LaFTer: Label-Free Tuning of Zero-shot Classifier using Language and Unlabeled Image Collections

- **Sources:** 37th Conference on Neural Information Processing Systems (NeurIPS 2023)
- **Citations:** 17
- **Institutions:** Institute for Computer Graphics and Vision, TU Graz, Austria and 2MIT-IBM Watson AI Lab, USA.
- **Implementation:** https://github.com/jmiemirza/LaFTer

Saman Khamesian

08/21/2024

INTRODUCTION

- Traditional methods for visual classification rely heavily on supervised learning, where models are trained using large datasets of labeled images.

- This approach has proven effective, particularly when plenty of labeled data is available.

- However, obtaining labeled datasets is often expensive, time-consuming, and manually, especially in specialized domains or when dealing with legacy systems like traffic control or medical imaging.

INTRODUCTION

- To overcome these limitations, zero-shot classification has emerged as an alternative.
- In this approach, models like CLIP leverage large-scale Vision-Language (VL) models to classify images into categories based purely on textual descriptions without needing labeled images.
- These models are trained on vast amounts of image-text pairs, enabling them to generalize to new, unseen categories.
- Despite their flexibility, zero-shot classifiers generally underperform compared to fully supervised models because they lack the fine-tuning that supervision provides, leading to a performance gap.

INTRODUCTION

- This gap arises because zero-shot models are not specifically adapted to the target task or domain, making them less accurate in practice.
- They often have difficulty handling specific details unique to certain domains that supervised models can capture more effectively.
- Additionally, the absence of labeled data makes it impossible to use traditional fine-tuning methods, which would otherwise help improve their performance.
- It seems there is gap here and this paper is about this. Let’s see.

INTRODUCTION

- LaFTer addresses this gap by proposing a method to fine-tune zero-shot classifiers without any labeled data.

- Instead of relying on labeled images, LaFTer uses a novel label-free approach that combines unlabeled images with auto-generated text descriptions, leveraging the shared embedding space between text and images.

- This enables the model to achieve performance levels closer to those of fully supervised models, without the need for expensive and time-consuming data labeling.

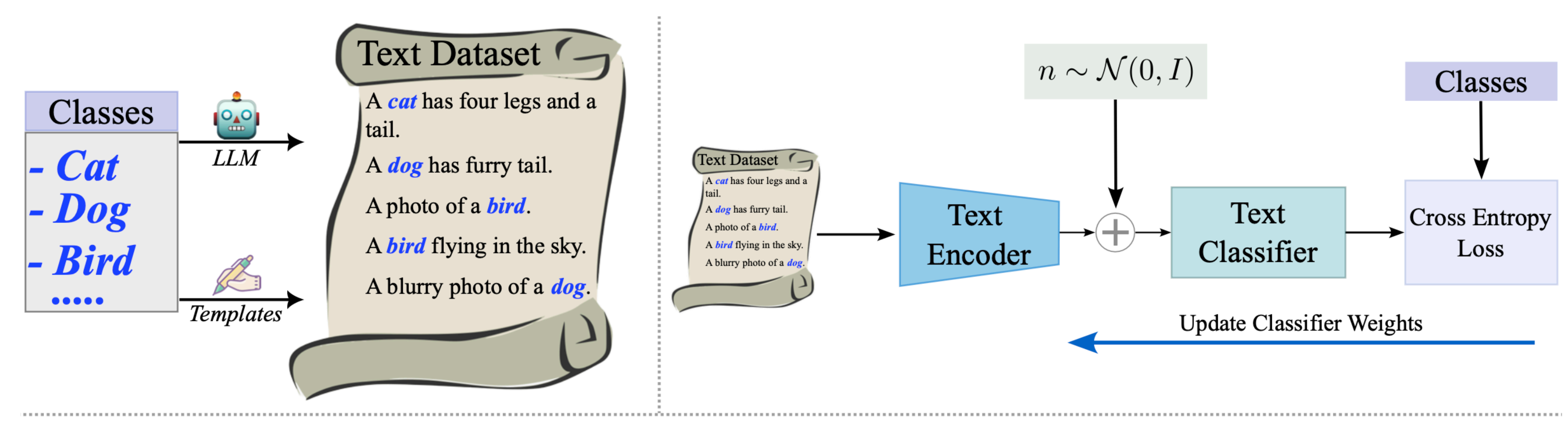
METHODOLOGY

- The process begins with identifying the target classes, such as “cat,” “dog,” or “bird,” which are known beforehand, although the images themselves are unlabeled.

