SELFIE: Refurbishing Unclean Samples for Robust Deep Learning

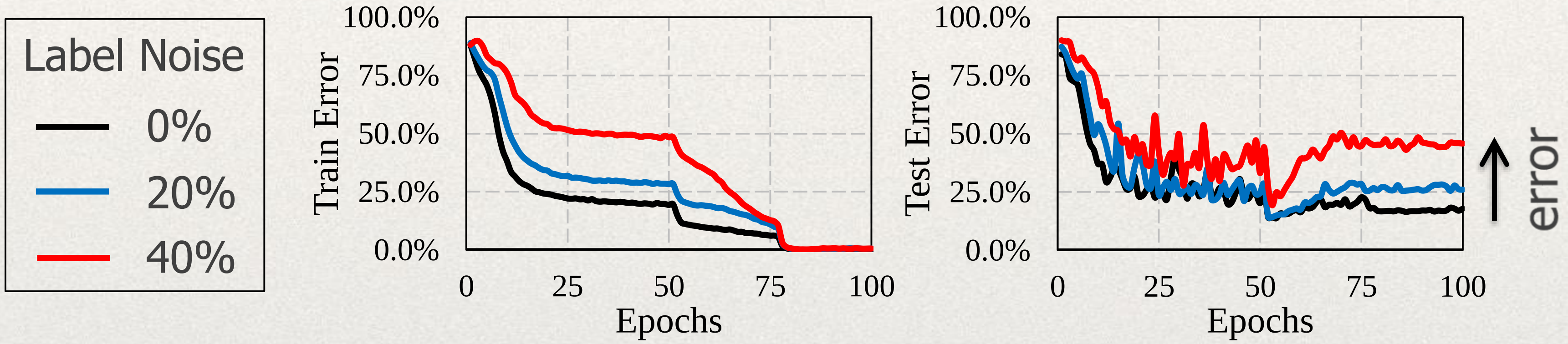
Hwanjun Song[†], Minseok Kim[†], Jae-Gil Lee[†*]

[†] Graduate School of Knowledge Service Engineering, KAIST
[*] Corresponding Author

Noisy Label Problem

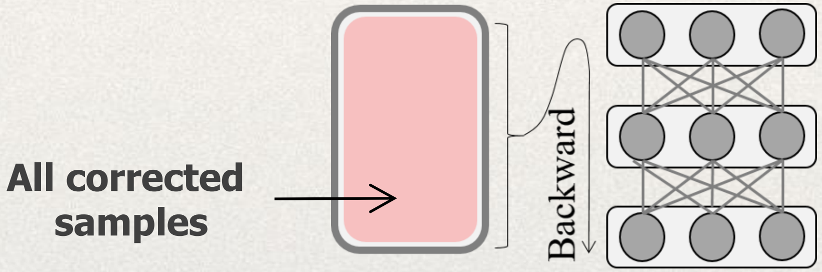
- Standard Supervised Learning Setting
  - Assume: training data $\{(x_i, y_i)\}_{i=1}^{N}$, y_i: **True label**
  - In practical setting, $y_i \to \tilde{y}_i$, $\tilde{y}_i$: **Noisy label**
    - High cost and time consuming
    - Expert knowledge
    - Unattainable at scale

    } Difficulties of label annotation
- Learning with Noisy Label
  - Suffer from poor generalization on test data (VGG-19 on CIFAR-10)

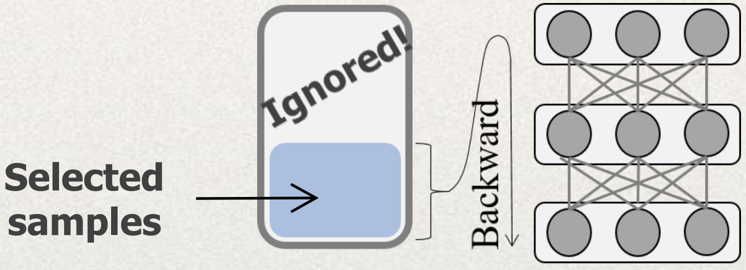


Existing Work: Two Directions

- Loss Correction
  - Modify the loss $\mathcal{L}$ of **all** samples before backward step
  - Suffer from **accumulated noise** by the false correction
    → Fail to handle heavily noisy data
- Sample Selection (Recent direction)
  - Select low-loss (easy) samples as **clean** samples $\mathcal{C}$ for SGD
  - Use only **partial exploration** of the entire training data
    → Ignore useful hard samples classified as unclean



(a) Loss correction

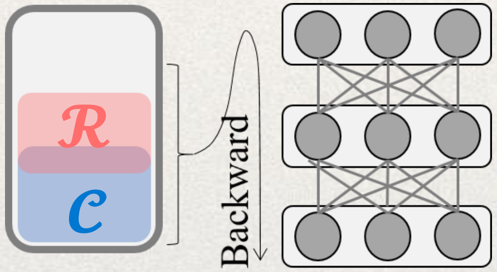


(b) Sample selection

Proposed Method

- **SELFIE** (**SEL**ectively re**F**urb**I**sh uncl**E**an samples)
  - Hybrid of loss correction and sample selection
  - Introduce **refurbishable samples** $\mathcal{R}$
    - The samples can be "corrected with high precision"
  - Modified update equation on mini-batch $\{(x_i, \tilde{y}_i)\}_{i=1}^{b}$
    - Correct the losses of **samples in** $\mathcal{R}$
    - Combine them with the losses of **samples in** $\mathcal{C}$
    - Exclude the samples not in $\mathcal{R} \cup \mathcal{C}$

$$\theta_{t+1} = \theta_t - \alpha\nabla \frac{1}{|\mathcal{R} \cup \mathcal{C}|}\left(\underbrace{\sum_{x\in\mathcal{R}} \mathcal{L}(x, y^{refurb})}_{\text{Corrected losses}} + \underbrace{\sum_{x\in\mathcal{C}\cap\mathcal{R}^{-1}} \mathcal{L}(x, \tilde{y})}_{\text{Selected clean losses}}\right)$$



