 ± 0.11%</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>63.36 ± 1.37%</td>
<td>94.97 ± 4.22%</td>
<td>95.84 ± 3.82%</td>
<td>62.81 ± 1.50%</td>
</tr>
<tr>
<td></td>
<td>M1</td>
<td>63.19 ± 0.27%</td>
<td>88.61 ± 2.90%</td>
<td>96.81 ± 2.27%</td>
<td>61.76 ± 0.23%</td>
</tr>
<tr>
<td></td>
<td>M2</td>
<td>60.16 ± 1.51%</td>
<td>85.18 ± 3.07%</td>
<td>94.43 ± 1.06%</td>
<td>62.83 ± 0.59%</td>
</tr>
<tr>
<td></td>
<td>M3</td>
<td>63.29 ± 0.27%</td>
<td>88.47 ± 3.59%</td>
<td>96.81 ± 2.53%</td>
<td>64.11 ± 0.20%</td>
</tr>
<tr>
<td>BM</td>
<td>C</td>
<td>90.98 ± 0.52%</td>
<td>94.05 ± 4.82%</td>
<td>97.84 ± 2.57%</td>
<td>90.57 ± 0.14%</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>90.35 ± 0.36%</td>
<td>99.58 ± 0.37%</td>
<td>92.50 ± 3.83%</td>
<td>90.83 ± 0.23%</td>
</tr>
<tr>
<td></td>
<td>M1</td>
<td>92.68 ± 0.78%</td>
<td>99.89 ± 0.10%</td>
<td>70.68 ± 8.54%</td>
<td>92.70 ± 0.81%</td>
</tr>
<tr>
<td></td>
<td>M2</td>
<td>92.31 ± 0.35%</td>
<td>99.80 ± 0.12%</td>
<td>92.45 ± 3.61%</td>
<td>90.15 ± 0.96%</td>
</tr>
<tr>
<td></td>
<td>M3</td>
<td>91.94 ± 0.56%</td>
<td>99.90 ± 0.11%</td>
<td>84.32 ± 5.31%</td>
<td>90.31 ± 0.53%</td>
</tr>
<tr>
<td>GM</td>
<td>C</td>
<td>78.91 ± 0.28%</td>
<td>95.18 ± 5.69%</td>
<td>100.00 ± 0.00%</td>
<td>78.37 ± 0.14%</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>75.04 ± 0.30%</td>
<td>84.64 ± 6.17%</td>
<td>96.10 ± 1.70%</td>
<td>77.96 ± 0.25%</td>
</tr>
<tr>
<td></td>
<td>M1</td>
<td>76.80 ± 0.36%</td>
<td>93.13 ± 4.51%</td>
<td>98.37 ± 0.52%</td>
<td>77.04 ± 0.58%</td>
</tr>
<tr>
<td></td>
<td>M2</td>
<td>77.39 ± 0.28%</td>
<td>95.70 ± 6.98%</td>
<td>96.17 ± 1.24%</td>
<td>77.20 ± 0.32%</td>
</tr>
<tr>
<td></td>
<td>M3</td>
<td>77.54 ± 0.55%</td>
<td>95.77 ± 6.14%</td>
<td>97.99 ± 1.49%</td>
<td>76.69 ± 0.45%</td>
</tr>
</tbody>
</table>
Impact of Hyperparameters

- Impact of poisoning rate.
- Impact of server & participant models.
- Impact of learning rate.
- Impact of trigger size.
- Impact of trigger magnitude.
- Impact of number of candidates.

The backdoor attack still works even with a low poisoning rate of only 0.5%.
Impact of Hyperparameters

- Impact of poisoning rate.
- Impact of server & participant models.
- Impact of learning rate.
- Impact of trigger size.
- Impact of trigger magnitude.
- Impact of number of candidates.

VILLAIN is robust to different server structures.

