Convex Optimization Project: Robust Prompt Learning via Optimal Transport and Adaptive GCE

Zihan Wang, Zhichen Zhong, Yixuan Liu, Haoyu Li
School of Information Science and Technology
ShanghaiTech University
{wangzh12023, zhongzhch2023, liuyx2023, lihy2023}@shanghaitech.edu.cn

Abstract

Vision-Language Models (VLMs) like CLIP have revolutionized representation learning but remain vulnerable to noisy labels in downstream tasks. This project focuses on the reproduction and extension of *NLPrompt* [1], a framework that utilizes Optimal Transport (OT) to purify noisy data. We provide a rigorous convex optimization perspective on the OT formulation used in NLPrompt, specifically analyzing it through Bregman Projections. Furthermore, we identify a limitation in the original "hard partition" strategy and propose a novel incremental improvement: **OT-Guided Adaptive Generalized Cross Entropy (GCE)**. Instead of a binary separation of clean and noisy data, our method dynamically adjusts the robustness parameter q of the GCE loss based on OT confidence scores. Our experimental results on benchmarks including Flowers102, DTD, and EuroSAT demonstrate the efficacy of the reproduction and the potential of the proposed soft-weighting mechanism.

1 Introduction

The advent of Vision-Language Models (VLMs) such as CLIP has bridged the gap between visual and textual data [1]. Prompt learning has emerged as a parameter-efficient fine-tuning method for these models [1]. However, real-world datasets are inherently noisy, and standard Cross-Entropy (CE) loss is known to overfit to incorrect labels, leading to performance degradation [1].

The paper *NLPrompt: Taming Noisy Labels in Vision-Language Models* [1] addresses this by employing Optimal Transport (OT) to align image features with text prototypes, thereby identifying and "purifying" noisy labels. The original method partitions data into "clean" and "noisy" subsets, applying CE loss to the former and Mean Absolute Error (MAE) loss to the latter [1, 2].

In this project, we aim to:

1. **Replicate** the NLPrompt framework and verify its performance on standard benchmarks.
2. **Analyze** the theoretical underpinnings of the OT formulation from a convex optimization perspective, specifically focusing on the transport polytope and entropic regularization.
3. **Propose** an incremental innovation: an OT-Guided Adaptive GCE loss that replaces the heuristic hard thresholding with a smooth, instance-dependent weighting mechanism.

2 Theoretical Analysis: A Convex Optimization Perspective

A core contribution of this project is the analysis of the optimization problem underlying the data purification process.

Preprint.

2.1 The Transport Polytope

The purification process relies on finding an optimal transport plan Q. We define the feasible region, the Transport Polytope $\mathcal{U}$, as the intersection of non-negative constraints and marginal constraints:

$$\mathcal{U}(\alpha, \beta) = \left\{ Q \in \mathbb{R}_+^{C \times N} \mid Q\mathbf{1}_N = \alpha,\, Q^\top \mathbf{1}_C = \beta \right\} \tag{1}$$

where α represents the target distribution of classes (prior) and β represents the weight of each sample (typically uniform $1/N$).

2.2 Entropic Regularization and Bregman Projection

The original NLPrompt paper minimizes the transport cost with entropic regularization. We formulate this as:

$$\min_{Q \in \mathcal{U}} \langle C, Q \rangle - \epsilon H(Q) \tag{2}$$

where $C = -\log(\text{Softmax}(S))$ represents the cost matrix derived from the similarity S between image and text features, and $H(Q)$ is the entropy.

From a convex optimization perspective, this problem can be viewed as a **Bregman Projection**. The objective is equivalent to minimizing the Kullback-Leibler (KL) divergence between Q and a Gibbs kernel $K = \exp(-C/\epsilon)$:

$$Q^* = \arg\min_{Q \in \mathcal{U}} D_{KL}(Q \,||\, K) \tag{3}$$

This formulation allows the problem to be solved efficiently using the Sinkhorn-Knopp algorithm, which iteratively projects onto the marginal constraints.