- Using a Large Language Model (LLM) like GPT-3, descriptive text is generated for each class, capturing various aspects and contexts (e.g., “A photo of a cat” or “A cat with a long tail”).

- These descriptions are used to train a text classifier, which learns to associate each text with the appropriate class label.

METHODOLOGY



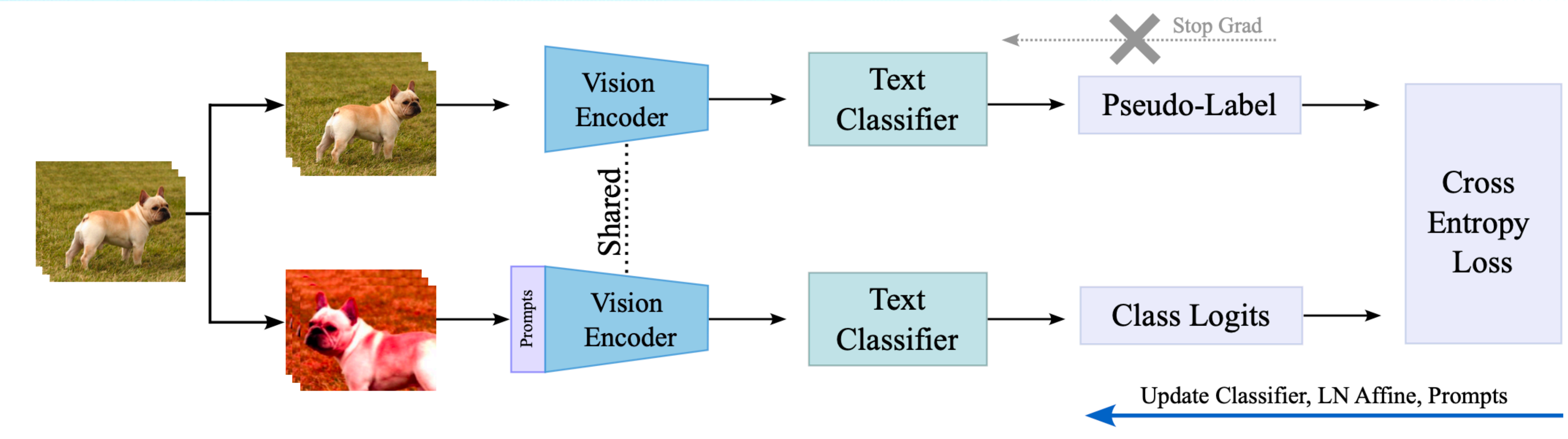
METHODOLOGY

- Once the text classifier is trained, it is used to assign pseudo-labels to the unlabeled images.
- This is done by passing each image through the visual encoder of the VL model to generate image embeddings.
- The text classifier then compares these image embeddings to the text embeddings and assigns the most likely class to each image based on this comparison. But how could be this possible?

METHODOLOGY

- Matching visual embeddings to text embeddings is possible because of how Vision-Language (VL) models like CLIP are trained.
- These models are designed to create a shared embedding space for both images and text.
- During training, the model learns to align the embeddings of an image with the embeddings of its corresponding textual description.

METHODOLOGY



METHODOLOGY

- The next step involves fine-tuning the visual encoder of the VL model using these pseudo-labeled images.
- However, to avoid overfitting and ensure efficiency, only a small portion of the model's parameters is updated during this process.
- This includes parameters such as visual prompts and the scale and shift settings of normalization layers.
- Visual Prompt Tuning is applied here to help the model adapt more effectively to augmented versions of the images.

METHODOLOGY

- Finally, after fine-tuning, the model is tested on new images.
- The fine-tuning process, guided by the pseudo-labels created from the text descriptions, significantly improves the model’s performance, enabling it to classify images with a level of accuracy closer to that of fully supervised models, but without the need for any labeled data.
- This methodology provides an efficient and scalable solution for image classification, especially in scenarios where labeled data is scarce or unavailable.

