CLIP Adaptation by Intra-Modal Overlap Reduction

Alexey Kravets
ak3095@bath.ac.uk
Vinay Namboodiri
vpn22@bath.ac.uk

Department of Computer Science
University of Bath
Bath, UK

Abstract

Numerous methods have been proposed to adapt a pre-trained foundational CLIP model for few-shot classification. As CLIP is trained on a large corpus, it generalises well through adaptation to few-shot classification. In this work, we analyse the intra-modal overlap in image space in terms of embedding representation. Our analysis shows that, due to contrastive learning, embeddings from CLIP model exhibit high cosine similarity distribution overlap in the image space between paired and unpaired examples affecting the performance of few-shot training-free classification methods which rely on similarity in the image space for their predictions. To tackle intra-modal overlap we propose to train a lightweight adapter on a generic set of samples from the Google Open Images dataset demonstrating that this improves accuracy for few-shot training-free classification. We validate our contribution through extensive empirical analysis and demonstrate that reducing the intra-modal overlap leads to a) improved performance on a number of standard datasets, b) increased robustness to distribution shift and c) higher feature variance rendering the features more discriminative for downstream tasks.

1 Introduction

Vision-language models (VLMs) represent a novel approach in artificial intelligence integrating the understanding of both visual and textual information. An exemplary model within VLMs is CLIP [27]. The fundamental strategy with the advent of large foundation models is to train models with a large number of parameters using vast amounts of data. The training of CLIP follows the same approach. Its task involves learning to match images with their corresponding textual descriptions through contrastive learning. This strategy has resulted in exceptional performance in zero-shot scenarios and requires minimal adaptation across various tasks including semantic segmentation [40], out-of-distribution detection [22], and classification [10, 34]. However, when we try to use this foundational model on a dataset whose distribution is significantly different from the pre-trained data, we observe that the performance is not so good. For instance, zero-shot classification performance of CLIP model on EuroSAT dataset is only 48.38%. To address this the conventional solution involves collecting a training dataset. However, collecting a large training dataset is often impractical and expensive leading to a collection of only a few examples per class. As few examples are available some approaches [10, 24, 30, 34, 36, 38, 39, 41] suggest adapting CLIP by

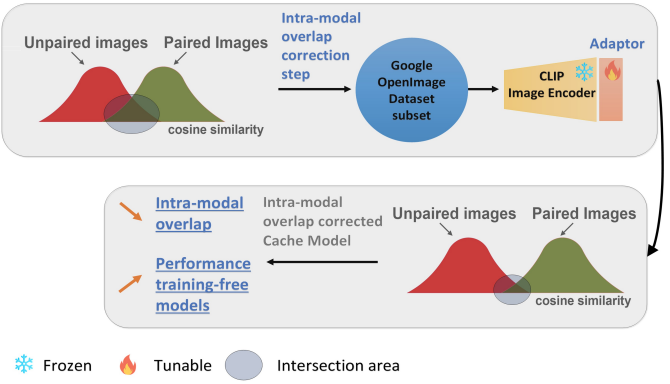


Figure 1: Overview of the process. First, we perform a intra-modal overlap correction step of CLIP image encoder through adaptation. Then, this new image encoder is used to create intra-modal overlap corrected cache model that can be used in any training-free method improving its performance.

adjusting additional parameters while keeping the original ones frozen. Sometimes, training even small adapters can be infeasible. Thus, alternative approaches [31, 35, 37] propose a training-free adaptation of CLIP. This involves creating a cached model [37] using CLIP encoded few-shot training images for which labels are available. This cached model can be used to compare a test image to cached images in image space determining the similarity of the test image to few-shot training examples which in combination with zero-shot CLIP logits determines the correct prediction. However, comparing images in the image space with embeddings from CLIP is problematic due to the contrastive training that maximizes the cosine similarity between paired image and text (inter-modal), but ignores the image-image similarity (intra-modal). This results in a substantial intra-modal overlap (IMO) between unpaired (images of different classes) and paired images (images of the same class) compromising the results of training-free methods that use the cached model.

