Towards Federated Learning against Noisy Labels via Local Self-Regularization

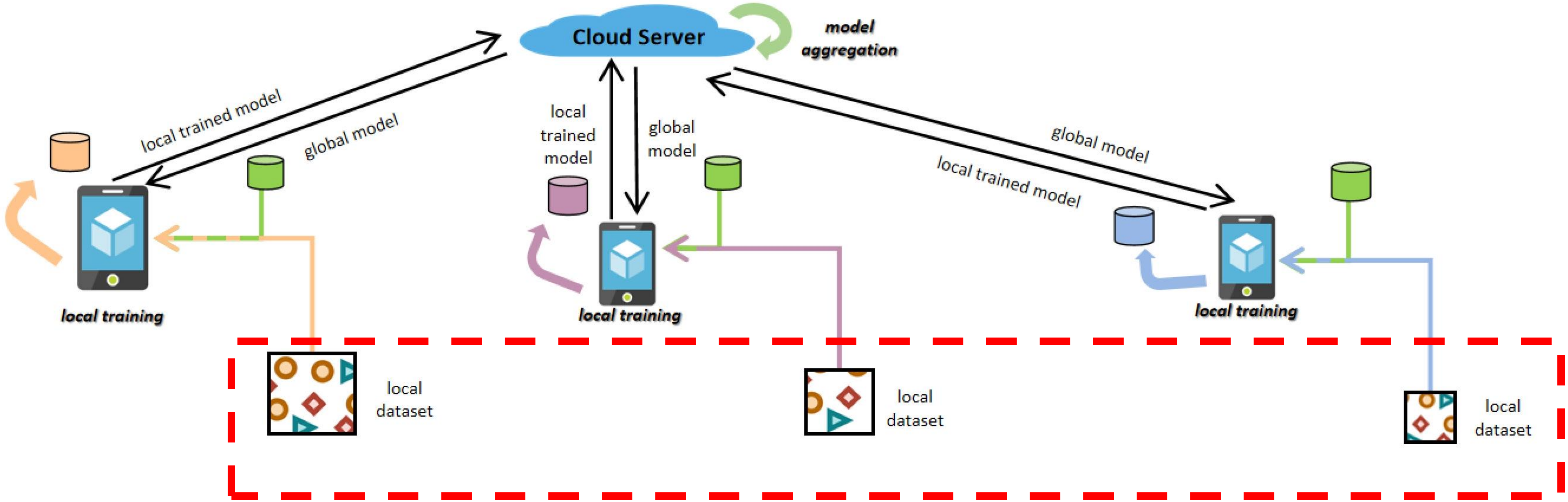
Xuefeng Jiang, Sheng Sun, Yuwei Wang, Min Liu

