CrossSplit: Mitigating Label Noise Memorization through Data Splitting

Jihye Kim[1], Aristide Baratin[2], Yan Zhang[2], and Simon Lacoste-Julien[2,3,4]

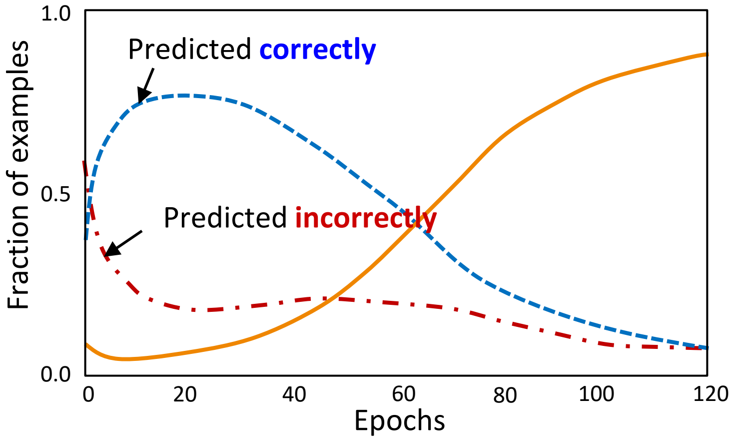
[1]Samsung Advanced Institute of Technology (SAIT), Korea
[2]SAIT AI Lab, Montreal
[3]University of Montreal
[4]Canada CIFAR AI Chair

DNNs suffer from noisy labels: Memorization

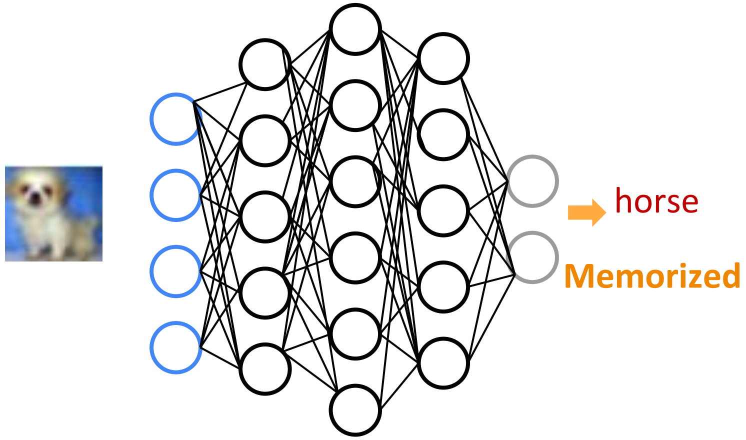
- Deep neural networks easily overfit noisy labels[1].
  - Large learning capacities and memorization power of DNNs.
  - It leads to poor generalization performance.

deer (cat) ship (frog) dog (horse)

Wrong (Noisy) Labels --> Source: [1]



Memorized (the prediction equals the wrong label)



→ **Poor generalization performance**

→ An important issue in the filed is therefore to adapt the training process to **improve robustness under label noise**.

[1] Liu et al., "Early-Learning Regularization Prevents Memorization of Noisy Labels", NeurIPS 2020

