2023 IEEE/CVF International Conference on Computer Vision (ICCV) | 979-8-3503-0718-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCV51070.2023.01420

Why Is Prompt Tuning for Vision-Language Models Robust to Noisy Labels?

Cheng-En Wu[1*] Yu Tian[2] Haichao Yu[2] Heng Wang[2] Pedro Morgado[1]
Yu Hen Hu[1] Linjie Yang[2]

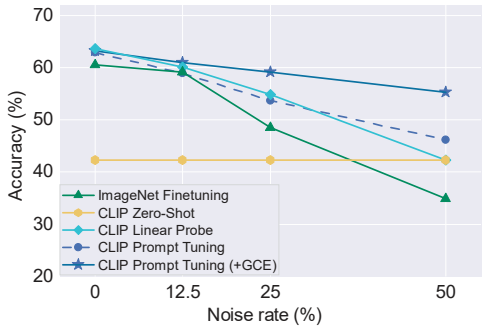
[1]University of Wisconsin-Madison [2]ByteDance Inc.
{cwu356, pmorgado, yhhu}@wisc.edu {yutian.yt, haichaoyu, heng.wang, linjie.yang}@bytedance.com

Abstract

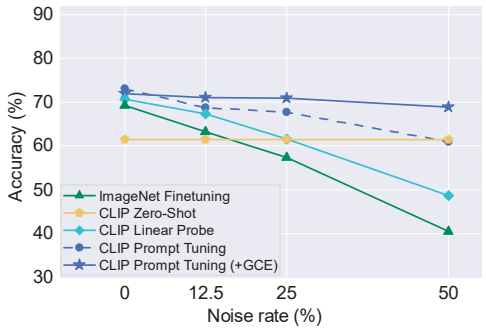
Vision-language models such as CLIP [27] learn a generic text-image embedding from large-scale training data. A vision-language model can be adapted to a new classification task through few-shot prompt tuning. We find that such a prompt tuning process is highly robust to label noises. This intrigues us to study the key reasons contributing to the robustness of the prompt tuning paradigm. We conducted extensive experiments to explore this property and find the key factors are: 1) the fixed classname tokens provide a strong regularization to the optimization of the model, reducing gradients induced by the noisy samples; 2) the powerful pre-trained image-text embedding that is learned from diverse and generic web data provides strong prior knowledge for image classification. Further, we demonstrate that noisy zero-shot predictions from CLIP can be used to tune its own prompt, significantly enhancing prediction accuracy in the unsupervised setting. The code is available at `https://github.com/CEWu/PTNL`.

1. Introduction

Large-scale vision-language models such as CLIP [27], ALIGN [13], and CoCa [43] are transforming how we learn and interact with visual representations. Since these models learn to align the representations of a broad set of natural images with their textual descriptions, they have shown an exceptional ability to solve a wide range of tasks in a data-efficient manner. For example, using the pre-trained text encoder, one can obtain a set of class embeddings by encoding a canonical sentence such as "A photo of a <CLS>" and use them to recognize objects without a labeled dataset. While promising, Zhou et al. [50] showed that these human-defined sentences (also known as class prompts) can be unstable, with seemingly equivalent descriptions leading to different predictions. To address this issue, researchers have focused on prompt tuning [50], where a learnable prompt is learned from a small target dataset by back-propagation.



(a) DTD



(b) UCF101

Figure 1: Comparison with transfer learning approaches on two datasets with training labels that have incremental noisy rates. ImageNet Finetuning is finetuning pre-trained model on ImageNet. For the CLIP pre-trained model, Prompt Tuning is much more robust to the Linear Probe manner. By combining the generalized cross-entropy (GCE) [46], we further improve the robustness of Prompt Tuning to noisy labels. ResNet-50 is used for all approaches as their image encoders.

*Work mostly done during an internship at ByteDance Inc.

2380-7504/23/$31.00 ©2023 IEEE
DOI 10.1109/ICCV51070.2023.01420

Since only the prompt needs to be trained, this framework is very data-efficient. As a result, prompt-tuning has gained popularity for adapting vision-language models to downstream tasks like few-shot learning [50, 49], continual learning [38], and object segmentation [28].

While prompt tuning has proven effective when training on downstream tasks with accurately annotated datasets, their robustness to noisy labels has been neglected. Since the quality of annotations for many applications can be low, learning with noisy labels is critical to solving real-world problems. In this work, we demonstrate that prompt tuning is robust to noisy labels, and investigate the mechanisms that enable this robustness. We hypothesize that the joint text and image embeddings of vision-language models can provide a well defined structure to the classification space (e.g., which classes are most similar and most distinct from each other). This model-informed structure compensates for the degradation of the structure present in the data due to label noise. To verify this hypothesis, we conducted extensive experiments to study the impact of each component of a prompt tuning task with noisy labeled data. Beyond the robustness conferred by the structured label space, we show that this robustness can be further enhanced when the learnable prompts are trained using a robust loss function that mitigates the impact of outliers. Our study has revealed several interesting findings.