Institute of Computing Technology, Chinese Academy of Sciences

Oct, 2022

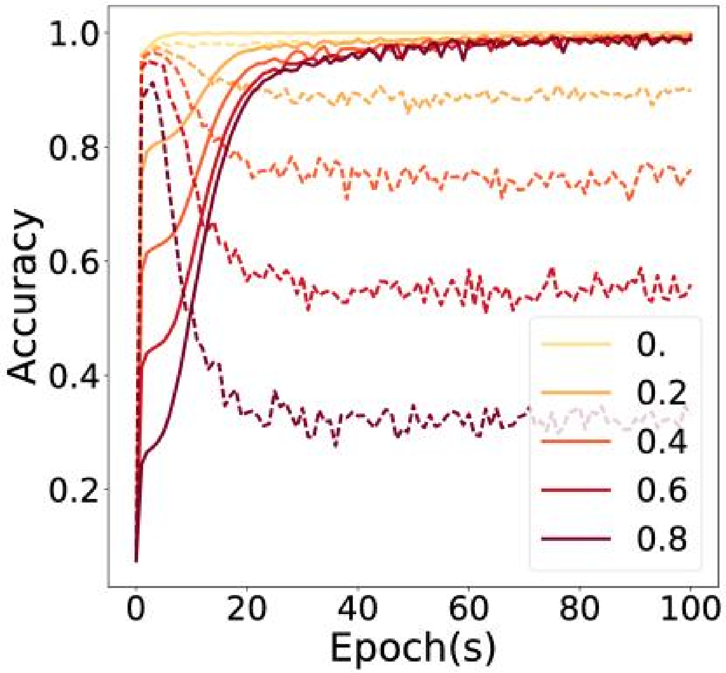
Introduction

From **Centralized Learning** to **Federated Learning**

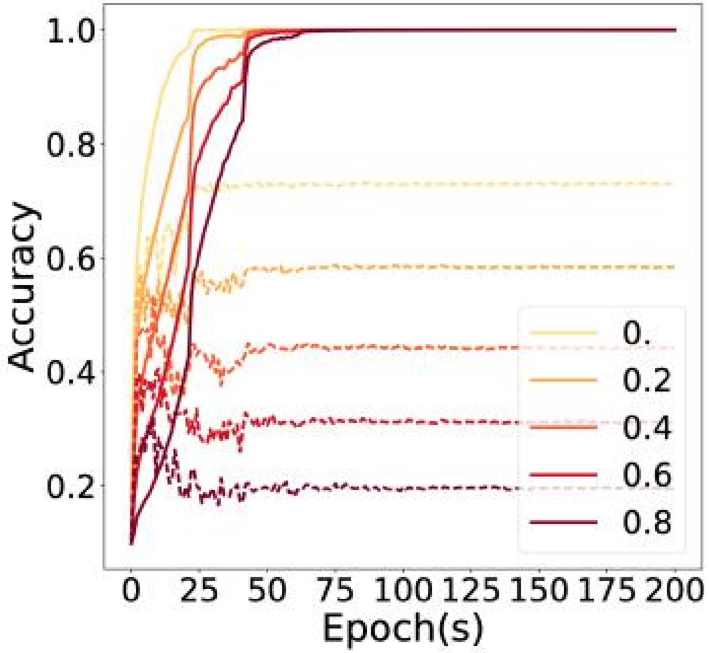


Note that training on data with high-quality labels is a strong assumption in FL!