We propose a simple approach to address this issue as illustrated in Fig. 1. The approach is to train a lightweight adapter on a subset of Google Open Images dataset [1] for one epoch. This subset has a different distribution from most of the downstream datasets we test on measured using Proxy-A-Distance [2] measure of divergence. We observe that this simple adaptation step successfully solves the IMO such that the distance between the similarity distributions of paired and unpaired image embeddings successfully increases for many downstream datasets. This approach is thus generalizable and also results in substantially improved performance (for instance performance improvement of around 5% for one-shot performance on EuroSat dataset taking it to more than 68% with a single example compared to 48.38% with zero-shot, cf - detailed table in supplementary material) in many of the downstream datasets. This improvement is complementary to existing approaches and by combining our contribution with [37] and [31] results in a consistent overall improvement in accuracy. In this work we mainly focus on fine-grained datasets where the samples are related making classification particularly challenging, but for completeness we perform experiments on some not fine-grained classification datasets whose results will be included in the appendix.

To summarize, through this paper we make the following **contributions**:

- We propose a novel method based on lightweight adaptation that reduces IMO in CLIP directly in the image space with new features being compatible with any training-free method that utilizes a cached model. These new features improve overall performance in all the training-free methods examined.
- We show that there is a positive relation between direct IMO reduction and performance.
- We explore the possibility to reduce the IMO by training a lightweight adapter in both supervised and self-supervised manners.

2 Related Work

Lightweight Adaptation Lightweight adaptation is a fine-tuning approach where the majority of parameters in pre-trained models remain fixed and only a small fraction undergoes tuning. While some lightweight adaptation techniques, like prefix-tuning [20], are specific to Natural Language Processing (NLP), many are versatile and applicable to both NLP and vision models. In [16] authors add sequentially two additional adapter modules inserted in each transformer layer after the projection following the Multi-Head Attention (MHA) layer and the second Multilayer Perceptron (MLP) layer. Each adapter comprises a bottleneck MLP with non-linearity and a residual connection. [26] simplify it further by inserting bottleneck adapter only after the second MLP layer, specifically after the LayerNorm. Low-Rank Adaptation (LoRA) [17] utilizes low-rank factorization to fine-tune attention weights, significantly reducing the number of parameters during adaptation. AdaptFormer [5] introduces a bottleneck MLP layer after the MHA of a transformer layer. This newly added MLP layer is parallel with the original MLP and the two are connected via a residual connection with a scale factor.
In this study we utilize adapters not for a downstream task adaptation but specifically to address IMO. Furthermore, our focus is on vision adaptation for CLIP vision encoder which is affected by IMO. We are not interested in reducing the intra-modal overlap in text space as text to text matching is not utilized to perform few-shot classification.

Few-shot Classification Methods We can categorize methods utilizing CLIP for few-shot classification into three different groups. Firstly, there are methods like [10, 24, 30, 34, 36, 38, 39, 41] that involve training. These methods use few-shot examples to adjust additional parameters while keeping the original CLIP parameters fixed. Secondly, there are zero-shot methods, such as [12, 27], which do not introduce any extra parameters to CLIP and do not necessitate training. Lastly, there are training-free methods or hybrid methods that are training-free but also might have a training counterpart. In this work we specifically focus on training-free methods [31, 35, 37], excluding their training counterparts. As all of them utilize the cached model component for prediction which is affected by the IMO [31], we show that replacing it with our IMO corrected cache model component improves performance in all the training-free methods.

Self-supervised Learning in Images Self-supervised learning (SSL) involves learning representations from unlabeled data without explicit annotations which is especially valuable when obtaining data labels is costly. While supervised models generally perform better, self-supervised trained models, particularly those based on the contrastive learning paradigm

have shown superiority in tasks like segmentation and detection and have been closing the gap in other tasks [4, 13]. Notable methods include SimCLR [6] which relies on contrastive learning and requires a large batch size to incorporate a sufficient number of negative examples, MoCo [13] which utilizes a queue mechanism to store negative samples, and BYOL [11] which introduces a novel paradigm eliminating the need for negative samples. DINO [4], like BYOL, relies on positive samples but utilizes cross-entropy loss rather than L2 loss. While SSL methods for training entire networks have been extensively studied there is no exploration training adapters using these methods. We utilize the state-of-the-art DINO method for this purpose and investigate the possibility of training adapters in a self-supervised manner to reduce IMO in CLIP.

