GOPro: Generate and Optimize Prompts in CLIP using Self-Supervised Learning

Mainak Singha Ankit Jha Biplab Banerjee

Indian Institute of Technology Bombay

BMVC 2023

Abstract

Large-scale foundation models, such as CLIP, have demonstrated remarkable success in visual recognition tasks by embedding images in a semantically rich space. Self-supervised learning (SSL) has also shown promise in improving visual recognition by learning invariant features. However, the combination of CLIP with SSL is found to face challenges due to the multi-task framework that blends CLIP's contrastive loss and SSL's loss, including difficulties with loss weighting and inconsistency among different views of images in CLIP's output space. To overcome these challenges, we propose a prompt learning-based model called GOPro, which is a unified framework that ensures similarity between various augmented views of input images in a shared image-text embedding space, using a pair of learnable image and text projectors atop CLIP, to promote invariance and generalizability. To automatically learn such prompts, we leverage the visual content and style primitives extracted from pre-trained CLIP and adapt them to the target task. In addition to CLIP's cross-domain contrastive loss, we introduce a visual contrastive loss and a novel prompt consistency loss, considering the different views of the images. GOPro is trained end-to-end on all three loss objectives, combining the strengths of CLIP and SSL in a principled manner. Empirical evaluations demonstrate that GOPro outperforms the state-of-the-art prompting techniques on three challenging domain generalization tasks across multiple benchmarks by a significant margin.

Motivation

- We should leverage the pre-trained CLIP backbone while introducing a small set of learnable parameters to learn an SSL-influenced joint image-text embedding space.
- We should replace ad-hoc prompts with learnable prompts to increase generalizability and jointly ensure a better alignment of image-text features.

Contributions

The present study investigates the following objectives:

- In this paper, we strategically enhance CLIP's prompt learning by using an SSL objective together with the notion of disentangled image-domain-conditioned prompt learning.
- Our key contributions involve updating newly-introduced light-weight vision and text projectors atop frozen CLIP using a combination of visual-space SSL contrastive loss, CLIP's image-text contrastive loss, and a novel prompt consistency loss that takes into account the various views of the images. Furthermore, we propose learning the prompt distributions leveraging the multi-scale visual content and style information extracted from CLIP.
- To evaluate the effectiveness of our proposed approach, we conduct extensive experiments across three different settings, including base-to-new class generalization, cross-dataset transfer, and single-source multi-target domain generalization on multiple benchmark datasets. Our GOPro method significantly outperforms other state-of-the-art comprehensively in all the cases.

