Weak Distribution Detectors Lead to Stronger Generalizability of Vision-Language Prompt Tuning

Kun Ding\textsuperscript{1,3}, Haojian Zhang\textsuperscript{2}, Qiang Yu\textsuperscript{1}, Ying Wang\textsuperscript{1,3*}, Shiming Xiang\textsuperscript{1,3}, Chunhong Pan\textsuperscript{3}

\textsuperscript{1}State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences
\textsuperscript{2}Engineering Laboratory for Intelligent Industrial Vision, Institute of Automation, Chinese Academy of Sciences
\textsuperscript{3}Research Center of Aerospace Information, Institute of Automation, Chinese Academy of Sciences

\{kun.ding, zhanghaojian2014, qiang.yu\}@ia.ac.cn, \{ywang, smxiang, chpan\}@nlpr.ia.ac.cn

Abstract

We propose a generalized method for boosting the generalization ability of pre-trained vision-language models (VLMs) while fine-tuning on downstream few-shot tasks. The idea is realized by exploiting out-of-distribution (OOD) detection to predict whether a sample belongs to a base distribution or a novel distribution and then using the score generated by a dedicated competition based scoring function to fuse the zero-shot and few-shot classifier. The fused classifier is dynamic, which will bias towards the zero-shot classifier if a sample is more likely from the distribution pre-trained on, leading to improved base-to-novel generalization ability. Our method is performed only in test stage, which is applicable to boost existing methods without time-consuming retraining. Extensive experiments show that even weak distribution detectors can still improve VLMs' generalization ability. Specifically, with the help of OOD detectors, the harmonic mean of CoOp (Zhou et al. 2022b) and ProGrad (Zhu et al. 2023) increase by 2.6 and 1.5 percentage points over 11 recognition datasets in the base-to-novel setting.

Introduction

Many large vision-language models (VLMs) trained on web-scale image-text pairs have emerged continuously in recent years (e.g., CLIP (Radford et al. 2021), ALIGN (Jia et al. 2021)) and they are proved to be very useful on many pure vision and multi-modal tasks, such as few-shot learning (Zhou et al. 2022b), image segmentation (Rao et al. 2022) and image captioning (Mokady, Hertz, and Bermano 2021). The huge parameter count of these models poses extreme challenges for classical fine-tuning techniques, which are both parameter inefficient and of low generalization capability. Vision-language prompt tuning (VLPT) has recently been proposed as an alternative technique (Zhou et al. 2022b,a; Zhu et al. 2023; Yao, Zhang, and Xu 2023), which is demonstrated to be more parameter efficient.

However, it is hard for VLPT to balance the performance between base and novel classes. The first VLPT method CoOp (Zhou et al. 2022b) achieves superior performance on base classes of 11 recognition datasets. Compared to zero-shot CLIP, the mean accuracy on novel classes drops 11 points (from 74.22% to 63.22%). CoCoOp (Zhou et al. 2022a) alleviates the above problem by conditioning prompts on image feature. CoCoOp improves the novel class accuracy in the sacrifice of base class accuracy and the model design is quite heavy weighted. ProGrad (Zhu et al. 2023) prevents prompt tuning from forgetting the general knowledge in VLMs by only updating the prompt whose gradient is aligned to the general direction. KgCoOp (Yao, Zhang, and Xu 2023) introduces an auxiliary loss to reduce the discrepancy between the textual embeddings generated by learned prompts and hand-crafted prompts.

Different from the above methods, we propose a new viewpoint for addressing the trade-off problem between base and novel classes. This viewpoint is motivated by the fact: zero-shot CLIP excels on novel classes while CoOp excels on base classes (Zhou et al. 2022a). Therefore, if it is possible to know a sample comes from the base distribution or the novel distribution beforehand and choose the right classifier accordingly, such a trade-off problem will be addressed perfectly. The relation between the base-to-novel evaluation metric (harmonic mean) and the in-distribution (ID) prediction accuracies is studied in Fig. 1, which implies that the
harmonic mean of the fused classifier with binary scoring function (in Eq. 7) increases with ID detection accuracies ($r_\alpha$ and $1 - r_\alpha$ on base and novel set respectively). The problem of discriminating a sample from which distribution is similar to the out-of-distribution (OOD) detection problem. As such, we could convert the base-to-novel trade-off problem to an OOD detection problem.

