南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

CLIPood: Generalizing CLIP to Out-of-Distributions

Yang Shu [*1] Xingzhuo Guo [*1 2] Jialong Wu [1] Ximei Wang [3] Jianmin Wang [1] Mingsheng Long [1]

ICML 2023

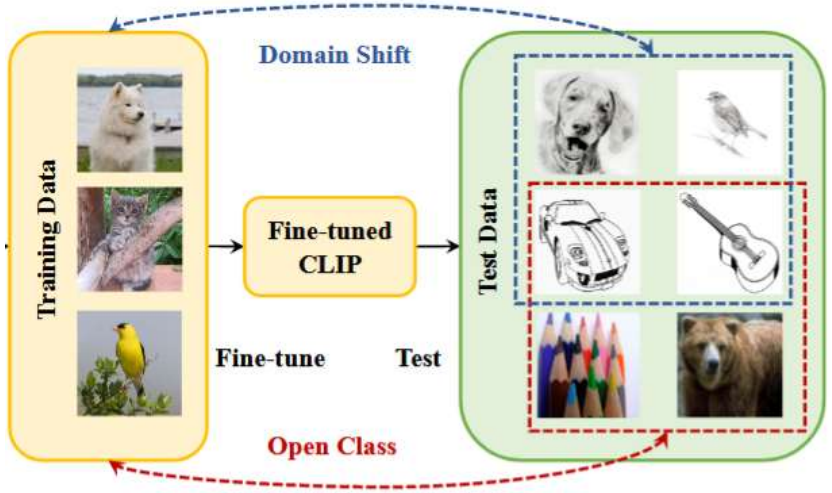
Background

Out-of-Distribution (OOD) Generalization

Domain generalization: Most domain generalization methods focus on the training strategies on source domains, include:

- cross-domain feature alignment
- decomposing domain-specific and domain-common knowledge
- Meta-learning over domains
- designing data-augmentation tasks
- weight ensemble

Open classes



Background

CLIP in Out-of-Distribution (OOD) Generalization

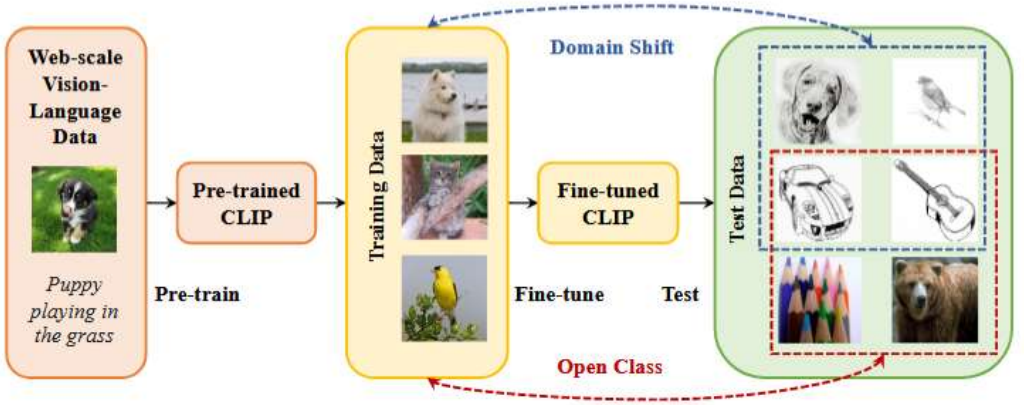
Instead of learning from human-labeled data, recent advances in vision-language pre-training seek to learn from naturally formed supervision of web-scale image-language pairs, which enables learning from diverse domains, and recognizing concepts from an open world.

As a result, vision-language pre-trained models demonstrate impressive zero-shot learning performance and outperform models trained from only labeled images, which reveals a promising approach toward OOD generalization.

Despite the good zero-shot performance, CLIP achieves OOD generalization in a task-agnostic way.

In order for more satisfactory performance on downstream tasks of interest, the pre-trained models still need to utilize task-specific data to make adaptations such as fine-tuning.

