Supplementary Material for Active Learning for Vision-Language Models

Bardia Safaei
Johns Hopkins University, USA
bsafaei1@jhu.edu

Vishal M. Patel
Johns Hopkins University, USA
vpatel36@jhu.edu

In the supplementary materials, we evaluate our proposed AL method on a large image classification dataset (CIFAR-10). We also provide further details on the comparison baselines, dataset statistics, and the implementation of our method.

1. Evaluations on CIFAR-10

In addition to the experiments on six real-world datasets included in our paper, in Fig.1, we conduct experiments on the CIFAR-10 dataset to show the effectiveness of our approach on large datasets in the AL setting (10 samples/cycle). We observe that our approach outperforms all other competitive baselines in almost all AL cycles, which shows the scalability of our method.

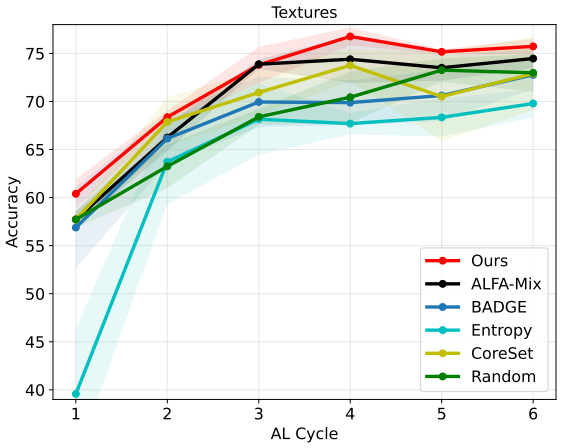

Figure 1. CoOp accuracy results over 6 AL cycles on the CIFAR-10 dataset.

2. Other Prompt Tuning AL Curves

From Fig. 2 and Fig. 3, we observe that our approach outperforms all other baselines in almost all AL cycles using VPT and MaPLe prompt tuning approaches, which shows the generalizability of our method regardless of the prompt tuning approach used.

3. Additional Implementation Details

For VPT experiments, we use a pretrained vision transformer ViT-B16 [5] model as the CLIP's backbone where the feature dimension is set to 512. We initialize prompts with 'a photo of a {}'. The number of context vectors in the vision branch is set to 2. We run all experiments for 50 epochs using the SGD optimizer [8] with an initial learning rate of 0.0025, a momentum of 0.9, a weight decay of 0.005, and a cosine annealing scheduler. The batch size is set to 4 for all experiments. For data augmentations, we perform RandomResizedCrop, RandomFlip, and Normalization.

4. Explanation for N/A Results in Tables

Some of the compared AL baselines require a small randomly initialized labeled set before the first AL cycle. For these methods, we initially perform random sampling and select 1% of the unlabeled data as the initial labeled data. As a result, the performance of these methods is identical to that of random sampling at a budget equal to 1%. We show these duplicate results via '-' in our tables.

5. Comparison baselines

We compare our method against a suite of state-of-the-art AL approaches: **1) ALFA-Mix [11].** This method employs a mixup technique and slightly perturbs unlabeled samples in the feature space. If the small added noise leads to inconsistent model predictions, the unlabeled sample is selected for annotation. **2) BADGE [1].** BADGE employs a hybrid AL strategy that combines both diversity and uncertainty criteria. The idea behind BADGE is that the magnitude of the model's gradient is a measure of uncertainty since it shows the amount of change in the model's weights to correctly classify the sample. **3) GCNAL [3].** This approach identifies dissimilar unlabeled samples to labeled ones using a Sequential Graph Convolution Network. **4) CoreSet [12].** CoreSet selects diverse samples based on the core-set concept. **5) Entropy [14].** It utilizes the model's entropy to choose informative samples. **6) Random.** It queries unlabeled data randomly.

