FedRN: Exploiting k-Reliable Neighbors Towards Robust Federated Learning

SangMook Kim KAIST Republic of Korea sangmook.kim@kaist.ac.kr
Wonyoung Shin* NA VER Shopping Republic of Korea wonyoung.shin@navercorp.com
Soohyuk Jang* POSTECH Republic of Korea jang001@postech.ac.kr Hwanjun Song†
NA VER Corp. Republic of Korea hwanjun.song@navercorp.com
Se-Young Yun† KAIST Republic of Korea yunseyoung@kaist.ac.kr

CIKM 2022

Introduction

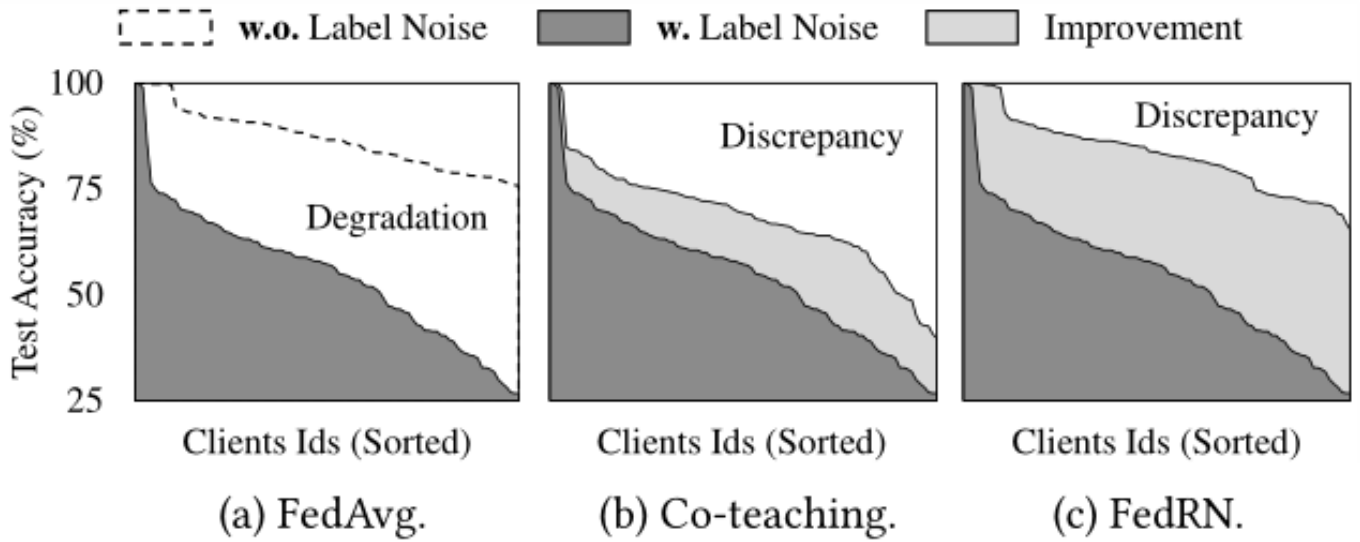
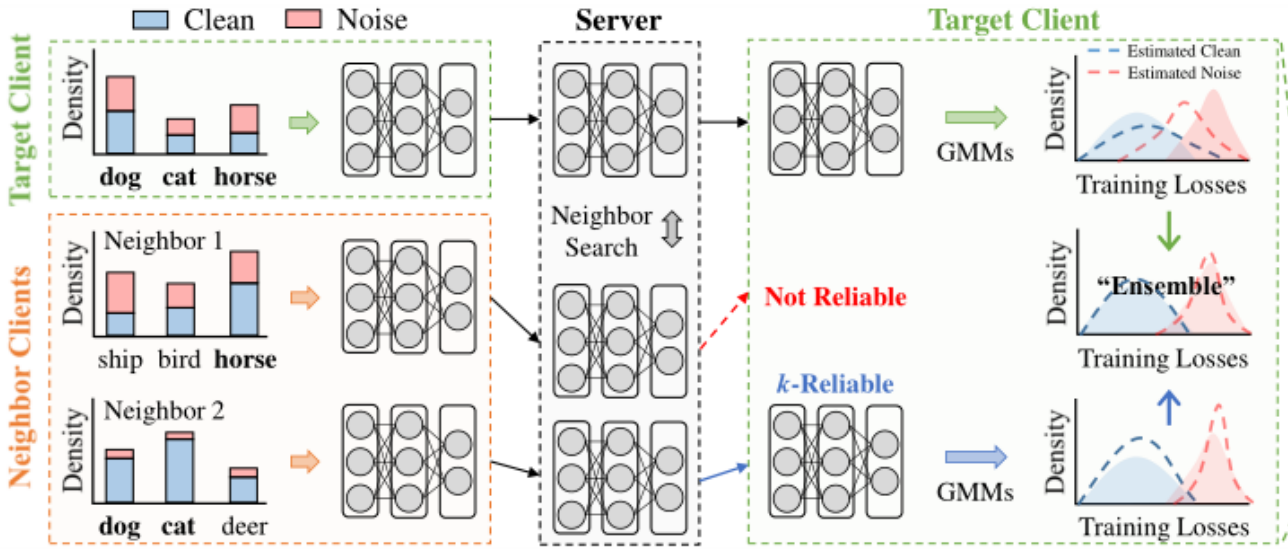
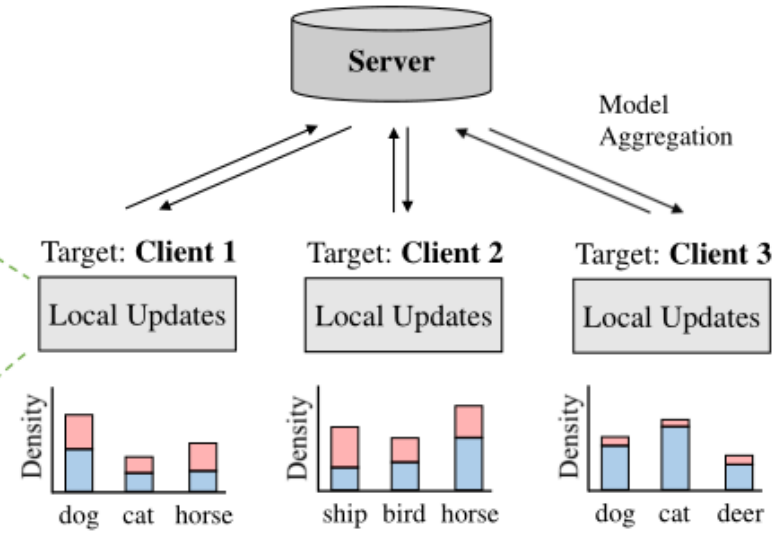


