南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Towards Robust and Generalized Parameter-Efficient Fine-Tuning for Noisy Label Learning

Yeachan Kim[1*], Junho Kim[1*], SangKeun Lee[1,2]

[1]Department of Artificial Intelligence, Korea University, Seoul, South Korea

[2]Department of Computer Science and Engineering, Korea University, Seoul, South Korea

{yeachan,monocrat,yalphy}@korea.ac.kr

ACL 2024

Background

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

- The ever-growing size of pre-trained language models (PLMs) has presented significant challenges in adapting these models to desired tasks. In response to this practical limitation, parameter-efficient fine-tuning (PEFT) has emerged as a promising strategy for real-world environments. Instead of fine-tuning all weights, PEFT optimizes only a minimal set of parameters, thereby drastically cutting down the computation and storage costs. Such efficiency has led PEFT methods to become the preferred standard approaches for applying PLMs in real-world contexts, such as federated learning and continual learning.

- While PEFT enables the efficient optimization of PLMs in real-world settings, datasets in such environments often contain noisy labels, which adversely affects the generalization capabilities of PLMs. Given such distinct characteristics of the practical environments, PEFT methods are inevitably exposed to noisy labels during the optimization phase. Despite this significant challenge, there is a lack of prior research on the general adaptability of PEFT methods to noisy label learning (NLL) scenarios.

Motivation

- Our results reveal that PEFT struggles in memorizing noisy labels due to its inherently limited capacity, which interestingly provides robustness to noisy labels. However, we also find that such limited capacity simultaneously makes PEFT more susceptible to interference of noisy labels, which impedes learning ability for clean samples, potentially leading to sub-optimal performance. This characteristic markedly contrasts with the behaviors in full fine-tuning, presenting the necessity of PEFT that steers its limited learning capacity towards clean samples.

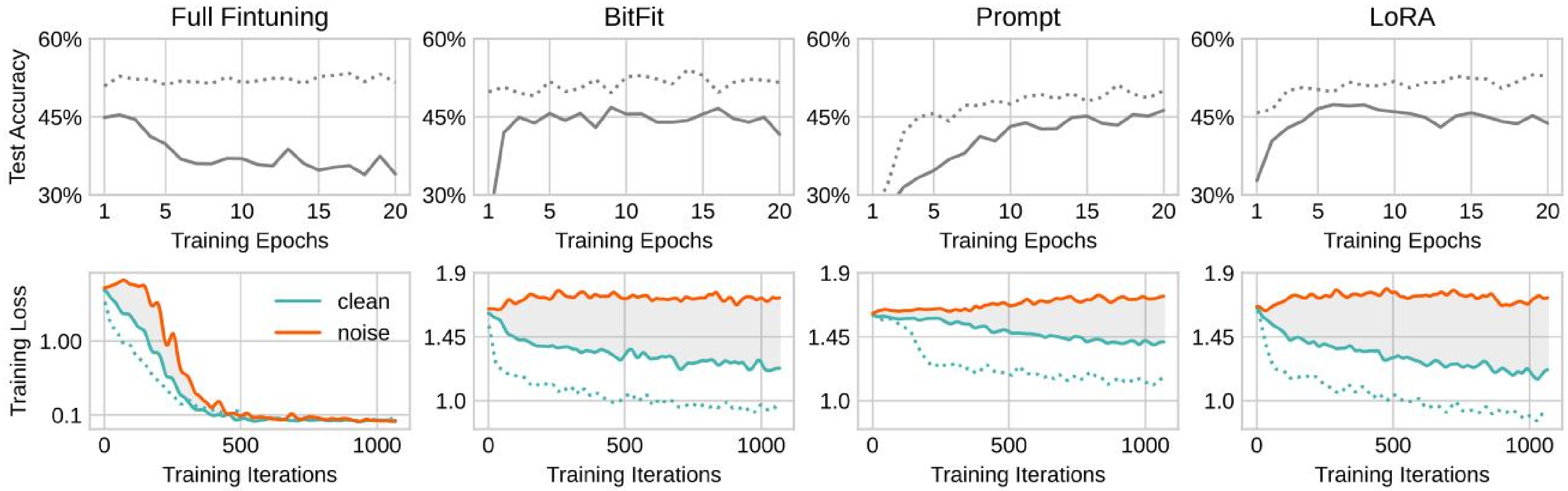


Figure 1: Comparison between PEFT methods and full fine-tuning on SST-5 with symmetric noise (60%). Dashed lines represent the training accuracy and loss of clean samples on uncorrupted datasets (i.e. only clean samples).