3 Background on Training-free Adaptation

In this section we provide an overview of training-free adaptation methods for CLIP.

3.1 Tip-Adapter: the Main Building Component in Training-free Methods

Zero-Shot CLIP Given N classes, CLIP encodes them inside a contextual prompt such as *A photo of a {class}* with the text encoder obtaining $W \in \mathbb{R}^{N\times d}$ classifier weight matrix where d is the embedding dimension. Then, given a test image I_i, it is encoded with CLIP image encoder VE:

$$T_i = VE(I_i),\ T_i \in \mathbb{R}^d \tag{1}$$

After that, we calculate the dot product between W and T_i to obtain the zero-shot classification logits:

$$\text{CLIPlogits} = T_i W^T,\ \text{CLIPlogits} \in \mathbb{R}^N \tag{2}$$

Tip-Adapter Given N classes K shots training dataset with images $I_k, k \in \{1, NK\}$, we encode them with CLIP image encoder. Such encoded images act as keys and their corresponding one-hot encoded labels $L_k, k \in \{1, NK\}$ as values to form the key-value cached model:

$$\begin{aligned} T_k &= VE(I_k), k \in [1, NK], T_k \in \mathbb{R}^d \\ F_{train} &= \text{Concat}([T_1, T_2, .., T_{NK}]), F_{train} \in \mathbb{R}^{NK\times d} \end{aligned} \tag{3}$$

$$\begin{aligned} L_k &= \text{OneHot}(L_k), k \in [1, NK], L_k \in \mathbb{R}^N \\ L_{train} &= \text{Concat}([L, L_2, .., L_{NK}]), L_{train} \in \mathbb{R}^{NK\times N} \end{aligned} \tag{4}$$

The cached model contains the new knowledge extracted from the few-shot training examples and its purpose is to enhance the prior knowledge of the CLIP model. During the testing phase, when presented with a test image denoted as I_i, which serves as a query, this image is encoded using the CLIP image encoder VE resulting in a vector representation $T_i \in \mathbb{R}^d$. Subsequently, an affinity matrix is computed. This matrix represents the similarity between the test image and all the NK training images:

$$A = exp(-\beta(1 - T_i F_{train}^T)), A \in \mathbb{R}^{NK} \tag{5}$$

The exponential function makes affinity matrix values non-negative and β is a hyper-parameter that modules its sharpness.
After obtaining the affinity matrix and zero-shot CLIP logits we can compute the Tip-Adapter logits by combining the new knowledge of the cached model represented by the product between the affinity matrix and labels matrix L_{train} and the prior knowledge of CLIP:

$$\text{TAlogits} = \alpha A L_{train} + T_i W^T,\ \text{TAlogits} \in \mathbb{R}^N \tag{6}$$

With α being a hyper-parameter that weights the importance of the new and prior knowledge.

3.2 Tip-X: Inter-modal Bridge for Intra-modal Overlap Correction

Authors in [31] propose to use inter-modal distances as a bridge to handle intra-modal overlap (IMO) between paired and unpaired samples in the image space. They construct an affinity matrix similarly to Tip-Adapter but in the image-text space where the similarity measure between two images is given by Kullback-Leibler (KL) divergence instead of the cosine similarity like in Tip-Adapter.
Given test image embedding $T_i \in \mathbb{R}^d$, classifier weight matrix $W \in \mathbb{R}^{N\times d}$, CLIP encoded few-shot training images $F_{train} \in \mathbb{R}^{NK\times d}$ and their one-hot encoded training labels $L_{train} \in \mathbb{R}^{NK\times N}$ we compute classes probability distribution for train images and the test image:

$$\begin{aligned} S &= \text{SoftMax}(F_{train} W^T), S \in \mathbb{R}^{NK\times N} \\ s_i &= \text{SoftMax}(T_i W^T), s_i \in \mathbb{R}^N \end{aligned} \tag{7}$$