Specifically, we design a competition based scoring function which exploits the ID scores generated by existing OOD detection methods to generate a score between 0 and 1 for a test sample and then use this score to dynamically fuse the pre-trained zero-shot classifier and the few-shot classifier tuned on limited training data. Different from the scoring functions in prior OOD detection methods, the proposed scoring function is based on the comparison between the outputs of two classifiers. The proposed fusion method is dynamic, i.e. the fusion weights vary from different samples, which is better than static method. Also, the fusion is implemented in test stage, making it easy for the integration in existing methods without re-training.

We conduct extensive experiments on 15 recognition datasets under different settings. In the base-to-novel setting, our method improves the harmonic mean accuracy of CoOp and ProGrad by 2.6% and 1.5%, respectively. In the domain generalization setting, the mean accuracies over target domains are competitive, either. The main contributions of this work are listed as follows:

- We propose a new viewpoint for dealing with the base-to-novel generalization problem of VLPT by introducing OOD detection.
- We propose a competition based scoring function for fusing the few-shot and zero-shot classifier dynamically.
- We demonstrate by experiment even weak distribution detectors can still improve the generalizability of existing VLPT methods.

**Related Work**

**Vision-Language Models:** Recent years have witnessed lots of vision-language models, such as CLIP (Radford et al. 2021), ALIGN (Jia et al. 2021), BLIP (Li et al. 2022). Benefited from massive image-text pairs, large vision-language models have demonstrated huge potential on many vision tasks, including but not limited to zero-shot learning (Radford et al. 2021; Shu et al. 2022), few-shot learning (Zhou et al. 2022b), image segmentation (Rao et al. 2022), image generation (Ramesh et al. 2022), etc. The key problem of exploiting so many pre-trained VLMs is how to fine-tune them on downstream tasks in a parameter efficient way.

**Vision-Language Prompt Tuning:** As one of the most popular parameter efficient tuning techniques, prompt tuning is first proposed in NLP and subsequently introduced into vision-language domain by Zhou et al. (2022b). Instead of manually designing prompts, prompt tuning learns soft context embeddings from data, which avoids cumbersome handcrafts. Based on CoOp, lots of efforts have been made to solve the problems in certain aspects or expand application scope. For example, CoCoOp (Zhou et al. 2022a), ProGrad (Zhu et al. 2023) and KgCoOp (Yao, Zhang, and Xu 2023) are devised to reduce the loss of generalization on unseen tasks. DualCoOp (Sun, Hu, and Saenko 2022) extends CoOp to the multi-label recognition task. SoftCPT (Ding et al. 2022) exploits the cross-task knowledge to further enhance the generalization ability. PLOT (Chen et al. 2023) introduces optimal transport to achieve fine-grained matching between multiple prompts and visual regions. Prompt tuning is also introduced in image segmentation (Rao et al. 2022) and video relation detection (Gao et al. 2023).

**OOD Detection:** OOD detection is to identify outliers in an open sample space not belonging to any seen training classes, which is especially vital for safety-critical circumstances. According to Yang et al. (2021), OOD detection methods can be categorized into classification-based, density-based, distance-based and reconstruction-based methods. Post-hoc detection is an important member of classification-based methods that does not need extra training data. Given a trained model, post-hoc methods calculate a score to indicate the ID-ness or OOD-ness. Hendrycks and Gimpel (2017) proposed the maximum softmax probabilities (MSP), which is a widely-used baseline. Considering that MSP is problematic for large-scale settings with hundreds or thousands of classes, MaxLogit (Hendrycks et al. 2022) that uses the maximum unnormalized logit as anomaly score was proposed. Liu et al. (2020) proposed an energy based indicator that has good theoretical interpretability. More recently, Zhang and Xiang (2023) decoupled MaxLogit into cosine similarity and logit norm and introduced a tunable factor between them for better performance. Based on these works we propose a novel scoring function tailored for VLPT, which considers two classifiers jointly, i.e. zero-shot and few-shot classifier. This is different from the above methods that only use one trained classifier to design scoring function.