Figure 1: Performance difference of local models over 100 clients after training on CIFAR-10 with symmetric noise of $0-80\%$ in the Non-IID setting [24]: (a) shows the performance degradation incurred by label noise in training data; (b) and (c) contrasts the improvement by Co-teaching and our proposed FedRN. Client Ids are sorted in ascending order by the noise ratio of their local data.

Method



(a) Local Update Phase.



(b) Model Aggregation Phase.

$$\Theta_{\text{global}} \leftarrow \sum_{c \in \mathcal{M}} w_c \Theta_c, \text{ where } w_c = \frac{|\mathcal{D}_c|}{\sum_{c' \in \mathcal{M}} |\mathcal{D}_{c'}|} \tag{1}$$

$$\Theta_{\text{global}} \leftarrow \Theta_{\text{global}} - \eta \nabla \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \ell(x, \tilde{y}; \Theta_{\text{global}}), \tag{2}$$

Method

(a) Local Update Phase. (b) Model Aggregation Phase.

$$\mathcal{R}_c(k) = \text{argmax}_{|(\mathcal{R}':n\in\mathcal{R}')|=k}\ \mathrm{R}(c,n),\ \text{where}$$
$$\mathrm{R}(c,n) = \alpha\cdot\mathrm{Exp}(n) + (1-\alpha)\cdot\mathrm{Sim}(c,n) \tag{3}$$

$$\mathrm{Exp}(c) = \frac{\mathrm{Acc}(c) - \mathrm{minAcc}(\{c\}\cup\mathcal{N}_c)}{\mathrm{maxAcc}(\{c\}\cup\mathcal{N}_c) - \mathrm{minAcc}(\{c\}\cup\mathcal{N}_c)}, \tag{4}$$

$$\mathrm{Sim}(c,n) = \mathrm{Cosine}\big(p(\tilde{x};\Theta_c), p(\tilde{x};\Theta_n)\big). \tag{5}$$

Method

(a) Local Update Phase.

(b) Model Aggregation Phase.

$$p\big(g|\ell(x,\tilde{y};\Theta)\big) = p(g)p\big(\ell(x,\tilde{y};\Theta)|g\big)/p\big(\ell(x,\tilde{y};\Theta)\big), \tag{6}$$

$$p\big(\text{clean}|x;\mathcal{R}_{\text{c}}(k)\big) = \sum_{\text{n}\in\{\text{c}\}\cup\mathcal{R}_{\text{c}}(k)} \text{R}'(\text{c},\text{n}) \times p\big(g|\ell(x,\tilde{y};\Theta_{\text{n}})\big),$$
$$\text{where } \text{R}'(\text{c},\text{n}) = \text{R}(\text{c},\text{n}) \,/\, \sum \quad \text{R}(\text{c},\text{n}'). \tag{7}$$

$$\mathcal{S}_{\text{c}} = \{x \in \tilde{\mathcal{D}}_{\text{c}} : p(\text{clean}|x;\mathcal{R}_{\text{c}}(k)) > 0.5\}, \tag{8}$$

Experiments

	# of Train	# of Val.	# of Classes	Noise Ratio
CIFAR-10	50,000	10,000	10	$\approx 0\%$
CIFAR-100	50,000	10,000	100	$\approx 0\%$
mini-WebVision	65,944	2500	50	$\approx 20\%$

Table 1: Summary of datasets.

Experiments

Non-IID Type	Shard ($S = 2$)				Shard ($S = 5$)				Dirichlet ($\beta = 0.5$)			
Noise Type	Symmetric		Asym	Mixed	Symmetric		Asym	Mixed	Symmetric		Asym	Mixed
Noise Rate	0.0–0.4	0.0–0.8	0.0–0.4	0.0–0.4	0.0–0.4	0.0–0.8	0.0–0.4	0.0–0.4	0.0–0.4	0.0–0.8	0.0–0.4	0.0–0.4
Oracle[1]	69.10	67.19	69.10	68.89	78.00	75.94	78.00	77.7	81.79	80.38	81.79	81.39
FedAvg [24]	46.39	38.69	49.09	46.82	66.44	56.46	67.23	67.95	75.92	70.60	77.77	77.17
Co-teaching [14]	63.20	52.72	62.48	61.68	74.66	67.17	74.18	75.18	79.97	75.39	79.53	79.94
Joint-optm [30]	50.44	42.23	38.04	47.28	69.22	64.02	67.29	68.84	75.46	70.43	74.81	75.43
SELFIE [27]	62.64	53.69	64.21	62.63	74.57	66.90	74.77	74.44	78.57	72.92	78.58	79.02
DivideMix [20]	62.35	58.18	62.07	63.38	68.73	65.82	68.95	69.32	74.26	72.25	73.29	73.47
Robust FL [36]	56.25	45.59	55.52	57.58	70.30	62.89	69.04	70.02	75.75	70.63	74.00	75.49
FedRN (k=1)	67.33	60.33	67.51	67.92	76.37	72.30	76.74	76.92	79.99	75.92	80.05	79.79
FedRN (k=2)	**67.62**	**62.94**	**68.33**	**68.11**	**76.81**	**72.85**	**77.33**	**76.99**	**80.34**	**76.49**	**80.38**	**80.28**

Table 2: Test accuracy (%) on CIFAR-10 with symmetric, asymmetric (Asym), and mixed noise (Mixed).