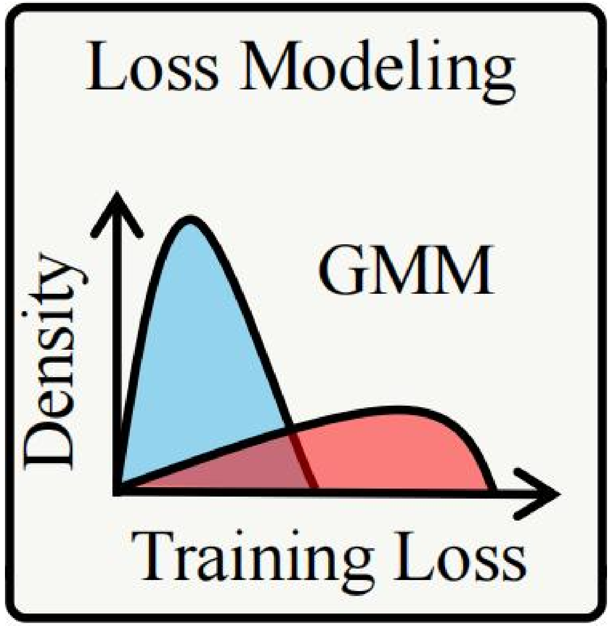
Methods

- Estimating clean probability for routing. $p(g|\ell)$

Deep networks prefer to learn clean samples first before fitting noisy ones.
Noisy samples tend to have a higher loss than clean samples in the early training stage.

This enables to distinguish potentially clean samples from the datasets based on loss deviation. Taking advantage of such phenomena, we adopt the widely-used Gaussian Mixture Model (GMM) in noise label learning, in which the probability of samples being clean is estimated by the per-sample loss.

Train model for the k epochs warm-up to measure the loss of samples, and then estimate the clean probability for each training sample on every subsequent epoch.



Methods

- Sampling routing decision

$$r \sim \textsc{Bernoulli}(\gamma p)$$

r is an independent Bernoulli random variable;
the coefficient $\gamma \in [0, 1]$ limits the range of clean probability, setting its upper bound at γ.

- Activating PEFT based on the decision

$$h^{(l+1)} = \begin{cases} \text{Trans}^{(l)}(h^{(l)}, \delta^{(l)} + \theta^{(l)}), & \text{if } r^{(l)} = 1 \\ \text{Trans}^{(l)}(h^{(l)}, \theta^{(l)}), & \text{if } r^{(l)} = 0 \end{cases}$$

Methods

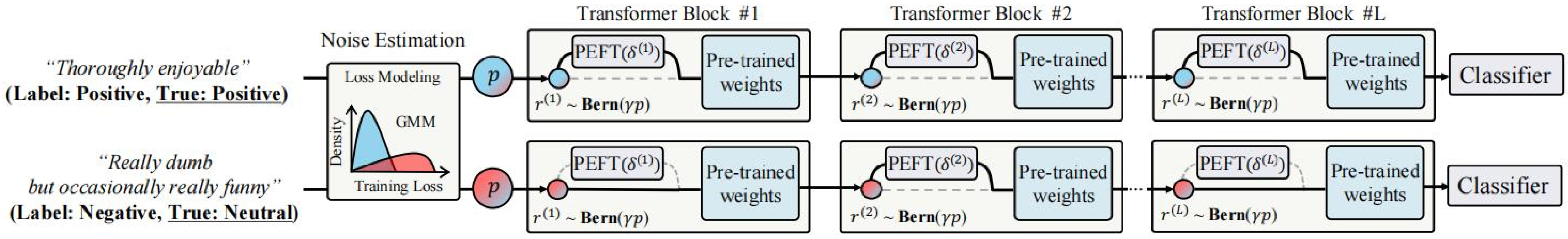


Figure 2: Overview of the **Clea**n **R**outing. CleaR first estimate the probability of each sample being clean based on the training losses. Based on the estimated probability, CleaR adaptively activates PEFT modules by favoring the potentially clean samples.