Architecture of GOPro

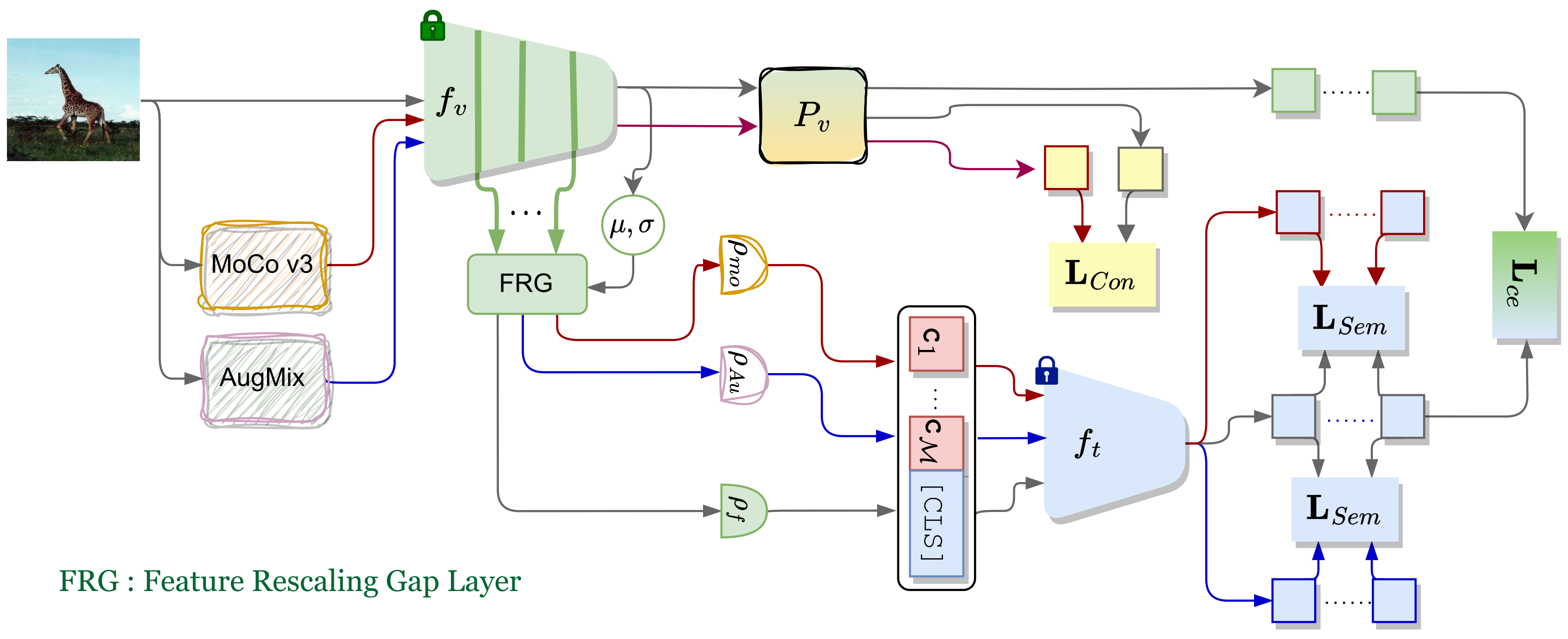


Figure 1. **The design of GOPro** entails utilizing the fixed image f_v and text encoders f_t of CLIP. In addition, GOPro incorporates several distinct trainable meta-networks that generate tokens for the original image and augmented images created by MoCo v3 and AugMix, denoted as ρ_f, ρ_{mo} and ρ_{Au} respectively. To rescale the features from intermediate layers of f_v, the architecture employs a combination of feature rescaling and the global average pooling (GAP) operation, which we collectively referred to as the FRG layer.

Formulation of Metric Objectives

- Cross entropy loss:

$$\mathbf{L}_{ce} = \underset{P_v,\rho}{\arg\min} \underset{(x,y)\in\mathcal{P}(\mathcal{D}_s)}{\mathbb{E}} - \sum_{k=1}^{\mathcal{Y}_{Seen}} y_k log(p(y_k|x)_{f_v,f_t}) \tag{1}$$

- Semantic consistency loss:

$$\begin{aligned}\mathbf{L}_{Sem} = \underset{P_v,\rho}{\text{argmin}} \underset{\mathcal{P}(\mathcal{D}_s)}{\mathbb{E}} &||f_t(\text{Pr}_y(\rho(F(x)))) - f_t(\text{Pr}_y(\rho(F(x_1))))||_2 \\ &+ ||f_t(\text{Pr}_y(\rho(F(x)))) - f_t(\rho(\text{Pr}_y(F(x_2))))||_2\end{aligned} \tag{2}$$

- Total loss:

$$\mathbf{L}_{Total} = \mathbf{L}_{Sem} + \mathbf{L}_{ce} + \mathbf{L}_{Con} \tag{3}$$

A. Results

Table 1. Comparison of GOPro with state-of-the-art methods on B2N generalization on the average metrics over 11 visual recognition datasets. HM represents the harmonic mean.

Method	Base	Novel	HM
CLIP	69.34	74.22	71.70
SLIP	69.77	74.28	71.96
CoOp	82.69	63.22	71.66
CoCoOp	80.47	71.69	75.83
MaPLe	82.28	75.14	78.55
StyLIP	83.22	75.94	79.41
GOPro	**84.21**	**77.32**	**80.62**

Results and discussion

Table 2. Comparison of GOPro with the prompt benchmark methods for domain generalization across datasets. We train the model on ImageNet using 16-shots with CLIP ViT-B/16 and test on 4 other datasets.