Experiments

Non-IID Type	Shard ($S = 20$)			
Noise Type	Symmetric		Asym	Mixed
Noise Rate	0.0–0.4	0.0–0.8	0.0–0.4	0.0–0.4
Oracle	47.83	44.82	47.83	47.33
FedAvg [24]	33.36	26.15	34.90	34.02
Co-teaching [14]	43.43	37.05	42.90	43.83
Joint-optm [30]	35.89	30.92	33.15	35.29
SELFIE [27]	44.14	37.65	43.25	43.78
DivideMix [20]	46.21	42.83	45.59	46.94
Robust FL [36]	35.33	29.20	33.08	34.27
FedRN (k=1)	47.30	42.27	**47.62**	46.62
FedRN (k=2)	**47.46**	**43.07**	47.52	**47.17**

Table 3: Test accuracy (%) on CIFAR-100.

Experiments

Shard ($S = 10$)	mini-WebVision	
Method	Top-1 Accuracy	Top-5 Accuracy
FedAvg [24]	20.80	45.00
Co-teaching [14]	21.76 (+ 0.96)	46.64 (+ 1.64)
Joint-optm [30]	13.56 (− 7.24)	35.56 (− 9.44)
SELFIE [27]	22.20 (+ 1.40)	48.76 (+ 3.76)
DivideMix [20]	20.84 (+ 0.04)	45.72 (+ 0.72)
Robust FL [36]	14.12 (− 6.68)	35.24 (− 9.76)
FedRN (k=1)	22.52 (+ 1.72)	48.48 (+ 3.48)
FedRN (k=2)	**22.76 (+ 1.96)**	**49.16 (+ 4.16)**

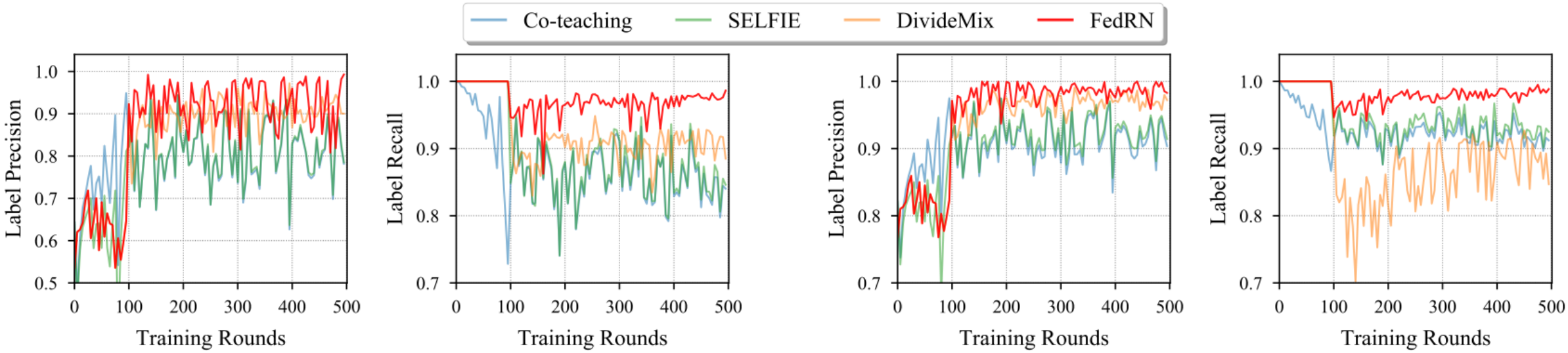
Table 4: Validation accuracy (%) on mini-WebVision. The values in parentheses are the improvement over FedAvg.

