

Supplementary Material

A CAFE vs DLG

Suppose that $N = 3$ and (4) can be rewritten as

$$\hat{\mathcal{X}}^* = \arg \min_{\hat{\mathcal{X}}} \left\| \frac{1}{3} \sum_{n=1}^3 \nabla_{\theta} \mathcal{L}(\theta, \mathbf{x}_n, y_n) - \frac{1}{3} \sum_{n=1}^3 \nabla_{\theta} \mathcal{L}(\theta, \hat{\mathbf{x}}_n, \hat{y}_n) \right\|^2 \quad (13)$$

We assume that there is a global optimal solution for (13) as

$$\hat{\mathcal{X}}^* = [\{\mathbf{x}_1, y_1\}; \{\mathbf{x}_2, y_2\}; \{\mathbf{x}_3, y_3\}] \quad (14)$$

However, besides the optimal solution, there might be other undesired solutions, such as $\hat{\mathcal{X}}^*$ shown in (15), whose gradients satisfy (16).

$$\begin{aligned} \hat{\mathcal{X}}^* &= [\{\hat{\mathbf{x}}_1^*, \hat{y}_1^*\}; \{\hat{\mathbf{x}}_2^*, \hat{y}_2^*\}; \{\mathbf{x}_3, y_3\}] \quad (15) \\ \sum_{n=1}^2 \nabla_{\theta} \mathcal{L}(\theta, \mathbf{x}_n, y_n) &= \sum_{n=1}^2 \nabla_{\theta} \mathcal{L}(\theta, \hat{\mathbf{x}}_n^*, \hat{y}_n^*) \\ \nabla_{\theta} \mathcal{L}(\theta, \mathbf{x}_n, y_n) &\neq \nabla_{\theta} \mathcal{L}(\theta, \hat{\mathbf{x}}_n^*, \hat{y}_n^*) \quad (16) \end{aligned}$$

Although solution (14) and (15) have the same loss value in (13), solution (15) is not an ideal solution for data recovery, which needs to be eliminated by introducing more constraints. When the number N increases, the number of both optimal and undesired solutions explodes. It is hard to find an approach which can converge to a certain solution through only one objective function.

However, in CAFE, the number of objective functions can be as many as $\binom{N}{N_b}$. As the case above, suppose $N_b = 2$. Then we can list all the objective functions

$$\begin{cases} \hat{\mathcal{X}}^{0*} &= \arg \min_{\hat{\mathcal{X}}^0} \left\| \frac{1}{2} \sum_{n=1}^2 \nabla_{\theta} \mathcal{L}(\theta, \mathbf{x}_n, y_n) - \frac{1}{2} \sum_{n=1}^2 \nabla_{\theta} \mathcal{L}(\theta, \hat{\mathbf{x}}_n, \hat{y}_n) \right\|^2 \\ \hat{\mathcal{X}}^{1*} &= \arg \min_{\hat{\mathcal{X}}^1} \left\| \frac{1}{2} \sum_{n=2}^3 \nabla_{\theta} \mathcal{L}(\theta, \mathbf{x}_n, y_n) - \frac{1}{2} \sum_{n=2}^3 \nabla_{\theta} \mathcal{L}(\theta, \hat{\mathbf{x}}_n, \hat{y}_n) \right\|^2 \\ \hat{\mathcal{X}}^{2*} &= \arg \min_{\hat{\mathcal{X}}^2} \left\| \frac{1}{2} \sum_{n=1, n \neq 2}^3 \nabla_{\theta} \mathcal{L}(\theta, \mathbf{x}_n, y_n) - \frac{1}{2} \sum_{n=1, n \neq 2}^3 \nabla_{\theta} \mathcal{L}(\theta, \hat{\mathbf{x}}_n, \hat{y}_n) \right\|^2. \end{cases} \quad (17)$$

Comparing with (13), (17) has more constraint functions which restrict $\hat{\mathcal{X}}$ and dramatically reduces the number of undesired solutions. Solution (15) thus can be eliminated by the second and the third equations in (17). It suggests that CAFE helps the fake data converge to the optimal solution.