Construction of $\mathcal{C}$ and $\mathcal{R}$

- Clean Samples $\mathcal{C}$ from $\mathcal{M}$ (mini-batch)
  - Adopt loss-based separation (Han et al., 2018)
  - $\mathcal{C} \leftarrow (100 - noise\ rate)\%$ of low-loss samples in $\mathcal{M}$
- Refurbishable Samples $\mathcal{R}$ from $\mathcal{M}$
  - $\mathcal{R} \leftarrow$ the samples with **consistent** label predictions
  - Replace its label into the **most frequently** predicted label $\tilde{y}_i \rightarrow y_i^{refurb}$

Consistent label predictions

x  … $\underbrace{\textbf{dog}\quad \text{cat}\quad \textbf{dog}\quad \textbf{dog}\quad \textbf{dog}\quad \textbf{dog}\quad \textbf{dog}}$

$\tilde{y}$ ~~cat~~

dog (refurbished label)

Evaluation: Noise Type

- Synthetic Noise: pair and symmetric
  - Injected two widely used noises
- Realistic Noise
  - Built **ANIMAL-10N** dataset with real-world noise
    - Crawled 5 pairs of **confusing animals**
      E.g., {(cat, lynx), (jaguar, cheetah),…}
    - Educated 15 participants for one hour
    - Asked the participants to annotate the label
  - Summary

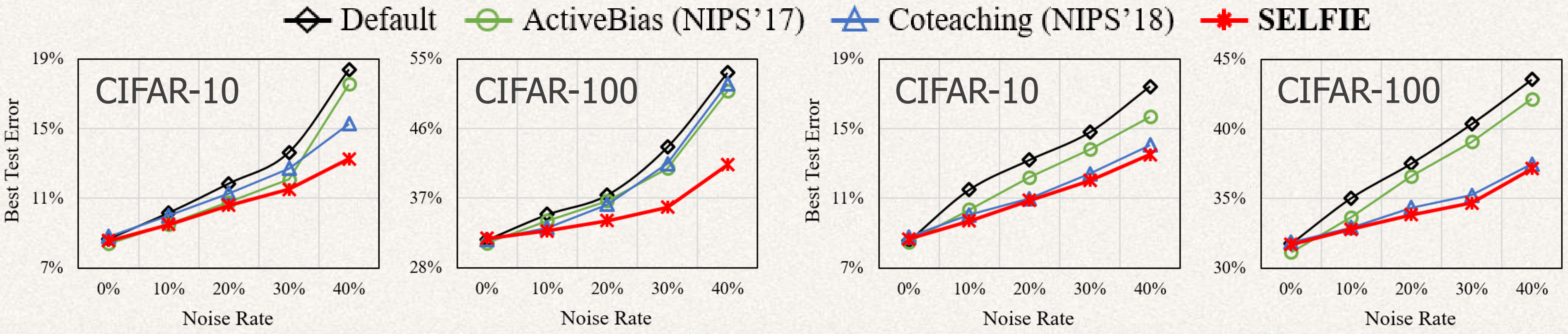
Scan me

# Training	50,000	Resolution	64x64 (RGB)
# Test	5,000	**Noise Rate**	**8% (estimated)**
# Classes	10	Data Created	April 2019

https://dm.kaist.ac.kr/datasets/animal-10n

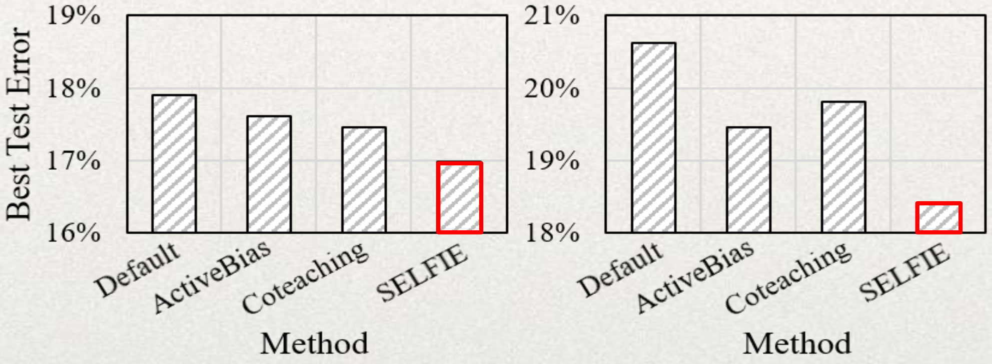
Evaluation: Performance

- Results with two synthetic noises (CIFAR-10, CIFAR-100)



(a) Varying pair noises (b) Varying symmetric noises

- Results with realistic noise (ANIMAL-10N)



(a) DenseNet (L=25, k=12) (b) VGG-19

Thank you

Further Details or Questions

Poster Session: Pacific Ballroom #157

https://dm.kaist.ac.kr/datasets/animal-10n

Scan me