Wortsman et al., 2022: the performance of fine-tuned models may be even worse than zero-shot models on related tasks with distribution shifts.

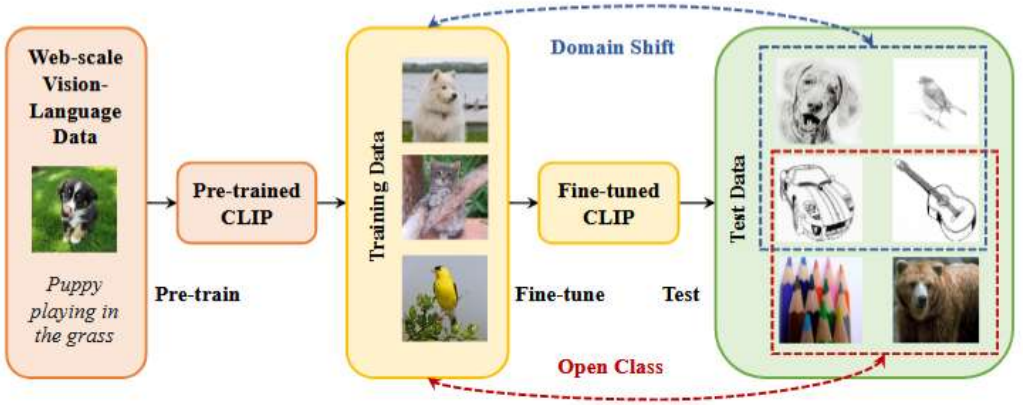


Motivation

From the view of fine-tuning and seek to handle the dilemma during the adaptation of CLIP models.

On the one hand, the pre-trained model should be given the flexibility to fine-tune with the downstream data thus mitigating the gap between upstream and downstream task distributions.

On the other hand, since the downstream data are limited and the concrete relationship between the specific training task and the OOD task is unconstrained, the generalization property from large-scale vision-language pre-training should be exploited or maintained to enable safe model adaptation and finally boosts OOD generalization.



Motivation

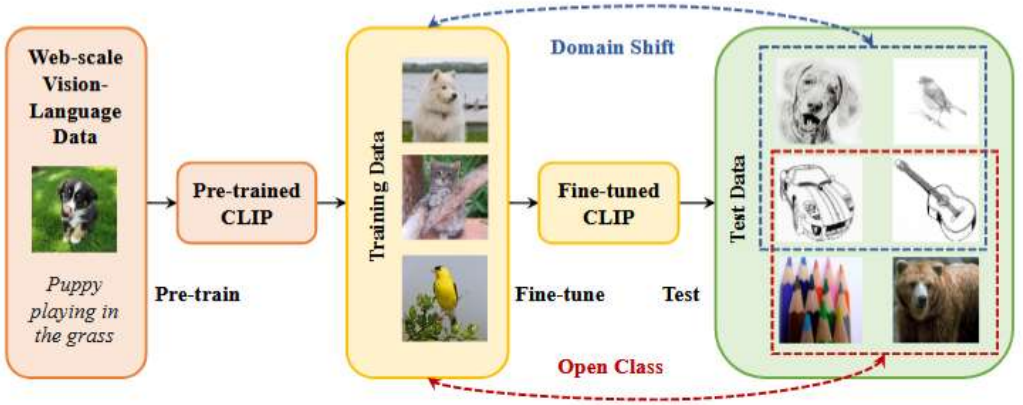
Previous works (standard fine-tuning, a linear classifier):

$$P(y|\mathbf{x}) = \frac{\exp\left(\mathbf{w}_y \cdot g_I\left(\mathbf{x}\right)\right)}{\sum_{c=1}^{C} \exp\left(\mathbf{w}_c \cdot g_I\left(\mathbf{x}\right)\right)}. \qquad (1)$$

Discards the knowledge in the text modality and breaks the connection between them. This may decrease the generalization ability benefiting from image-text alignment.

Besides, the added classifier is tailored to the training dataset, making it hard to generalize to unseen classes.