The affinity matrix M is then constructed by calculating the KL divergence between the test image s_i and the training images S. It tells us how closely the distribution of a given test image aligns with the distribution of the training images in the image-text space:

$$M_{i,j} = KL(s_i || S_j), j \in [1, NK] \tag{8}$$

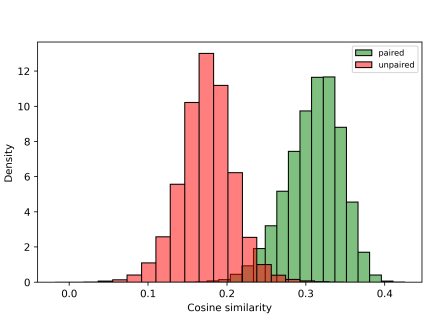
Next, we take the negative of the affinity matrix M because KL divergence is close to 0 for similar images and increases for dissimilar images. It is also rescaled to ensure that it falls within the same range as the Tip-Adapter's affinity matrix. Finally, Tip-X logits are computed by taking the product of the rescaled affinity matrix and the labels matrix L_{train} weighted by a scaler γ which is combined with Tip-Adapter logits weighted by a scaler α and CLIP logits to arrive to the final *TXlogits*:

$$\text{TXlogits} = T_i W^T + \alpha A L_{train} + \gamma \phi(-M) L_{train}, \text{TXlogits} \in \mathbb{R}^N \tag{9}$$

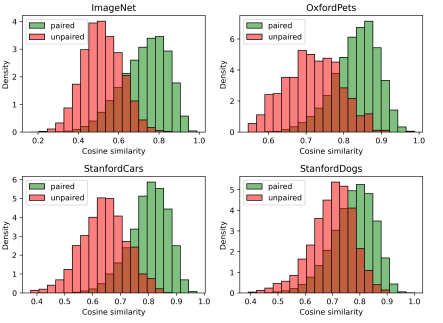
While the authors of Tip-X have achieved superior results compared to the original Tip-Adapter, they still incorporate Tip-Adapter logits into the final prediction, which are influenced by the IMO. We later show that replacing this component with IMO-corrected features further improves the results of Tip-X.

3.3 Adaptive-Prior Refinement

A recent work [35] proposes an alternative training-free method to select more discriminative features by eliminating certain feature channels based on a prior refinement module. This method, however, does not reduce the IMO. Hence, we discuss it and provide comparisons in the supplementary material.



(a) Inter-modal similarity



(b) Intra-modal similarity

Figure 2: Fig. (a) shows the inter-modal cosine similarities on the ImageNet validation set. Fig. (b) demonstrates the intra-modal cosine similarities for different datasets on the validation set.

4 Approach

4.1 Analysis of Intra-modal Overlap - Intra vs Inter

We analyse the IMO due to contrastive learning that maximizes the cosine similarity between paired image and text (inter-modal) but ignores the image-image similarity (intra-modal) as illustrated in Fig. 2. We argue that this hampers the performance of few-shot classification. We next proceed to solve this problem.

4.2 Intra-Modal Overlap Correction via Adaptation

We provide two methods to correct IMO via adaptation.

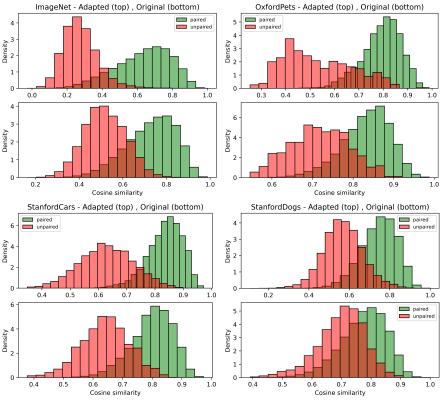
Supervised Adapter Fine-tuning To correct IMO in CLIP vision encoder we incorporate bottleneck adapters [5] into CLIP visual encoder layers which are fine-tuned in a supervised manner on a small sample of images from Google Open Images dataset (ablations on other standard datasets and number of samples in the Appendix E). Adapters are lightweight components that add 0.80% (approx. 1M) new parameters to the model with the bottleneck of size 64. All the original weights of CLIP remain frozen. Following the fine-tuning of CLIP Vision Encoder (*VEimo*) through adapters, we utilize it to create an improved cached model like Tip-Adapter but with IMO-corrected encoded training images $G_{train} \in \mathbb{R}^{NK \times d}$. Then, given a test image encoded with *VEimo*, $U_i \in \mathbb{R}^d$, the affinity matrix Y and logits of Tip-Adapter++ (TA++) are calculated as follows:

$$Y = exp(-\beta(1 - U_i G_{train}^T)), Y \in \mathbb{R}^{NK} \tag{10}$$

$$\text{TA++logits} = T_i W^T + \alpha Y L_{train}, \text{TA++logits} \in \mathbb{R}^N \tag{11}$$

