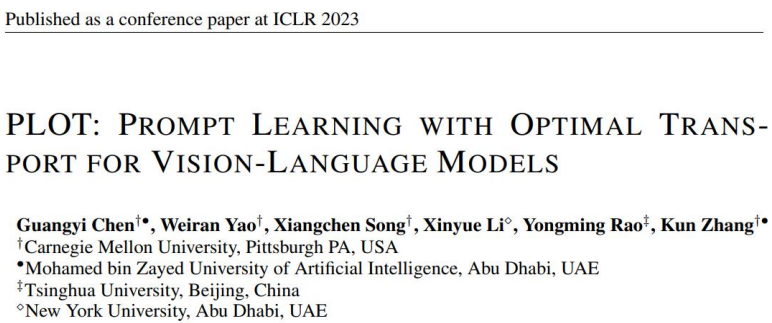
PLOT: Prompt Learning with Optimal Transport for Vision-Language Models

ICLR 2023 Paper Review by Ravialdy

Paper's Information :



Published as a conference paper at ICLR 2023

PLOT: PROMPT LEARNING WITH OPTIMAL TRANSPORT FOR VISION-LANGUAGE MODELS

Guangyi Chen[†•], **Weiran Yao**[†], **Xiangchen Song**[†], **Xinyue Li**[◇], **Yongming Rao**[‡], **Kun Zhang**[†•]
[†]Carnegie Mellon University, Pittsburgh PA, USA
[•]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[‡]Tsinghua University, Beijing, China
[◇]New York University, Abu Dhabi, UAE

Figure 1. PLOT's paper with its title and authors information.

- This paper is about adapting a large-scale pre-trained vision-language model, i.e., CLIP, to be applicable to downstream datasets with prompting technique.
- This paper is published in ICLR 2023 conference and even got notable top 25% (spotlight) award.
- The authors are coming from many universities, such as Carnegie Mellon University, Tsinghua University, etc.

Main source : [Chen'23;ICLR] Chen *et al.*, "PLOT: Prompt Learning with Optimal Transport for Vision-Language Models", ICLR, 2023

Motivation :

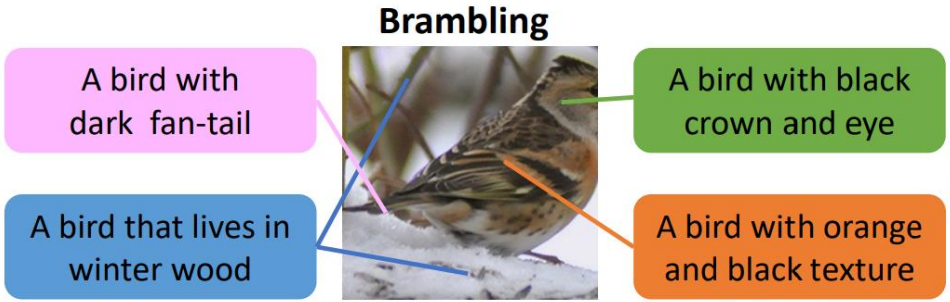


Figure 2. The idea is that a single category, such as "Brambling", can be described from multiple complementary perspectives.

- Existing methods only learn a single prompt to represent a class. This is not enough as one image can be described in different views.
- Directly implementing learnable multiple prompts is problematic because it will lead the model to learn similar characteristics across prompts.
- Thus, we need novel method for aligning visual features and multiple textual prompts in a fine-grained manner.

Main source : [Chen'23;ICLR] Chen *et al.*, "PLOT: Prompt Learning with Optimal Transport for Vision-Language Models", ICLR, 2023

Prompt Learning for Vision-Language Models :