First, the classification performance obtained by tuning the prompt through a pre-trained CLIP model is significantly more robust to noisy labels than the traditional fine-tuning or linear probing paradigms (see Figure 1). The robustness of prompt tuning is evident not only due to their smaller performance degradation with higher noise rates, but also due to its ability to diminish the gradients induced by noisy samples. Second, while priming each class with a shared learnable prompt is necessary for adaptation, ensuring that the class name remains in the prompt strongly regularizes the class embeddings and prevents overfitting to the noisy labels. Finally, we demonstrate the benefits of this robustness by showing that CLIP zero-shot (noisy) predictions can be used to tune its own prompt, and significantly enhance CLIP prediction accuracy. In fact, we show that, instead of focusing on samples with confident predictions (as proposed in prior unsupervised prompt tuning approaches [12]), prompt tuning benefits more from an increased diversity of training samples as it can tolerate the noisier predictions associated with them.

The main contributions of our work are as follows:

- We demonstrate that prompt tuning for pre-trained vision-language models (e.g., CLIP) is more robust to noisy labels than traditional transfer learning approaches, such as model fine-tuning and linear probes.
- We further demonstrate that prompt tuning robustness can be further enhanced through the use of a robust training objective.
- We conduct an extensive analysis on why prompt tuning is robust to noisy labels to discover which components contribute the most to its robustness.
- Motivated by this property, we propose a simple yet effective method for unsupervised prompt tuning, showing that randomly selected noisy pseudo labels can be effectively used to enhance CLIP zero-shot performance. The proposed robust prompt tuning outperformed prior work [12] on a variety of datasets, even though noisier pseudo-labels are used for self-training.

2. Related Work

Prompt tuning for Vision-Language models. Over the past few years, there has been huge progress in Vision-Language Pre-Trained Models (VL-PTMs) [27, 13, 40, 43]. CLIP [27] is considered a representative model of VL-PTMs. Unlike the conventional, finetuning paradigm, CLIP applies prompt engineering to incorporate the category information in the text input such that its pre-trained model can adapt to various image classification tasks without further training. However, the design of a proper prompt is challenging and requires heuristics. CoOp [50] introduces learnable prompts optimized on target datasets to address this problem. To further extend the generalization of CoOp, CoCoOp [49] introduces a lightweight network to add additional information from image inputs into learnable prompts. CoOp has also faced criticism for disregarding the diversity of visual representations. In contrast, ProDA [18] tackles this issue by utilizing diverse prompts to capture the distribution of varying visual representations.

In contrast to the supervised tuning methods above, UPL [12] proposes a framework to perform prompt tuning without labeled data. TPT [22] achieves zero-shot transfer by dynamically adjusting prompts using only a single test sample.

In addition to downstream tasks for image classification, recent works have applied prompt tuning on CLIP to various computer vision tasks such as object detection [28, 6], video understanding [16, 14], and multi-label recognition [35]. These works reveal the further potential of prompt tuning.

Label noise-robust learning. Deep neural networks (DNNs) have been well-studied for classification tasks without label noises. However, if the training data contains label noise, DNNs would easily overfit to the noisy labels [44]. To overcome this issue, several works have attempted to improve the noise robustness of DNNs by approaches including robust losses that tolerate noisy labels [8, 46, 37, 19], loss correction approaches that estimate a transition matrix to correct the predictions [25, 11, 3, 30, 1, 20, 33, 42, 41], meta-learning frameworks that learn to

correct the label noise in training examples [17, 31, 15, 32, 47, 48] and regularization techniques that are customized to lower the negative impact of noise [45, 26, 10, 39].

In this work, we demonstrate that prompt tuning on CLIP naturally holds powerful noise robustness. We explore the key factors behind such robustness and show its application on unsupervised prompt tuning.

3. Prompt Tuning

CLIP [27] can perform zero-shot transfer by prompt engineering – the practice of designing text inputs for downstream tasks. Specifically, in the case of image classification, a normalized image embedding $\mathbf{f}^v$ is obtained by passing an image $\mathbf{x}$ through CLIP's visual encoder, and a set of normalized class embeddings $\{\mathbf{f}_i^t\}_{i=1}^K$ by feeding template prompts of the form "A photo of a <CLS>" into CLIP's text encoder. The class posterior is then estimated as

$$Pr(y=i|\mathbf{x}) = \frac{\exp\Big(\texttt{sim}(\mathbf{f}^v, \mathbf{f}_i^t)/\tau\Big)}{\sum_{j=1}^K \exp\Big(\texttt{sim}(\mathbf{f}^v, \mathbf{f}_j^t)/\tau\Big)}, \quad (1)$$

where τ is a temperature factor learned by CLIP and $\texttt{sim}$ denotes cosine similarity.

Prompt Tuning Although CLIP is capable of zero-shot transfer, its performance is sensitive to designed text prompts. To avoid the need for hand-crafted prompts and improve transfer performance, CoOp [50] showed that text prompts can be replaced with continuous soft prompts that can be optimized on a target dataset. Specifically, the name of a class c is first converted into a classname embedding $\mathbf{w}_c \in \mathbb{R}^d$ and prepended with a sequence of M learnable tokens $\mathbf{p}_m \in \mathbb{R}^d$ shared across all classes. The full prompt $\mathrm{P_c} = [\mathbf{p}_1, \mathbf{p}_2, ...\mathbf{p}_M, \mathbf{w}_c]$ for each class c is then processed by CLIP's text encoder to compute the corresponding text embedding $\mathbf{f}_c^t$, and the class posteriors $Pr(y=i|\mathbf{x})$ are obtained once again through Eq. 1. To adapt the prompt to the target dataset, CoOp [50] optimizes the shared learnable tokens $\mathbf{p}_1, \mathbf{p}_2, ...\mathbf{p}_M$ on a small labeled dataset $\mathcal{D} = \{(\mathbf{x}_i, c_i)_{i=1}^N\}$ to minimize the cross-entropy loss

$$\mathcal{L}_{CE} = -\mathbb{E}_{(\mathbf{x},c)\in\mathcal{D}}\left[\log Pr(y=c|\mathbf{x})\right]. \quad (2)$$