Similarly, we improve standard Tip-X by replacing the Tip-X affinity matrix A with IMO corrected Y, obtaining this way Tip-X++ (TX++) logits:

$$\text{TX++logits} = T_i W^T + \alpha Y L_{train} + \gamma\phi(-M) L_{train}, \text{TX++logits} \in \mathbb{R}^N \tag{12}$$



Dataset	Adapted	Original
ImageNet	0.1839	0.3277
OxfordPets	0.3577	0.3856
StanfordCars	0.2147	0.3231
StanfordDogs	0.3375	0.6208

Figure 3: Intra-modal overlap measured as intersection area between cosine similarity distribution of paired and unpaired images using adapted and original CLIP image encoder (the lower the better)

Note that when computing CLIP logits in the image-text space we use CLIP without adapters, which are only integrated into CLIP visual encoder when we need to compute similarity in the image space, thus the zero-shot learning capability of the original CLIP model is not affected.

Self-supervised Adapter Fine-tuning via DINO We also explore the possibility of training adapters in an unsupervised manner to investigate whether we can reduce the IMO through self-supervised training. While self-supervised methods for training entire neural networks have been extensively studied, there is less exploration into training adapters using these methods. We utilize the state-of-the-art DINO [4] method for this purpose, although we also experimented with SimCLR [6] and BYOL [11] both of which yielded inferior results. We observe that while the self-supervised training method proves effective, it falls short of the supervised alternative. We therefore defer the discussion about the performance and analysis of the same to the supplementary material.

5 Experiments - Supervised Training

Datasets We conduct extensive experiments on 11 fine-grained classification datasets: Caltech101 [9], EuroSAT [15], StanfordCars [19], OxfordPets [25], DescribableTextures [7], OxfordFlowers [23], Food101 [3], FGVCAircraft [21], StanfordDogs [18], PLANTDOC [29] and CUB [14]. To ensure completeness, we include results for not fine-grained datasets in some tables. Comprehensive results for not fine-grained datasets will be provided in the supplementary material.

Performance Comparison Fig. 3 illustrates the difference in IMO between the original CLIP visual encoder and the adapted one on the validation set of four different datasets - ImageNet, OxfordPets, StanfordCars and StanfordDogs (the results for all the datasets are in the Appendix D). The inclusion of the adapter contributes to reducing intra-modal overlap between paired and unpaired images. Tab. in Fig. 3 quantifies the intersection area between

Dataset	Zero-Shot	Tip-Adapter (TA)	Tip-Adapter++ (TA++)	Tip-X (TX)	Tip-X++ (TX++)	Δ (TA++, TA)	Δ (TX++,TX)	Δ (TA++, TX)
EuroSAT	48.383	71.754	**74.86**	71.985	**75.364**	3.106	3.379	2.875
StanfordCars	65.514	70.981	**73.546**	73.276	**74.744**	2.565	1.467	0.27
PLANTDOC	34.994	47.775	**50.25**	48.206	**50.893**	2.475	2.687	2.044
DescribableTextures	43.972	58.676	**60.922**	60.012	**61.151**	2.246	1.139	0.91
StanfordDogs	59.117	61.392	**63.385**	64.988	**65.438**	1.993	0.45	-1.603
SUN397	62.579	68.746	**70.047**	69.938	**70.733**	1.301	0.795	0.109
FGVCAircraft	24.752	33.167	**34.401**	34.945	**35.692**	1.234	0.746	-0.544
OxfordPets	89.071	90.382	**91.567**	91.569	**92.076**	1.185	0.507	-0.002
CUB	55.009	65.138	**66.042**	67.088	**68.135**	0.904	1.047	-1.046
ImageNet	68.802	69.91	**70.431**	70.039	**70.468**	0.521	0.429	0.392
Caltech101	93.306	94.315	**94.778**	94.299	**94.799**	0.462	0.5	0.479
Food101	85.888	**86.195**	86.165	86.253	**86.28**	-0.03	0.027	-0.088
UCF101	67.46	**75.041**	74.757	76.038	**76.098**	-0.284	0.06	-1.281
OxfordFlowers	70.767	**89.622**	88.575	**90.305**	89.687	-1.048	-0.617	-1.73
Average fine-grained	60.979	69.945	**71.317**	71.175	**72.205**	1.372	1.03	0.142
Average all	62.115	70.221	**71.409**	71.353	**72.254**	1.188	0.901	0.056