	Source	Target			
Method	ImageNet	ImageNetV2	ImageNet-Sketch	ImageNet-A	ImageNet-R
CLIP	66.73	60.83	46.15	47.77	73.96
SLIP	68.01	61.12	46.35	47.54	73.88
CoOp	71.51	64.20	47.99	49.71	75.21
CoCoOp	71.02	64.07	48.75	50.63	76.18
MaPLe	70.72	64.07	49.15	50.90	76.98
StyLIP	72.30	64.28	**50.83**	51.14	76.53
GOPro	**73.27**	**65.35**	50.36	**52.23**	**78.02**

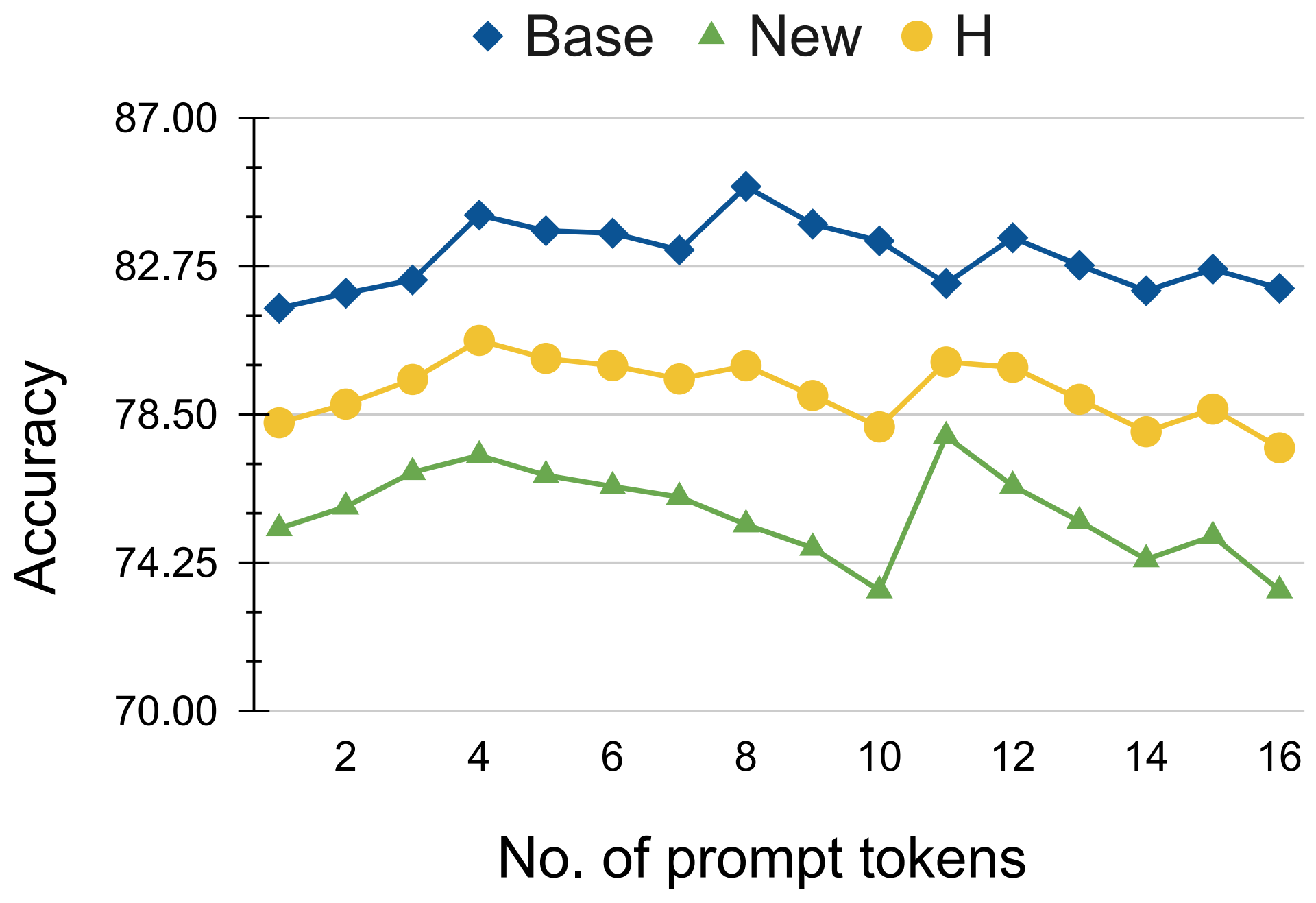
B. Ablation on different loss terms

Table 3. Ablation study of GOPro with different losses in B2N generalization setup.