Robust Prompt Tuning In this work, we show that the prompt tuning framework [50], describe above, displays surprising robustness to noisy labels. However, this robustness can be further enhanced by optimizing the learnable prompts using the generalized cross-entropy (GCE) loss [46], a robust generalization of cross-entropy loss. Formally, the GCE loss is defined as

$$\mathcal{L}_{GCE} = \mathbb{E}_{(\mathbf{x},c)\in\mathcal{D}}\left[\frac{1-Pr(y=c|\mathbf{x})^q}{q}\right]. \quad (3)$$

As shown in [46], GCE is equivalent to the standard cross-entropy loss of Eq. 2 when $x \to 0$, and equivalent to the (robust) mean absolute error (MAE) loss $\|1 - Pr(y=c|\mathbf{x})\|_1$ when $q = 1$. The hyper-parameter q can therefore control the tradeoff between the highly robust but less performing MAE loss and the less robust but highly performing CE loss. While the optimal value for q could be adjusted to the amount of noise by cross-validation, we found that $q = 0.7$ lead to overall good performance across several experimental settings.

4. Analysis of Prompt Tuning with Label Noise

Methods based on prompt tuning for CLIP [27] have been shown to be effective in few-shot learning [50, 49]. However, these methods have been studied on datasets with perfect labels. It remains unknown how prompt tuning performs under label noise. We explore this practical training setting and present our key findings.

4.1. Experimental Settings

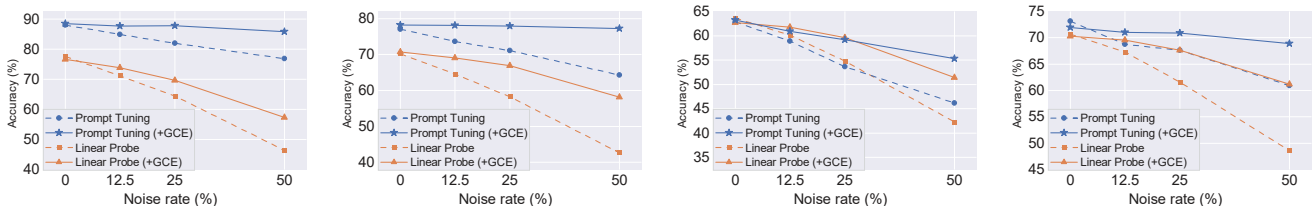
Datasets. We conduct in-depth studies on a diverse set of visual tasks, including generic object classification, fine-grained recognition, action recognition, and texture identification. We conduct our experimental analysis on eight datasets, OxfordPets [24], Food101 [2], DTD [4], UCF101 [34], Flowers102 [23], FGVCAircraf [21], Caltech101 [7] and ImageNet [29]. Since one of the main benefits of prompt tuning is its data efficiency [12], we focus our studies on a 16-shot image classification problem, i.e. for each dataset, we select 16 images per class as our training set. To examine the effect of noise in prompt tuning, we randomly perturb training labels with different levels of noise rate (12.5%, 25%, and 50%). Unless otherwise specified, noisy labels are drawn uniformly at random from other categories of the dataset. We report average results over four runs with different training sets in all experiments.

Backbone. We adopt pre-trained CLIP models, namely using the 63M parameter text Transformer [36] as the text encoder, and either a ResNet-50 [9] or a ViT-B/32 [5] as the visual encoder. Following CoOp [49], we use 16 learnable tokens in each prompt shared across all categories.

Optimization. Models are trained with a batch size of 32 for 50 epochs, using stochastic gradient descent (SGD) with momentum of 0.9 and an initial learning rate of 0.002, annealed to zero with a cosine decay schedule.

4.2. Prompt Tuning Is Robust to Noisy Labels

The core observation of this paper is that prompt tuning vision-language models, such as CLIP, is surprisingly robust to noisy labels. This can be observed by comparing prompt tuning for CLIP with two traditional transfer learning approaches: 1) training a linear classifier on CLIP's visual representations (denoted CLIP Linear Probe); and 2)



(a) OxfordPets (b) Food101 (c) DTD (d) UCF101

Figure 2: Incorporating the generalized cross-entropy (GCE) [46] loss with Prompt Tuning and Linear Probe methods, originally trained using cross-entropy, can enhance their noise robustness. At high noise rates, Prompt Tuning with GCE outperforms other methods by a significant margin across four datasets.