Therefore, we propose to perform a vision-language fine-tuning strategy on CLIP to enhance its OOD generalization ability.



Motivation

Previous works:

$$P(y|\mathbf{x}) = \frac{\exp\left(S\left(\mathbf{I}_{\mathbf{x}}, \mathbf{T}_y\right)/\tau\right)}{\sum_{c=1}^{C} \exp\left(S\left(\mathbf{I}_{\mathbf{x}}, \mathbf{T}_c\right)/\tau\right)}, \qquad (2)$$

Different from the pre-training stage with abundant and diverse image-text pairs, in downstream scenarios with images and class text prompts (a photo of a [CLASS]), image patterns are still rich, but the diversity of text corpus is limited.

Fine-tune the image encoder and freeze the text encoder to avoid representation collapse.

This aligns the image embedding with the correct text embedding but treats all the false equally, which ignores the potential semantic relations between classes.

On the other hand, the pre-trained text modality contains more detailed semantic knowledge, which quantifies the semantic relationships between texts in detail other than just discriminating between classes.

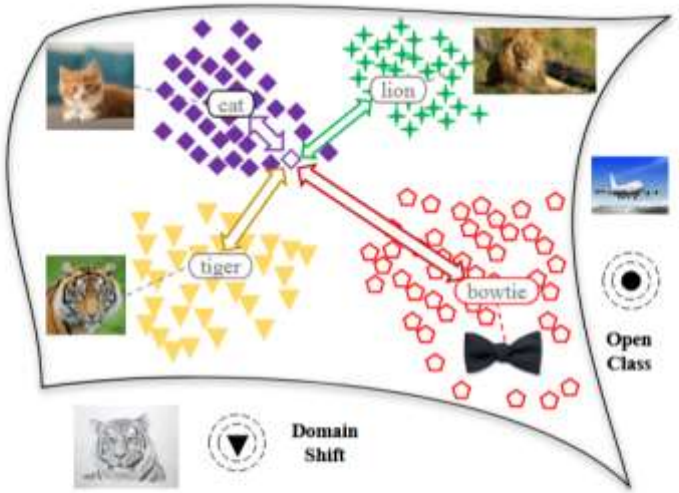


Figure 3: The illustration of Margin Metric Softmax (MMS), where the hollow diamond at the center represents the image embedding $\mathbf{I}_{\mathbf{x}}$. Since $D\left(\mathbf{T}_y, \mathbf{T}_c\right)$ varies across classes, the *adaptive margin* is attained, preserving the inherent unequal relations of classes.

Method

Margin Metric Softmax

$$\mathcal{L} = -\log \frac{\exp\left(S\left(\mathbf{I}_{\mathbf{x}}, \mathbf{T}_y\right)/\tau\right)}{\sum_{c=1}^{C} \exp\left(\left(S\left(\mathbf{I}_{\mathbf{x}}, \mathbf{T}_c\right) + \lambda \cdot D\left(\mathbf{T}_y, \mathbf{T}_c\right)\right)/\tau\right)}. \tag{3}$$

$$D\left(\mathbf{T}_y, \mathbf{T}_c\right) = 1 - S\left(\mathbf{T}_y, \mathbf{T}_c\right). \tag{4}$$

D serves as an adaptive margin for each *S* (Ix, Tc) in the loss.

In this way, MMS exploits the more detailed knowledge of semantic relations in the pre-trained text modality to achieve a better image-text cross-modal alignment and enhance the generalization of the model during vision-language fine-tuning.

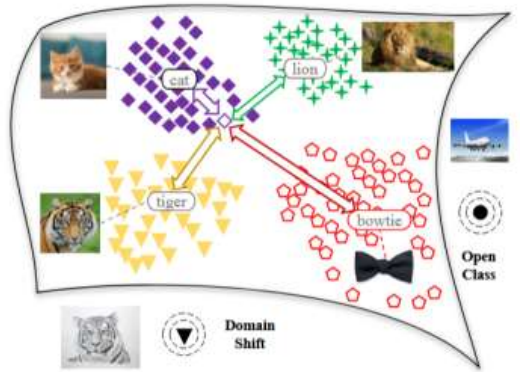


Figure 3: The illustration of Margin Metric Softmax (MMS), where the hollow diamond at the center represents the image embedding $\mathbf{I}_{\mathbf{x}}$. Since $D\left(\mathbf{T}_y, \mathbf{T}_c\right)$ varies across classes, the *adaptive margin* is attained, preserving the inherent unequal relations of classes.