Existing Approaches and Our Goal

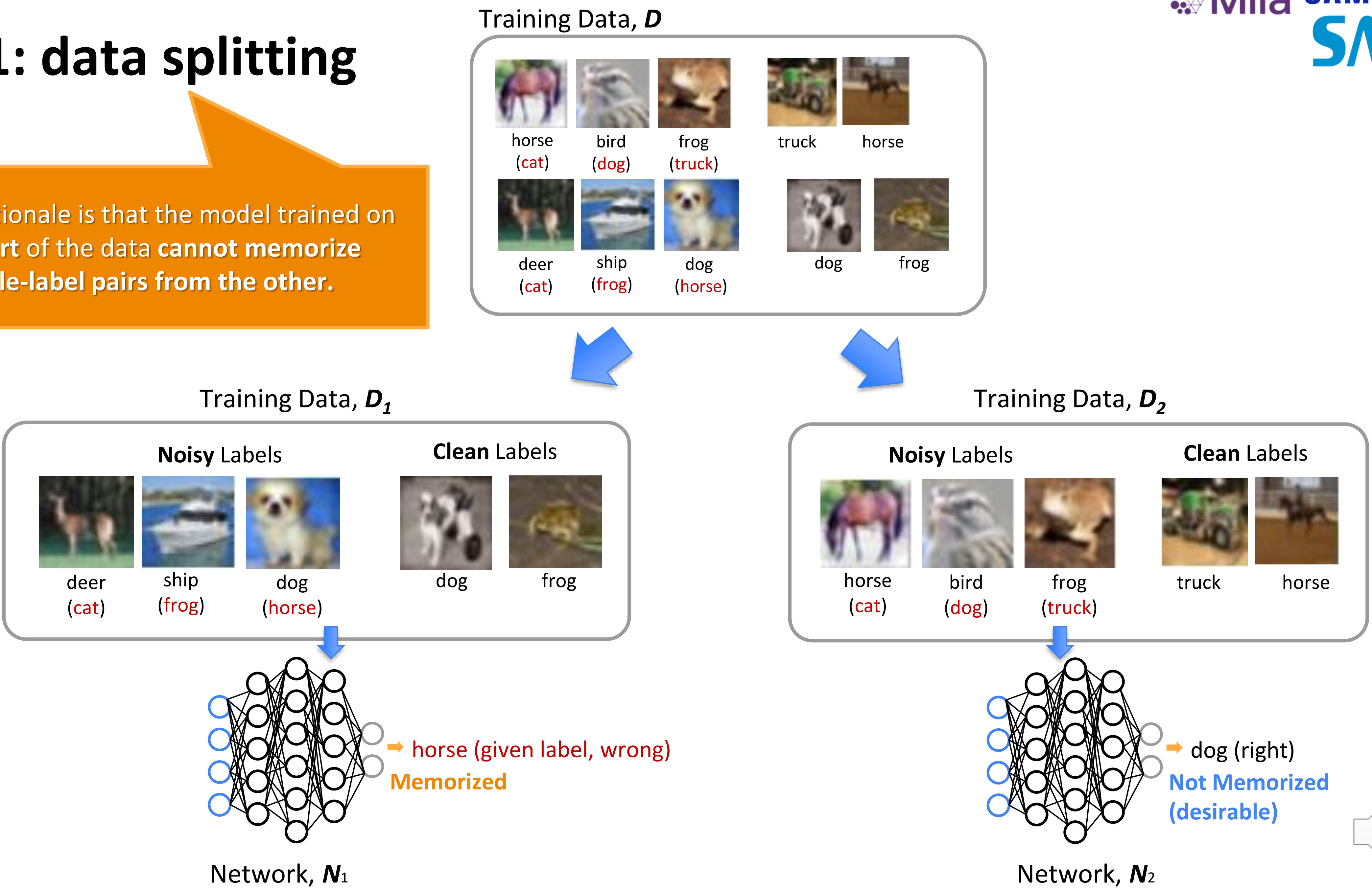
- Existing Learning with Noisy Labels (LNL) Methods[2,3,4]
  1. Label correction [2,3,4]
     → Define soft target labels in terms of their own prediction, which may become unreliable as training progresses and memorization occurs.
  2. Sampling selection [2,3]
     → Making an accurate distinction between mislabeled and inherently difficult examples is challenging.

