Research Survey 1

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

Zhihao Li

School of Computer Science and Technology
Xidian University

February 1, 2024

Weekly Work

1. Read the paper of *Why Is Prompt Tuning for Vision-Language Models Robust to Noisy Labels?*;
2. Learn about some concepts;

A prompt tuning process is highly robust to label noises.

1. **Interest**: Studying the key reasons contributing to the robustness of the prompt tuning. paradigm.
2. **Findings**:
    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.

Author's Contributions

- 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.**
- 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 on a variety of datasets, even though noisier pseudo-labels are used for self-training.

Mathematical Models

- CLIP
  In the case of image classification, a normalized image embedding $\boldsymbol{f}^v$ is obtained by passing an image through CLIP's visual encoder, and a set of normalized class embeddings $[\boldsymbol{f}_i^t]_{i=1}^K$ by feeding template prompts of the form "A photo of a" into CLIP's text encoder.

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

- Prompt Tuning
  The name of a class c is first converted into a classname embedding $\boldsymbol{w} \in R^d$ and prepended with a sequence of M learnable tokens $\boldsymbol{p_m} \in R^d$ shared across all classes.

$$P_c = [\boldsymbol{p_1}, \boldsymbol{p_2}, \cdots, \boldsymbol{p_M}, \boldsymbol{w_c}] \rightarrow \boldsymbol{f}_c^t \tag{2}$$

  CoOp optimizes the shared learnable tokens $\boldsymbol{p_1}, \boldsymbol{p_1}, \cdots, \boldsymbol{p_M}$ on a small labeled dataset $D = [(\boldsymbol{x_i}, c_i)_{i=1}^N]$ to minimize the cross-entropy loss:

$$L_{CE} = -E_{(\boldsymbol{x},c)\in D}[\log Pr(y=c|\boldsymbol{x})] \tag{3}$$

Mathematical Models

- Robust Prompt Tuning
  Further enhance this robustness by optimizing the learnable prompts using the generalized cross-entropy (GCE) loss:

$$L_{GCE} = E_{(\boldsymbol{x},c)\in D}[\frac{1 - Pr(y = c|\boldsymbol{x})^q}{q}] \tag{4}$$

- Author's Conclusion: $q = 0.7$ leads to overall good performance across several experimental settings.