<table>
<thead>
<tr>
<th>dep.</th>
<th>MNIST LIA</th>
<th>ASR</th>
<th>CIFAR-10 LIA</th>
<th>ASR</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>94.03 ± 2.56%</td>
<td>100.00 ± 0.00%</td>
<td>96.08 ± 4.28%</td>
<td>98.68 ± 0.59%</td>
</tr>
<tr>
<td>4</td>
<td>95.89 ± 2.95%</td>
<td>100.00 ± 0.00%</td>
<td>96.63 ± 3.55%</td>
<td>96.97 ± 0.45%</td>
</tr>
<tr>
<td>5</td>
<td>94.92 ± 2.63%</td>
<td>99.53 ± 0.24%</td>
<td>97.55 ± 3.97%</td>
<td>96.83 ± 0.24%</td>
</tr>
<tr>
<td>6</td>
<td>92.85 ± 4.10%</td>
<td>100.00 ± 0.00%</td>
<td>97.06 ± 1.73%</td>
<td>98.03 ± 0.58%</td>
</tr>
<tr>
<td>7</td>
<td>95.73 ± 2.66%</td>
<td>100.00 ± 0.00%</td>
<td>98.53 ± 2.66%</td>
<td>97.86 ± 0.13%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>dep.</th>
<th>CINIC-10 LIA</th>
<th>ASR</th>
<th>BM LIA</th>
<th>ASR</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>93.19 ± 3.05%</td>
<td>99.55 ± 0.62%</td>
<td>94.05 ± 4.82%</td>
<td>97.84 ± 2.57%</td>
</tr>
<tr>
<td>4</td>
<td>94.10 ± 2.56%</td>
<td>97.27 ± 1.43%</td>
<td>95.03 ± 5.93%</td>
<td>96.91 ± 0.92%</td>
</tr>
<tr>
<td>5</td>
<td>93.68 ± 1.41%</td>
<td>98.03 ± 0.20%</td>
<td>98.23 ± 0.96%</td>
<td>98.35 ± 0.47%</td>
</tr>
<tr>
<td>6</td>
<td>96.14 ± 3.02%</td>
<td>95.82 ± 3.94%</td>
<td>94.76 ± 2.59%</td>
<td>92.47 ± 1.69%</td>
</tr>
<tr>
<td>7</td>
<td>95.16 ± 3.97%</td>
<td>96.29 ± 3.46%</td>
<td>95.91 ± 2.49%</td>
<td>95.10 ± 0.82%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>dep.</th>
<th>ImageNette LIA</th>
<th>ASR</th>
<th>GM LIA</th>
<th>ASR</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>90.41 ± 2.18%</td>
<td>92.79 ± 1.58%</td>
<td>95.18 ± 5.69%</td>
<td>100.00 ± 0.00%</td>
</tr>
<tr>
<td>4</td>
<td>92.14 ± 3.06%</td>
<td>93.01 ± 1.65%</td>
<td>98.62 ± 0.63%</td>
<td>100.00 ± 0.00%</td>
</tr>
<tr>
<td>5</td>
<td>95.52 ± 3.45%</td>
<td>96.68 ± 0.94%</td>
<td>96.28 ± 3.10%</td>
<td>99.35 ± 0.20%</td>
</tr>
<tr>
<td>6</td>
<td>97.05 ± 7.49%</td>
<td>90.93 ± 3.69%</td>
<td>93.60 ± 4.60%</td>
<td>100.00 ± 0.00%</td>
</tr>
<tr>
<td>7</td>
<td>94.11 ± 2.46%</td>
<td>92.04 ± 0.75%</td>
<td>94.04 ± 3.63%</td>
<td>98.80 ± 0.94%</td>
</tr>
</tbody>
</table>
Possible Defenses

- **Label Inference Defense**
  - **DPSGD**
  - **Gradient compression**
  - **Privacy-preserving Deep Learning**

Villain can defeat existing label inference methods.
Possible Defenses

- **Backdoor Attack Defense**
  - Model reconstruction
  - Sample preprocessing
  - Trigger synthesis
  - Poison suppression

Both trends prove the defense can not keep high CDA while reducing the ASR.
Conclusion

➢ Design effective data poisoning strategies to strengthen the link between the trigger and the backdoor in the server model.

➢ Develop a new label inference algorithm to locate samples of the target label.

➢ Our attack is validated to be effective, robust, and efficient based on extensive experiments.
VILLAIN: Backdoor Attacks Against Vertical Split Learning

Thank you for your patience!

Contract us at: baiyj@zju.edu.cn

USSLAB Website: www.usslab.org