- Estimating clean probability for routing.
- Sampling routing decision
- Activating PEFT based on the decision

Methods

- Consistency Regularization for CleaR

$$f_{ens}(x, \bar{\delta}_r + \theta) = \frac{1}{N}\sum_{k=1}^{N} f(x, \bar{\delta}_{r,k} + \theta)$$

$$\min_{\delta} \mathcal{L}(x) = \mathcal{L}(f(x, \boldsymbol{\delta} + \boldsymbol{\theta}), y)$$

$$\min_{\delta_r} \mathcal{L}(x) = \mathcal{L}_{CE}(f(x, \delta_r + \theta), y) + \lambda \mathcal{L}_{CE}(f(x, \delta_r + \theta), f_{ens}(x, \bar{\delta}_r + \theta))$$

Methods

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

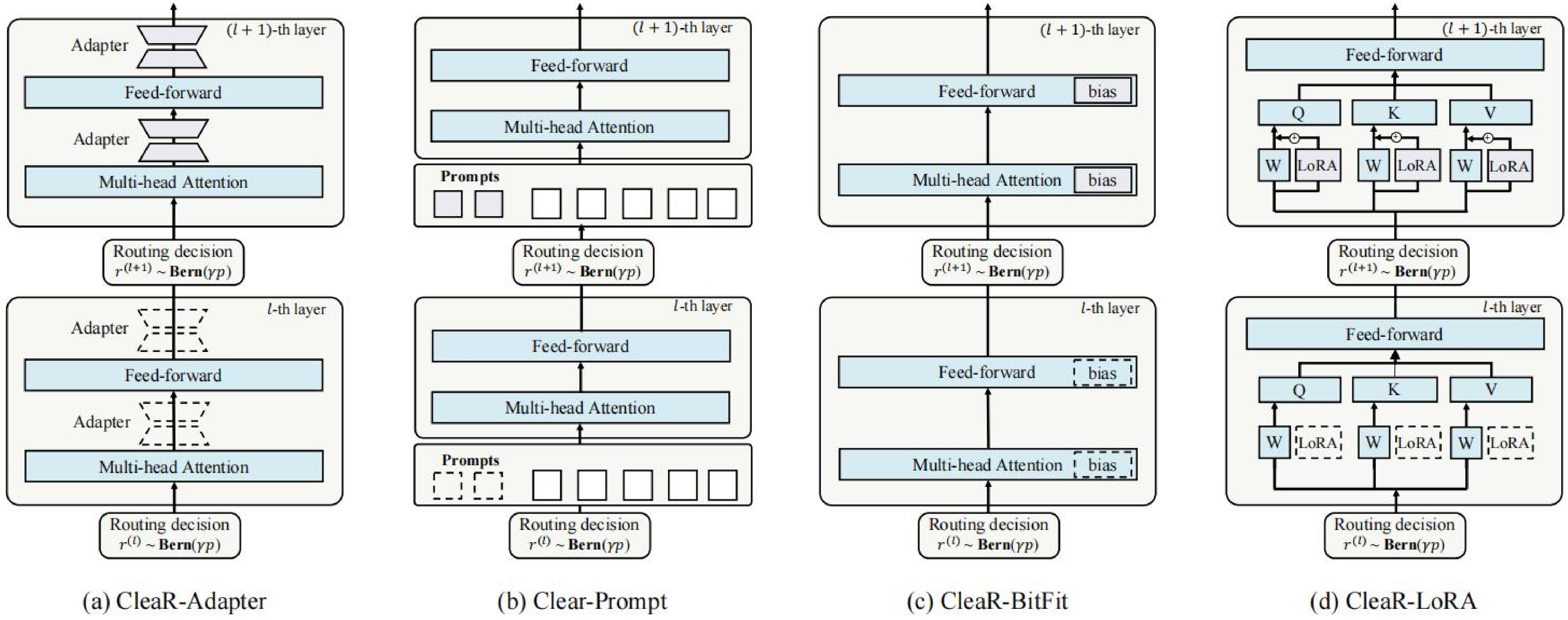


Figure 6: Detailed illustration of the CleaR adaptation to PEFT methods (e.g., Adapter, Prompt Tuning, BitFit, LoRA). Dashed lines indicate the unused modules, except for the $\text{CleaR}_{\text{BitFit}}$ that uses fixed pre-trained biases.

Experiments

Table 1: Evaluation results of Peak accuracy and Average accuracy on SST-5 test set under different levels of label noise. The best and second best results are highlighted in **boldface** and underlined, respectively.