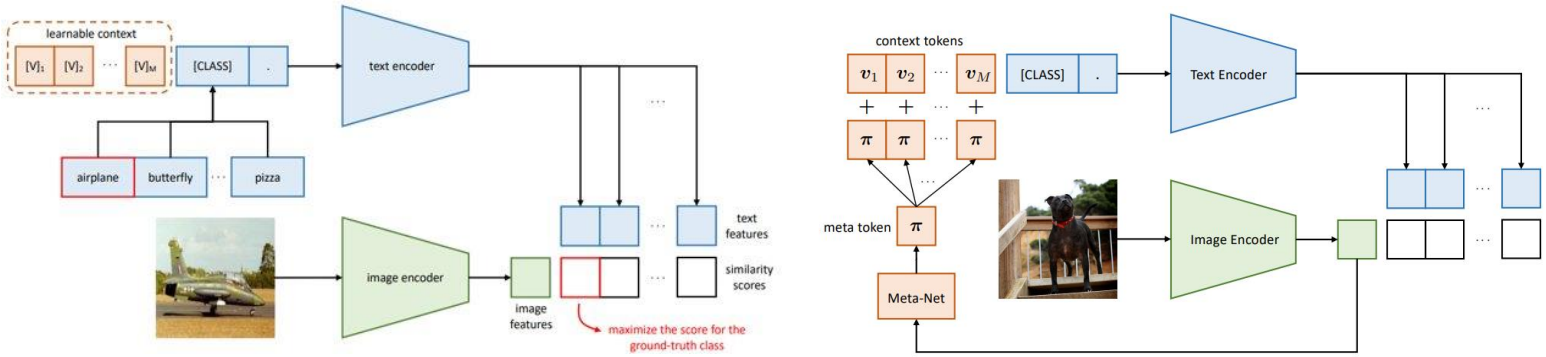


Figure 3. Model's architecture of Context Optimization (CoOp) and Conditional Context Optimization (CoCoOp).

- This topic is about how to automatically learns context prompts in vision-language models.
- There are in total two previous works mentioned in the paper, such as CoOp [1] and CoCoOp [2].
- CoOp uses single learnable vectors as prompts [1], while CoCoOp add that context prompts with image features to be better at domain generalization [2].

Main source: [Chen'23;ICLR] Chen *et al.*, "PLOT: Prompt Learning with Optimal Transport for Vision-Language Models", ICLR, 2023

Main Contributions :

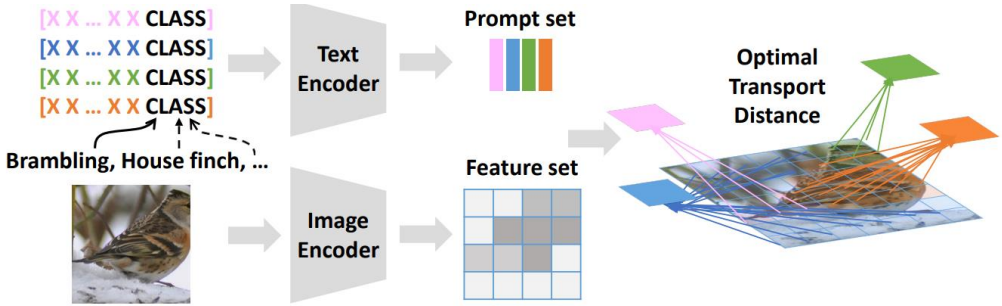


Figure 4. Illustration of PLOT's architecture (more detailed explanation will be delivered in the next slide).

- Introduce a novel method called Prompt Learning with Optimal Transport (PLOT) which applies optimal transport (OT) to align local visual features and multiple textual prompts.
- Introduce a two-stage optimization strategy to first learn transport plan matrix, and then learning the prompts.

Main source : [Chen'23;ICLR] Chen *et al.*, "PLOT: Prompt Learning with Optimal Transport for Vision-Language Models", ICLR, 2023

Brief Recap of Optimal Transport Theory

- Optimal Transport (OT) theory is an optimization problem that deals with transforming one distribution into another in the most efficient way (minimizing the total "cost" of this transportation).
- Specifically, the authors use Sinkhorn distance which introduces an entropy constraint for fast optimization.
- The algorithm above can be expressed as an optimization problem like below :

$$d_{\mathrm{OT},\lambda}(\boldsymbol{u},\boldsymbol{v}|\boldsymbol{C}) = \underset{\boldsymbol{T}}{\text{minimize}} < \boldsymbol{T},\boldsymbol{C} > -\lambda h(\boldsymbol{T})$$
$$\text{subject to} \quad \boldsymbol{T}\mathbf{1}_N = \boldsymbol{u},\ \boldsymbol{T}^\top \mathbf{1}_M = \boldsymbol{v},\ \boldsymbol{T} \in \mathbb{R}_+^{M\times N}$$

Notation meanings :

$\boldsymbol{f}$ is the feature points at distribution A.
$\boldsymbol{g}$ is the feature points at distribution B.
$h(\cdot)$ is an entropy.
$\lambda \geq 0$ is a hyper-parameter.
$\boldsymbol{u}$ and $\boldsymbol{v}$ are the discrete probability vectors that sum to 1.
$\boldsymbol{T}$ is the transport plan, which is learned to minimize total distance.
$\boldsymbol{C}$ is the cost matrix which each point denotes the cost between $\boldsymbol{f}$ and $\boldsymbol{g}$.