Pre-trained CLIP Generates Effective Class Embeddings

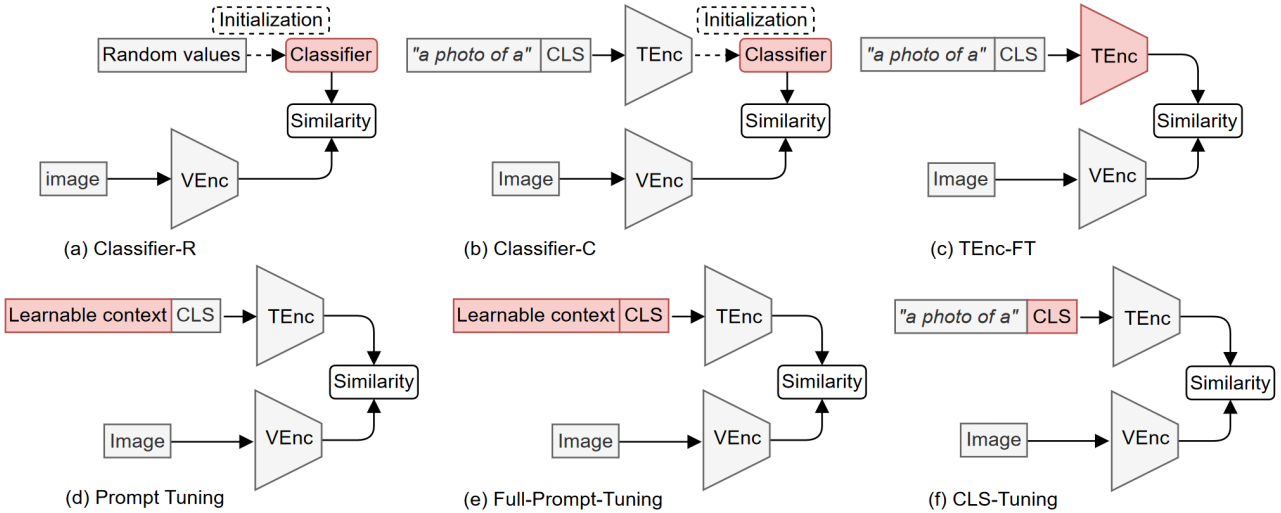

Figure 3: Illustration of different structures for studying the effect of image and text encoders on prompt tuning and prompt design. The blocks highlighted in red are to be trained, while those highlighted in gray are to be frozen.

- Classifier-R v.s. Classifier-C: CLIP class embeddings provide a strong initialization for few-shot learning.
- TEnc-FT v.s. Classifier-C: The highly expressive CLIP text encoder can easily overfit to the noisy labels.
- Prompt Tuning v.s. Classifiers: The text encoder is essential for providing a strong but informative regularization of the text embeddings to combat noisy inputs.
- Prompt Tuning v.s. TEnc-FT: The text encoder should be fixed to prevent overfitting.

Other Aspects of Robustness

- **Effectiveness of Prompt**
- **Prompt Tuning Suppresses Noisy Gradients**
- **Generalization Across Model Architectures**
- **Robustness to Correlated Label Noise**

Improve UPL in Unsupervised Prompt Tuning

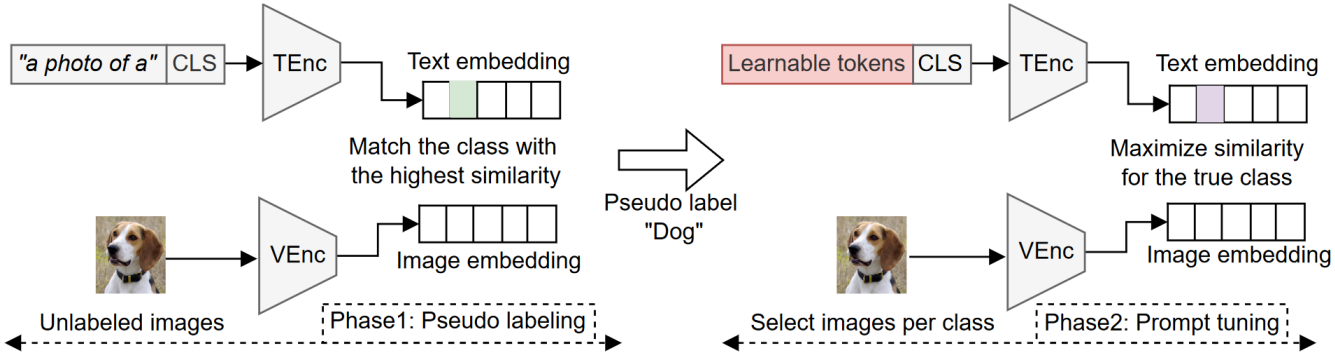

Figure 7: The pipeline of unsupervised prompt tuning. It consists of two main phases: Pseudo labeling and Prompt tuning. To begin, we generate pseudo labels for target datasets by utilizing CLIP with a template prompt for zero-shot transfer. Next, we randomly select samples per class from the pseudo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

- Baseline UPL
  - Phase 1: Leverage pre-trained CLIP to generate pseudo labels for unlabeled images.
  - Phase 2: Select **the K most confident samples per class** to optimize the learnable tokens through the typical prompt-tuning optimization process (described in CoOp).
- Robust UPL
  Based on UPL, **randomly sample K training samples** and optimize the prompt with the **robust GCE loss**.

New Plans for Next Week

1. Reproduce the most of results about this paper.
2. Survey other relavent methods in this domain.

Thanks for your listening!