EXPERIMENTAL SETUP

Dataset

- The experiments were conducted on 12 different datasets from various domains:

  - **Natural Image Datasets:** ImageNet, CIFAR-10, CIFAR-100, Caltech-101.

  - **Specialized Image Datasets:** EuroSat (satellite images), UCF-101 (action recognition), SUN-397 (scene recognition), Flowers-102 (flower classification).

  - **ImageNet Variants:** ImageNet-A (Adversarial), ImageNet-S (Sketch), ImageNet-R (Rendition).

- These datasets cover a wide range of image types and classification tasks.

EXPERIMENTAL SETUP

Baselines

- **CLIP:** Standard zero-shot classification using CLIP's visual and text encoders without fine-tuning.
- **UPL (Unsupervised Prompt Learning):** Fine-tunes CLIP using unsupervised text prompts and offline pseudo-labeling.
- **CLIP-PR:** Optimizes an adapter on top of the CLIP visual encoder using label distribution priors and offline pseudo-labels.
- **CoOp (Learning to Prompt):** A few-shot fine-tuning method that learns soft text prompts using k labeled images per class (1, 5, and 10 shots).
- **PEFT (Parameter Efficient Fine-Tuning):** Fine-tunes the same parameters as LaFTer (prompts, classifier, affine parameters) in a few-shot manner.

EXPERIMENTAL SETUP

Results

	ImageNet	CIFAR-10	CIFAR-100	EuroSat	DTD	CALTECH-101
CLIP	61.9	88.8	64.2	45.1	42.9	90.5
CLIP-PR	60.4	89.3	63.2	44.2	40.1	84.8
UPL	61.2	89.2	65.8	62.2	**48.0**	90.6
LaFTer	**64.2**	**95.8**	**74.6**	**73.9**	<u>46.1</u>	**93.3**
	UCF-101	**Flowers-102**	**SUN-397**	**ImageNet-A**	**ImageNet-S**	**ImageNet-R**
CLIP	61.0	66.6	60.8	<u>29.6</u>	40.6	65.8
CLIP-PR	57.9	57.7	54.7	11.6	38.6	54.1
UPL	63.9	**71.5**	**66.0**	26.9	<u>42.4</u>	65.6
LaFTer	**68.2**	<u>71.0</u>	<u>64.5</u>	**31.5**	**42.7**	**72.6**

Table 1: Top-1 Classification Accuracy (%) while using the CLIP pre-trained ViT-B/32 backbone for 12 image classification benchmarks. LaFTer represents results obtained by first pre-training the visual classifier on text-only data and then performing unsupervised finetuning on the unlabeled image data. Highest accuracy is shown in bold, while second best is underlined.

EXPERIMENTAL SETUP

Results

	ImageNet	CIFAR-10	CIFAR-100	EuroSat	DTD	CALTECH-101
LaFTer (no-shot)	64.2	95.8	74.6	73.9	46.1	93.3
CoOp (1-shot)	60.6	83.0	55.6	58.4	40.1	91.7
CoOp (5-shot)	61.3	86.6	63.2	71.8	41.1	93.2
CoOp (10-shot)	62.3	88.5	66.6	81.6	65.8	94.6
PEFT (1-shot)	50.7	62.7	50.2	37.5	42.6	90.6
PEFT (5-shot)	59.3	80.0	67.3	55.3	59.9	94.5
PEFT (10-shot)	62.8	87.9	74.1	67.9	67.3	96.1

	UCF-101	Flowers-102	SUN-397	ImageNet-A	ImageNet-S	ImageNet-R
LaFTer (no-shot)	68.2	71.0	64.5	31.5	42.7	72.6
CoOp (1-shot)	63.8	71.2	64.1	24.5	39.9	60.0
CoOp (5-shot)	74.3	85.8	67.3	30.0	46.5	61.6
CoOp (10-shot)	77.2	92.1	69.0	35.0	49.1	63.6
PEFT (1-shot)	60.5	66.9	58.3	20.9	38.5	57.2
PEFT (5-shot)	72.6	91.1	68.7	33.3	55.3	66.4
PEFT (10-shot)	79.8	95.2	72.3	40.2	61.1	71.0

Table 2: Top-1 Accuracy (%) for our LaFTer (no-shot) compared to few-shot methods. We compare to CoOp [30] in 1-, 5- and 10-shot supervised finetuning regimes. Parameter Efficient Finetuning (*PEFT*) represents tuning the same parameters as in LaFTer (prompts, classifier, affine) but in a few-shot manner. For each dataset/compared method, blue highlights the highest number of shots outperformed by *no-shot* LaFTer. Notably, LaFTer improves over 10-shot and all compared methods in 4 datasets, including ImageNet, where 10-shot = 10K labeled samples.

