OT^{3}P: Optimal-Transport guided Test-Time Adaptation for Vision-Language Models

Yunbei Zhang
Tulane University / New Orleans, LA
yzhang111@tualne.edu

Janet Wang
Tulane University / New Orleans, LA
swang47@tulane.edu

Abstract

We propose a novel algorithm for Test-Time Adaptation (TTA), Optimal Transport-guided Test-Time Visual Prompting (OT-VP), for Language and Vision-Language models. This method aims to enhance the performance of pre-trained models when applied to previously unseen target tasks. Specifically, OT-VP employs Optimal Transport (OT) distance to learn a universal visual prompt, effectively aligning the distribution of the unseen target data with the source distribution, thereby facilitating more accurate model predictions. Upon evaluation, our approach demonstrates a significant improvement in accuracy.

1 Introduction

The remarkable successes of Deep Neural Networks (DNNs) are often tempered by the challenges posed by discrepancies between training and testing data distributions (Recht et al., 2019; Hendrycks and Dietterich, 2019; Koh et al., 2021). Such discrepancies are not uncommon in real-world applications, where variations in data due to natural differences and stylistic changes can significantly impact model performance (Li et al., 2017). This problem has been well-documented across various studies, underscoring the fragility of DNNs in the face of distribution shifts (Hendrycks and Dietterich, 2019). Domain Generalization (DG) has been proposed as a solution to this challenge, aiming to equip models with the ability to generalize to unseen domains by leveraging information from source domains during the training phase (Blanchard et al., 2011; Zhou et al., 2022). Despite considerable efforts in this area, machine learning systems remain susceptible to domain shifts, with studies indicating minimal performance gains over conventional Empirical Risk Minimization (ERM) approaches (Gulrajani and Lopez-Paz, 2020; Mehra et al., 2022).

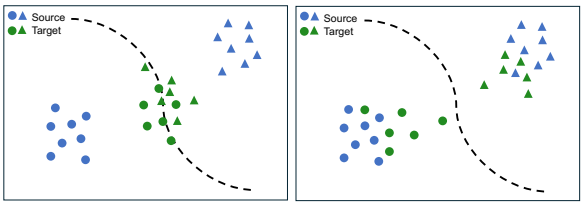

Figure 1: Motivation of our approach. (a) An ERM model trained on the source domain struggles to adapt to the target domain due to domain shifts. (b) Our method optimizes prompt tokens by minimizing the Optimal Transport distance to align the target domain with the source domain without changing the decision boundary.

In response to these limitations, a new line of research has emerged, focusing on enhancing model performance directly at test time (Wang et al., 2021; Iwasawa and Matsuo, 2021). This approach allows models to leverage unlabeled test data from target domains. This data offers insights into the target distribution, insights that are typically inaccessible in the DG framework. Test-time adaptation, as demonstrated in (Iwasawa and Matsuo, 2021), surpasses the capabilities of many existing DG strategies by utilizing immediate, real-world data to refine model accuracy. Inspired by these insights, our work pivots towards exploring test-time adaptation (TTA) as a strategic response to the challenges of domain shifts, aiming to harness the full potential of DNNs in unseen environments. Our motivations shows in Fig. 1

2 Related Work

Optimal Transport Optimal Transport (OT) theory, tracing back to the Monge problem in 1781, evolved significantly with the introduction of the Kantorovich relaxation (Kantorovich, 1942) in 1942. This advancement transformed OT into a robust framework for comparing distributions, shapes, and point clouds (Peyré et al., 2019), lever-

aging the geometry of the underlying space. OT operates on a complete and separable metric space $\mathcal{X}$, utilizing continuous or discrete probability measures $P, Q \in \mathcal{P}(\mathcal{X})$. The Kantorovich formulation defines the OT problem as:

$$\mathrm{OT}_c(P,Q) := \inf_{\pi \in \Pi(P,Q)} \int_{\mathcal{X}\times\mathcal{X}} c(\mathbf{x}_1,\mathbf{x}_2) d\pi(\mathbf{x}_1,\mathbf{x}_2), \tag{1}$$

where $c(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to R^+$ denotes a cost function, and $\Pi(P,Q)$ represents the set of all possible couplings or joint distributions over $\mathcal{X} \times \mathcal{X}$ with P and Q as their marginals. The term $W_p(P,Q) := \mathrm{OT}_c(P,Q)^{\frac{1}{p}}$ is referred to as the p-Wasserstein distance when the cost function $c(\mathbf{x}_1,\mathbf{x}_2) = d(\mathbf{x}_1,\mathbf{x}_2)^p$ for some $p \geq 1$ where d is a metric of $\mathcal{X}$.

