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

Abstract

Recently, large-scale pre-trained Vision and Language (VL) models have set a new state-of-the-art (SOTA) in zero-shot visual classification enabling open-vocabulary recognition of potentially unlimited set of categories defined as simple language prompts. However, despite these great advances, the performance of these zero-shot classifiers still falls short of the results of dedicated (closed category set) classifiers trained with supervised fine-tuning. In this paper we show, for the first time, how to reduce this gap without any labels and without any paired VL data, using an unlabeled image collection and a set of texts auto-generated using a Large Language Model (LLM) describing the categories of interest and effectively substituting labeled visual instances of those categories.

1 Introduction

Vision and Language (VL) models [1–3] recently became the de-facto standard for generalized zero-shot learning enabling recognition of arbitrary (open) set of categories provided with just their text descriptions and without requiring any additional data or training. However, this incredible flexibility comes at a performance cost and all the VL models, including the most widely used CLIP [1], still require additional supervised training (e.g. tuning its vision encoder) to compete with the closed set (of target categories) supervised training methods. Naturally, this incurs undesirable adaptation costs for those, otherwise very versatile and flexible, foundation models. Image annotations are often expensive to collect, especially for legacy vision systems, such as traffic surveillance, quality control, or security.

In this paper we propose LaFTer, an approach for completely label-free parameter efficient finetuning of VL models to a set of target classes. Our goal is to substitute the need for expensive-to-obtain image annotations with unsupervised finetuning of VL models. We show that since the VL models share a common text-image embedding space (due to their contrastive learning objective), there is a possibility to train using embeddings of samples from one modality (e.g., auto-labeled text) and then successfully apply what we trained to classify embeddings of samples from the other modality (e.g., unlabeled images). More specifically, we show that it is possible to train a neural network (e.g., a classifier) to classify text instances that can be successfully used to classify images, showing successful cross-modal transfer. Instead of collecting labeled visual instances, in our label-free LaFTer approach, we are mining text descriptions of the target categories by prompting an LLM (e.g., GPT-3 [4]) and combining them with handcrafted prompts, as shown in Figure 1. After creating such a text dataset, we train a neural network to classify each text instance (sentence) in it to the source class label that was used to produce the instance. This text classifier can be readily used to classify images when used on top of a CLIP visual encoder. Furthermore, we take advantage of this text-only pre-trained classifier by employing it in a pseudo-labeling pipeline (inspired by FixMatch [5]), to further finetune the CLIP vision encoder on an unlabeled image collection. To reduce overfitting and keep the finetuning parameter efficient, we make use of Visual Prompt Tuning [6] combined with adapting the affine transformations (scale and shift) of the normalization layers in the otherwise frozen network.

2 LaFTer

CLIP [1] consists of a vision encoder and a text encoder, which project images and texts to a common embedding space. It has been trained on a very large volume (400M) of image-text pairs to align the embedding of each image to the embedding of its corresponding text, at the same time pushing it away

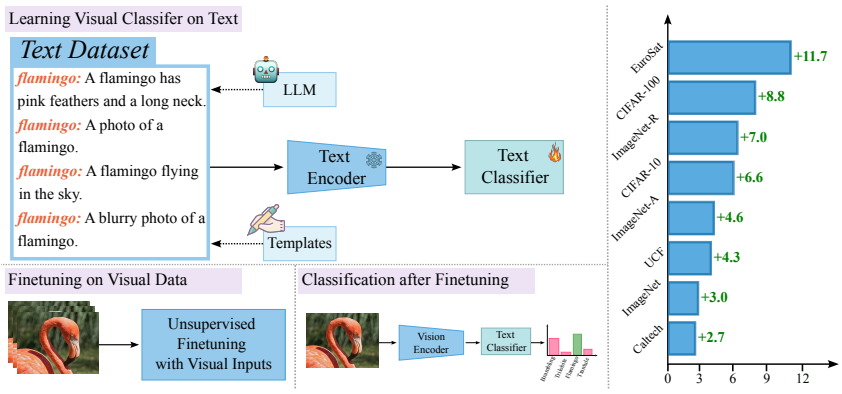

Figure 1: LaFTer proposes to first train a classifier on a natural language text dataset mined in a controlled manner from a set of target classes by generating descriptions for each class label using an LLM and mixing them with handcrafted templates. The training objective is to classify each description to the correct (source) class name (top-left). In the second stage, LaFTer employs the text-only classifier to generate pseudo-labels on the unlabeled data to further finetune the vision encoder in a parameter-efficient manner (bottom-left). Finally, the finetuned visual encoder and text classifier is used for eventual classification (bottom-middle).

from the embeddings of unrelated texts corresponding to other images. In [1], it was demonstrated that the text and vision encoders enable effective zero-shot image classification. Given the set of class names C, the text encoder u is used to generate the embedding u_c of a prompt for each class $c \in C$, typically of the form '*A photo of ...*', completed with the class name. Each test image x is encoded by the vision encoder v and classified using its cosine similarity $\cos$ to the text embedding of each class. The likelihood of a predicted class $\hat{c}$ is computed with a softmax,

$$l_{\hat{c}}(x) = \frac{e^{\cos(u_{\hat{c}}, v(x))/\tau}}{\sum_{c \in C} e^{\cos(u_c, v(x))/\tau}}, \tag{1}$$

where τ denotes the temperature constant.