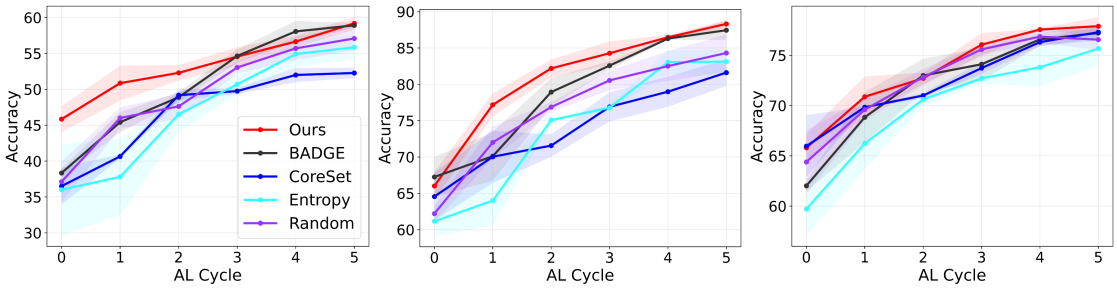

Figure 2. MaPLe accuracy results over 6 AL cycles. From left to right: Textures, Flowers102, and UCF101 datasets.

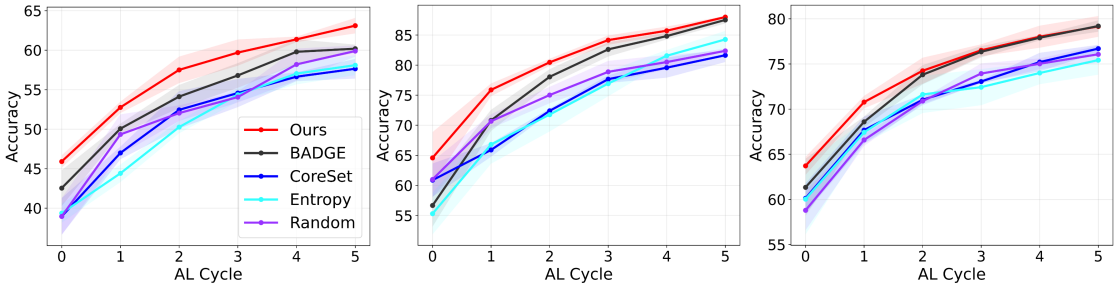

Figure 3. VPT accuracy results over 6 AL cycles. From left to right: Textures, Flowers102, and UCF101 datasets.

Table 1. **Datasets details.**

Dataset	#Categories	#Unlabeled data	#Test data
Textures	47	2,820	1,692
Caltech-101	100	4,128	2,465
EuroSAT	10	13,500	8,100
FGVC-Aircraft	100	3,334	3,333
Flowers102	102	4,093	2,463
UCF101	101	7,639	3,783

6. Datasets

Following prompt tuning works [9, 15], we select 6 different image classification datasets for our experiments, namely Describable Textures [4], Caltech-101 [2], EuroSAT [6], FGVC-Aircraft [10], Flowers-102 [7], and UCF-101 [13]. These datasets cover a range of specialized computer vision tasks that are suitable for evaluating a large pre-trained model like CLIP that has zero-shot capabilities. More information about these datasets is shown in Table 1.

References

[1] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019. 1

[2] Monika Bansal, Munish Kumar, Monika Sachdeva, and Ajay Mittal. Transfer learning for image classification using vgg19: Caltech-101 image data set. *Journal of ambient intelligence and humanized computing*, pages 1–12, 2021. 2

[3] Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. Sequential graph convolutional network for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9583–9592, 2021. 1

[4] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arxiv 2020. *arXiv preprint arXiv:2010.11929*, 2010. 1

[6] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 2

[7] Christopher Kanan and Garrison Cottrell. Robust classification of objects, faces, and flowers using natural image statistics. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2472–2479. IEEE, 2010. 2

[8] Nitish Shirish Keskar and Richard Socher. Improving generalization performance by switching from adam to sgd. *arXiv preprint arXiv:1712.07628*, 2017. 1

[9] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 2

[10] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 2

[11] Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Gholamreza Reza Haffari, Anton Van Den Hengel, and Javen Qinfeng Shi. Active learning by feature mixing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12237–12246, 2022. 1

[12] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017. 1

[13] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[14] Dan Wang and Yi Shang. A new active labeling method for deep learning. In *2014 International joint conference on neural networks (IJCNN)*, pages 112–119. IEEE, 2014. 1

[15] 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. 2