2.3 Handling Class Imbalance

We analyzed the impact of class imbalance on the optimization objective. Intuitively, one might attempt to enforce the class prior π in both the cost matrix and the constraints:

$$\min_{Q} \langle -\log(\Pi P), Q \rangle - \epsilon H(Q) \quad \text{s.t.} \quad Q\mathbf{1}_N = \pi,\, Q^\top \mathbf{1}_C = \frac{1}{N}\mathbf{1} \tag{4}$$

Expanding the objective reveals:

$$\langle -\log(\Pi P), Q \rangle = \langle -\log P, Q \rangle + \sum_{i,j} (-\log \pi_i) Q_{ij} = \langle -\log P, Q \rangle + \text{const} \tag{5}$$

Thus, class imbalance should be handled strictly through the constraint set $\mathcal{U}$ (via $\alpha = \pi$) rather than modifying the cost matrix.

3 Proposed Innovation: OT-Guided Adaptive GCE

3.1 Limitation of Hard Partitioning

The original NLPrompt uses a binary "hard partition": samples are classified as either Clean or Noisy based on whether the OT pseudo-label matches the given label [1]. This approach discards the nuance of the OT transport plan. A sample might be "mostly" clean or "ambiguous," yet it is forced into a binary bucket (MAE or CE).

3.2 Instance-Dependent Generalized Cross Entropy

To address this, we propose a soft adaptation using Generalized Cross Entropy (GCE) [2]. The GCE loss is defined as:

$$\mathcal{L}_{GCE}(f(x), y; q) = \frac{1 - f_y(x)^q}{q} \tag{6}$$

where $q \in (0, 1]$ controls robustness. As $q \to 0$, GCE approaches CE; as $q \to 1$, it approaches MAE.

We introduce an **adaptive parameter** q_i for each image i, derived from the Optimal Transport plan Q^*.

1. **Confidence Extraction:** We normalize the i-th column of the optimal plan Q^* to obtain the conditional probability of the label y_i:

$$a_i = \frac{Q^*_{y_i,i}}{\sum_c Q^*_{c,i}} \in [0, 1] \tag{7}$$

2. **Adaptive Mapping:** We map high confidence (likely clean) to low q (CE behavior) and low confidence (likely noisy) to high q (MAE behavior):

$$q_i = (1 - a_i)^k, \quad k \geq 1 \tag{8}$$

3. **Unified Loss:**

$$\mathcal{L}_i = \frac{1 - f_{y_i}(x_i)^{q_i}}{q_i} \tag{9}$$

4 Experiments and Results

We evaluate our reproduction and proposed method on several benchmark datasets with both synthetic label noise and real-world noisy annotations (Food101N).

4.1 Implementation Details

All experiments are conducted using a pre-trained CLIP (ViT-B/32) backbone. For optimization, we adopt the Sinkhorn algorithm to solve the optimal transport (OT) problem and use stochastic gradient descent (SGD) for prompt tuning. The hyperparameter k in the proposed adaptive GCE loss is set to 1.0 for all experiments.

To rigorously evaluate the efficacy of the proposed method and the baseline models, we conducted a systematic reproduction and integration of several state-of-the-art multimodal learning frameworks. Our implementation is primarily built upon the foundational codebases of **PromptSRC** [5], **MaPLe** [6], and **Visual Prompt Tuning (VPT)** [7].

4.2 Standardization of the Experimental Framework

A significant challenge in reproducing results across different papers lies in the variation of training pipelines. We addressed this by establishing a unified benchmark environment using the `Dassl.pytorch` toolbox, ensuring that all data augmentation, optimizer settings, and evaluation metrics were consistent across all reproduced methods.

- **Data Synthesis and Noise Injection:** Using the data loading infrastructure from **PromptSRC** , we implemented controlled noise injection modules. We generated **Symmetric Noise** by uniformly flipping labels and **Asymmetric Noise** by simulating class-specific confusion (e.g., flipping "dog" to "cat"), matching the settings used in the *NLPrompt* evaluation.
- **Backbone Configuration:** All models utilize a pre-trained CLIP (ViT-B/32) as the frozen backbone. We extracted visual and textual features following the multi-modal alignment protocols defined in *MaPLe* to ensure high-fidelity feature representation.