Main source : [Chen'23;ICLR] Chen *et al.*, "PLOT: Prompt Learning with Optimal Transport for Vision-Language Models", ICLR, 2023

Main Proposed Scheme :

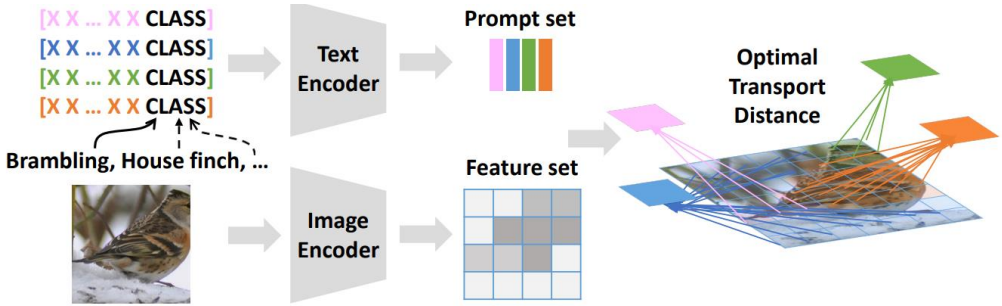


Figure 5. PLOT uses multiple prompts to define each category and creates prompt features with a text encoder. Then, optimal transport method measures the distance between prompts and visual features.

- Two-stage strategy consists of inner loop where the model aims to learn transport plan matrix, and outer loop for learning prompts further based on distance from the supervised data.
- The use of a transport matrix allows each visual feature to be assigned to a weighted combination of prompts, rather than a single prompt -> prompts have a broader semantic capacity.

Main source: [Chen'23;ICLR] Chen *et al.*, "PLOT: Prompt Learning with Optimal Transport for Vision-Language Models", ICLR, 2023

PLOT's Performance :

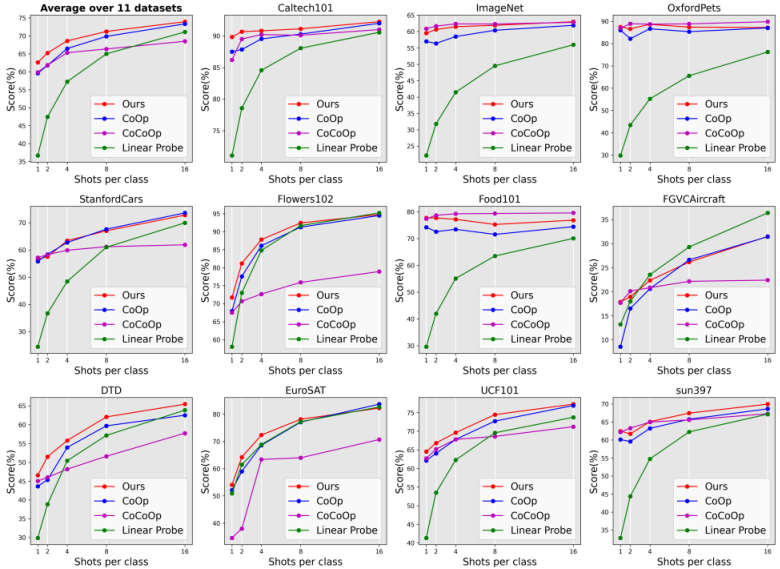


Figure 6. The few-shot learning results on 11 datasets.

- The proposed method PLOT successfully outperform baseline techniques, such as CoOp, CoCoOp, and Linear Probe in many different datasets.

Main source : [Chen'23;ICLR] Chen *et al.*, "PLOT: Prompt Learning with Optimal Transport for Vision-Language Models", ICLR, 2023

Review summary in OpenReview :

- This paper is overall well-structured and easy to follow.
- The motivations behind the work are clear and straightforward, and the proposed optimal transport-based method aligns with these motivations.
- The paper includes comprehensive ablation studies and analysis, with clear illustrations of the pipeline.
- Despite more complicated training, PLOT performs worse than previous prompting methods in some datasets.
- The PLOT's performance compared with more recent work, e.g., CoCoOp, is not clear.

- Conclusion : The paper proposes a technically sound and well-motivated approach, but more comprehensive comparison with state-of-the-art works are needed.

Main source : [Chen'23;ICLR] Chen *et al.*, "PLOT: Prompt Learning with Optimal Transport for Vision-Language Models", ICLR, 2023

My Review :

- The idea of using a diverse set of prompts to generate different text classifiers for a single class prediction is brilliant.
- The results consistently outperform the CoOp model, demonstrating the benefits of using a large set of prompts for better generalization to new classes.
- The paper lacks a discussion comparing their approach with other similar methods, such as CoCoOp.
- The paper also not discuss reasons on why the proposed model PLOT achieve worse performance in some scenarios compared to the baselines.

Main source : [Chen'23;ICLR] Chen *et al.*, "PLOT: Prompt Learning with Optimal Transport for Vision-Language Models", ICLR, 2023

Thank You