**Dynamic Neural Networks:** Designing instance-aware dynamic neural networks is an important research topic in the literature. Dynamic convolution (Chen et al. 2020; Hou and Kung 2021; Li, Zhou, and Yao 2022) learns a linear combination of multiple static kernels with input-dependent dynamic attention. Such a design effectively increases model capability without increasing FLOPs a lot. Dynamic quantization (Liu et al. 2022) learns quantization configuration for each image and layer individually, which further reduces computation cost. The Mixture-of-Expert (MoE) layer contains many expert sub-networks and learns a gating function to route individual inputs to some of the experts (Shazeer et al. 2017). It substantially increases the model capacity without introducing large computation overhead. Our proposed method dynamically fuses two classifiers, it can also be seen as a dynamic network. However, it is a test-time method as it does not need re-training.

**Background**

Radford et al. (2021) demonstrated the contrastive learning based vision-language model CLIP has strong zero-shot recognition ability. The good generalizability is believed to be benefited from natural language supervision and diverse large-scale training data. CLIP has a two-stream structure:
one image encoder for representing image as visual embedding and one text encoder for representing text as text embedding. Due to the contrastive learning based design, CLIP can be easily adapted for zero-shot and few-shot recognition.

Zero-shot CLIP. We denote the image and text encoder in CLIP as $E_{img}$ and $E_{txt}$, respectively. For a zero-shot classification task of $N$ classes, Radford et al. (2021) used the classifier generated by the text encoder for classification. The $N$ class names are assumed to be accessible and the classifier weight (an unnormalized column vector) associated to the $i$-th class ($i \in \{1, 2, \ldots, N\}$) is obtained by encoding its class name, i.e., $w_i^e = E_{txt}([\text{"class name"}])$. Adding a prefix string to class name is demonstrated to be useful, that is, $w_i^e = E_{txt}([\text{"a photo of a [class name]"}])$.

Collecting the weights of all classes into a matrix results in $W^e = [w_1^e, \ldots, w_N^e]^T$. Given an image $I$, its visual embedding $x$ is extracted by $x = E_{img}(I)$, which is an unnormalized vector with a same length to $w_i^e$. To classify $I$, zero-shot CLIP computes the posterior probability of a certain class $y$ as follows:

$$p^e(y|x) = \exp(o^e_y(x))/\sum_{y=1}^N \exp(o^e_y(x)), \tag{1}$$

where $o^e_y(x) = d(x, w_i^e)/\tau$ is the logit of class $y$ with $\tau$ a temperature parameter and $d(\cdot)$ the cosine similarity.

Context Optimization (CoOp). Vision-language prompt tuning (VLPT) methods like CoOp and its successors (Pre-Grad, KgCoOp, etc.) extend zero-shot CLIP for few-shot classification. Meanwhile, the fixed prompt prefix is upgraded as a learnable prompt context. They define a sequence of $M$ vectors as the context denoted as a matrix $V = [v_1, \ldots, v_M]$. For the $i$-th class, the class token embeddings can be represented as a matrix $C_i$. The classifier weight associated to class $i$ is computed by encoding $[V, C_i]$ as a text embedding: $w_i^f = E_{txt}([V, C_i])$. Collecting the weights of all classes into a matrix results in $W^f = [w_1^f, \ldots, w_N^f]^T$. To classify an image of encoded feature $x$, the following posterior probability of a certain class $y$ is calculated:

$$p^f(y|x) = \exp(o^f_y(x))/\sum_{y=1}^N \exp(o^f_y(x)), \tag{2}$$

where $o^f_y(x) = d(x, w_i^f)/\tau$ is the logit of the $y$-th class. VLPT methods optimize the context $V$ while fixing the image and text encoder. To align with upstream task, the negative log likelihood loss is widely adopted.