B CAFE IN HORIZONTAL FL

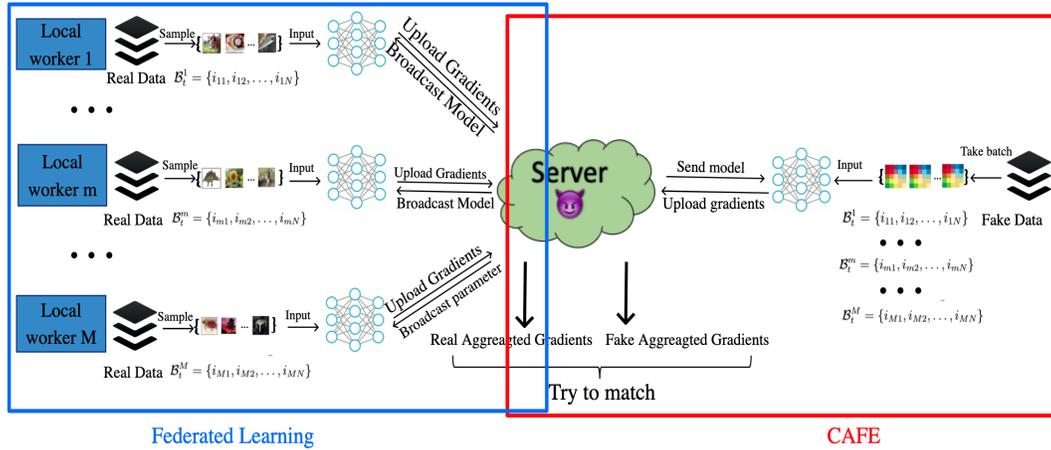


Figure 7: Overview of CAFE in HFL

Algorithm 2 CAFE in HFL (regular HFL protocol and CAFE protocol)

```

1: for  $t = 1, 2, \dots, T$  do
2:   Server broadcasts all models to all the local workers (a total of  $M$  workers)
3:   for  $m = 1, 2, \dots, M$  do
4:     Real worker  $m$  takes batched data  $\mathcal{X}_m^t$ 
5:     Real worker  $m$  uses the received model to compute  $\mathcal{L}(\theta, \mathcal{X}_m^t)$  and gradients  $\nabla_{\theta}\mathcal{L}(\theta, \mathcal{X}_m^t)$ 
6:     Real worker  $m$  uploads real local aggregated gradients to the server
7:   end for
8:   Server computes real global aggregated gradients  $\nabla_{\theta}\mathcal{L}(\theta, \mathcal{X}^t)$ 
9:   for  $m = 1, 2, \dots, M$  do
10:    the server takes corresponding batched data  $\hat{\mathcal{X}}_m^t$ 
11:    the server uses the received model to compute  $\mathcal{L}(\theta, \hat{\mathcal{X}}_m^t)$  and  $\nabla_{\theta}\mathcal{L}(\theta, \hat{\mathcal{X}}_m^t)$ 
12:   end for
13:   Server computes fake global aggregated gradients  $\nabla_{\theta}\mathcal{L}(\theta, \hat{\mathcal{X}}^t)$ 
14:   Server computes CAFE loss:  $\mathbb{D}(\mathcal{X}^t; \hat{\mathcal{X}}^t)$  and  $\nabla_{\hat{\mathcal{X}}^t}\mathbb{D}(\mathcal{X}^t; \hat{\mathcal{X}}^t)$ 
15:   for  $m = 1, 2, \dots, M$  do
16:     Server updates the batch data  $\hat{\mathcal{X}}_m^t$ 
17:   end for
18: end for

```

Figure 7 shows the overview of CAFE in HFL settings. The left blue part indicates a normal HFL process and the right red part represents the attack. According to our simulation, more than 2000 private images from 4 local workers can be leaked by CAFE.

C COMPARISON WITH GIVEN LABELS

From Table 5a and Table 5b, recovery results on dataset with more categories are more likely to be effected if the labels are given. However, recoveries on datasets with few categories (10 or 5) have little influence.

Table 5: Impact by given labels

PSNR \ Datasets	CIFAR-10	CIFAR-100	Linnaeus	PSNR \ Datasets	CIFAR-10	CIFAR-100	Linnaeus
Setting				Setting			
Not given labels	35.03	36.90	36.37	Not given labels	41.80	44.42	38.96
Given labels	35.93	39.51	38.07	Given labels	40.20	40.29	39.50
Number of categories	10	100	5	Number of categories	10	100	5

(a) HFL (b) VFL

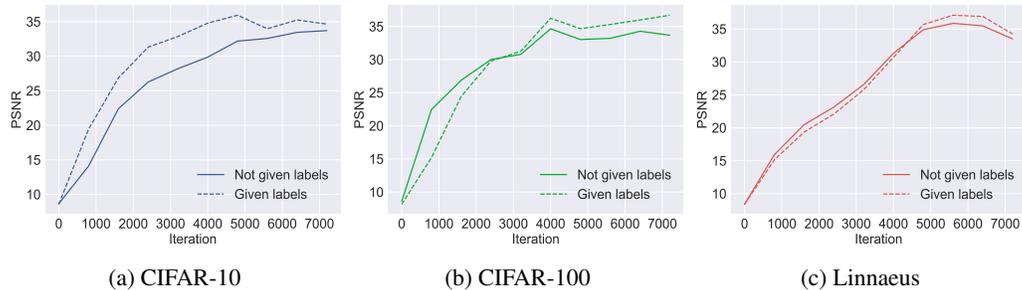


Figure 8: Impact by given labels (HFL)

D CONTRIBUTION OF AUXILIARY REGULARIZERS

Table 6: Effect of auxiliary regularizers

PSNR \ Datasets Algorithm	CIFAR-10	CIFAR-100	Linnaeus
	CAFE	35.03	36.90
CAFE ($\xi = 0$)	26.53	27.43	28.99
CAFE ($\beta = 0$)	23.19	22.09	31.67
CAEE ($\gamma = 0$)	25.41	18.14	24.17

(a) HFL

(4 workers, batch size = 10 per worker, 800 epochs)

PSNR \ Datasets Algorithm	CIFAR-10	CIFAR-100	Linnaeus
	CAFE	43.31	48.10
CAFE ($\xi = 0$)	42.03	36.69	34.40
CAFE ($\beta = 0$)	30.37	38.38	31.29
CAEE ($\gamma = 0$)	12.67	12.48	11.72

(b) VFL

(4 workers, batch size = 40, 800 epochs)