The zero-shot classifier (1) does not require any training data, but is typically outperformed by networks trained on data from the target domain. In the next section 2.1, we describe a technique to train a visual classifier on purely textual data, conveniently generated by Large Language Models. In Section. 2.2, we propose an unsupervised training setup that lets us benefit from unlabelled data from the target domain, when such data is available. Combining these techniques significantly reduces the performance gap to the supervised classifier and is even competitive with few-shot approaches despite not using any supervision. These results are provided in the Supplementary.

2.1 Learning an Image Classifier using Text

Aligning the text and image embedding spaces is the key idea underlying Vision Language models. It gives CLIP and similar methods, their main advantages: the capacity to be trained on an extremely large volume of data available on the Internet and the resulting effectiveness as zero-shot classifiers. Recently, Zhang et al. [7] explore yet another advantage of the alignment between text and image embedding spaces - it allows diagnosing and rectifying vision models spawn from the VL model vision encoder by using the other modality. In particular, it hints that it might be possible to finetune a zero-shot image classifier on textual data.

While acquiring and annotating images requires manual labor, a training set of annotated texts can be constructed automatically, using a Large Language Model, like the GPT-3 [4]. Generating text using an LLM constitutes a convenient alternative to text mining, as LLMs represent extremely large text corpora on which they were trained. Moreover, prompting LLMs is much more efficient than searching a large body of text for the words of interest. To construct our training set, we take inspiration from [8] and for each class c, we prompted GPT-3 with queries of the following kind: '*Describe what a [c] looks like.*' We repeated the prompting for each class c, in the dataset and complemented the resulting set of synthetic texts with text generated procedurally, using the same hand-crafted templates as for constructing prompts for the zero-shot classifier, for example, '*A photo of a [c].*'. We defer the full list of prompts (to LLMs) and templates to the supplementary material.

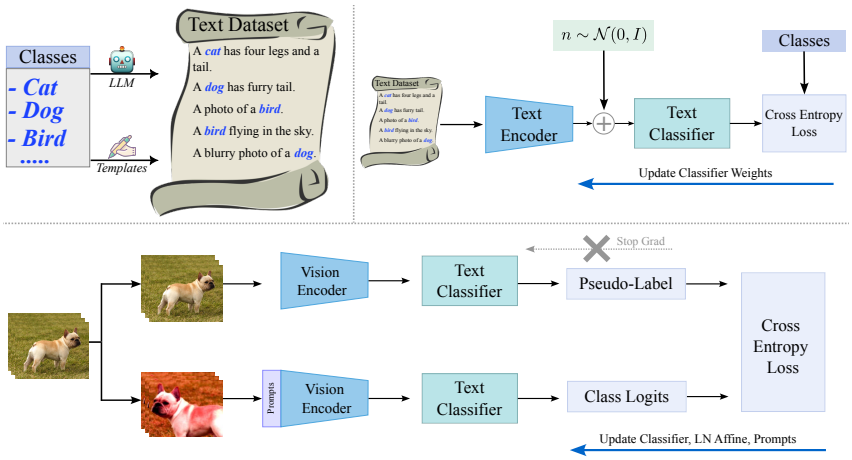

Figure 2: Overview of our LaFTer. (top) Given a set of class labels, we generate a data set of short texts by prompting a Large Language Model (LLM) multiple times with each class name. We compute embeddings of these texts using CLIP text encoder. This lets us train a neural network, the *Text Classifier*, to infer the class used to prompt the LLM from the embedding of the text it generated. Even though the *Text Classifier* has been trained exclusively on text, it performs well in classifying image embeddings generated by CLIP vision encoder. (bottom) We further take advantage of the *Text Classifier* by leveraging it in a pseudo-labeling setup to finetune the VL model.

The capacity to finetune the classifier in a supervised text classification setting lifts the architectural constraints imposed by the zero-shot setup. More precisely, the class likelihoods no longer need to be produced according to (1), but can instead be computed by a trainable classifier f, of arbitrary architecture. We train f with the Smoothed Cross Entropy loss. To regularize training, we add Gaussian Noise $n \sim \mathcal{N}(0, I)$ to the l_2-normalized feature vector from the clip text encoder. Formally, given a text training set T consisted of pairs of text fragments t and class labels c, the training objective is:

$$\min_{\theta} \sum_{\substack{(t,c)\in T \\ n\sim\mathcal{N}(0,I)}} \mathcal{L}_{\text{SCE}}\big(f_\theta(\frac{u(t)}{\|u(t)\|} + n), c\big), \tag{2}$$

where θ is the parameter of the classifier. Training of the text-only classifier is very efficient. For example, 3000 epochs of training the classifier on the data set of 130000 text sentences, representing the 1000 classes of the ImageNet [9] dataset is completed in $\sim$ 120 seconds on an NVIDIA 3090 graphics card. As demonstrated in Supplementary, training the classifier on our dataset of synthetic texts yields a network that matches or outperforms the zero-shot classifier, even though the classifier is initialized randomly.