Noisy labels in **centralized learning (CL)**



(a)MNIST in various noise levels



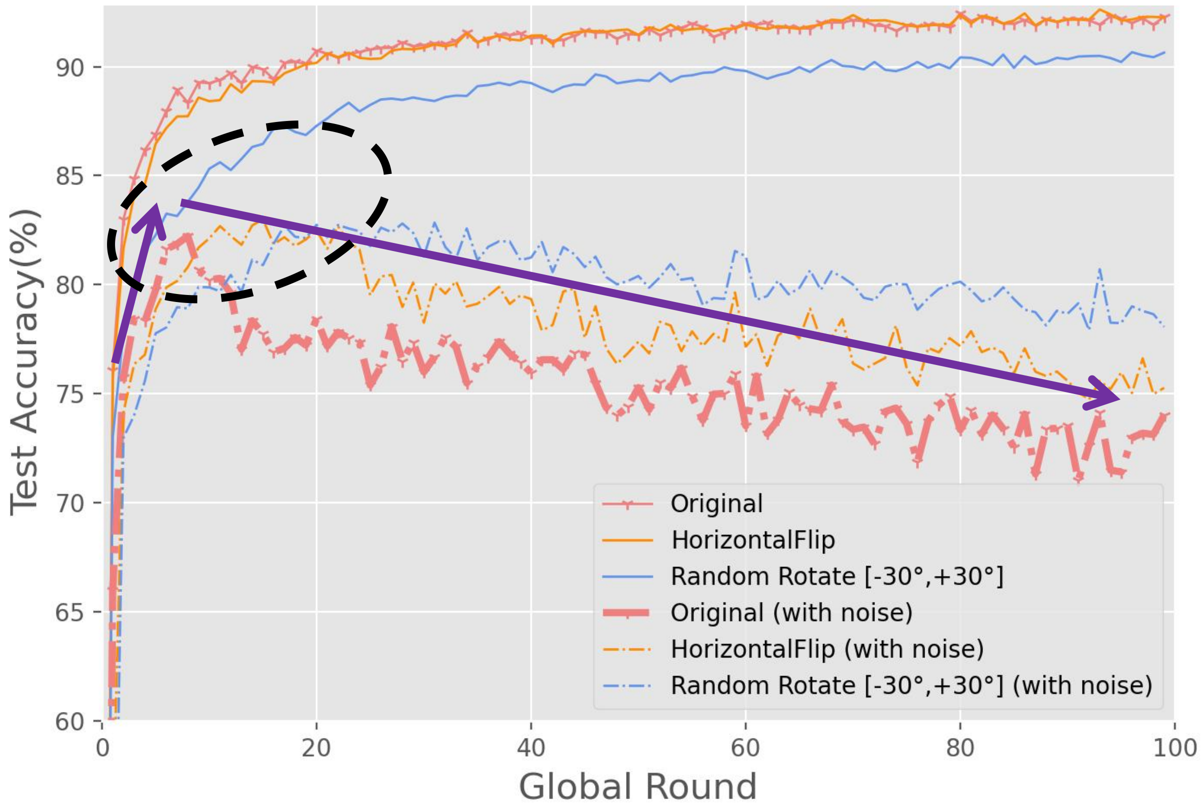
(b)CIFAR-10 in various noise levels

solid line:
training acc
dashed line:
testing acc

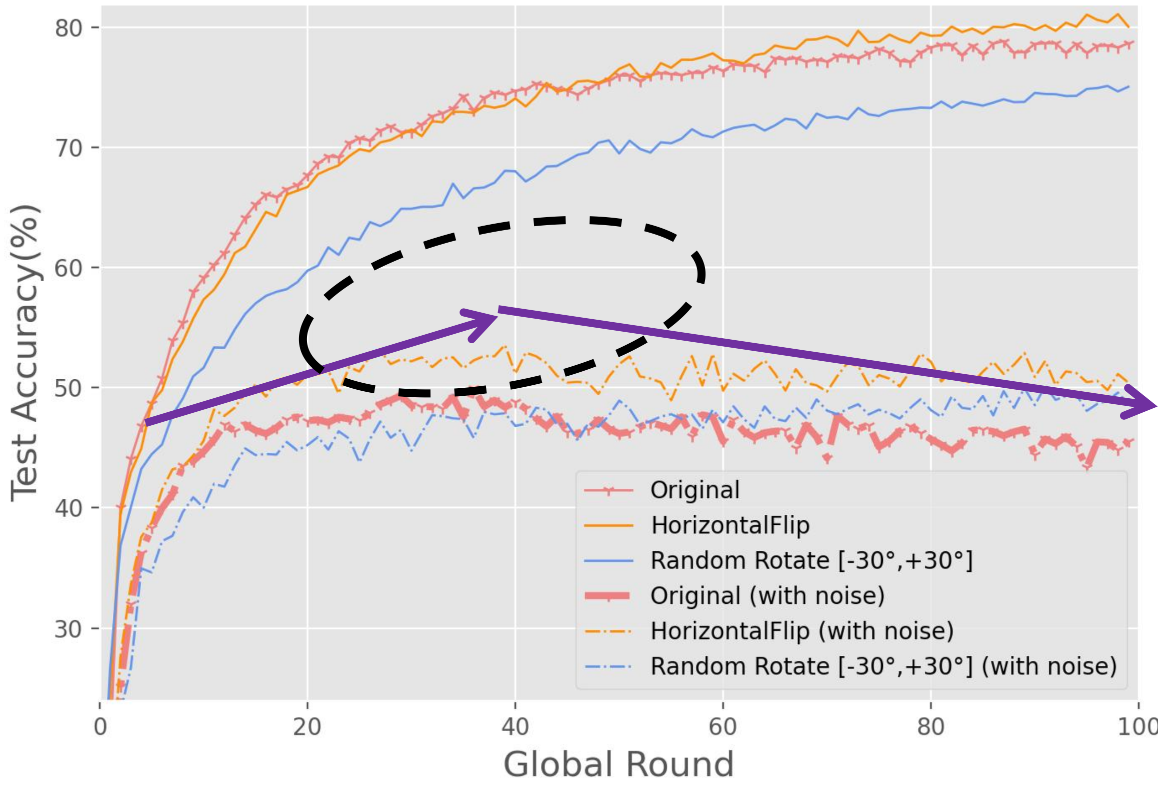
Correctly-labeled data fits before noisy-labeled data (Deep Network Memorization Effect).