Loss	Base	Novel	HM
$\mathbf{L}_{ce}$	81.34	72.16	76.48
$\mathbf{L}_{ce} + \mathbf{L}_{Con}$	82.15	75.02	78.42
$\mathbf{L}_{ce} + \mathbf{L}_{Sem}(x_1)$	83.23	74.64	78.70
$\mathbf{L}_{ce} + \mathbf{L}_{Sem}(x_2)$	81.65	73.97	77.62
$\mathbf{L}_{ce} + \mathbf{L}_{Sem}(x_1 + x_2)$	83.87	75.15	79.27
$\mathbf{L}_{ce} + \mathbf{L}_{Sem} + \mathbf{L}_{Con}$	84.21	77.32	80.62

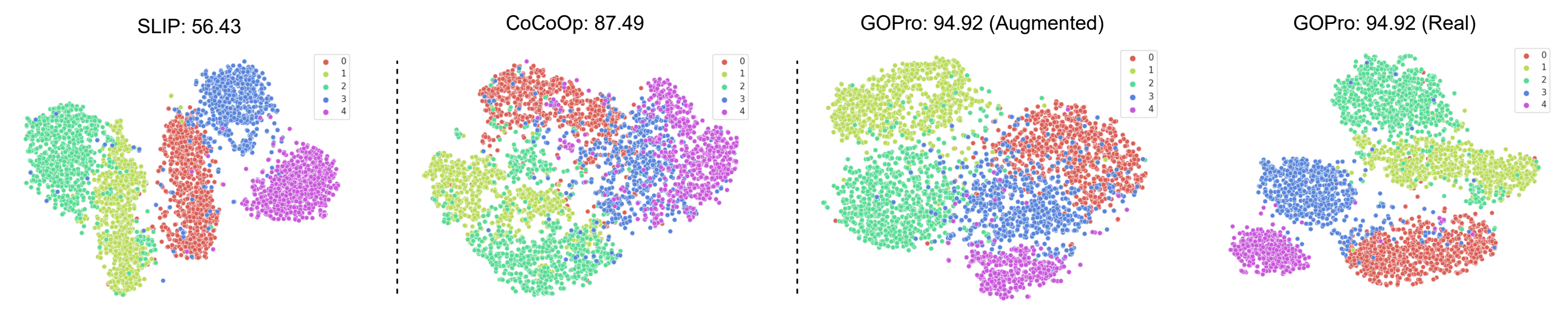
C. Sensitivity to the number of prompt tokens

Figure 2. Comparison of results of GOPro with different numbers of prompt tokens in B2N generalization setup.



D. t-SNE visualization

Figure 3. The t-SNE visualizations of visual embeddings from SLIP, CoCoOp and our proposed GOPro, on the base classes of Eurosat dataset. GOPro archives better discriminativeness.



Conclusions

- We present a comprehensive analysis of how self-supervised learning can enhance vision-language models. We propose a novel approach called GOPro that ensures consistency among the augmented views of input images in both the visual and semantic space of CLIP, using innovative loss functions.
- We introduce a new prompt learning framework in GOPro that leverages visual features by disentangling content and style information and incorporates them into prompt learning through a learnable encoder-decoder-based text projector.
- We are excited to explore the potential of GOPro for more specific applications, such as medical imaging and remote sensing, among others, in the future.

References

[1] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI, pages 529–544. Springer, 2022.
[2] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In Proceedings of the IEEE/CVF international conference on computer vision, pages 1476–1485, 2019.
[3] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16816–16825, 2022.

Authors

Mainak Singha

Ankit Jha

Biplab Banerjee