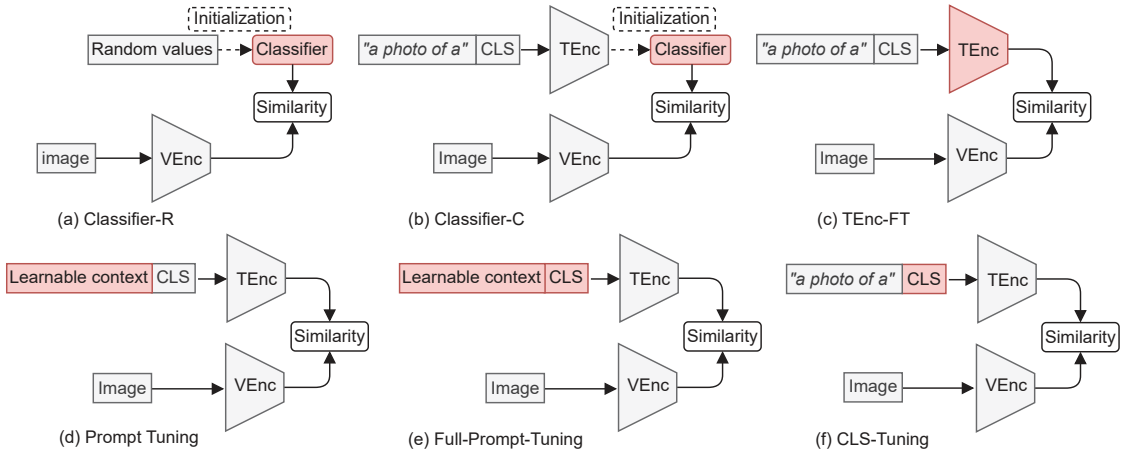


Figure 3: Illustration of different structures for studying the effect of image and text encoders on prompt tuning and prompt design. The blocks highlighted in red are to be trained, while those highlighted in gray are to be frozen.

fine-tuning the same visual backbone pre-trained on ImageNet. The results on two datasets, DTD and UCF101, are shown in Figure 1 (a) and (b), respectively. As can be seen, although linear probes and fine-tuning achieve competitive performance with perfectly labeled data (0% noise rate), both procedures suffer from a significant accuracy drop with higher noise rates of 25% and 50%. This result shows that prompt tuning is naturally more resistant to noisy labels than the alternatives. We show nevertheless that its robustness can be further enhanced by training the prompt using the robust generalized cross-entropy loss (denoted CLIP Prompt Tuning (GCE) in Figure 1). As can be seen, when combining Prompt Tuning and GCE, the model's performance remains highly competitive, even for noise rates as high as 50%. Furthermore, we observe that this robustness stems from the combination between Prompt Tuning and GCE, and not from GCE alone. This can be seen in Figure 2, which depicts the noise robustness of Prompt Tuning and Linear Probes both trained under the cross-entropy and GCE losses on four datasets. While the robustness of the linear probe also improves with a GCE loss, the performance drop at high noise rates is significantly smaller when learning through prompt tuning.

Now that we have established the noise robustness of prompt tuning, the remainder of this Section is dedicated to providing intuitions and experimental analysis to answer the why question.

> ***Question:*** *Why is prompt tuning for CLIP-like vision-language models more robust than traditional transfer learning against noisy labels?*

4.3. Robustness Attribution

To answer this question, we begin by analyzing two key components of CLIP in isolation, namely the generated class embeddings and learnable prompts.

Pre-trained CLIP generates effective class embeddings. We first analyse the impact of the class embeddings generated by the CLIP text encoder. To this end, in addition to the class embeddings generated through prompt tuning, we assess the noise robustness of three different models:

Dataset	Method	Noise rate			
		0	12.5	25	50
OxfordPets	Classifier-R	74.82	64.10	55.96	36.63
	Classifier-C	81.47	70.29	61.87	44.21
	TEnc-FT	84.38	70.73	61.11	41.21
	Prompt Tuning	**87.89**	**84.62**	**81.20**	**73.13**
Food101	Classifier-R	63.80	54.66	46.23	28.97
	Classifier-C	69.36	60.46	51.85	34.37
	TEnc-FT	71.30	61.60	52.64	34.74
	Prompt Tuning	**76.99**	**73.63**	**71.07**	**64.30**
DTD	Classifier-R	48.02	44.30	40.32	30.10
	Classifier-C	**63.83**	57.14	50.36	34.86
	TEnc-FT	63.61	55.47	48.21	33.12
	Prompt Tuning	62.86	**58.90**	**53.62**	**46.19**
UCF101	Classifier-R	67.16	58.33	50.34	31.07
	Classifier-C	71.87	64.12	54.79	38.01
	TEnc-FT	**73.74**	64.52	56.10	37.88
	Prompt Tuning	73.12	**68.73**	**67.66**	**60.93**

Table 1: Comparison of transfer performance at incremental noise rates between different variants.

Classifier-R Trains a linear probe on the output of CLIP's pre-trained visual encoder. The class embeddings (i.e., the classifier weights) are initialized at *random*, and learned without constrains. See Figure 3 (a).

Classifier-C Similar to Classifier-R, but the classifier weights are initialized using the text embeddings $\mathbf{f}_c^t$ obtained from CLIP's pre-trained text encoder for the handcrafted prompt. Note that Classifier-C only uses the CLIP text encoder for initializing its weights. See Figure 3 (b).

TEnc-FT Trains a CLIP classifier, by associating the image embedding $\mathbf{f}^v$ with the CLIP text embedding $\mathbf{f}^t$ of the correct class through the posterior of eq. (1). In this case, the entire CLIP text encoder is *fine-tuned* on an hand-crafted prompt of the form "A photo of a <CLS>". See Figure 3 (c).