In real-world applications, the true marginal distributions P, Q are often unknown, leading to reliance on discrete empirical distributions $\hat{P} = \sum_{i=1}^m \mathbf{a}_i \delta_{\mathbf{x}_1^i}$ and $\hat{Q} = \sum_{i=1}^n \mathbf{b}_i \delta_{\mathbf{x}_2^i}$, with $\mathbf{a}, \mathbf{b}$ as vectors in the probability simplex. The cost function then simplifies to an $m \times n$ cost matrix $\mathbf{C}$, where $\mathbf{C}_{ij} = c(\mathbf{x}_1^i, \mathbf{x}_2^j)$. For computational efficiency, the Sinkhorn algorithm (Cuturi, 2013) introduces an entropic regularizer to the OT problem, facilitating practical applications such as domain adaptation (Courty et al., 2016) and the evaluation of distances between datasets (Alvarez-Melis and Fusi, 2020). This regularized approach, which can be computed using `POT` (Flamary et al., 2021), allows for the computation of an optimal mapping from source to target domains.

3 Methods

3.1 Preliminaries

Problem Definitions Denote the source (target) domain as $\mathcal{D}^s = \{\mathbf{x}_i^s, y_i^s\}_{i=1}^{n_s}$ ($\mathcal{D}^t = \{\mathbf{x}_i^t, y_i^t\}_{i=1}^{n_t}$, where $\mathbf{x} \in \mathcal{X}$ represents the input image and $y \in \mathcal{Y}$ is the label. The dataset $\mathcal{D}^s$ ($\mathcal{D}^t$) comprises samples that are identically and independently distributed (i.i.d.), characterized by some probability distribution $P^s(X,Y)$ ($P^t(X,Y)$).

In the context of Test-Time Adaptation (TTA), the model f is initially trained on the source domain, e.g. minimizing the empirical risk,

$$\arg\min_f \frac{1}{n_s} \sum_{i=1}^{n_s} \ell(f(\mathbf{x}_i^s), y^s) \tag{2}$$

where ℓ is a loss function. Throughout this paper, we refer to optimizing the model with Eq. 2 as ERM. Generally, the model f is structured as a composition $h \circ \phi$, with the feature extractor $\phi : \mathcal{X} \to \mathcal{Z}$ learning the input's representation, and the classifier $h : \mathcal{Z} \to \mathcal{Y}$ predicting the class label.

For any unlabelled target domain $\mathcal{D}^t$, TTA aims to adapt model f or/and target input x^t to bridge the performance gap under the assumption that the source domain and target domain share the same label set. In our approach, we employ a Vision Transformer as the model f, which remains fixed during adaptation.

3.2 Test-time Adaptation with OT^{3}P

In the test-time scenario, labeled target data is not available, thereby the prompt cannot be optimized. Thus, we propose an unsupervised prompt adaptation method: Optimal Transport-guided Test-Time Visual Prompt Adaptation (OT-VP).

