

# Supplementary Materials

## VeCAF: Vision-language Collaborative Active Finetuning with Training Objective Awareness

Anonymous Authors

In the supplementary material, we provide additional information for the main paper. We start by providing details on the methodology of reproducing the previous active learning baselines in Appendix B, and explain the missing results indicated by “-” in the experiment section of the paper. Next, the insights about loss convergence on the CIFAR dataset are meticulously documented in Appendix C.1. We delve into the efficiency of our approach in Appendix C.2, where we provide compelling reasons for the VeCAF usage. Lastly, an extensive evaluation of the model’s accuracy with unrestricted count of training batches is detailed in Appendix C which solidify the robustness of our experimental findings.

We also include the source code of VeCAF in the supplementary material. Please check the README file for details.

### A DATASETS

CIFAR-10 [5] consists of 60,000 images with a resolution of  $32 \times 32$  pixels divided into 10 categories. The training set contains 50,000 images while the test set has 10,000 images. The Caltech101 [2] dataset consists of images from 101 object categories with 40 to 800 images per category. Most classes have around 50 images, and the image resolution is approximately  $300 \times 200$  pixels. ImageNet-1K [1] is a larger dataset with 1,331,167 images belonging to 1,000 classes. The training set consists of 1,281,167 images while the validation set contains 50,000 images. All the training sets from these datasets are considered candidate pools for selection. We also leverage the ImageNet-C [3] as an OOD test set to evaluate our VeCAF under out-of-distribution scenarios. ImageNet-C consists of algorithmically generated corruptions such as blur and noise which are applied to the ImageNet test set. It is used for evaluating the robustness and generalization capabilities of computer vision models.

### B ACTIVE LEARNING BASELINES

In our study, we incorporate three well-established active learning baselines (LearnLoss [8], TA-VAAL [4], and ALFA-Mix [6]) within the pretraining-finetuning paradigm for image classification task. To ensure a systematic and consistent evaluation, all three methods employ a batch selection strategy for sample acquisition during the active learning process.

In Table 1 of our paper, the presence of “-” is attributed to the nature of traditional active learning methods, which require a small initial set randomly sampled at the beginning of the process. It is important to note that the performance of these active learning algorithms on this initial set is comparable to random sampling. Therefore, to avoid redundancy, we have omitted reporting duplicate results for these random initial sets. For instance, we exclude reporting CIFAR-10 results for the sampling ratio of 0.5% and Caltech101 results for the sampling ratio of 1%. Moreover, the ImageNet-1K results with sampling ratios of 1% and 2% are omitted as the size

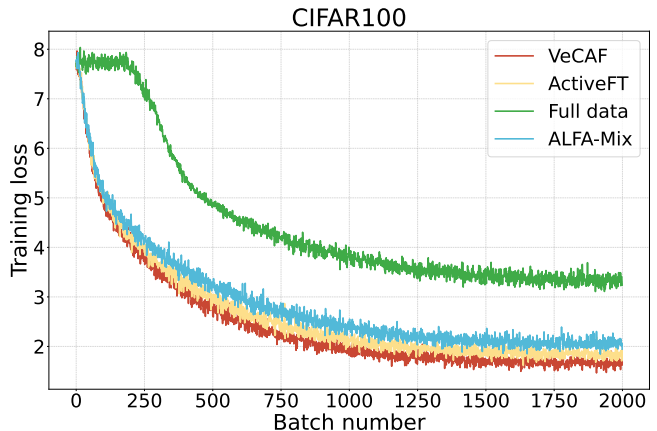


Figure 7: Training curve of VeCAF and baselines including ActiveFT, ALFA-Mix and Full Data FT with 5% of CIFAR-100.

of the initial set is 2.5% according to the reported setting of the corresponding papers. By excluding duplicate results for random initial sets in the smaller sampling ratios, we aim to present clear and concise information in Table 1 and to focus on the most relevant and informative performance metrics for the active learning methods applied in our study.

### C ADDITIONAL EXPERIMENT RESULTS

#### C.1 Loss Convergence Analysis

The training loss on CIFAR-10 converges rapidly for each baseline and, thus, cannot effectively highlight the advantages of the proposed VeCAF. Instead, we present the loss convergence on CIFAR-100, which possesses similar to CIFAR-10 domain characteristics as illustrated in Figure 7. Then, Figure 7 exhibits a comparable to the Figure 4 of the main paper trend in training loss convergence, where the VeCAF not only converges faster but also to a lower loss value in comparison to other baselines. This demonstrates the improved behavior of VeCAF convergence both in terms of speed and performance.

#### C.2 Time Complexity of Data Selection

Efficiency is a crucial aspect of the VeCAF and it is desirable to operate in a time-efficient manner to reduce the overhead of data selection in each training loop. In our study, we evaluate the time required to select various proportions of training samples from Caltech101 with selection ratio 2% and 10% as shown in Table 9. Here we consider the image captions of each training data point are readily available as they can be generated offline and only once, while the time for performing ODS, text embedding generation and CEA