4.3 Implementation Details of Reproduced Methods

We re-implemented four distinct approaches to provide a comprehensive comparison against our proposed OT-Guided Adaptive GCE:

CoOp (Context Optimization). We reproduced CoOp by introducing $M = 16$ learnable prompt tokens in the textual branch. The implementation follows the "unified context" design from *MaPLe*, where prompts are optimized using standard Cross-Entropy loss. This serves as our primary baseline to demonstrate the performance degradation caused by noisy labels.

GCE (Generalized Cross Entropy). We integrated the GCE loss function into the training loop as a robust alternative to standard CE. By setting the noise-robust parameter $q = 0.7$, we verified its ability to mitigate the impact of outliers, though it lacks the instance-specific adaptability of our proposed method.

JoAPR (Joint Agreement-based Purification). Drawing inspiration from the **PromptSRC** self-regulation mechanism, we implemented JoAPR to utilize the consensus between different views of the data. We leveraged the deep prompting architecture from *VPT* to increase the model's capacity to learn clean patterns before the noise dominates the gradients.

NLPrompt (Original Framework). The core reproduction of NLPrompt involved implementing the **Optimal Transport (OT) purification layer**. We utilized the *Sinkhorn-Knopp algorithm* to solve the entropic-regularized OT problem. The process involves:

1. Computing a cost matrix C based on the cosine similarity between visual features and learnable text prototypes.
2. Iteratively projecting the transport plan Q onto the marginal constraints to obtain pseudo-labels.
3. Performing a "hard partition" to separate the dataset into clean and noisy subsets based on the matching results.

4.4 Verification of the Reproduction

We validated our implementation by comparing the results on **Flowers102** and **EuroSAT**[cite: 31, 35]. As shown in Table **??**, our reproduced NLPrompt achieves 91.13% accuracy under 12.5% symmetric noise, which is highly consistent with the trends reported in the original CVPR paper[cite: 43, 54]. Furthermore, our implementation on the real-world **Food101N** dataset reached 72.9%, confirming the robustness of our code across different domains[cite: 42, 50].

4.5 Quantitative Results

We provide a comprehensive quantitative evaluation of our reproduction. Table 1 details the performance of various methods across six datasets under symmetric and asymmetric noise. Table 2 presents the few-shot learning performance of the CoOp baseline.

Main Results. The following table summarizes the Top-1 accuracy (%) for CoOp, GCE, JoAPR, and our proposed SoftNLP.

Table 1: Top-1 accuracy (%) under Symmetric (Sym) and Asymmetric (Asym) noise at varying ratios (0.125 to 0.75). Missing values are denoted by '-'.