Given an unlabelled target dataset $\mathcal{D}^t$, we pass them through the visual encoder with learnable prompts to get the representation. Source representations are computed offline and utilized directly at test time. Then we compute the OT distance between source and target representations. Here, we consider two cost functions for OT: The first is the cost between two representations without label information, i.e., for any two representations $\mathbf{z}^s$, $\mathbf{z}^t$,

$$c(\mathbf{z}^s, \mathbf{z}^t) = \|\mathbf{z}^s - \mathbf{z}^t\|_2 \tag{3}$$

The other includes label/pseudo-label information:

$$c((\mathbf{z}^s, y^s), (\mathbf{z}^t, \hat{y}^t)) = \|\mathbf{z}^s - \mathbf{z}^t\|_2 + \infty \cdot 1_{\{y^s \neq \hat{y}^t\}} \tag{4}$$

where $\hat{y}^t$ is the pseudo label from the pre-trained model. The OT distance will be used to update the prompts for the target dataset as follows:

$$\gamma^* = \arg\min_{\gamma} OT_c(P_{\#}^s, P_{\#}^t) \tag{5}$$

where $P_{\#}^s$ is a distribution over $(\phi(\mathbf{x}^s), y^s)$ and $P_{\#}^t$ is a distribution over $(\phi(\mathbf{x}^t), \hat{y}^t)$ with pseudo label $\hat{y}^t := f(\mathbf{x}^t; \gamma))$

3.3 Vision Transformers

A Vision Transformer (ViT) (Dosovitskiy et al., 2021; Liu et al., 2021) processes an input image x by initially dividing it into k patches $\{I_i\}_{i=1}^k$. An encoding layer E is employed to transform the input patches into patch tokens, to which positional embedding are subsequently added to retain spatial

information. The inputs to the transformer layers consist of these encoded patch tokens augmented with a special classification token `[CLS]`. The ViT is composed of several sequential blocks, and each block contains an attention layer and a Multi-Layer Perceptron (MLP) layer. The prediction of the vision transformer can be formulated as follows:

$$\texttt{[CLS]} = \phi([\texttt{[CLS]}, E(I_1), ..., E(I_k)]), y = h(\texttt{[CLS]}) \tag{6}$$

where $[\cdot]$ represents concatenation of tokens.

Incorporating a visual prompt into the ViT represents a parameter-efficient approach for fine-tuning or adapting the model, particularly when it is fixed (Jia et al., 2022; Ge et al., 2023). By introducing l prompt tokens $\{\texttt{[Prompt]}_i\}_{i=1}^{l} =: \gamma$, the prediction process can be reformulated as follows:

$$\texttt{[CLS]} = \phi([\texttt{[CLS]}, \{E(I_i)\}_{i=1}^{k}, \gamma])y = h(\texttt{[CLS]}) \tag{7}$$

The optimal prompts can be optimized as follows when the labels are available:

$$\gamma^* = \arg\min_{\gamma} E[\ell(f(\mathbf{x}; \gamma), y)] \tag{8}$$

3.4 Optimal Transport

Optimal Transport (OT) theory, tracing back to the Monge problem in 1781, evolved significantly with the introduction of the Kantorovich relaxation (Kantorovich, 1942) in 1942. This advancement transformed OT into a robust framework for comparing distributions, shapes, and point clouds (Peyré et al., 2019), leveraging the geometry of the underlying space. OT operates on a complete and separable metric space $\mathcal{X}$, utilizing continuous or discrete probability measures $P, Q \in \mathcal{P}(\mathcal{X})$. The Kantorovich formulation defines the OT problem as:

$$\mathrm{OT}_c(P, Q) := \inf_{\pi \in \Pi(P,Q)} \int_{\mathcal{X}\times\mathcal{X}} c(\mathbf{x}_1, \mathbf{x}_2) d\pi(\mathbf{x}_1, \mathbf{x}_2), \tag{9}$$

where $c(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to R^+$ denotes a cost function, and $\Pi(P, Q)$ represents the set of all possible couplings or joint distributions over $\mathcal{X} \times \mathcal{X}$ with P and Q as their marginals. The term $W_p(P, Q) := \mathrm{OT}_c(P, Q)^{\frac{1}{p}}$ is referred to as the p-Wasserstein distance when the cost function $c(\mathbf{x}_1, \mathbf{x}_2) = d(\mathbf{x}_1, \mathbf{x}_2)^p$ for some $p \geq 1$ where d is a metric of $\mathcal{X}$.