Experiments

α	w.o. Fine-tuning 0.0–0.4	w.o. Fine-tuning 0.0–0.8	w. Fine-tuning 0.0–0.4	w. Fine-tuning 0.0–0.8
0.0	**65.03**	**58.20**	67.11	61.81
0.2	63.77	57.24	67.45	62.69
0.4	62.45	53.89	67.03	62.94
0.6	58.74	51.36	**67.62**	62.94
0.8	56.60	43.61	67.19	**62.97**
1.0	56.06	44.87	66.73	61.61

Table 5: Effect of fine-tuning when trained on CIFAR-10 with symmetric noise and shard ($S = 2$) settings.

Experiments



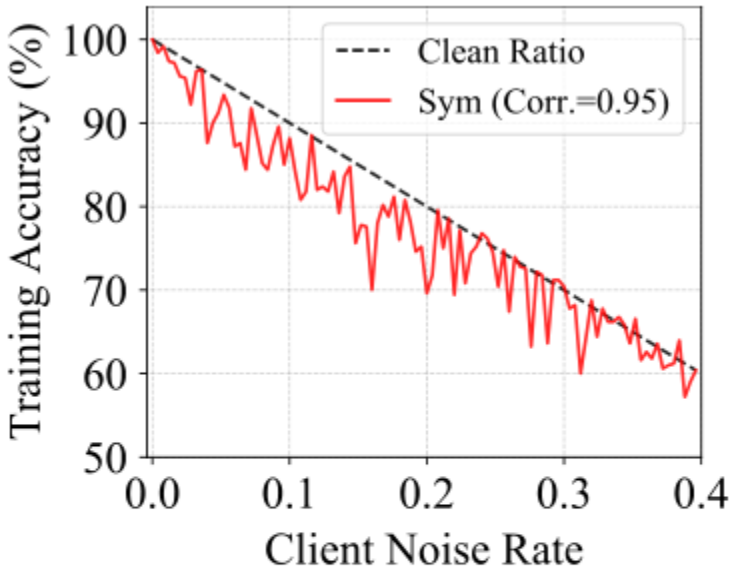
(a) Symmetric Noise of 0.0–0.8.

(b) Asymmetric Noise of 0.0–0.4.

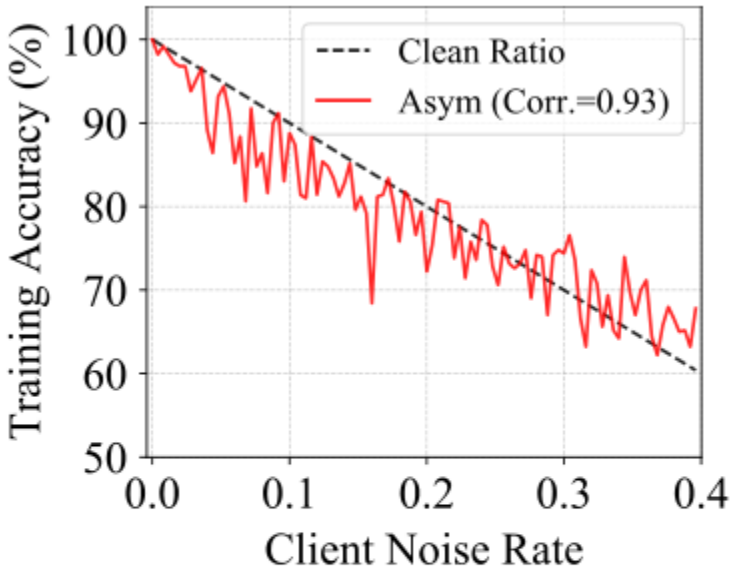
Figure 3: Label precision and label recall curves on CIFAR-10 with the shard ($S = 2$) setting.