Methods	Clean	Symmetric 20% Peak.	Symmetric 20% Avg.	Symmetric 40% Peak.	Symmetric 40% Avg.	Symmetric 60% Peak.	Symmetric 60% Avg.	Asymmetric 10% Peak.	Asymmetric 10% Avg.	Asymmetric 20% Peak.	Asymmetric 20% Avg.	Asymmetric 40% Peak.	Asymmetric 40% Avg.
Full Fine-tuning	<u>53.4</u>	51.3	47.0	50.6	42.9	47.9	35.5	**52.5**	49.1	50.8	46.5	46.1	37.4
PEFT methods													
Adapter (2019)	53.3	51.9	48.1	50.5	45.8	47.2	38.1	<u>52.2</u>	51.0	50.9	47.0	48.1	38.0
BitFit (2022)	53.0	51.7	51.0	50.8	48.1	48.1	43.5	52.1	50.5	<u>52.1</u>	49.2	<u>48.9</u>	42.1
Prompt (2022b)	52.7	51.1	48.6	50.7	49.1	47.7	45.7	51.7	50.8	49.4	48.2	46.1	41.7
LoRA (2022)	**53.6**	<u>52.0</u>	49.5	50.2	47.5	48.2	46.1	51.9	51.1	50.5	47.4	47.2	41.8
PEFT methods with CleaR (ours)													
$\text{CleaR}_{\text{Adapter}}$	<u>53.4</u>	**52.4**	**51.8**	<u>51.5</u>	<u>50.4</u>	<u>50.4</u>	<u>49.7</u>	**52.5**	50.8	51.4	47.4	48.1	44.6
$\text{CleaR}_{\text{BitFit}}$	53.1	51.9	<u>51.1</u>	**51.6**	**51.2**	**51.4**	**51.1**	52.0	**51.4**	**52.3**	<u>50.4</u>	**49.2**	**48.3**
$\text{CleaR}_{\text{Prompt}}$	52.6	51.0	50.5	51.4	49.5	49.4	47.2	52.1	<u>51.2</u>	52.0	**51.4**	47.8	<u>46.5</u>
$\text{CleaR}_{\text{LoRA}}$	53.3	51.4	50.1	51.2	49.0	50.0	48.9	52.0	51.1	51.0	<u>50.4</u>	47.6	43.2

Experiments

Table 2: Evaluation results of Peak accuracy and Average accuracy on BANKING77 test set under different levels of label noise. The best and second best results are highlighted in **boldface** and underlined, respectively.

Methods	Clean	Symmetric						Asymmetric					
		20%		40%		60%		10%		20%		40%	
		Peak.	Avg.	Peak.	Avg.	Peak.	Avg.	Peak.	Avg.	Peak.	Avg.	Peak.	Avg.
Full Fine-tuning	92.9	88.8	83.4	84.3	72.6	78.2	58.5	90.8	87.6	87.3	79.4	66.9	54.6
PEFT methods													
Adapter (2019)	92.7	88.5	85.4	86.6	78.4	80.9	67.1	90.3	88.6	86.7	78.5	65.3	56.2
BitFit (2022)	92.5	88.9	88.7	86.7	85.9	80.1	76.5	90.2	89.8	86.3	83.1	66.7	62.4
Prompt (2022b)	91.9	87.8	87.4	85.6	84.5	83.2	77.2	89.7	88.4	85.4	84.9	61.6	58.9
LoRA (2022)	93.0	89.2	88.3	86.8	85.8	81.9	77.5	90.1	88.6	86.9	83.1	64.5	61.8
PEFT methods with CleaR (ours)													
$\text{CleaR}_{\text{Adapter}}$	**93.1**	**90.1**	89.7	**88.2**	**87.3**	82.3	80.2	**91.4**	90.3	**87.6**	**86.1**	67.3	66.1
$\text{CleaR}_{\text{BitFit}}$	92.4	89.8	89.2	87.3	86.9	82.9	82.2	90.7	**90.4**	87.5	86.1	67.1	63.4
$\text{CleaR}_{\text{Prompt}}$	92.1	88.1	87.6	85.8	84.9	83.7	81.0	89.9	89.2	85.7	84.8	64.5	62.3
$\text{CleaR}_{\text{LoRA}}$	92.8	90.0	**89.8**	**87.4**	86.9	**84.2**	**83.5**	91.3	90.3	87.2	85.9	**68.9**	**68.1**