EXPERIMENTAL SETUP

Results

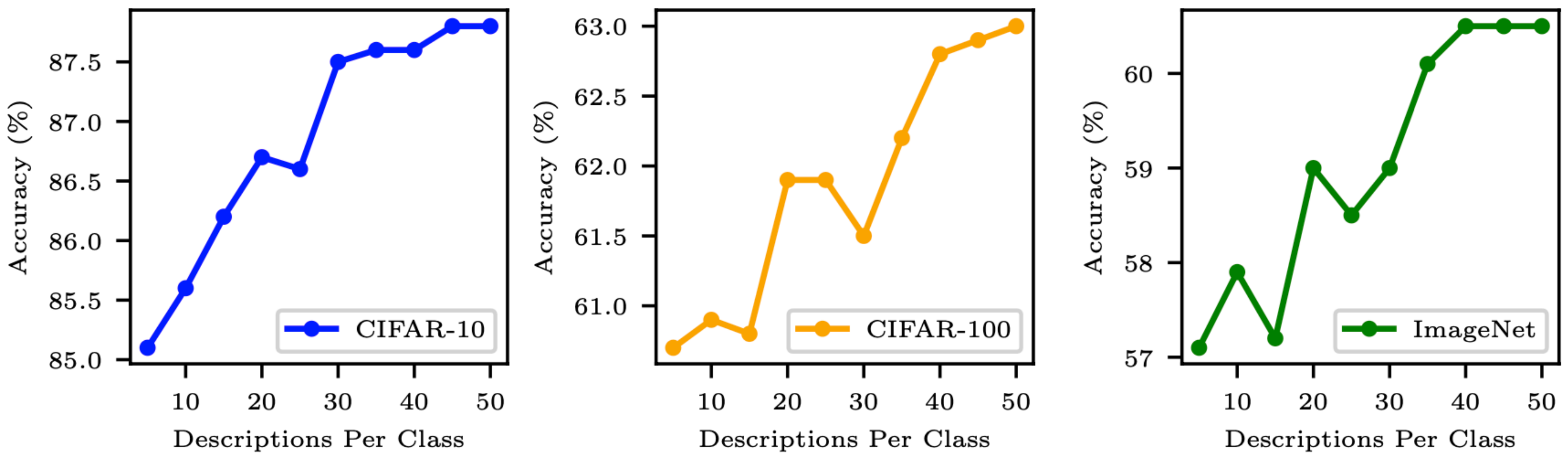


Figure 4: Effect of diversity of descriptions on the Top-1 Accuracy (%). We train the text classifier by randomly choosing (with an increment of 5) a certain number of descriptions per class for each evaluation step. For all the main evaluations, we use a maximum of 50 descriptions per class.

LIMITATIONS

- **Simplicity of the Classifier:** LaFTer uses a single linear layer as the classifier for cross-modal transfer between text and visual data.
- This choice was made to prevent overfitting due to the sparse nature of natural language data. However, this simplicity might limit the model's capacity to capture more complex relationships.
- **Limited Exploration of Complex Structures:** The method did not explore more complex classifier architectures or expand the text dataset further, which could potentially enhance performance. These are left as areas for future research.

LIMITATIONS

- **Dependence on Text Quality:** The effectiveness of LaFTer relies heavily on the quality and diversity of the text descriptions generated by the LLM.
- If the descriptions are not varied or accurate enough, the model's performance might suffer.
- **Application Scope:** While LaFTer shows promise in reducing the performance gap between zero-shot and supervised learning, its application has primarily been tested on specific datasets and scenarios.
- Further experimentation across a broader range of tasks and domains is needed to fully understand its generalizability and limitations.

Thank you for your attention