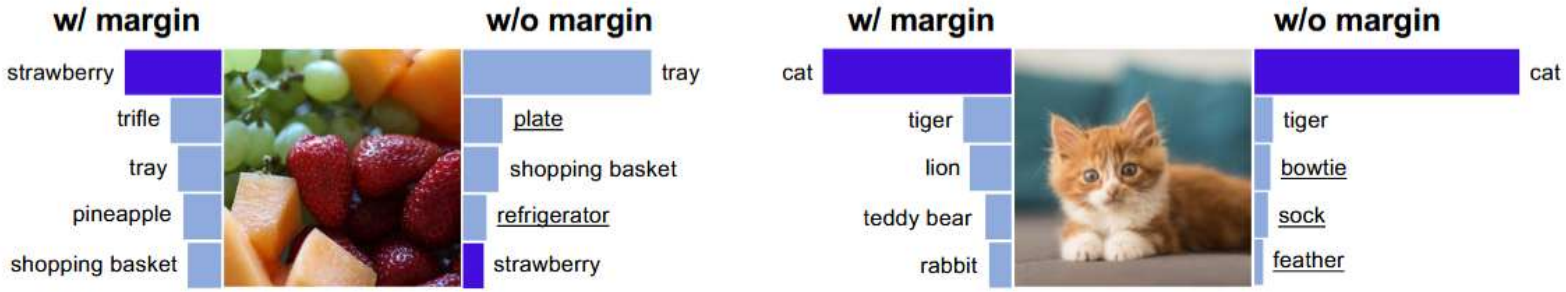


Figure 6: Predictions from models trained with and without adaptive margin.

Method

Beta Moving Average

Despite generally better performance on downstream tasks, fine-tuning pushes the model far away from the pre-trained one at the risk of catastrophic forgetting and representation collapse.

Consider a fine-tuning procedure of T training steps, we can get a trajectory of models $\{\theta_t\}_{t=0}^{T}$

$$\theta^{\mathrm{TE}} = \sum_{t=0}^{T} \frac{\alpha_t}{\sum_{k=0}^{T} \alpha_k} \cdot \theta_t, \qquad (5)$$

$$\alpha_t = \mathrm{Beta}(\beta, \beta)\left(\frac{t+0.5}{T+1}\right). \qquad (6)$$

αt determines the contribution of each model θt

To mitigate greatly increases the storage cost.

$$\theta_t^{\mathrm{BMA}} = \frac{\sum_{k=0}^{t-1} \alpha_k}{\sum_{k=0}^{t} \alpha_k} \cdot \theta_{t-1}^{\mathrm{BMA}} + \frac{\alpha_t}{\sum_{k=0}^{t} \alpha_k} \cdot \theta_t. \qquad (7)$$

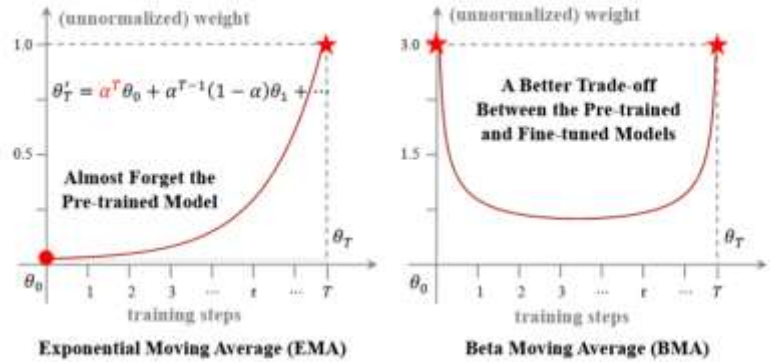


Figure 4: A comparison between Exponential Moving Average (EMA) and Beta Moving Average (BMA), in which the first term of EMA is $\alpha^T\theta_0$ over the T training step and θ_0 is the pre-trained model. Since $\alpha^T \to 0$ when $0 < \alpha < 1$, the fine-tuned model with EMA will almost forget the knowledge of the pre-trained model.