$$\mathrm{LP} = \frac{|\{(x, \tilde{y}) \in \mathcal{S}_c : \tilde{y} = y^*\}|}{|\mathcal{S}_c|}, \quad \mathrm{LR} = \frac{|\{(x, \tilde{y}) \in \mathcal{S}_c : \tilde{y} = y^*\}|}{|\{(x, \tilde{y}) \in \mathcal{D}_c : \tilde{y} = y^*\}|},$$

Experiments



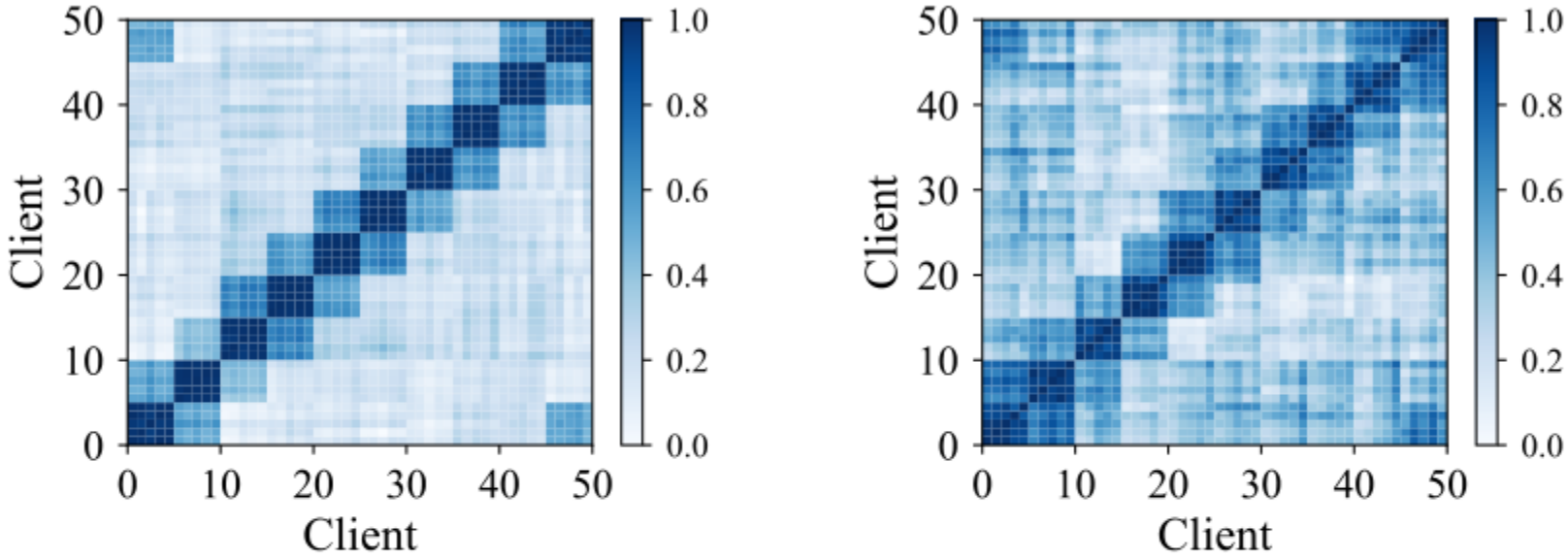
(a) Symmetric Noise.



(b) Asymmetric Noise.

Figure 4: Correlation between the training accuracy and noise rate on CIFAR-10 with shard ($S = 2$) of symmetric and asymmetric noises 0.0–0.4.

Experiments



(a) w.o Label Noise. (b) w. Label Noise.

Figure 5: Similarity matrix between the client's softmax output for a Gaussian random noise when trained on CIFAR-10 with shard ($S = 2$) without noise (a) and with symmetric noise 0.0–0.4 **(b).**

Experiments

	Shard ($S = 2$)		Dirichlet ($\beta = 0.5$)	
k	0.0–0.4	0.0–0.8	0.0–0.4	0.0–0.8
1	67.33	60.33	79.99	75.92
2	67.62	62.94	80.34	76.49
3	68.10	62.77	80.31	76.01
4	67.38	62.75	80.36	76.16
5	67.60	63.10	80.18	76.00

Table 6: Test accuracy (%) on CIFAR-10 using FedRN with different number of reliable neighbors.