- Our Goal
  - To propose a novel robust training scheme that addresses some of drawbacks of existing LNL methods.
    - **Data splitting:** The idea is to bypass the sample selection process by using a random splitting of the data into **two disjoint parts**, and **train a separate network on each of these splits.**
    - **Cross-split label correction:** We propose to **correct the labels by using the peer prediction.**

[2] Karim et al., "UNICON: Combating Label Noise through Uniform Selection and Contrastive Learning," CVPR 2022.
[3] Li et al., "Dividemix: Learning with Noisy Labels as Semi-supervised Learning," ICLR 2020.
[4] Lu et al. "SELC: Self-ensemble Label Correction Improves Learning with Noisy Labels," IJCAI 2022.

Part 1: data splitting

The rationale is that the model trained on **one part** of the data **cannot memorize example-label pairs from the other.**



Part 2: cross-split label correction

- Soft label, $\boldsymbol{s}_i$

= Convex combination of $\boldsymbol{y}_i$ and the cross-split probability (softmax) vector, $\hat{\boldsymbol{y}}_{\text{peer},i} = N_2(\boldsymbol{x}_i)$:

$$\boldsymbol{s}_i = \beta_i \hat{\boldsymbol{y}}_{\text{peer},i} + (1-\beta_i)\boldsymbol{y}_i$$

Peer Prediction
(Cross-split Probability)

Assigned Label

$$\beta_i = \gamma(\text{JSD}_{\text{norm}}(\hat{\boldsymbol{y}}_{\text{peer},i}, \boldsymbol{y}_i) - 0.5) + 0.5$$

$(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \boldsymbol{D}_1$
$\boldsymbol{x}_i$: an input image
$\boldsymbol{y}_i$: the one-hot vector associated to its (possibly noisy) class label.
$\boldsymbol{s}_i$: the soft label
$\hat{\boldsymbol{y}}_{\text{peer},i} = N_2(\boldsymbol{x}_i)$
JSD_{norm} : a normalized version of the Jensen-Shannon Divergence.