[1]Arpit D, Jastrzębski S, Ballas N, et al. A closer look at memorization in deep networks[C]//International conference on machine learning. PMLR, 2017: 233-242.

Noisy labels in federated learning (FL)



(a)Fashion-MNIST

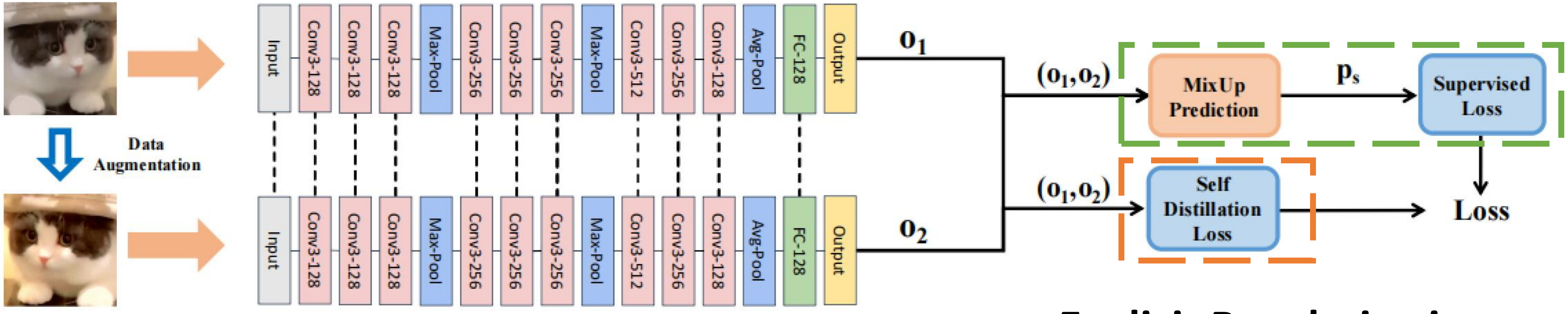


(b)CIFAR-10

The deep network memorization effect can still exist in FL!
We can't ignore noisy labels even in FL!

Our Approach

How to locally regularize the on-device training?



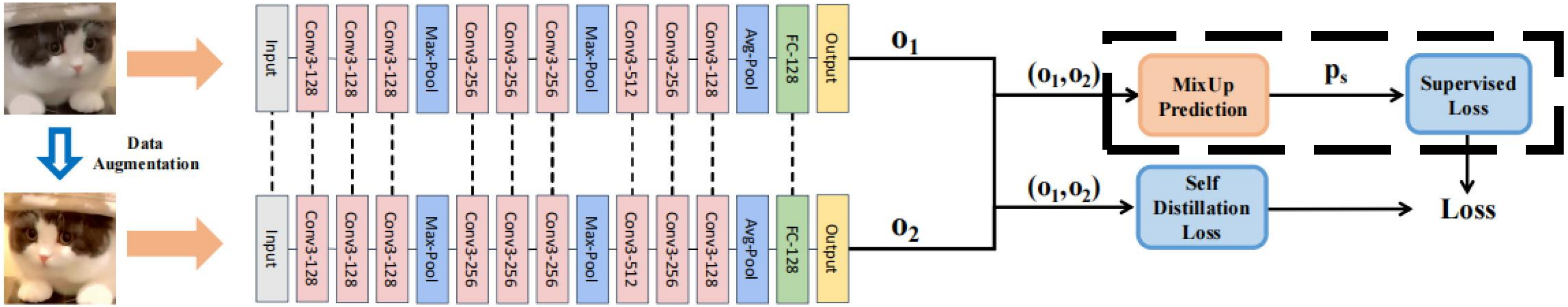
Implicit Regularization:

to enhance the model discrimination confidence via **sharpening operation and entropy regularization**

Explicit Regularization:

to narrow the discrepancy of the model output of oringinal and augmented instances via **self-distillation**