Table 1: Average performance across all shots on all datasets.

paired and unpaired images (the lower the better). The reduction of IMO is expected to correspond to an improvement in performance. In Tab. 1 we compare the performance of Tip-Adapter and Tip-Adapter++, observing that our method outperforms Tip-Adapter on 11 out of 14 datasets with 1 dataset (Food101) achieving similar results. Additionally, in the same we compare Tip-X and Tip-X++ achieving similar results with Tip-X++ outperforming Tip-X on 13 out of 14 datasets. It is also worth noting that Tip-Adapter++ is competitive or outperforms Tip-X, even with a smaller margin than Tip-X++, on 7 datasets. Overall, Tip-X++ achieves the best performance. These results indicate that our intra-modal overlap corrected encoder is able to extract better features for training-free models. Granular results by number of shots are shown in the Appendix in Fig. 7 and Tab. 6 where it can be seen that the improvement is usually consistent across different numbers of examples chosen for few-shot classification.

Relation Between Intra-modal Overlap and Performance We plot the relation between the difference in intersection area and the average performance difference between Tip-Adapter and Tip-Adapter++. This is to confirm our hypothesis: *the higher the difference in the intersection areas between the original and adapted visual encoders, the higher the performance difference between Tip-Adapter++ and Tip-Adapter* as the IMO reduction was higher. This is illustrated in Fig. 4 where we observe a positive relation between the two, thus reducing by 1% the IMO (increasing area intersection difference) leads to approx. 0.10% improvement of Tip-Adapter++ over Tip-Adapter performance. Furthermore, the two measures exhibit a strong correlation with a correlation coefficient of 0.67. There are, however, few outliers - Food101 has a relatively high difference in intersection areas but the performance of Tip-Adapter++ has not improved over Tip-Adapter. Also, StanfordDogs has a relatively high difference in intersection areas and we expected the performance difference to be higher.

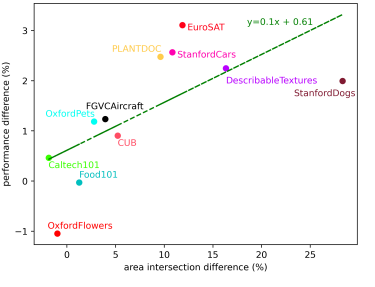


Figure 4: Relation between IMO reduction vs average performance difference between TA++ and TA on fine-grained datasets.

Models	Source	Target	
	ImageNet	ImageNet-V2	ImageNet-Sketch
Zero-Shot CLIP	68.804	60.83	46.14
Tip-Adapter	70.753	63.02	47.24
Tip-Adapter++	**71.505**	**63.96**	**48.38**
Tip-X	70.973	63.19	47.79
Tip-X++	**71.587**	**63.98**	**48.82**

Table 2: Robustness to distribution shift

Dataset	Δ (Adapted, Original)	Proxy-A-Distance
CUB	0.904	1.094
Caltech101	0.462	0.926
DescribableTextures	2.246	0.888
EuroSAT	3.106	1.992
FGVCAircraft	1.234	1.658
Food101	-0.03	1.524
ImageNet	0.521	0.632
OxfordFlowers	-1.048	1.67
OxfordPets	1.185	1.033
PLANTDOC	2.475	1.612
SUN397	1.301	0.906
StanfordCars	2.565	1.543
StanfordDogs	1.993	1.034
UCF101	-0.284	1.425

Table 3: Proxy-A-Distance for all datasets.