Table 1 compares the various models on four datasets under different levels of label noise. The linear classifier with CLIP initialization (Classifier-C) outperformed random initialization across all levels of noise. This shows that CLIP class embeddings provide a strong initialization for few-shot learning. Furthermore, although both Classifiers degrade considerably with high noise ratios, the CLIP initialization is also more robust to noise. As for TEnc-FT, it achieved competitive performance at zero noise rates, but its accuracy also dropped significantly as the noise rate increased. This highlights (unsurprisingly) that the highly expressive CLIP text encoder can easily overfit to the noisy labels. Finally, Prompt Tuning outperformed all alternative strategies across all noise rates. The advantage of prompt tuning was especially large for high noise levels. These observations confirm that (a) the text encoder is essential for providing a strong but informative regularization of the text embeddings to combat noisy inputs (Prompt Tuning v.s. classifiers); and (b) the text encoder should be fixed to prevent overfitting (Prompt Tuning v.s. TEnc-FT).

Dataset	Method	Noise rate			
		0	12.5	25	50
OxfordPets	Full-Prompt-Tuning	85.39	74.00	68.66	50.50
	CLS-Tuning	85.04	77.02	71.03	53.15
	Prompt Tuning	**87.89**	**84.62**	**81.20**	**73.13**
Food101	Full-Prompt-Tuning	72.36	63.14	55.29	38.69
	CLS-Tuning	72.07	63.91	56.97	41.73
	Prompt Tuning	**76.99**	**73.63**	**71.07**	**64.30**
DTD	Full-Prompt-Tuning	62.80	55.50	49.01	34.66
	CLS-Tuning	62.78	56.15	48.46	35.43
	Prompt Tuning	**62.86**	**58.90**	**53.62**	**46.19**
UCF101	Full-Prompt-Tuning	73.02	64.31	57.11	40.42
	CLS-Tuning	72.73	65.64	58.91	44.55
	Prompt Tuning	**73.12**	**68.73**	**67.66**	**60.93**

Table 2: Comparison of transfer performance at incremental noise rates between different prompt designs.

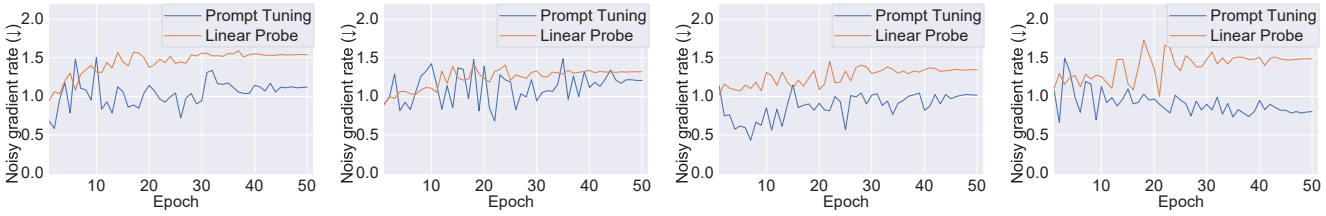
Effectiveness of prompt. The previous experiment showed that the class embeddings generated by CLIP pre-trained text encoder plays a critical role in noise robustness. Next, we keep the text encoder fixed, and attempt to answer another question: *Which components of the prompt provide noise robustness to prompt tuning?*

We hypothesize that the classname token $\mathbf{w}_c$ provides a strong regularization to the model, since it is leveraged by the text encoder to encode relationships between the different visual concepts (e.g. how similar or different classes are from each other). Respecting this structure could help the model avoid fitting noisy data during training. To verify our hypothesis, we assess the noise robustness of two additional models:

Full Prompt Tuning Learns the classname token jointly with the original learnable tokens (see Figure 3 (e)).

CLS Tuning Adopts a *fixed* template prompt "A photo of a <CLS>" and optimizes only the classname token (see Figure 3(f)).

Table 2 shows the analysis on four dataset for different noise levels. Compared to prompt-tuning, which optimizes only learnable tokens shared across all classes, both CLS-Tuning and Full-Prompt-Tuning models struggle at high noise rate. Even when the training data is clean, learning the classname tokens produces worse performance on two of the four datasets (OxfordPets and Food101). This analysis validates our assumption that the fixed classname token is indeed a critical regularization for the prompt tuning. Learnable classname tokens can be fitted to the noisy train-



(a) OxfordPets (b) Food101 (c) DTD (d) UCF101

Figure 4: We assess the ability of both methods to suppress noisy gradients by evaluating their noisy-to-clean gradient norm ratio (noisy gradient rate). This ratio is determined by taking the L2 norm of gradients with respect to the learnable parameters, which we compute by feeding 64 clean samples and 64 noisy samples to the model during each training epoch. Specifically, we train the models on data with a 50% noise rate. Results on four datasets show that Prompt Tuning achieves a lower noisy gradient rate compared to Linear Probe, indicating its superior ability to suppress noisy gradients.

ing data, perturbing the class embeddings and leading to worse performance.