Implicit Regularization: to enhance model prediction confidence



MixUp Prediction

1. Mix Prediction Up: $p = \lambda * p_1 + (1-\lambda) * p_2.$

2. Sharpening Operation: $p_{s,i} = \text{Sharpen}(p, T)_i := p_i^{\frac{1}{T}} / \sum_{j=1}^{M} p_j^{\frac{1}{T}}$

3. Use p_S to compute classification loss: Loss_{cls}

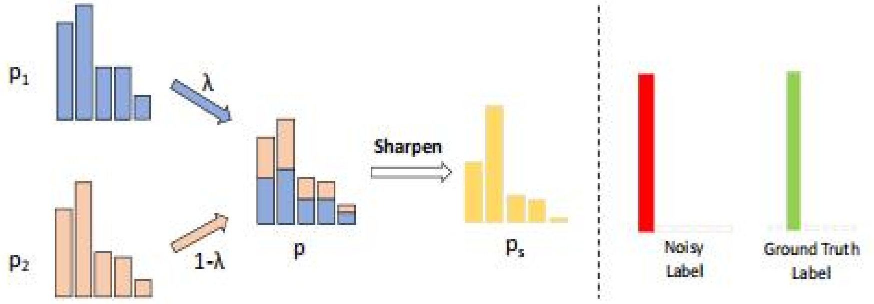
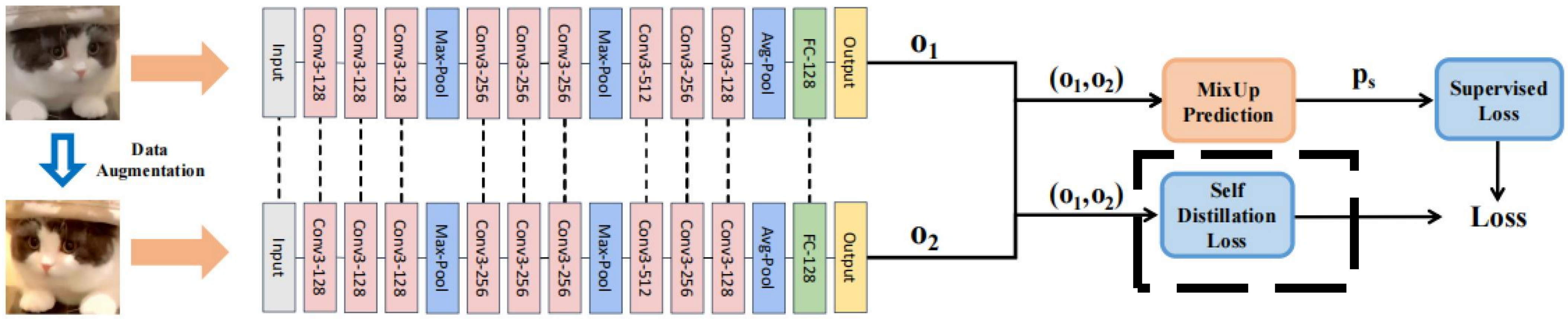


Figure 4: An intuitive understanding for MixUp prediction. In this scenario, the model gives correct prediction but the given label y is wrong (namely y is a noisy label). By conducting sharpening operation, $CE(p_s, y)$ is larger than $CE(p_1, y)$, $CE(p_2, y)$ and $CE(p, y)$, which implicitly adds the difficulty for overfitting the noisy label y.

Entropy Regularization:

- append optimizing term: L_e

Explicit Regularization: to enhance instance-level consistency



Instance-level Self-distillation

- Modify the model output with temperature T_d , with JS divergence or L1 loss as metric

$$q_{1,i}, q_{2,i} = \frac{\exp\left(o_{1,i}/T_d\right)}{\sum_j \exp\left(o_{1,j}/T_d\right)}, \frac{\exp\left(o_{2,i}/T_d\right)}{\sum_j \exp\left(o_{2,j}/T_d\right)}; \Rightarrow Loss_{reg} = JS(q_1, q_2) = \frac{1}{2}(KL(q_1 \| U) + KL(q_2 \| U))$$