2.2 Unsupervised Finetuning on Target Domain Images

Text-only training does not require any image data. However, in many applications, unlabeled images from the target domain are readily available, or can be acquired at a low cost. Given a set of unlabeled images we propose to take advantage of the text-only pre-trained classifier and use it in a pseudo-labeling pipeline on top of the vision encoder as demonstrated at the bottom part of Figure 2.

Inspired by Fixmatch [5], for each training image x, we generate two views: the weakly-augmented view $\alpha_w(x)$ and the strongly-augmented view $\alpha_h(x)$, where α denotes a stochastic augmentation function. Contrary to Fixmatch, in our unsupervised finetuning we set α_w as an identity transformation. For α_h we use the augmentations proposed in [10]. The weakly-augmented view serves to generate a pseudo-label. To that end, it is passed through the vision encoder v and the text classifier f, yielding a vector of class probabilities p. Class probabilities give rise to pseudo labels, that we denote by $\hat{c}(x)$. Formally,

$$\hat{c}(x) = \arg\max_{c\in C} p_c, \qquad \text{where} \qquad p = f\Big(v\big(\alpha_w(x)\big)\Big). \tag{3}$$

$\hat{c}(x)$ is not differentiable, and we do not backpropagate the error signal through it during training. The pseudo labels $p_{\hat{c}}$, are used as ground truth for training the network on the heavily-augmented views. Since the vision encoders in the two branches share the weights, the pseudo-labels are generated in an *online* manner and are constantly refined as adaptation is progressing.

When processing the heavily augmented views $\alpha_h(x)$, we augment the vision encoder by visual prompt tuning [6]. That is, we append randomly initialized, learnable parameters (a matrix of size $N_P \times d_v$ where N_P is the number of prompts and d_v is the channel dimension of the vision encoder) to the input of the vision transformer after the initial embedding layer. This helps the network account for the heavy augmentation of the input and, as we show in Supplementary, boosts the performance of the finetuned classifier. We denote the vision encoder with prompting by v^p. The prediction for the heavily-augmented image is obtained as $f(v^p(\alpha_h(x)))$.

The network is trained with the smoothed cross entropy loss $\mathcal{L}_{\text{SCE}}$. We denote the set of unlabelled target domain images by D, and formalize the training objective as

$$\min_{\theta,\eta} \sum_{x \in D} \mathcal{L}_{\text{SCE}}\Big(f_\theta\Big(v^p_\eta\big(\alpha_h(x)\big)\Big), \hat{c}(x)\Big), \tag{4}$$

where θ and η denote the finetuned parameters of the classifier and of the vision encoder, respectively. The parameters η are the visual prompts and the scale and shift (affine) parameters of the normalization layers of the vision encoder. The selection of η is motivated by keeping the adaptation *parameter-efficient*. For LaFTer, the number of trainable parameters are less than 0.4% of the entire model parameters, making adaptation extremely lightweight.

3 Conclusion

We propose a completely label-free finetuning method for Vision-Language models by first showing cross-modality transfer and learning a classifier on natural language inputs which can successfully classify visual data. Later, we leverage this text-only pre-trained classifier in our pseudo-labeling pipeline to further finetune the VL models in a parameter efficient manner. We extensively evaluate our LaFTer and achieve state-of-the-art results when comparing to other methods in an unsupervised finetuning paradigm, while also performing favorably in comparison to methods relying on few-shot supervised learning routines. We refer to the supplementary material for related work, detailed experimental results and discussion.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning Transferable Visual Models from Natural Language Supervision," in *Proc. ICML*, 2021.

[2] Y. Li, F. Liang, L. Zhao, Y. Cui, W. Ouyang, J. Shao, F. Yu, and J. Yan, "Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm," *arXiv:2110.05208*, 2021.

[3] N. Mu, A. Kirillov, D. Wagner, and S. Xie, "Slip: Self-supervision Meets Language-image Pre-training," *arXiv:2112.12750*, 2021.

[4] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," *arXiv:2005.14165*, 2020.

[5] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying Semi-supervised Learning with Consistency and Confidence," in *NeurIPS*, 2020.

[6] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual Prompt Tuning," in *Proc. ECCV*, 2022.

[7] Y. Zhang, J. Z. HaoChen, S.-C. Huang, K.-C. Wang, J. Zou, and S. Yeung, "DrML: Diagnosing and Rectifying Vision Models using Language," in *Proc. ICLR*, 2023. [Online]. Available: `https://openreview.net/forum?id=losu6IAaPeB`.

[8] S. Pratt, R. Liu, and A. Farhadi, "What does a Platypus Look Like? Generating Customized Prompts for Zero-shot Image Classification," *arXiv:2209.03320*, 2022.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A Large-scale Hierarchical Image Database," in *Proc. CVPR*, 2009.

[10] X. Chen and K. He, "Exploring Simple Siamese Representation Learning," in *Proc. CVPR*, 2021.