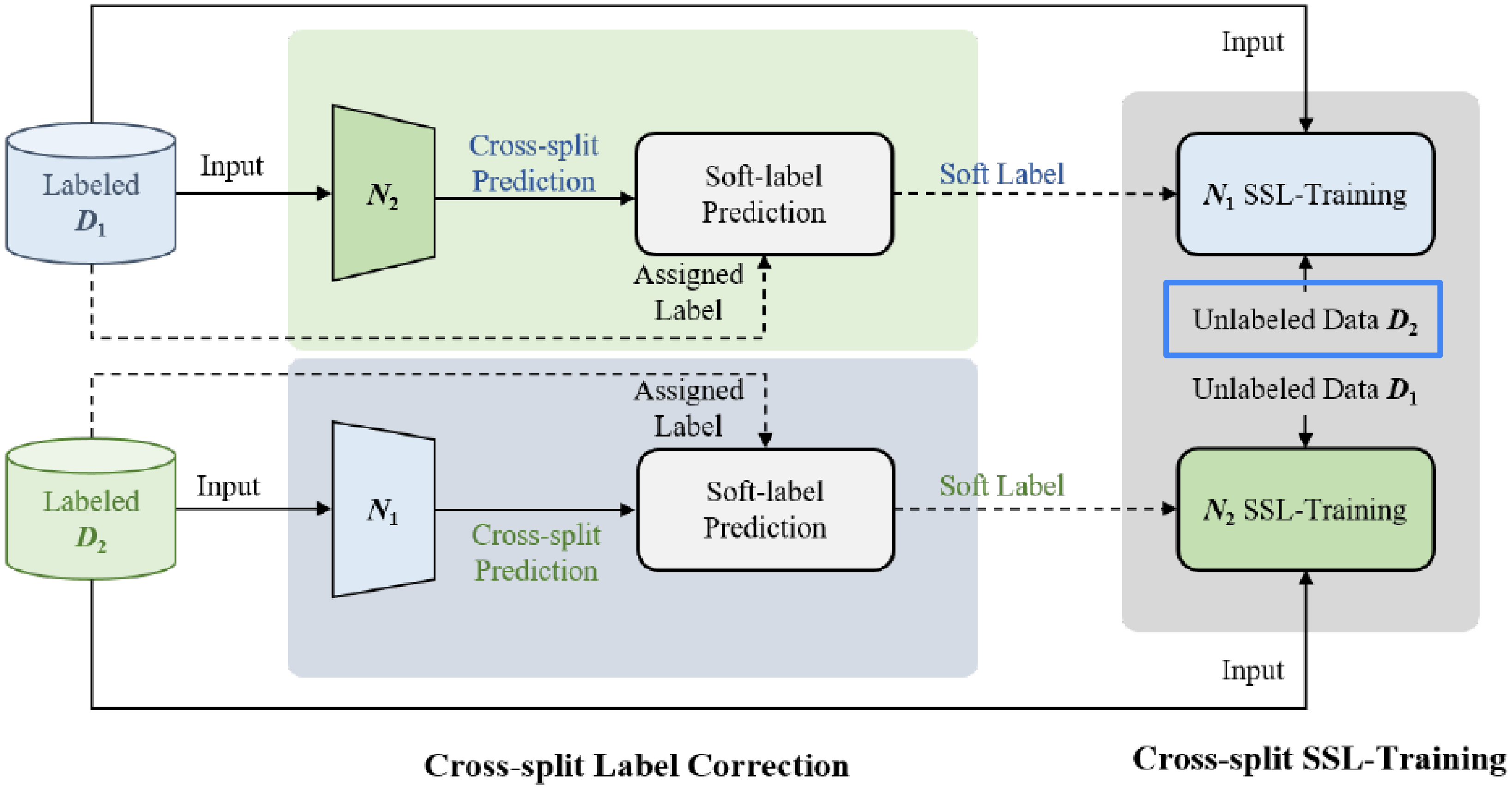
- Class-balancing coefficient normalization
  - Importance of class-wise difficulty consideration [UNICON]
    - If there is no consideration, model is biased towards selecting samples from easy classes to be clean, while rejecting clean samples from harder classes as noisy.
  - We normalize the JSD the standard JSD, within each class, it ranges from 0 to 1.

$$\text{JSD}_{\text{norm}}(\hat{\mathbf{y}}_{\text{peer},i}, \mathbf{y}_i) := \frac{\text{JSD}(\hat{\mathbf{y}}_{\text{peer},i}, \mathbf{y}_i) - \text{JSD}^{\min}_{\mathbf{y}_i}}{\text{JSD}^{\max}_{\mathbf{y}_i} - \text{JSD}^{\min}_{\mathbf{y}_i}}$$

Class-wise statistics

Part 3: cross-split SSL training

- A network trained on one part of the data also uses the unlabeled inputs of the other part.



Mila SAMSUNG SAIT

Results

Table 1. Test accuracy (%) comparison on CIFAR-10 (left) and CIFAR-100 (right) without label noise and with symmetric and asymmetric label noise. Our model achieves state-of-the-art performance on almost every dataset-noise combination. The best scores are **boldfaced**, and the second best ones are underlined. The baseline results are imported from (Karim et al., 2022; Li et al., 2020; 2022). For CrossSplit, mean and standard deviation of best accuracy are calculated over 3 repetitions of the experiments. The results are sorted according to their performance in the case of a 20% symmetric noise ratio.