**Method**

The flowchart of the proposed method is presented in Fig. 2. For an image, we extract the visual embedding vector. With hand-crafted prompts, we compute the zero-shot classifier. With learnable prompts, we compute the few-shot classifier. Subsequently, two posterior probability distributions, which serve as the inputs of OOD detector, are computed using these two classifiers. The OOD detector outputs in-distribution score for a test sample. The two in-distribution scores, $ids^e(x)$ and $ids^f(x)$, are computed by a scoring function $s(x)$. This final score is used to dynamically weight the two classifiers. In the following, we detail the key ingredients of the proposed method. Besides, we further analyze the relation between harmonic mean accuracy and OOD detection accuracies.

**OOD Detectors**

Existing OOD detection methods calculate a score to reflect ID-ness or OOD-ness. By choosing a valid threshold and thresholding the score with it, binary prediction can be obtained. This work adopts post-hoc methods due to their simplicity. The original scores proposed in these methods are converted to values reflecting ID-ness. We use three existing methods (MSP, MaxLogit and Energy) and also consider a new entropy based method. Actually, many other OOD detection methods can also be used.

MSP (Hendrycks and Gimpel 2017) uses the maximum softmax probability of a classifier as the ID score, that is

$$ids^e(x) = \max_i p^e(i|x), \tag{3}$$

where $s$ is zs or $f$. $ids^e(x)$ and $ids^f(x)$ denote the in-distribution score corresponding to the zero-shot and few-shot classifier, respectively.

MaxLogit (Hendrycks et al. 2022) uses the maximum logit of a classifier as the ID score, that is

$$ids^e(x) = \tau \cdot \max_i o^e_i(x), \tag{4}$$

where $\tau$ is multiplied for normalizing the final score into the range $[-1, 1]$.

In Liu et al. (2020), the free energy function over $x$ in terms of the denominator of the softmax function is used as the OOD score. Accordingly, the negative free energy is used as the ID score:

$$ids^e(x) = (\log N + 1/\tau)^{-1} \cdot \tau \cdot \log \sum_i^N \exp(o^e_i(x)), \tag{5}$$

where $(\log N + 1/\tau)^{-1}$ is a normalizing factor.

The entropy of a softmax probability distribution reflects the OOD-ness of a sample. Larger entropy means the sample is more likely from a different distribution. We thus use the negative entropy as the ID score:

$$ids^e(x) = (\log N)^{-1} \cdot \sum_i^N p^e(i|x) \log p^e(i|x), \tag{6}$$

where $(\log N)^{-1}$ is a normalizing factor.


**Competition Based Scoring Function**

We propose a scoring function to indicate a sample is more likely from the base or novel distribution. The intuition is that if \( ids^{bs}(x) - ids^{ns}(x) \) is larger, it has a higher probability that the sample is from the base distribution (also the distribution fine-tuned on). To obtain a valid weight, the score should be normalized into \([0,1]\), thus, we introduce a sigmoid function \( \sigma(x) \) to normalize the above difference. Formally, the scoring function is defined as

\[
s(x) = \sigma(\alpha \cdot (ids^{bs}(x) - ids^{ns}(x))),
\]

where \( \alpha \) is a tunable scaling factor, which controls the dynamic range of the score (ref. Fig. 2). Not that, applying sigmoid function over the difference equals to applying the softmax function over \( ids^{bs}(x) \) and \( ids^{ns}(x) \), implying the existence of competition. Note that, with a sufficiently large \( \alpha \), the sigmoid function equals to a Heaviside step function. In this case, it actually makes a binary prediction.

**Dynamic Fusion**

When a sample is more likely from the novel distribution, it is better to use the zero-shot classifier; conversely, when it is more likely from the base distribution, it is better to use the few-shot classifier. A more general model is weighted combining the few-shot and zero-shot classifier with \( s(x) \):

\[
W = s(x) \cdot W^{fs} + (1 - s(x)) \cdot W^{zs}.
\]

By this way, the fused classifier is instance-conditioned, which dynamically changes with different inputs. Note that other strategies such as weighted combination of the posterior probabilities from the two classifiers are also possible.