Final Local on-device Optimizing Objective:

$$Loss = Loss_{cls} + \gamma * Loss_{reg} + \lambda_e * L_e.$$

Experiments

Noise transition matrix

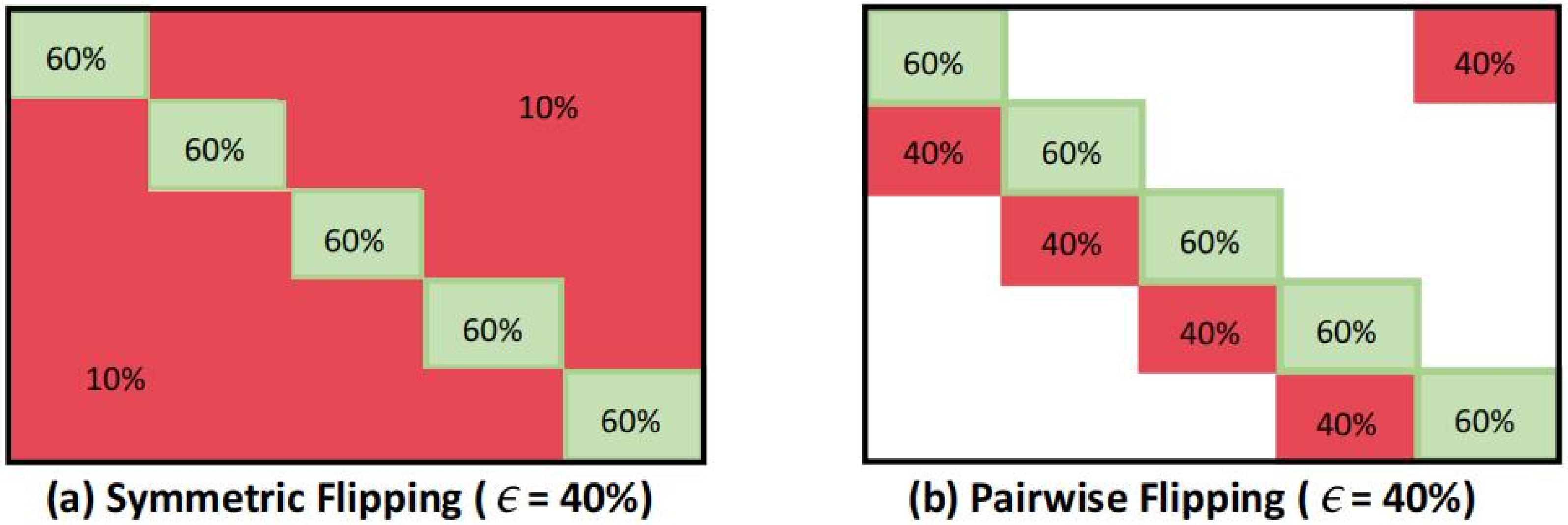


Figure 5: Transition matrices of different noise types (using 5 classes as an example). The green and red grids represent the percentage of samples that are correctly labeled to the ground truth class, and the percentage of samples that are incorrectly labeled to other classes, respectively.

Basic experimental settings:

Dataset and base model

- benchmark datasets:
  MNIST、Fashion-MNIST、CIFAR-10 （9-Layer CNN）
- real-world dataset:
  Clothing-1M (pretrained Resnet-50)

Clothing-1M[1]

Main baseline

- existing off-the-shelf methods:
  FedAvg， Symmetric CE， Co-teaching
- pioneering method to tackle noisy labels in FL:
  Robust Federated Learning

Data augmentation (mild)

- Random rotation within 30 degrees for MNIST and Fashion-MNIST
- Color distortion for CIFAR-10 and Clothing-1M

[1]Xiao T, Xia T, Yang Y, et al. Learning from massive noisy labeled data for image classification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 2691-2699.

Main experimental results

Table 3: Test accuracy on benchmark datasets with various noise levels. LSR is the proposed Local Self-Regularization method.