Method

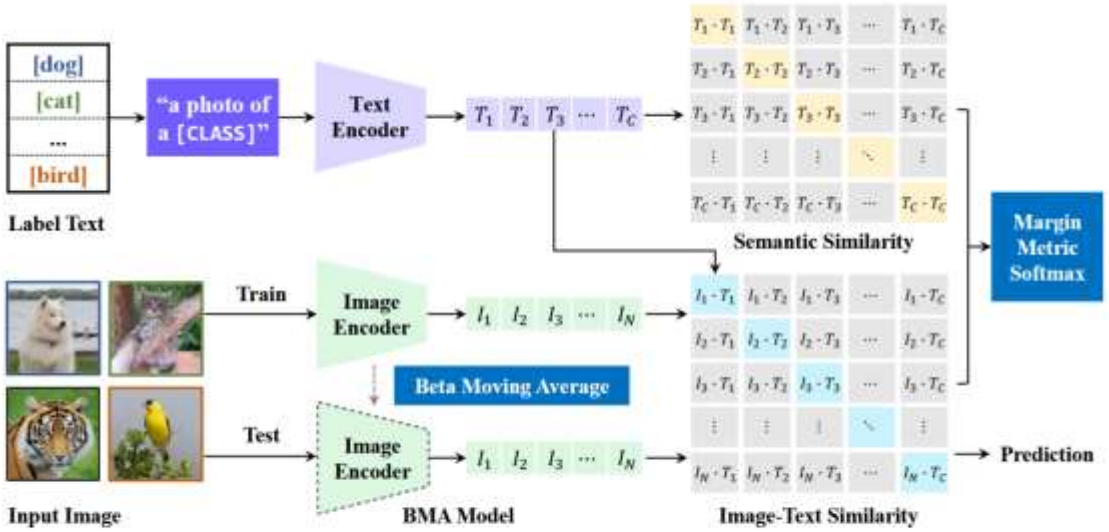


Figure 2: Overview of the proposed CLIPood method. CLIPood compares image embeddings with class text embeddings. Margin Metric Softmax is introduced to exploit semantic relationships between classes. Moreover, a Beta Moving Average model is maintained for prediction, which incorporates both the pre-trained zero-shot model and the fine-tuned model.

Algorithm 1 Training Procedure of CLIPood

Input: Pre-trained CLIP model θ_0, learning rate η
  Initialize the BMA model $\theta_0^{\text{BMA}} \leftarrow \theta_0$
  for $t \in [1, T]$ **do**
    Sample data $\{(\mathbf{x}, y)\}$ from the training set $\mathcal{S}$
    Calculate MMS loss $\mathcal{L}$ as in Eq. (3)
    Update model parameters $\theta_t \leftarrow \theta_{t-1} - \eta \nabla_{\theta_{t-1}} \mathcal{L}$
    Calculate α_t of the current model as in Eq. (6)
    Update the BMA model θ_t^{BMA} as in Eq. (7)
  end for
Output: The final BMA model θ_T^{BMA}

Experiment

Table 1: Accuracy on the DomainBed benchmark with domain shift.