**Analysis of Harmonic Mean**

The harmonic mean performance of the classifier in Eq. 8 has direct relation to OOD detection performance. In the base-to-novel evaluation setting of VLPT, there are two test sets of non-overlapping classes: base set and novel set. Let us denote \( r_b \) and \( r_n \) as the probability of predicting a sample in the base and novel set as a sample from base distribution, respectively. The classification accuracy of the few-shot classifier \( W^{fs} \) and the zero-shot classifier \( W^{zs} \) on base set are denoted as \( p_1 \) and \( p_0 \), respectively. The classification accuracy of the few-shot classifier \( W^{fs} \) and the zero-shot classifier \( W^{zs} \) on novel set are denoted as \( q_1 \) and \( q_0 \), respectively. We have the following proposition:

**Proposition 1.** The harmonic mean accuracy \( H \) of the fused classifier \( W \) with the Heaviside step function used on base and novel set can be represented as:

\[
H = 2P_bP_n/(P_b + P_n),
\]

\[
P_b = p_1 r_b + p_0 (1 - r_b),
\]

\[
P_n = q_1 r_n + q_0 (1 - r_n),
\]

Proof: The proposition can be proved by the law of total probability.

We plot the contour map of \( H \) with respect to \( r_b \) and \( r_n \) in Fig. 1 with \( p_0 = 0.6, p_1 = 0.9, q_0 = 0.8, q_1 = 0.6 \). Each point in the figure represents an OOD detector of certain performance. The black dashed line corresponds to random OOD detectors. When moving towards the lower right corner, \( H \) increases gradually. A relatively weak classifier (e.g., the classifiers on the 0.750 contour) can still improve the harmonic mean accuracy of the zero-shot and few-shot classifier, i.e., \( H_{zs} = 0.69 \) and \( H_{fs} = 0.72 \), respectively. This implies the feasibility of the proposed method.

**Experiment**

Similar to prior work, we evaluate our method under two settings: 1) base-to-novel generalization within a dataset; 2) domain generalization across datasets with the same classes.

**Setup**

**Dataset.** Following prior works (Zhou et al. 2022b), for the base-to-novel experiment, we use 11 image classification datasets: ImageNet (Deng et al. 2009) and Caltech101 (Li, Fergus, and Perona 2007) for generic object classification; Food101 (Bossard, Guillaumin, and Gool 2014), StanfordCars (Krause et al. 2013), OxfordPets (Parkhi et al. 2012) and Flowers102 (Nilsback and Zisserman 2008) and FGVC-Aircraft (Maji et al. 2013) for fine-grained visual classification; UCF101 (Soomro, Zamir, and Shah 2012) for action recognition; EuroSAT (Helber et al. 2019) for satellite image classification; DTD (Cimpoi et al. 2014) for texture classification; SUN397 (Xiao et al. 2010) for scene recognition. For the domain generalization experiment, we use ImageNet as the source domain and other four datasets, i.e. ImageNetV2 (Recht et al. 2019), ImageNet-Sketch (Wang et al. 2019), ImageNet-A (Gao et al. 2022) and ImageNet-R (Hendrycks et al. 2021) as the target domains.

**Compared Methods.** We propose to use an OOD detector to enhance the generalization ability of the existing VLPT methods. To validate the versatility of our method, we use four existing methods as the baselines: CoOp (Zhou et al. 2022b), CoCoOp (Zhou et al. 2022a), ProGrad (Zhu et al. 2023) and KgCoOp (Yao, Zhang, and Xu 2023). The results of these methods with and without OOD detector will be compared. Besides, zero-shot CLIP is also compared.

**Implementation Details.** With the public code of CoOp, CoCoOp, ProGrad and KgCoOp, we reproduce the results on the aforementioned datasets using the suggested parameters. For CoOp, the context length is 16 and random initialization is adopted. The batch size is 32 and 50 epochs are trained. For CoCoOp, the context is initialized by “a photo of a”, batch size is 1 and 10 epochs are trained. For ProGrad, the training setting is identical to CoOp and the two extra hyper-parameters are set to 1.0. For KgCoOp, the context is initialized by “a photo of a” and the additional hyper-parameter \( \lambda \) is set to be 8.0. Due to limited GPU memory, the batch size is set to be 32. Besides, 100 epochs are trained. All methods use the same learning rate scheduler and optimizer. Our reproduced results are slightly different from previously reported, however, the differences are minor. For our method, we insert the proposed test-time technique into these methods and conduct model inference on test sets. The entropy-based ID score and \( \alpha = 64 \) are used as default.
datasets into two groups: base classes (\textit{Base}) and novel classes (\textit{Novel}). The two sets do not share any identical classes. We train VLPT methods on base classes and evaluate them on novel classes. When a VLPT method is enhanced by OOD detection, the corresponding method is annotated by a plus sign (+) in the name. The detailed results of 9 methods are reported in Table 1.