4.4. Prompt Tuning Suppresses Noisy Gradients

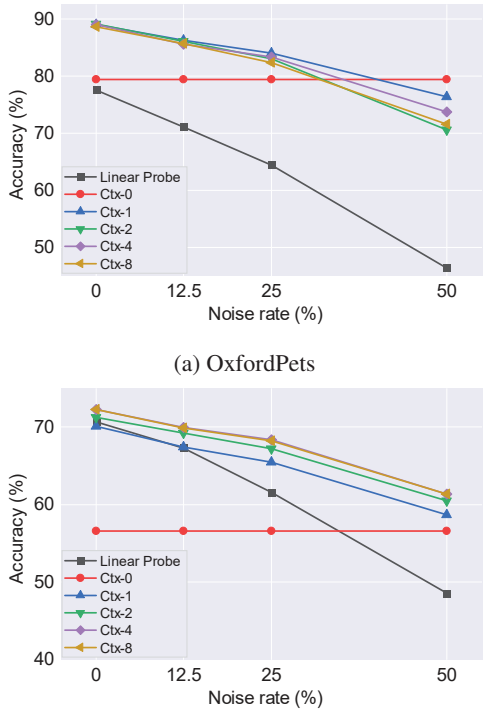
The previous section provided clear evidence of the robustness of the prompt tuning framework in comparison to other alternatives. These findings suggest that, by learning only shared prompt tokens, prompt tuning focuses better on clean samples than noisy samples. In other words, prompt tuning can suppress gradient updates from noisy samples, while aggregating gradients from clean samples. To verify this hypothesis, we measure the gradients with respect to the learnable parameters of both CLIP prompt tuning and linear probing using 50% noise rate. Specifically, we measure the ratio between the gradient norm induced by noisy samples and that induced by clean samples. A ratio above one indicates that noisy samples play a bigger role in the optimization than clean samples.

Figure 4 shows the noisy-to-clean gradient norm ratio as models are trained on four datasets. As can be seen, prompt tuning displays significantly lower ratios than linear probing. This indicates that noisy samples play a comparatively small role with prompt tuning compared to linear probes. This property likely arises from the highly constrained prompt tuning optimization, which restricting the model to fit the noisy labels.

4.5. Generalization Across Model Architectures

Previous sections have focused on four datasets (OxfordPets, Food101, DTD, and UCF101) and a ResNet-50 image encoder. We now show that these findings generalize across model architectures and datasets.

Context length. We first assess the noise robustness of prompt tuning with increasing numbers of learnable tokens. We also evaluate a baseline without any learnable tokens by directly feeding the classname into the model (denoted as Ctx-0). Figure 5 shows that the optimal context lenght



(a) OxfordPets

(b) UCF101

Figure 5: Investigation on noise robustness of prompt tuning accompanied by various context lengths. Ctx-x denotes the model with x learnable tokens.

is dataset dependent, but all context lengths achieve superior performance compared to traditional linear probing. Ctx-0 outperforms some prompt tuning variants under large noise rates at 50%, suggesting fixed prompts may be a good choice when the labeling noise is too strong on the downstream task.

Image encoders. To validate whether the noise robustness of prompt tuning is backbone-agnostic, we also assess CLIP

Dataset	Method	Noise rate			
		0	12.5	25	50
ImageNet	RN50-PT	62.83	61.98	60.60	57.97
	ViT-B/32-PT	66.48	65.82	64.50	61.75
Caltech101	RN50-PT	90.65	82.51	78.70	70.13
	ViT-B/32-PT	93.63	90.34	84.99	77.16
OxfordPets	RN50-PT	87.89	84.62	81.20	73.13
	ViT-B/32-PT	89.10	86.59	83.65	75.50
Flowers102	RN50-PT	92.57	86.85	81.73	71.80
	ViT-B/32-PT	93.26	87.90	85.34	72.83
Food101	RN50-PT	76.99	73.63	71.07	64.30
	ViT-B/32-PT	80.16	77.60	76.06	68.77
FGVCAircraft	RN50-PT	27.13	25.07	23.34	19.05
	ViT-B/32-PT	28.37	27.57	25.47	19.57
DTD	RN50-PT	62.86	58.90	53.62	46.19
	ViT-B/32-PT	64.88	59.57	57.09	45.22
UCF101	RN50-PT	73.12	68.73	67.66	60.93
	ViT-B/32-PT	78.12	75.97	72.83	65.75

Table 3: Noise robustness of prompt tuning (PT) with ResNet50 or ViT-B/32 as the image encoder on eight datasets.

Dataset	Method	Random	Confusion
OxfordPets	Linear Probe	$46.42_{\pm 0.88}$	$41.39_{\pm 1.87}$
	Prompt Tuning	$73.13_{\pm 3.76}$	$66.55_{\pm 2.02}$
Food101	Linear Probe	$42.63_{\pm 0.89}$	$37.71_{\pm 0.52}$
	Prompt Tuning	$64.30_{\pm 2.58}$	$63.93_{\pm 1.45}$
DTD	Linear Probe	$42.29_{\pm 2.12}$	$37.69_{\pm 1.70}$
	Prompt Tuning	$46.19_{\pm 2.12}$	$45.76_{\pm 1.23}$
UCF101	Linear Probe	$54.05_{\pm 1.19}$	$50.90_{\pm 1.45}$
	Prompt Tuning	$60.93_{\pm 0.94}$	$59.11_{\pm 0.70}$

Table 4: The impact of random and confusion label noise at a 50% noise rate on Linear Probing and Prompt Tuning strategies.

with ViT-B/32 for prompt tuning (denoted ViT-B/32-PT). Table 3 shows the comparison with RN50-PT. ViT-B/32-PT outperforms RN50-PT under most settings. Moreover, both methods do not suffer from a large performance drop and maintain competitive accuracy at high noise rates.