METHOD	BACKBONE	PACS	VLCS	OFFICEHOME	TERRAINC	DOMAINNET	AVG.
ERM	RESNET	85.5	77.5	66.5	46.1	40.9	63.3
CORAL (2016)	RESNET	86.2	78.8	68.7	47.6	41.5	64.6
ZERO-SHOT	CLIP	96.2	81.7	82.0	33.4	57.5	70.2
ERM	CLIP	$96.1_{\pm 0.5}$	$83.0_{\pm 0.2}$	$83.3_{\pm 0.3}$	$\mathbf{60.9}_{\pm 0.2}$	$59.9_{\pm 0.1}$	$76.7_{\pm 0.2}$
MIRO (2022)	CLIP	95.6	82.2	82.5	54.3	54.0	73.7
DPL (2022)	CLIP	**97.3**	84.3	84.2	52.6	56.7	75.0
CLIPOOD	CLIP	$\mathbf{97.3}_{\pm 0.1}$	$\mathbf{85.0}_{\pm 0.4}$	$\mathbf{87.0}_{\pm 0.2}$	$60.4_{\pm 0.7}$	$\mathbf{63.5}_{\pm 0.1}$	$\mathbf{78.6}_{\pm 0.1}$

Table 2: Accuracy on ImageNet with various domain shifts.

METHOD	IN-DISTRIBUTION	OUT-OF-DISTRIBUTIONS				
	IMAGENET	IMAGENET-V2	IMAGENET-S	IMAGENET-A	IMAGENET-R	AVG.
ZERO-SHOT	66.7	60.8	46.1	47.8	74.0	57.2
FINE-TUNE	$68.2_{\pm 0.1}$	$61.9_{\pm 0.1}$	$46.8_{\pm 0.1}$	$46.4_{\pm 0.1}$	$75.1_{\pm 0.1}$	$57.6_{\pm 0.1}$
COOP (2022B)	71.5	64.2	48.0	49.7	75.2	59.3
COCOOP (2022A)	71.0	64.2	48.8	**50.6**	76.2	59.9
CLIPOOD	$\mathbf{71.6}_{\pm 0.1}$	$\mathbf{64.9}_{\pm 0.1}$	$\mathbf{49.3}_{\pm 0.1}$	$50.4_{\pm 0.1}$	$\mathbf{77.2}_{\pm 0.1}$	$\mathbf{60.4}_{\pm 0.1}$

one domain is chosen as the test domain for evaluating OOD generalization, and other domains are chosen as the training domains.

Experiment

Table 3: Generalization performance on 11 downstream datasets with open classes.

(a) Average over 11 datasets

	BASE	NEW	H
CLIP	69.3	74.2	71.7
COOP (2022B)	82.7	63.2	71.7
COCOOP (2022A)	80.5	71.7	75.8
CLIPOOD	**83.9**$_{\pm 0.1}$	**74.5**$_{\pm 0.1}$	**78.9**$_{\pm 0.1}$

(b) ImageNet

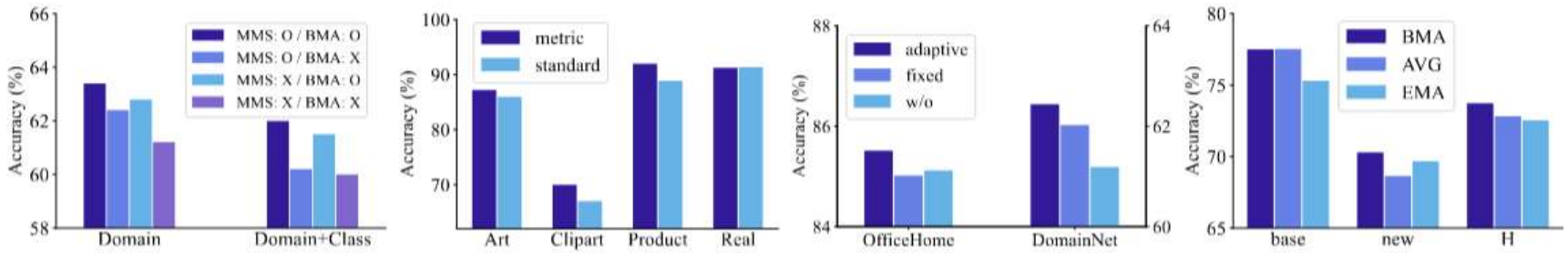
	BASE	NEW	H
CLIP	72.4	68.1	70.2
COOP (2022B)	76.5	67.9	71.9
COCOOP (2022A)	76.0	**70.4**	73.1
CLIPOOD	**77.5**$_{\pm 0.1}$	70.3$_{\pm 0.1}$	**73.7**$_{\pm 0.1}$

Table 4: Accuracy on OfficeHome and DomainNet with both domain shift and open classes.