Table 1(a) shows the average results over 11 datasets, where zero-shot CLIP acquires the worst accuracy on base classes. This is because it uses fixed prompts which are sub-

We report the final performance averaged over three random seeds. The same evaluation metrics adopted in CLIP are used. All experiments are conducted on RTX A4000 GPU. Without further specification, ViT-B/16 is adopted as the default vision backbone.

**Base to Novel Generalization**

In this setting, we split the train and test samples in 11 datasets into two groups: base classes (\textit{Base}) and novel classes (\textit{Novel}).
optimal compared to learned prompts on base set. However, CLIP’s accuracy on novel classes is quite appealing (better than CoOp, CoCoOp, ProGrad and KgCoOp), demonstrating its strong generalization ability across different classes.

CoOp obtains the highest Base accuracy, implying its strong fitting capability on base classes. In contrast, its Novel accuracy is the worst, showing its low base-to-novel generalization ability. By introducing OOD detection into CoOp, the Novel accuracy is improved significantly with sacrificing the Base accuracy a little. As such, the harmonic mean increases a lot, as large as 2.6%. As for CoCoOp, CoCoOp+ improves the harmonic mean by 0.7% with a slight drop of Base accuracy (i.e. 0.86%) and a remarkable increase of Novel accuracy (i.e. 2.01%). Besides, ProGrad+ improves the harmonic mean of ProGrad by 1.5% and the Novel accuracy by 3.85% with a slight sacrifice of Base accuracy, i.e. 1.53%. It obtains a new state of the art of harmonic mean, i.e. 77.20%, outperforming KgCoOp’s best result, 76.97%. Finally, KgCoOp+ and KgCoOp acquire the similar harmonic mean. One of the reasons the proposed method has little effect on KgCoOp is that it adds an extra loss function to enforce the few-shot and zero-shot classifier approach each other, which is less friendly for model fusion. Another reason is KgCoOp already achieves excellent harmonic mean, i.e. 76.97%, which is very near to the upper bound harmonic mean of fusing KgCoOp and zero-shot CLIP, i.e. 77.67%. This small gap makes further improvement harder. By summarizing the above observations, OOD detection is useful in improving base-to-novel generalization ability of existing VLPT methods.

Table 1(b)-(l) show results on each dataset. First, CoOp+, CoCoOp+, ProGrad+ and KgCoOp+ reduce the Base performance compared to CoOp, CoCoOp, ProGrad and KgCoOp on 6/11 datasets, 9/11 datasets, 10/11 datasets and 11/11 datasets, respectively. However, as shown in Table 1(a), the average Base performance drops a little especially for CoOp and CoCoOp. Second, they improve the Novel performance on 11/11 datasets, 9/11 datasets, 11/11 datasets and 7/11 datasets, respectively, implying boosted performance on new classes. Finally, they improve the harmonic mean of ProGrad by 1.5% and the harmonic mean, i.e. 77.20%, outperforming KgCoOp’s best result, 76.68%. As for CoCoOp, CoCoOp+ obtains an obvious improvement of harmonic mean of 1.81%, 2.25%, 7.12%, 2.28%, 2.69%, 2.07% and 6.89% on ImageNet, StandfordCars, Flowers102, FGVC/Aircraft, SUN397, EuroSAT and UCF101 upon CoOp, respectively. In summary, the results demonstrate OOD detection can improve base-to-novel generalization ability.