Noise type	CIFAR-10 -	CIFAR-10 Symmetric				CIFAR-10 Asymmetric			CIFAR-100 -	CIFAR-100 Symmetric				CIFAR-100 Asymmetric		
Method/Noise ratio	0%	20%	50%	80%	90%	10%	30%	40%	0%	20%	50%	80%	90%	10%	30%	40%
CE	95.4	86.8	79.4	62.9	42.7	88.8	81.7	76.1	77.3	62.0	46.7	19.9	10.1	68.1	53.3	44.5
Bootstrapping (Reed et al., 2015)	-	86.8	79.8	63.3	42.9	-	-	-	-	62.1	46.6	19.9	10.2	-	-	-
JPL (Kim et al., 2021)	-	93.5	90.2	35.7	23.4	94.2	92.5	90.7	-	70.9	67.7	17.8	12.8	72.0	68.1	59.5
M-Correction (Arazo et al., 2019)	-	94.0	92.0	86.8	69.1	89.6	92.2	91.2	-	73.9	66.1	48.2	24.3	67.1	58.6	47.4
MOIT (Ortego et al., 2021)	-	94.1	91.1	75.8	70.1	94.2	94.1	93.2	-	75.9	70.1	51.4	24.5	77.4	75.1	74.0
SELC (Lu & He, 2022)	-	95.0	-	78.6	-	-	-	92.9	-	76.4	-	37.2	-	-	-	73.6
Sel-CL (Li et al., 2022)	-	95.5	93.9	89.2	81.9	95.6	95.2	93.4	-	76.5	72.4	59.6	48.8	78.7	76.4	74.2
MixUp (Zhang et al., 2018)	95.8	95.6	87.1	71.6	52.2	93.3	83.3	77.7	78.9	67.8	57.3	30.8	14.6	72.4	57.6	48.1
ELR (Liu et al., 2020)	-	95.8	94.8	93.3	78.7	95.4	94.7	93.0	-	77.6	73.6	60.8	33.4	77.3	74.6	73.2
UNICON (Karim et al., 2022)	-	96.0	95.6	93.9	90.8	95.3	94.8	94.1	-	78.9	**77.6**	63.9	44.8	78.2	75.6	74.8
DivideMix (Li et al., 2020)	-	96.1	94.6	93.2	76.0	93.8	92.5	91.7	-	77.3	74.6	60.2	31.5	71.6	69.5	55.1
CrossSplit (PRN-18)	**97.0** ±0.16	**96.9** ±0.05	**96.3** ±0.05	**95.4** ±0.64	**91.3** ±0.79	**96.9** ±0.04	**96.4** ±0.16	**96.0** ±0.12	**81.7** ±0.25	**79.9** ±0.19	75.7 ±0.18	**64.6** ±1.43	**52.4** ±1.78	**80.7** ±0.05	**78.5** ±0.19	**76.8** ±0.66
CrossSplit (PRN-34)	**97.3** ±0.16	**97.1** ±0.16	**96.5** ±0.24	**95.2** ±0.59	85.3 ±3.61	**97.2** ±0.09	**96.6** ±0.11	**96.1** ±0.08	**83.0** ±0.15	**81.4** ±0.38	77.2 ±0.25	**67.0** ±0.49	**52.6** ±3.43	**82.6** ±0.15	**80.5** ±0.27	**79.1** ±0.40

Table 2. Tiny-ImageNet

Noise type / Noise ratio	Symmetric 20%		50%	
Method	Best	Avg.	Best	Avg.
CE	35.8	35.6	19.8	19.6
coupling (Malach & Shalev-Shwartz, 2017)	37.0	36.3	22.8	22.6
MentorNet (Jiang et al., 2018)	45.7	45.5	35.8	35.5
Co-teaching+ (Yu et al., 2019)	48.2	47.7	41.8	41.2
M-Correction (Arazo et al., 2019)	57.2	56.6	51.6	51.3
NCT (Sarfraz et al., 2021)	58.0	57.2	47.8	47.4
UNICON (Karim et al., 2022)	**59.2**	58.4	**52.7**	**52.4**
CrossSplit (ours)	59.1	**58.8**	52.4	52.0