SPLIT	METHOD	OFFICEHOME				DOMAINNET					
		A	C	P	R	C	I	P	Q	R	S
BASE	CLIP	86.8	75.5	89.5	92.6	72.8	51.7	66.0	13.5	83.4	66.9
	COOP	87.0$_{\pm 0.4}$	78.3$_{\pm 1.2}$	92.4$_{\pm 0.2}$	91.4$_{\pm 0.6}$	75.7$_{\pm 0.2}$	58.8$_{\pm 0.5}$	68.5$_{\pm 1.3}$	13.1$_{\pm 1.0}$	84.0$_{\pm 0.5}$	70.0$_{\pm 0.1}$
	CLIPOOD	**90.1**$_{\pm 0.2}$	**79.7**$_{\pm 0.2}$	**93.1**$_{\pm 0.1}$	**94.8**$_{\pm 0.1}$	**79.0**$_{\pm 0.2}$	**62.2**$_{\pm 0.1}$	**73.0**$_{\pm 0.2}$	**20.2**$_{\pm 0.2}$	**86.2**$_{\pm 0.1}$	**73.8**$_{\pm 0.1}$
NEW	CLIP	76.6	59.4	88.1	86.2	70.2	44.1	66.4	14.1	83.5	61.0
	COOP	76.5$_{\pm 1.1}$	56.6$_{\pm 2.4}$	88.0$_{\pm 1.9}$	**86.8**$_{\pm 0.7}$	**71.5**$_{\pm 0.2}$	47.2$_{\pm 0.3}$	67.3$_{\pm 0.7}$	14.8$_{\pm 0.7}$	**83.7**$_{\pm 0.7}$	**63.1**$_{\pm 0.3}$
	CLIPOOD	**77.8**$_{\pm 0.2}$	**60.0**$_{\pm 0.2}$	**88.3**$_{\pm 0.1}$	86.7$_{\pm 0.1}$	71.2$_{\pm 0.1}$	**48.1**$_{\pm 0.1}$	**68.2**$_{\pm 0.2}$	**18.0**$_{\pm 0.4}$	83.4$_{\pm 0.1}$	62.9$_{\pm 0.1}$
TOTAL	CLIP	82.6	67.3	88.8	89.5	71.4	47.1	66.2	13.8	83.4	63.4
	COOP	82.7$_{\pm 0.5}$	67.2$_{\pm 0.7}$	90.2$_{\pm 1.0}$	89.2$_{\pm 0.6}$	73.4$_{\pm 0.3}$	51.8$_{\pm 0.3}$	67.9$_{\pm 1.0}$	13.7$_{\pm 0.8}$	83.9$_{\pm 0.5}$	66.0$_{\pm 0.2}$
	CLIPOOD	**85.1**$_{\pm 0.1}$	**69.6**$_{\pm 0.2}$	**90.8**$_{\pm 0.1}$	**91.0**$_{\pm 0.1}$	**74.8**$_{\pm 0.1}$	**53.6**$_{\pm 0.1}$	**70.6**$_{\pm 0.1}$	**19.1**$_{\pm 0.3}$	**84.8**$_{\pm 0.1}$	**67.4**$_{\pm 0.1}$

Split the classes in each dataset equally into two parts, one as base classes and the other as new classes. We train the model on base-class data and test on base classes and new classes separately to evaluate the generalization ability.

Experiment



(a) Ablation Study (b) Metric Softmax (c) Adaptive Margin (d) Beta Moving Average

$$\theta_{\text{AVG}} = \frac{1}{T}\sum_{t=1}^{T}\theta_t$$

Figure 5: Analysis experiments for CLIPood.

Thank you