最後一組 Final Report

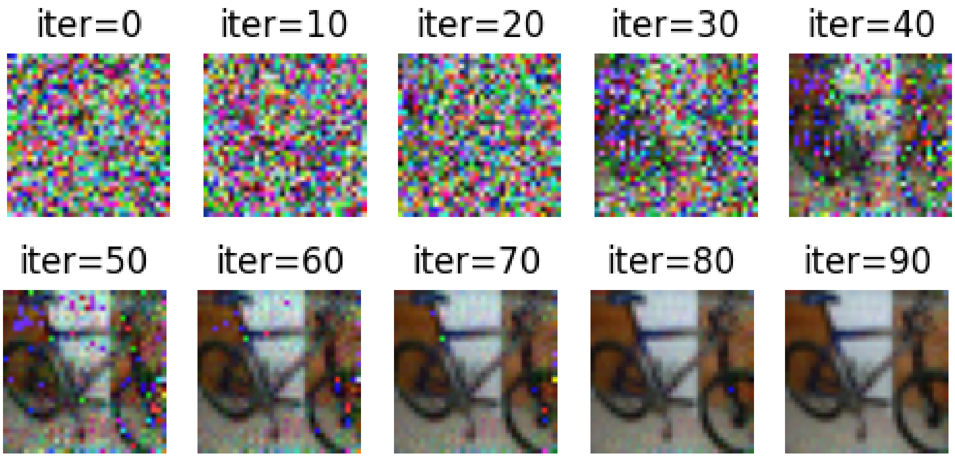
M11215052陳奕帆 M11215066 鄭宜珊 M11215032 葉品和

1. Selected paper & source code

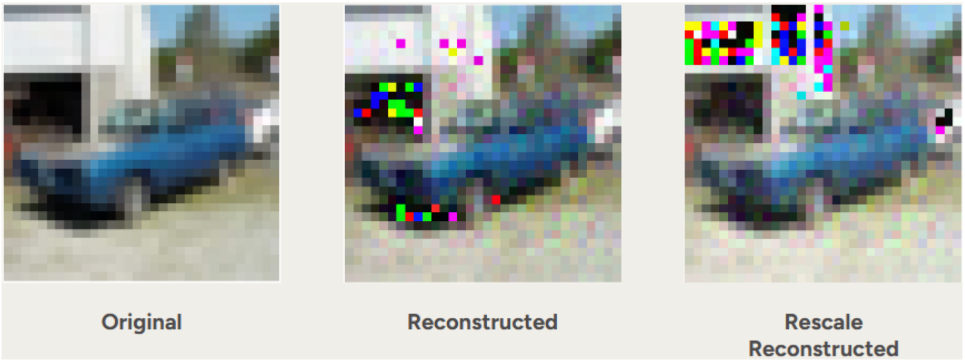
- Deep Leakage from Gradients (optimization-based)
  https://proceedings.neurips.cc/paper/2019/file/60a6c4002cc7b29142def8871531281a-Paper.pdf
  https://github.com/mit-han-lab/dlg
- R-GAP: Recursive gradient attack on privacy (analytics-based)
  https://arxiv.org/pdf/2010.07733
  https://github.com/JunyiZhu-AI/R-GAP

2. Result:

- DLG



- R-GAP



3. Comparison

- DLG (Optimization-based)
  Randomly generate fake data & fake labels, and then calculate the loss and gradient using the prediction results from the fake data and fake labels.
  Compute the difference between the real gradient and the fake gradient, and use this to update the fake data and fake labels to restore the original data.

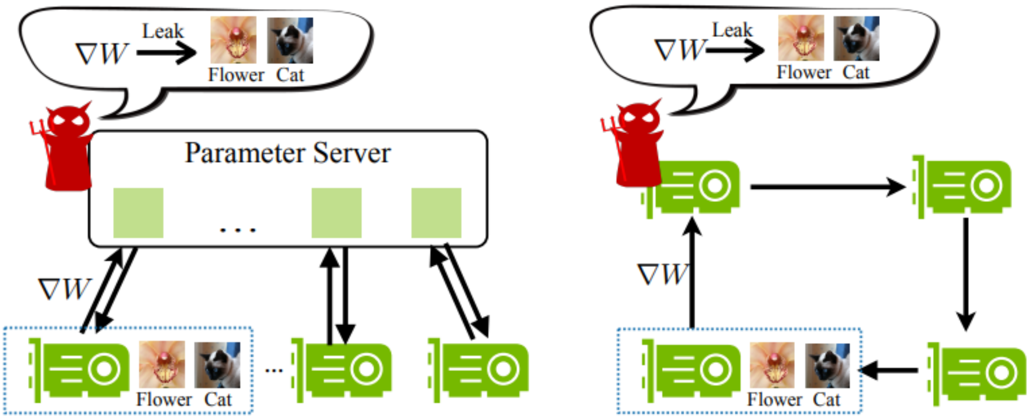
- R-GAP(Analytics-based)
  Reverse-engineer the feature map based on the gradients and weight, and then use the least square solution to trace back the input of each layer.
  After reaching the first layer, the original data can be obtained.

Optimization-based attack methods simulate the process of approximating the fake gradient to the real gradient to restore the original data.

Analytics-based attack methods use the mathematical properties to retrieve the input of each layer to restore the original data from the first layer.

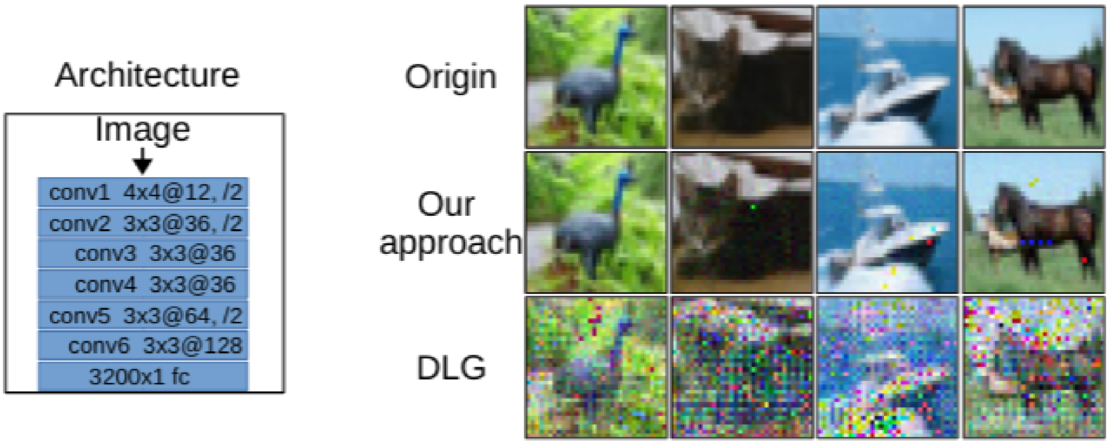
4. Observations

a. DLG is not always able to work on clients. As the figure shown below, only the parameter server can get clients' gradients when there is a centralized server in distributed training.



(a) Distributed training with a centralized server (b) Distributed training without a centralized server

b. CNN implemented in R-GAP doesn't include pooling layers. Adding pooling layers in the model may decrease the quality of the restoration image.



5. Responsibilities

陳奕帆	report, implementation
鄭宜珊	report, implementation
葉品和	implementation