	Shard ($S = 20$)			
	Symmetric		Asym	Mixed
k	0.0–0.4	0.0–0.8	0.0–0.4	0.0–0.4
1	47.30	42.27	47.62	46.62
2	47.46	43.07	47.52	47.17
3	47.65	42.33	48.18	47.01
4	47.66	42.48	47.55	47.72
5	48.10	42.58	47.43	47.40

Table 7: Test accuracy (%) of on CIFAR-100 using FedRN with different number of reliable neighbors.

Experiments

	Communication			Computation		
	Server→Client	Client→Server	**Total Cost**	Forward	Backward	**Total Cost**
FedAvg [24]	$\mathbf{M}$	$\mathbf{M}$	$2\mathbf{M}$	$e\mathbf{F}$	$e\mathbf{B}$	$e\mathbf{F}+e\mathbf{B}$
Co-Teaching [14]	$2\mathbf{M}$	$2\mathbf{M}$	$4\mathbf{M}$	$2e\mathbf{F}$	$2e\mathbf{B}$	$2e\mathbf{F}+2e\mathbf{B}$
Joint-optm [30]	$\mathbf{M}$	$\mathbf{M}$	$2\mathbf{M}$	$e\mathbf{F}$	$e\mathbf{B}$	$e\mathbf{F}+e\mathbf{B}$
SELFIE [27]	$\mathbf{M}$	$\mathbf{M}$	$2\mathbf{M}$	$e\mathbf{F}$	$e\mathbf{B}$	$2e\mathbf{F}+e\mathbf{B}$
DivideMix [20]	$2\mathbf{M}$	$2\mathbf{M}$	$4\mathbf{M}$	$(4m+2)e\mathbf{F}$	$2e\mathbf{B}$	$(4m+2)e\mathbf{F}+2e\mathbf{B}$
Robust FL [36]	$\mathbf{M}$	$\mathbf{M}$	$2\mathbf{M}$	$e\mathbf{F}$	$e\mathbf{B}$	$e\mathbf{F}+e\mathbf{B}$
FedRN	$(k+1)\mathbf{M}$	$\mathbf{M}$	$(k+2)\mathbf{M}$	$(e+2k+1)\mathbf{F}$	$(k+e)\mathbf{B}$	$(e+2k+1)\mathbf{F}+(k+e)\mathbf{B}$

Table 8: Analysis of the communication and computation costs in federated learning setting: M is the communication cost to send the model; F and B are the computational costs of forward and backward propagation, respectively; e is the number of local epochs for each communication round; k is the number of reliable neighbors.

Experiments

Type	Symmetric Noise					
	Shard ($S = 2$)		Shard ($S = 5$)		Dirichlet	
	0.0–0.4	0.0–0.8	0.0–0.4	0.0–0.8	0.0–0.4	0.0–0.8
k-random	66.82	60.11	76.37	71.56	79.79	75.11
k-reliable	**67.62**	**62.94**	**76.81**	**72.85**	**80.34**	**76.49**

Table 9: Performance comparison with k-random neighbors on CIFAR-10 when $k = 2$.

Experiments

	Shard ($S = 2$)		Dirichlet ($\beta = 0.5$)		
α	0.0–0.4	0.0–0.8	0.0–0.4	0.0–0.8	Mean
0.0	67.11	61.81	80.18	75.94	71.26 ± 8.33
0.2	67.45	62.69	80.35	75.95	71.61 ± 8.00
0.4	67.03	62.94	79.95	76.27	71.55 ± 7.90
0.6	67.62	62.94	80.34	76.49	**71.85 ± 7.97**
0.8	67.19	62.97	80.18	76.28	71.76 ± 7.80
1.0	66.73	61.61	80.00	76.07	71.10 ± 8.43

Table 10: Test accuracy (%) with different α values.

THANKS