In real-world applications, the true marginal distributions P, Q are often unknown, leading to reliance on discrete empirical distributions $\hat{P} = \sum_{i=1}^{m} \mathbf{a}_i \delta_{\mathbf{x}_1^i}$ and $\hat{Q} = \sum_{i=1}^{n} \mathbf{b}_i \delta_{\mathbf{x}_2^i}$, with $\mathbf{a}, \mathbf{b}$ as vectors in the probability simplex. The cost function then simplifies to an $m \times n$ cost matrix $\mathbf{C}$, where $\mathbf{C}_{ij} = c(\mathbf{x}_1^i, \mathbf{x}_2^j)$. For computational efficiency, the Sinkhorn algorithm (Cuturi, 2013) introduces an entropic regularizer to the OT problem, facilitating practical applications such as domain adaptation (Courty et al., 2016) and the evaluation of distances between datasets (Alvarez-Melis and Fusi, 2020). This regularized approach, which can be computed using `POT` (Flamary et al., 2021), allows for the computation of an optimal mapping from source to target domains.

4 Results

For the purpose of computational efficiency, we employ simplified versions of the pre-trained BERT and RoBERTa models (Sanh et al., 2020). We focus on two datasets, Stanford Sentiment Treebank (SST) and Yelp (Zhang et al., 2015), each containing five distinct classes. Our baseline approach, denoted as Empirical Risk Minimization (ERM), involves using one of these datasets as the source domain for fine-tuning the pre-trained model, followed by evaluation on the other dataset to assess transferability. Our proposed method introduces the concept of learning prompt tokens tailored to the target task subsequent to the initial pre-training phase. Furthermore, we conduct a comparative analysis with the current Test-Time Adaptation (TTA) method, known as Tent (Wang et al., 2021), to evaluate the efficacy of our approach.

Table 1 demonstrates our method across different pre-trained models and target tasks. Table 2 show the comparison between our method and SOTA on CLIP models.

Model	Algorithm	SST	Yelp
BERT	ERM	34.8	37.3
	Tent	41.2	43.5
	OT^{3}P	47.2	49.3
RoBERTa	ERM	35.1	36.3
	Tent	46.7	41.5
	OT^{3}P	51.3	48.7

Table 1: Comparison of Algorithm Performance on Sentiment Analysis for the Target Domain

5 Discussion and Conclusion

We propose OT^{3}P, a novel test-time adaptation approach that leverages prompt learning for effective

Algorithm	ImageNet V2	ImageNet Sketch	ImageNet A	ImageNet R	OOD Avg.
MaPLe	64.07	49.15	50.90	76.98	60.28
MaPLe+TPT	64.87	48.16	58.08	78.12	62.31
OT^{3}P	65.29	50.23	59.37	79.33	63.55

Table 2: Comparison of Algorithm Performance on Objective Detection Tasks for CLIP Model

test-time adaptation. OT^{3}P stands out by adapting without altering the pre-trained model and effectively aligning the source and target domains, offering a more practical solution than established methods. Our experiments reveal OT^{3}P strengths: consistently outperforming existing TTA methods for ViT and DG prompt learning, and reducing prediction entropy to increase model confidence. By optimizing universal prompts for the target domain, OT^{3}P simplifies the adaptation process, enhancing the applicability of deep learning models in real-world settings.

References

David Alvarez-Melis and Nicolo Fusi. 2020. Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems*, 33:21428–21439.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. 2011. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. 2016. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. 2021. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8.

Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. 2023. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Ishaan Gulrajani and David Lopez-Paz. 2020. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*.

Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Yusuke Iwasawa and Yutaka Matsuo. 2021. Test-time classifier adjustment module for model-agnostic domain generalization. In *Advances in Neural Information Processing Systems*, volume 34, pages 2427–2440. Curran Associates, Inc.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer.

Leonid V Kantorovich. 1942. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. 2017. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Akshay Mehra, Bhavya Kailkhura, Pin-Yu Chen, and Jihun Hamm. 2022. Do domain generalization methods generalize well? In *NeurIPS ML Safety Workshop*.

Gabriel Peyré, Marco Cuturi, et al. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2021. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2022. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415.