Dataset	Method	Test Accuracy (%)								
	Noise Type	Symmetric					Pairwise			Avg.
	Noise Ratio	0.30	0.40	0.50	0.60	0.70	0.20	0.30	0.40	
MNIST	FedAvg [42]	91.34	83.53	73.60	57.86	42.12	94.27	85.52	70.35	74.82
	Symmetric CE [58]	99.10	98.91	98.54	97.77	95.10	99.10	98.63	94.13	97.66
	Co-teaching [15]	98.80	98.11	97.38	95.94	93.84	98.83	97.97	94.43	96.91
	Robust Federated Learning [66]	99.07	98.92	98.84	98.44	**98.40**	99.08	99.01	**98.98**	98.84
	FedAvg + LSR	**99.23**	**99.14**	**98.91**	**98.63**	98.01	**99.36**	**99.22**	98.49	**98.87**
Fashion-MNIST	FedAvg [42]	80.02	72.75	62.29	49.22	35.87	86.44	77.38	63.05	65.88
	Symmetric CE [58]	88.86	85.96	80.32	69.87	49.34	89.99	84.51	68.44	77.16
	Co-teaching [15]	89.22	88.11	86.82	84.43	81.03	90.37	87.77	83.03	86.35
	Robust Federated Learning [66]	88.26	87.41	85.54	84.04	79.22	89.67	89.12	88.17	86.43
	FedAvg + LSR	**90.42**	**89.73**	**88.67**	**87.00**	**82.72**	**90.84**	**90.27**	**88.34**	**88.50**
CIFAR-10	FedAvg [42]	53.78	46.06	36.93	28.45	19.80	66.93	58.47	48.04	44.81
	Symmetric CE [58]	64.80	56.40	47.45	34.11	23.97	67.56	59.48	45.91	49.96
	Co-teaching [15]	70.23	66.84	62.54	**56.25**	**45.28**	71.44	66.41	57.21	62.03
	Robust Federated Learning [66]	66.29	60.38	54.05	43.18	32.38	69.01	61.18	49.71	54.52
	FedAvg + LSR	**72.10**	**68.53**	**64.27**	55.10	40.61	**73.79**	**70.66**	**59.40**	**63.06**

Training detail: SGD optimizer with momentum = 0.9 & wd = 0.0001, batch size = 64, learning rate = 0.15

Ablation study & Combination with existing works

Noise Type	Noise Ratio	Ours w/o MixUp Pred.	Ours (fix $\lambda = 1$)	Ours
Symmetric	0.3	84.37(-6.05)	90.16(-0.26)	90.42
	0.4	78.33(-11.40)	89.78(+0.05)	89.73
	0.5	69.24(-19.43)	88.69(+0.02)	88.67
	0.6	55.93(-31.07)	86.97(-0.03)	87.00
	0.7	41.53(-41.19)	81.82(-0.90)	82.72
Pairwise	0.2	87.39(-3.45)	90.82(-0.02)	90.84
	0.3	79.53(-10.74)	90.23(-0.04)	90.27
	0.4	64.88(-23.46)	87.33(-1.01)	88.34

W and w/o Implicit Reg.

Table 7: Test accuracy (%) of existing works combined with our method on Fashion-MNIST dataset.

Method	Symmetric (ϵ=0.7)	Pairwise (ϵ=0.4)
Co-teaching	81.03	83.03
Co-teaching + LSR	**85.76 (+ 4.73)**	**89.35 (+ 6.32)**
Symmetric CE	49.34	68.44
Symmetric CE + LSR	**72.81 (+ 23.47)**	**77.33 (+ 8.89)**

Combine with existing works

Noise Type	Noise Ratio	Self Distillation Loss Term: w/o	JS Div	L1 Loss	L2 Loss	Cosine Similarity
Symmetric	0.3	89.86	**90.42**	89.86	89.90	89.47
	0.4	89.19	**89.73**	89.62	89.51	88.83
	0.5	88.31	**88.67**	88.49	88.31	87.89
	0.6	86.83	87.00	**87.07**	86.72	85.33
	0.7	78.70	82.72	**83.22**	80.61	80.87
Pairwise	0.2	90.51	**90.84**	90.37	90.56	88.59
	0.3	89.61	**90.27**	89.71	89.86	86.42
	0.4	83.10	**88.34**	88.00	85.48	82.31

W various metrics and w/o Explicit Reg.

#	Method	Setting	Test Accuracy(%)
1	Cross Entropy	C. L.	68.94
2	Symmetric CE	C. L.	71.02
3	Forward	C. L.	69.84
4	Generalized Cross Entropy	C. L.	69.75
5	FedAvg	F. L.	68.56
6	FedAvg + LSR	F. L.	69.30
7	Symmetric CE	F. L.	69.63
8	Symmetric CE + LSR	F. L.	70.46
9	Robust Federated Learning	F. L.	70.32

Experiments on Clothing-1M

Thanks for listening :)

Github link