4.6. Robustness to Correlated Label Noise

So far, we assumed white label noise (i.e., noisy labels are uniformly drawn from the label space). However, label noise produced by either human annotators or machine-generated pseudo labels often displays correlations between similar visual concepts. For example, UPL [12] observed that pre-trained CLIP prefers some classes over others during zero-shot transfer. Inspired by this observation, we examine whether CLIP inherent preferences affect the performance of prompt tuning when confronted with CLIP-generated label noise.

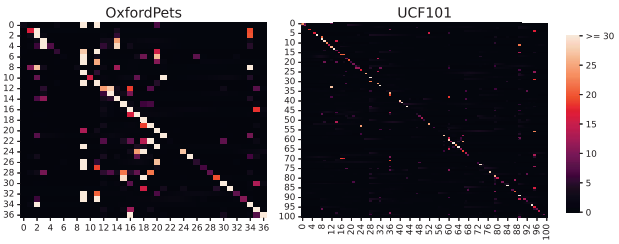


Figure 6: Confusion matrix generated by averaging the zero-shot performance over 100 runs using random prompt tokens.

We begin by measuring the confusion matrix of CLIP zero-shot predictions with *randomly initialized* learnable tokens on the OxfordPets and UCF101 datasets (see Figure 6). Next, we introduce a challenging type of label noise, named *Confusion* noise, where each mislabeled sample is labeled as the incorrect class that is most favored by zero-shot CLIP. Finally, we examine the transfer performance of prompt tuning with both random and confusion noise at a 50% noise rate. Table 4 presents the results on four datasets. As can be seen, confusion noise presents a bigger challenge to transfer learning, leading to larger degradation of classification accuracy at high noise ratios compared to random noise. Such degradation is visible both for prompt tuning and linear probes. However, among the two, prompt tuning still achieves the best overall performance, providing further evidence for its robustness even to more challenging types of noise.

5. Application to Unsupervised Prompt Tuning

Prior work UPL [12] demonstrated that unsupervised prompt tuning can outperform the transfer performance of zero-shot transfer based on CLIP. However, UPL does not fully utilize the noise robustness of prompt tuning.

Baseline UPL. UPL [12] proposed a framework to adapt CLIP for downstream tasks without any labeled images. An overview of the framework is shown in Figure 7. This framework is divided into two phases. In phase 1, UPL leverages pre-trained CLIP to generate pseudo labels for unlabeled images. Then, in phase 2, a set of K pseudo-labels are chosen to optimize the learnable tokens through the typical prompt-tuning optimization process (described in CoOp [50]). To increase the quality of training examples, UPL ranks all pseudo-labeled images based on their confidence score (Eq. 1) and selects the K most confident samples per class. Furthermore, inspired by prompt ensembling in CLIP [27], UPL improved transfer performance by ensembling multiple predictions generated by models with different learnable prompts.

Robust UPL. In Section 4, we showed that prompt tuning

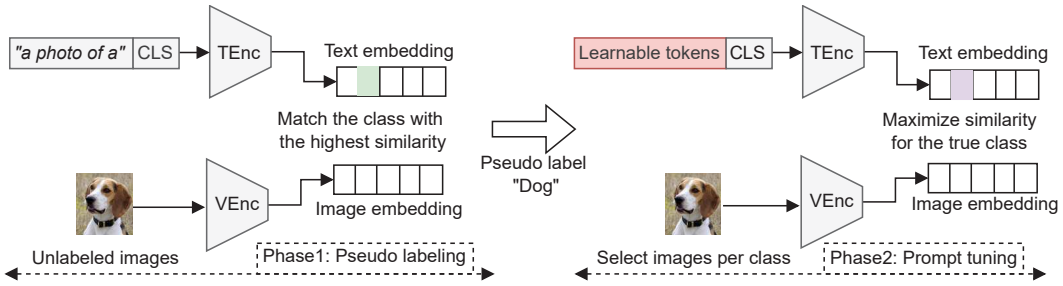


Figure 7: The pipeline of unsupervised prompt tuning. It consists of two main phases: Pseudo labeling and Prompt tuning. To begin, we generate pseudo labels for target datasets by utilizing CLIP with a template prompt for zero-shot transfer. Next, we randomly select samples per class from the pseudo labels for subsequent training. Finally, we optimize the learnable prompt representation using the selected pseudo-labeled samples.

can be robust to noisy labels. Furthermore, we showed that prompt tuning robustness can be further strengthened using the generalized cross-entropy loss (GCE). Given these observations, we propose to perform unsupervised prompt tuning by 1) randomly sample training samples and 2) optimizing the prompt with the robust GCE loss. Random sampling has two effects. On the one hand, it increases the diversity of training samples which benefits learning. On the other hand, it increases the amount of label noise. However, we expect the label noise to be tolerable by our robust prompt tuning framework.