Robustness to Distribution Shift We assess the model's robustness to distribution shift. It consists of creating a cached model using one dataset and evaluating it on another. We use ImageNet [8] as the source dataset, employing a 16-shot training set, and test on two target datasets: ImageNet-V2 [28] and ImageNet-Sketch [33]. These datasets contain similar categories to ImageNet but exhibit semantic gaps. Our findings, shown in Tab. 2 reveal that addressing IMO not only contributes to improved performance when cached model is evaluated on the same dataset but also showcases increased resilience to distribution shift.

Increase in Features Variance We observe that the visual features obtained from CLIP exhibit low variance. Evaluating on ImageNet validation set, as illustrated in Fig. 5, it is apparent that over 50% of the features exhibit a low variance close to 0. This trend is consistent across all datasets. Low variance across multiple dimensions suggests that these features lack discriminative power and are less effective. However, upon addressing the IMO, we observed an increase in variance within the visual feature space. This is translated into an enhanced class separability as visually demonstrated in Fig. 6 where we show the t-SNE visualization of the original and adapted CLIP visual features.

Measuring the Distance Between Training and Target Data We also investigated whether the data samples from Google Open Images closely matched the distributions of the downstream datasets we tested on. We aimed to determine if our adapters were potentially overfitting to datasets that resemble each other rather than effectively addressing the broader IMO issue. We use Proxy-A-Distance (PAD) [2] as a measure of the divergence between these datasets. To compute Proxy-A-Distance we create an SVM classifier that is trained to distinguish between the source domain (Google Open Images) and the target domains (other datasets). The PAD is calculated based on the error of this domain classifier:

$$\text{PAD} = 2 \cdot (1 - 2 \cdot \varepsilon) \tag{13}$$

where ε is the domain classifier error. The PAD score falls within the range of 0 to 2 - PAD close to 0 corresponds to a classifier accuracy of 50% indicating that the domain classifier is unable to distinguish between the source and target domains. Conversely, a PAD value of 2 indicates that the classifier is capable of completely discriminating between the two domains, thus they do not follow the same distribution, achieving 100% accuracy or equivalently with the error rate $\varepsilon = 0$. After computing PAD we measure the correlation between the average difference in performance of the Tip-Adapter and Tip-Adapter++ to determine if there is any connection between improved performance and the proximity of source and target data distributions. The correlation between the two is 0.14 suggesting that there is a weak relation

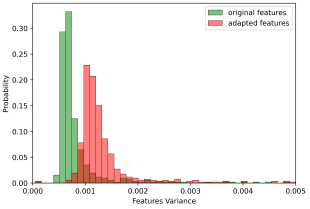


Figure 5: Variance of features on ImageNet validation set of the original and adapted visual encoders.

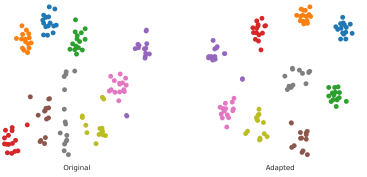


Figure 6: T-SNE visualization of randomly chosen classes from ImageNet validation dataset using original (on the left) and adapted (on the right) visual features.

between them. Surprisingly, EuroSAT which has a very different distribution from the training data exhibits the most substantial performance enhancement following the adaptation. In contrast, ImageNet which has a relatively closer resemblance to the training dataset displays a comparatively smaller performance improvement. We thus conclude that we reduced IMO generalizing to datasets that are relatively different from the training adaptation data. PAD for all the datasets can be found in Tab. 3.

6 Conclusions

This paper examines the relationship between performance and the intra-modal overlap in training-free methods demonstrating a positive relation between the reduction in intra-modal overlap and improved performance. We show that it's possible to directly correct it within the image space, as opposed to using image-text space as a bridge, by introducing bottleneck adapters to the CLIP vision encoder fine-tuned on a subset from the Google Open Images dataset. We further show that such fine-tuning can be done in both a supervised and self-supervised manner. The supervised intra-modal overlap correction improved the performance by 1.38% across all the datasets.

Acknowledgements The authors gratefully acknowledge Microsoft's support in providing GPU compute resources through the Microsoft's Accelerating Foundation Models Research grant. We'd also like to acknowledge the support from the University of Bath for studentship.