Table 3. Mini-WebVision

Method	Best	Last
Decoupling (Malach & Shalev-Shwartz, 2017)	62.54	-
MentorNet (Jiang et al., 2018)	63.00	-
Co-teaching (Han et al., 2018)	63.58	-
Iterative-CV (Chen et al., 2019)	65.24	-
ELR (Liu et al., 2020)	73.00	71.88
SELC (Lu & He, 2022)	74.38	-
MixUp (Zhang et al., 2018)	74.96	73.76
DivideMix (Li et al., 2020)	76.08	74.64
UNICON(Karim et al., 2022)	77.60	-
CrossSplit (ours)	**78.48**	**78.07**

Memorization behavior

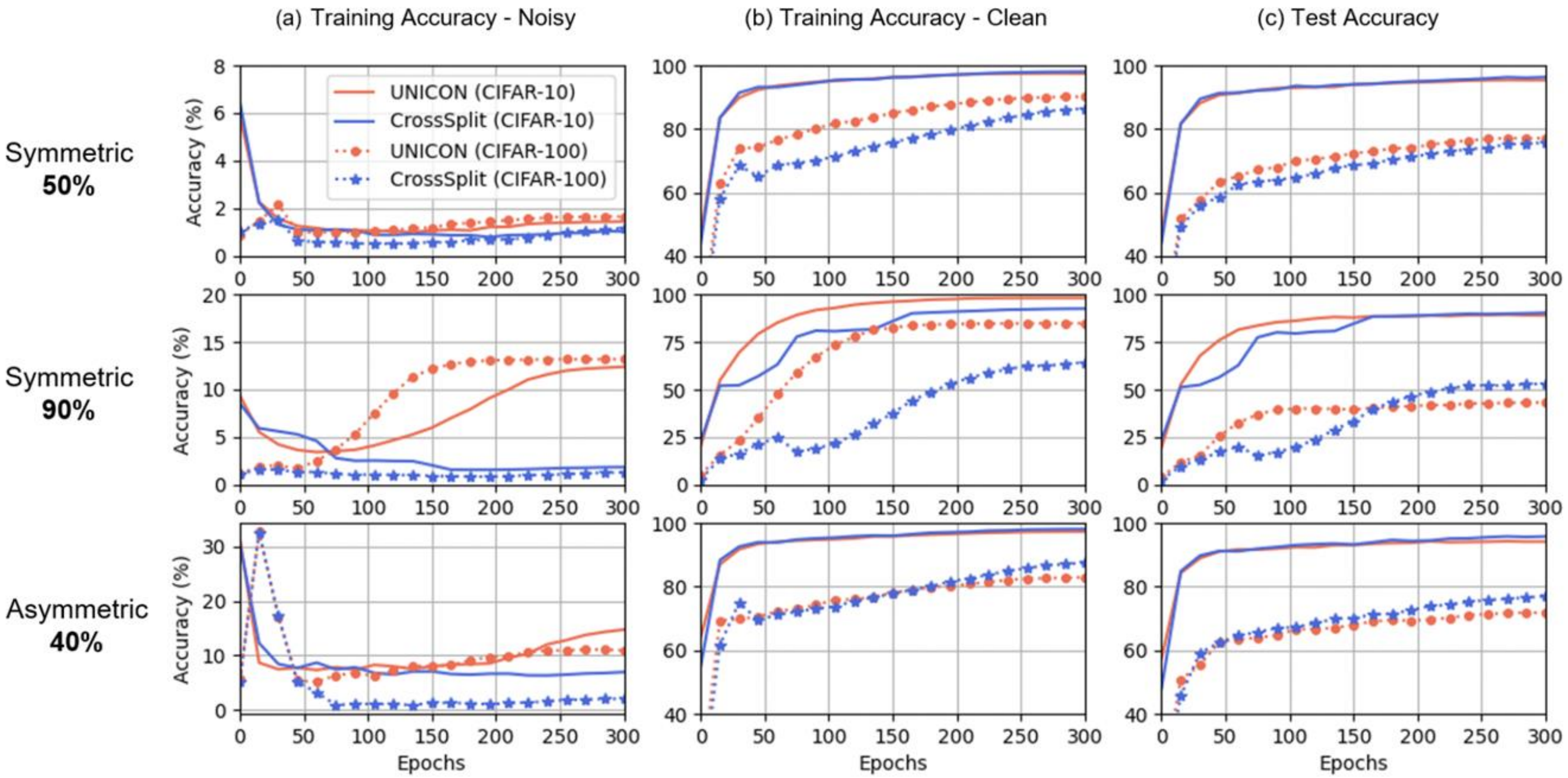


Figure 3. Memorization of clean and noisy training samples of CIFAR-10 and CIFAR-100 for different types of noise and noise ratio. Compared to UNICON [12], *CrossSplit* induces less memorization of the noisy labels. It is interesting to note that in the case of very a high noise ratio (90%), *CrossSplit* has a lower training accuracy also on clean data, yet yields a higher test performance.

Ablation study

Noise type	Symmetric				Asymmetric			
Noise ratio	50%		90%		10%		40%	
Method	Best	Last	Best	Last	Best	Last	Best	Last
CrossSplit	96.34±0.05	96.23±0.07	**91.25**±0.79	**91.02**±0.77	96.85±0.04	96.74±0.07	96.01±0.12	95.88±0.13
CrossSplit w/o data splitting	96.10±0.04	95.96±0.00	90.30±0.13	89.93±0.24	96.76±0.05	96.63±0.06	92.16±0.09	86.24±0.37
CrossSplit w/o class-balancing normalization	**96.73**±0.13	**96.61**±0.07	75.54±2.82	74.88±2.50	**97.33**±0.02	**97.20**±0.02	**96.22**±0.07	**96.04**±0.12
CrossSplit w/o cross-split label correction	96.12±0.05	95.99±0.03	90.83±0.25	90.08±0.40	**97.33**±0.08	97.15±0.09	96.12±0.14	95.95±0.10