Experimental Settings. We experiment with the unsupervised prompt tuning following the same training setting of Section 4. Pseudo-labels are generated by CLIP zero-transfer with ResNet50 image encoder. We follow the prompt engineering used by CLIP. There are three types of hand-crafted prompts, with more details listed in the supplementary material. K is set to 16 in all experiments. During the inference stage, we employ the ensemble-average approach following UPL [46] to generate predictions combining the outputs of four distinct models. Each model has a distinct learnable prompt that was initialized with a unique random seed.

Experimental Results. We compared UPL [12] and the proposed Robust UPL on a diverse set of visual tasks, including generic object classification, fine-grained recognition, and texture identification. We also assessed Robust UPL using both a cross-entropy (CE) and generalized cross-entropy (GCE) losses. Table 5 shows that all three unsupervised prompt tuning methods can improve transfer learning over zero-shot predictions, at no additional labeling cost. Among the three methods, Robust UPL trained under GCE loss obtains the best performance on average. We highlight once again that Robust UPL randomly samples pseudo labeled images for training, instead of using high-confidence samples as in UPL. As a result, UPL training pseudo-labels are less diverse, but have less noise. For example, the pseudo-labels used to train UPL on Caltech were 93% correct, while the pseudo-labels used to train Robust UPL were only 83% correct. Nevertheless, these errors did not harm final performance of Robust UPL; on the contrary, learning from a more diverse set, while being robust to the noise enhanced prompt tuning.

Dataset	0-Shot	UPL [12]	Robust UPL	
			CE	GCE
ImageNet	58.18	60.22	61.11	**62.14**
Caltech101	86.29	**90.10**	87.14	88.07
OxfordPets	85.77	87.60	86.89	**87.71**
Flowers102	66.14	69.31	70.04	**70.52**
Food101	77.31	77.30	77.84	**78.51**
FGVCAircraft	**17.28**	15.93	16.35	16.29
DTD	42.32	37.47	44.80	**46.69**
UCF101	61.46	65.00	66.01	**67.12**

Table 5: Comparison between CLIP zero-shot classification and three strategies for unsupervised prompt tuning: UPL [12], and our robust UPL framework trained with cross-entropy and generalized cross-entropy losses.

6. Conclusion

In this paper, we provide a comprehensive study on the robustness to label noise of prompt tuning large vision-language models (namely, CLIP). Through a series of experiments, we demonstrated that the noise robustness of prompt tuning can be attributed to the structure imposed on class embeddings by CLIP's pre-trained text encoder. We further demonstrate that prompt tuning can ease overfitting to mislabeled samples by reducing the gradients induced by label noise. We extensively experimented with different model configurations such as backbones and context length, obtaining consistent results and conclusions. Finally, inspired by our findings, we presented a new robust unsupervised prompt tuning approach that favors diversity over correct predictions, to improve the transfer performance.

Acknowledgement

This work was partially supported by the National Science Foundation under Grant No. 2006394.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin Mcguinness. Unsupervised label noise modeling and loss correction. In *ICML*, 2019.

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014.

[3] Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. In *NeurIPS*, 2017.

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

[6] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, 2022.

[7] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR-W*, 2004.

[8] Aritra Ghosh, Himanshu Kumar, and P Shanti Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI*, 2017.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[10] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *ICML*, 2019.

[11] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *NeurIPS*, 2018.

[12] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint*, 2022.

[13] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.

[14] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *ECCV*, 2022.

[15] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In *CVPR*, 2019.

[16] Muheng Li, Lei Chen, Yueqi Duan, Zhilan Hu, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Bridge-prompt: Towards ordinal action understanding in instructional videos. In *CVPR*, 2022.

[17] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *ICCV*, 2017.

[18] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *CVPR*, pages 5206–5215, 2022.

[19] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *ICML*, 2020.

[20] Xingjun Ma, Yisen Wang, Michael E Houle, Shuo Zhou, Sarah Erfani, Shutao Xia, Sudanthi Wijewickrema, and James Bailey. Dimensionality-driven learning with noisy labels. In *ICML*, 2018.

[21] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint*, 2013.

[22] Shu Manli, Nie Weili, Huang De-An, Yu Zhiding, Goldstein Tom, Anandkumar Anima, and Xiao Chaowei. Test-time prompt tuning for zero-shot generalization in vision-language models. In *NeurIPS*, 2022.

[23] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008.

[24] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012.

[25] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, 2017.

[26] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint*, 2017.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[28] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, 2022.

[29] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019.

[30] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint*, 2014.

[31] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *ICML*, 2018.

[32] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *NeurIPS*, 2019.

[33] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Selfie: Refurbishing unclean samples for robust deep learning. In *ICML*, 2019.

[34] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint*, 2012.

[35] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *arXiv preprint*, 2022.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.

[37] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, 2019.

[38] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *CVPR*, 2022.

[39] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *ICLR*, 2020.

[40] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *ICLR*, 2021.

[41] Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual t: Reducing estimation error for transition matrix in label-noise learning. *NeurIPS*, 2020.

[42] Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *CVPR*, 2019.

[43] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint*, 2022.

[44] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICML*, 2017.

[45] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

[46] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, 2018.

[47] Zizhao Zhang, Han Zhang, Sercan O Arik, Honglak Lee, and Tomas Pfister. Distilling effective supervision from severe label noise. In *CVPR*, 2020.

[48] Guoqing Zheng, Ahmed Hassan Awadallah, and Susan Dumais. Meta label correction for noisy label learning. In *AAAI*, 2021.

[49] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022.

[50] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022.