References

[1] Open images dataset v5, 2019. URL https://storage.googleapis.com/openimages/web/index.html.

[2] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. *Computer Vision – ECCV 2014*, pages 446–461, 2014. doi: 10.1007/978-3-319-10599-4_29.

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 05 2021. URL https://arxiv.org/abs/2104.14294.

[5] S. Chen, C. GE, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 16664–16678, 2022.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *ICML 2020*, 06 2020. URL https://arxiv.org/abs/2002.05709.

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai L., and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[9] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106:59–70, 04 2007. doi: 10.1016/j.cviu.2005.09.012.

[10] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 2023. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85171365193&doi=10.1007%2fs11263-023-01891-x&partnerID=40&md5=0564904b99ae0d72c6dceecdeba3e92d.

[11] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent a new approach to self-supervised learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

[12] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: zero-shot enhancement of clip with parameter-free attention. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI)*, AAAI'23/IAAI'23/EAAI'23, 2023. ISBN 978-1-57735-880-0. URL https://doi.org/10.1609/aaai.v37i1.25152.

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. doi: 10.1109/CVPR42600.2020.00975.

[14] Xiangteng He and Yuxin Peng. Fine-grained visual-textual representation learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 30:520–531, 02 2020. doi: 10.1109/tcsvt.2019.2892802.

[15] 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.

[16] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 2790–2799, 09–15 Jun 2019.

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

[18] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

[19] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops*, ICCVW '13, page 554–561, 2013. ISBN 9781479930227.

[20] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Meeting of the Association for Computational Linguistics*, Online, August 2021.

[21] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft, 06 2013. URL https://arxiv.org/abs/1306.5151.

[22] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. In *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023.

[23] M. Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. *2008 Sixth Indian Conference on Computer Vision, Graphics Image Processing*, 2008. URL https://www.semanticscholar.org/paper/

Automated-Flower-Classification-over-a-Large-Number-Nilsback02b28f3b71138a06e40dbd614abf8568420ae183.

[24] Omiros Pantazis, Gabriel Brostow, Kate Jones, and Oisin Mac Aodha. Svl-adapter: Self-supervised adapter for vision-language pretrained models. In *British Machine Vision Conference (BMVC)*, 2022.

[25] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505, 2012.

[26] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. AdapterFusion: Non-destructive task composition for transfer learning. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 487–503, April 2021.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8748–8763. PMLR, 18–24 Jul 2021.

[28] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, pages 5389–5400, 2019.

[29] Davinder Singh, Naman Jain, Pranjali Jain, Pratik Kayal, Sudhakar Kumawat, and Nipun Batra. Plantdoc: A dataset for visual plant disease detection. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, CoDS COMAD 2020, page 249–253, 2020.

[30] Jingchen Sun, Jiayu Qin, Zihao Lin, and Changyou Chen. Prompt tuning based adapter for vision-language model adaption, 03 2023. URL https://arxiv.org/abs/2303.15234.

[31] Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2725–2736, October 2023.

[32] Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip H. S. Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. *arXiv preprint arXiv:2404.04125*, 2024.

[33] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, 2019.

[34] T. Yu, Z. Lu, X. Jin, Z. Chen, and X. Wang. Task residual for tuning vision-language models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10899–10909, Los Alamitos, CA, USA, jun 2023. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.01049.

[35] Xiang yu Zhu, Renrui Zhang, Bowei He, A-Long Zhou, Dong Wang, Bingyan Zhao, and Peng Gao. Not all features matter: Enhancing few-shot clip with adaptive prior refinement. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2605–2615, 2023. URL https://api.semanticscholar.org/CorpusID:257913684.

[36] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *CoRR*, abs/2210.07225, 2022. doi: 10.48550/ARXIV.2210.07225. URL https://doi.org/10.48550/arXiv.2210.07225.

[37] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *Computer Vision – ECCV 2022: 17th European Conference*, page 493–510, 2022.

[38] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130:2337 – 2348, 2021. URL https://api.semanticscholar.org/CorpusID:237386023.

[39] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16795–16804, 2022. URL https://api.semanticscholar.org/CorpusID:247363011.

[40] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11175–11185, June 2023.

[41] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15659–15669, October 2023.