**Table 8: Classification accuracy with the unlimited number of batches. The percentage value on top reports the ratio of data selected during each loop. Top-1 accuracy with a standard error of 3 repetitions is reported, %.**

| Method        | Loop       | CIFAR-10                         |                                  |                                  | Caltech101                       |                                  |                                  | ImageNet-1K                      |                                  |                                  |
|---------------|------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|
|               |            | 1%                               | 2%                               | 5%                               | 2%                               | 5%                               | 10%                              | 1%                               | 2%                               | 4%                               |
| Full Data FT  | single-run | 99.31 $\pm$ 0.01                 |                                  |                                  | 88.24 $\pm$ 0.02                 |                                  |                                  | 82.76 $\pm$ 0.01                 |                                  |                                  |
| LearnLoss [8] | single-run | 90.07 $\pm$ 0.02                 | 93.67 $\pm$ 0.03                 | 95.99 $\pm$ 0.02                 | 62.88 $\pm$ 0.02                 | 73.09 $\pm$ 0.03                 | 83.04 $\pm$ 0.04                 | 52.97 $\pm$ 0.03                 | 60.14 $\pm$ 0.03                 | 61.93 $\pm$ 0.03                 |
|               | multi-run  | 90.25 $\pm$ 0.03                 | 94.21 $\pm$ 0.03                 | 96.32 $\pm$ 0.04                 | 63.74 $\pm$ 0.02                 | 73.25 $\pm$ 0.03                 | 83.31 $\pm$ 0.03                 | 53.66 $\pm$ 0.05                 | 60.49 $\pm$ 0.03                 | 62.32 $\pm$ 0.04                 |
| ActiveFT [7]  | single-run | 92.31 $\pm$ 0.02                 | 95.46 $\pm$ 0.02                 | 98.18 $\pm$ 0.04                 | 73.69 $\pm$ 0.02                 | 81.33 $\pm$ 0.03                 | 86.78 $\pm$ 0.02                 | 56.87 $\pm$ 0.04                 | 63.19 $\pm$ 0.03                 | 66.01 $\pm$ 0.03                 |
|               | multi-run  | 92.95 $\pm$ 0.01                 | 95.87 $\pm$ 0.03                 | 98.54 $\pm$ 0.02                 | 74.22 $\pm$ 0.02                 | 81.88 $\pm$ 0.03                 | 87.04 $\pm$ 0.02                 | 57.11 $\pm$ 0.03                 | 63.46 $\pm$ 0.03                 | 66.21 $\pm$ 0.02                 |
| VeCAF (ours)  | multi-run  | <b>93.87<math>\pm</math>0.02</b> | <b>96.47<math>\pm</math>0.01</b> | <b>98.97<math>\pm</math>0.01</b> | <b>75.36<math>\pm</math>0.01</b> | <b>83.62<math>\pm</math>0.02</b> | <b>87.72<math>\pm</math>0.01</b> | <b>59.41<math>\pm</math>0.02</b> | <b>65.64<math>\pm</math>0.01</b> | <b>68.52<math>\pm</math>0.02</b> |

**Table 9: Running time to select various percentages of samples from the Caltech101 training set for each data selection loop.**

| Sel. ratio | ALFA-Mix | LearnLoss | ActiveFT | VeCAF  |
|------------|----------|-----------|----------|--------|
| 2%         | 6m45s    | 1m42s     | 12.02s   | 16.38s |
| 10%        | 52m31s   | 23m17s    | 13.36s   | 18.87s |

are included in the reported estimates. Conventional active learning algorithms such as LearnLoss [8] and ALFA-Mix [6] require multiple trial model updates to gradually adjust the selected data, where these trial updates constitute the majority of the time in the data selection process which make them relatively inefficient. In contrast, ActiveFT and VeCAF perform sample selection in a single pass at the beginning of each data selection loop which eliminates the need to perform trial model updates. This results in significant time savings compared to conventional approaches. The minor increase in time of VeCAF over ActiveFT is caused by the text embedding CEA process. Considering the > 150s model training time in each loop, the 4-5 second additional time overhead (3%) is negligible and also justifiable due to the advantages introduced by VeCAF.

### C.3 Accuracy with Unlimited Training Batches

This work focuses on an efficient training paradigm, and accordingly, we have presented most of our experimental outcomes in the main paper using the fixed training cost (i.e., number of batches). This approach inherently benefits methodologies that enable quicker convergence. To thoroughly assess VeCAF's convergence efficacy, we have lifted the constraints on the number of training batches in this section and conducted a comparative analysis of VeCAF's final convergence metrics against established active learning baselines. The results in Table 8 demonstrate that VeCAF not only achieves expedited convergence but also surpasses previous active learning methods in terms of final performance metrics. Remarkably, VeCAF attains performance on par with comprehensive finetuning by utilizing only 5% of the data for CIFAR-10 and 10% for Caltech101. These findings suggest that VeCAF is capable of significantly enhancing both computational and data efficiency throughout the PVM finetuning procedure.

## REFERENCES

- [1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [2] Li Fei-Fei, Rob Fergus, and Pietro Perona. 2004. Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.
- [3] Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- [4] Kwanyoung Kim, Dongwon Park, Kwang In Kim, and Se Young Chun. 2021. Task-Aware Variational Adversarial Active Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [5] A. Krizhevsky and G. Hinton. 2009. Learning multiple layers of features from tiny images. *MS thesis, University of Toronto* (2009).
- [6] Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Gholamreza Reza Haffari, Anton Van Den Hengel, and Javen Qinfeng Shi. 2022. Active Learning by Feature Mixing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [7] Yichen Xie, Han Lu, Junchi Yan, Xiaokang Yang, Masayoshi Tomizuka, and Wei Zhan. 2023. Active Finetuning: Exploiting Annotation Budget in the Pretraining-Finetuning Paradigm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [8] Donggeun Yoo and In So Kweon. 2019. Learning Loss for Active Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.