		Symmetric Noise						Asymmetric Noise					
Dataset	**Method**	0.125	0.25	0.375	0.5	0.625	0.75	0.125	0.25	0.375	0.5	0.625	0.75
Flowers102	CoOp	88.0	82.3	75.8	70.6	56.5	34.7	84.2	73.7	58.1	42.1	25.3	12.4
	GCE	88.5	85.2	84.2	82.1	76.7	63.2	85.6	83.0	73.6	64.1	54.2	38.5
	JoAPR	85.7	80.1	74.6	70.5	68.1	50.4	83.1	79.5	72.6	68.1	40.8	15.4
	SoftNLP	-	-	-	-	-	-	-	-	-	-	-	-
DTD	CoOp	55.4	51.5	43.7	37.3	27.5	15.6	55.0	47.8	39.7	29.4	19.9	12.3
	GCE	60.3	58.7	56.1	52.1	44.6	32.3	59.9	57.7	52.2	45.2	30.2	21.9
	JoAPR	57.5	56.3	55.7	52.5	46.8	29.8	52.3	56.6	53.0	46.7	37.1	26.7
	SoftNLP	-	-	-	-	-	-	-	-	-	-	-	-
EuroSAT	CoOp	74.37	68.40	58.77	51.70	41.40	24.80	75.67	64.17	53.27	41.60	29.13	18.53
	GCE	80.03	78.60	72.20	63.07	47.33	33.03	78.37	72.57	60.53	45.23	24.13	11.87
	JoAPR	73.20	59.10	58.90	61.70	36.90	27.40	69.10	67.30	57.50	47.50	33.90	16.40
	SoftNLP	81.40	79.23	74.70	71.17	56.93	40.73	80.43	80.63	76.67	42.70	29.13	10.00
OxfordPets	CoOp	78.9	79.9	80.9	81.9	82.9	83.9	84.9	85.9	86.9	87.9	88.9	89.9
	GCE	85.93	85.10	83.30	77.23	73.80	54.90	85.27	83.53	77.50	67.33	53.47	31.07
	JoAPR	84.0	83.8	83.5	83.1	81.4	74.4	82.7	83.9	79.4	75.4	50.1	42.8
	SoftNLP	86.97	85.13	84.07	83.33	79.23	71.10	86.93	84.90	83.13	77.63	65.07	45.23
UCF101	CoOp	70.3	63.3	55.1	49.4	41.3	26.7	68.9	57.5	45.9	33.0	23.2	13.1
	GCE	74.7	75.2	72.7	68.2	64.5	54.9	74.5	72.0	69.6	60.9	50.4	39.5
	JoAPR	72.7	72.5	70.1	67.6	60.4	50.1	72.1	68.4	64.2	58.4	48.6	42.7
	SoftNLP	73.0	72.0	70.1	69.2	64.3	59.9	73.1	71.2	70.4	64.6	60.1	47.0
Caltech101	CoOp	81.5	78.4	72.3	61.7	52.2	40.6	81.5	72.6	63.7	47.5	33.1	19.3
	GCE	89.5	89.6	88.7	85.4	81.9	79.2	89.4	88.2	85.6	79.9	69.1	63.9
	JoAPR	88.3	88.5	87.5	84.1	81.4	80.1	88.9	88.5	83.2	81.4	80.1	70.5
	SoftNLP	88.6	87.8	87.2	85.1	81.3	80.8	89.4	87.4	86.7	83.9	80.3	71.4

Few-shot Learning. We also investigate the impact of the number of training shots on CoOp's performance under label noise.

Table 2: Few-shot performance (accuracy %) of CoOp under Symmetric (Sym) and Asymmetric (Asym) noise settings.

Dataset / Setting	1-Shot	2-Shot	4-Shot	8-Shot	16-Shot
Caltech101 (Sym)	40.8	36.6	52.2	55.8	61.7
Caltech101 (Asym)	30.0	43.0	40.3	44.6	47.5
DTD (Sym)	17.6	23.3	26.1	28.3	37.3
DTD (Asym)	20.8	20.7	24.7	28.3	29.4
UCF101 (Sym)	27.1	27.0	29.7	36.1	49.4
UCF101 (Asym)	27.3	28.2	33.8	33.9	33.0
Flowers102 (Sym)	10.0	17.7	27.5	43.4	70.6
Flowers102 (Asym)	14.7	23.7	26.3	34.2	42.1

Real-world Noise (Food101N) On the Food101N dataset, our implementation achieved an accuracy of **72.9%**, outperforming the CoOp baseline [1].

5 Conclusion

We successfully reproduced NLPrompt, validating its effectiveness in taming noisy labels for VLMs [1]. We provided a deeper convex optimization analysis and proposed an OT-Guided Adaptive GCE loss to handle label uncertainty beyond binary thresholding.

Acknowledgments and Disclosure of Funding

This project was completed for the SI251 Convex Optimization course at ShanghaiTech University, Fall 2025, under the instruction of Prof. Ye Shi.

References

[1] Bikang Pan, Qun Li, Xiaoying Tang, et al. NLPrompt: Taming Noisy Labels in Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[2] Zhilu Zhang and Mert Sabuncu. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[3] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision (IJCV)*, 2022.

[4] Alec Radford, et al. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, 2021.

[5] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Self-regulating Prompts: Foundational Model Adaptation without Forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[6] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal Prompt Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[7] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual Prompt Tuning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.