Table 6. **Ablation study on CIFAR-10**: Test accuracy (%) of different setting on CIFAR-10 with varying noise rates (50% - 90% for Sym. and 10% - 40% for Asym.). Mean and standard deviation of best and average of last 10 epochs are calculated over 3 repetitions of the experiments. The best results are highlighted in **bold** and scores that differ from them by more than 5% are marked in red.

Noise type	Symmetric				Asymmetric			
Noise ratio	50%		90%		10%		40%	
Method	Best	Last	Best	Last	Best	Last	Best	Last
CrossSplit	75.72±0.18	75.50±0.18	**52.40**±1.78	**52.05**±1.94	80.71±0.05	80.50±0.06	**76.78**±0.66	**76.56**±0.55
CrossSplit w/o data splitting	73.63±0.18	73.36±0.14	14.19±1.30	13.28±2.21	78.97±0.07	78.77±0.43	72.12±0.43	71.83±0.42
CrossSplit w/o class-balancing normalization	**77.67**±0.03	**77.17**±0.17	33.37±0.52	18.53±0.19	**82.86**±0.14	**82.57**±0.18	71.59±0.28	60.35±0.37
CrossSplit w/o cross-split label correction	70.20±0.16	65.74±0.10	31.77±0.32	15.93±0.21	82.38±0.16	82.10±0.23	69.61±0.65	59.67±0.11

Table 7. **Ablation study on CIFAR-100**: Test accuracy (%) of different settings on CIFAR-100 with varying noise rates (50% - 90% for Sym. and 10% - 40% for Asym.). Mean and standard deviation of best and average of last 10 epochs are calculated over 3 repetitions of the experiments. The best results are highlighted in **bold** and scores that differ from them by more than 5% are marked in red.

Thanks!

CrossSplit: Mitigating Label Noise Memorization through Data Splitting

Please check our paper for more details.

Wed 26 11 a.m. HST – 12: 30 p.m. HST
(Exhibit Hall 1 #210)