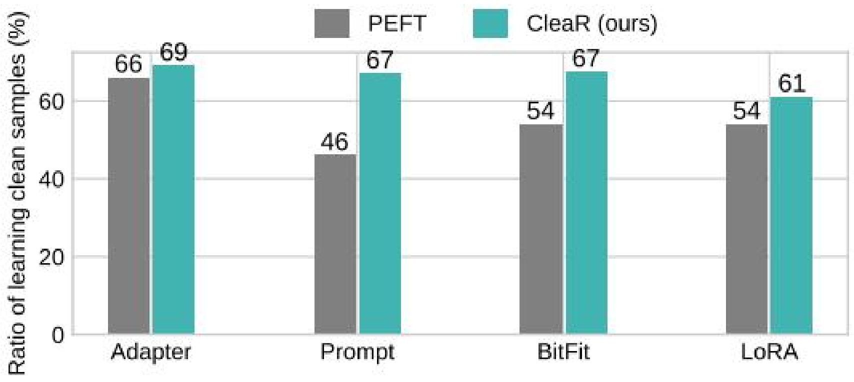
Experiments

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

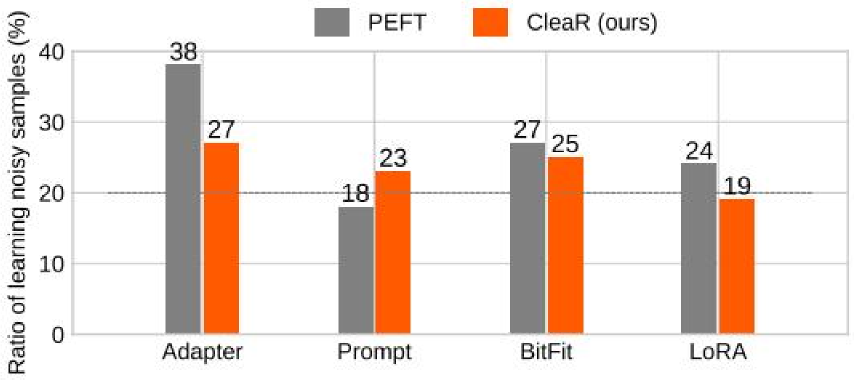
Table 3: Ablation study of CleaR on SST-5 (60% of symmetric noise). For the ablation of routing strategies, we remove the consistency regularization to solely evaluate each routing strategy.

Methods	Peak.	Avg.
$\text{CleaR}_{\text{Adapter}}$(ours)	50.4	49.7
Components in CleaR		
CleaR w/o Clean Routing	48.4	41.1
CleaR w/o Regularization	49.9	48.6
CleaR w/o Clean Routing & Regularization	47.2	40.0
Routing Strategy in CleaR		
CleaR w/ Clean Routing	49.9	48.6
CleaR w/ Deterministic Routing	48.1	44.2
CleaR w/ Random Routing	47.5	40.5
CleaR w/ Noisy Routing	46.9	34.3

Experiments



(a) Impact on learning clean samples



(b) Impact on learning noisy samples

Figure 4: Impact on two memorizations when applying CleaR to PEFT methods. Best viewed in color..

Experiments

Table 5: Peak and Average accuracy (%) on SST-5 under different levels of instance-dependent noise.

Method	Clean	40%		60%	
		Peak.	Avg.	Peak.	Avg.
Full Fine-tuning	53.4	49.0	43.9	44.8	38.9
PEFT methods					
Adapter	53.3	48.8	44.1	44.2	39.7
BitFit	53.0	50.0	45.4	44.8	41.6
Prompt	52.7	49.5	46.2	42.8	38.8
LoRA	**53.6**	49.1	46.9	44.2	39.8
PEFT methods with CleaR (ours)					
$\text{CleaR}_{\text{Adapter}}$	53.4	50.5	46.4	45.7	43.4
$\text{CleaR}_{\text{BitFit}}$	53.1	**51.0**	46.5	45.2	42.6
$\text{CleaR}_{\text{Prompt}}$	52.6	50.2	44.1	44.8	43.2
$\text{CleaR}_{\text{LoRA}}$	53.3	49.7	**47.1**	**46.5**	**44.8**

Thanks