The proposed dynamic fusion is also suitable for two few-shot classifiers with simple modification. By replacing zero-shot CLIP with ProGrad in KgCoOp+, the OOD detection based dynamic fusion (Pro+Kg) can obtain better results compared to both KgCoOp and ProGrad under the base-to-novel generalization setting, as can be seen from Table 1. Actually, Pro+Kg achieves 0.4% higher harmonic accuracy compared to the best result reported above, i.e. 77.20%. Such an effectiveness is due to the strong complementarity between KgCoOp and ProGrad.

### Domain Generalization

Domain generalization evaluates the effect of distribution shift under the same classes. The prompts are learned on source domain (i.e., ImageNet) and transferred to target domains (i.e., ImageNetV2, etc.) for evaluation. The results of zero-shot CLIP and four prompt tuning methods with and without OOD detection are reported in Table 2. From the results, OOD detection may reduce the performances on target domains slightly in some cases. Fortunately, the drops are marginal compared to the improvements in Table 1. In Table 2, almost all results of CoOp/CoCoOp/ProGrad/KgCoOp surpass CLIP. This is reasonable as the source and target domains share the classes, learning on source domain has positive effect on improving the recognition accuracy on target domains. As such, when fusing the zero-shot and few-shot classifier, more weights of the latter will gather around one. Meantime, the weaker zero-shot classifier may bring side effect for fusion. These can explain why OOD detection does not demonstrate significant improvements in this setting. Actually, OOD detection also works well in this setting if there is strong complementarity between the two classifiers being fused. From Table 2, KgCoOp and ProGrad have strong complementarity on different datasets. As such, we replace zero-shot CLIP with ProGrad in KgCoOp+. This method is denoted as Pro+Kg. From Table 2, Pro+Kg can achieve a new state of art, i.e. 60.36% average accuracy on target domains. Meantime, the accuracy on source domain is also appealing. The above results suggest that OOD detection is also helpful for domain generalization.
### Ablation Study

**Static Weighting:** We can also use static (fixed) weight $s(x)$ for all test samples. The results of using static weights ($s(x) = 0.05, 0.25, 0.50, 0.75, 0.95$) and dynamic weights ($\alpha = 64$) are compared in Table 3. Using fixed weights also implies the OOD detectors with 50% accuracy are adopted. The improved result of dynamic weighting demonstrates that weak OOD detectors help to boost generalizability across different classes.

**Different OOD Detection Methods:** The proposed method is flexible, it supports many different OOD detection methods. Besides the aforementioned ID scores, the maximum softmax probability (MSP) of the zero-shot and few-shot classifier can also serve as the weight, i.e., $s(x) = 1 - \text{softmax}(x)$ and $s(x) = \text{softmax}(x)$. The results of different methods are compared in Table 4. When using sigmoid function, the results with the best $\alpha$ in $(0.5, 1, 2, 4, 8, 16, 32, 64, 128)$ and infinite $\alpha$ are reported. From the table, MSP (Row 2) achieves the best result, slightly better than entropy based method (Row 8). For different ID scores, using sigmoid function with smaller $\alpha$ is better than Heaviside step function ($\alpha = \infty$). The underlying reason is that continuous weights are helpful for model fusion compared to binary weights. We also find that MaxLogit and Energy do not work well, which is contrary to the observations in prior work of OOD detection. We conjecture that, for these methods, the ID scores from different classifiers are not well normalized for comparison.

**Effect of Scaling Factor $\alpha$:** The $\alpha$ is the scaling factor of the score difference. Its effect for entropy and MSP based methods are shown in Fig. 3. Larger $\alpha$ damages the performance on novel classes, but improves the performance on base classes. Thus, it in fact plays a role in balancing the two performances. This makes the balance tunable compared to using Heaviside step function.

**Different Backbones:** The proposed method also works for other vision backbones. To validate this point, the base-to-novel and domain generalization results with ResNet-50 and ViT-B/32 are presented in Table 5. With OOD detection, the harmonic mean accuracies in base-to-novel setting are improved uniformly (0.07%~3.31%). Meantime, methods with OOD detection can achieve similar or better mean accuracies on target domains in most cases (6 out of 8).

### Conclusion

Existing methods for vision-language prompt tuning have poor generalization ability. This work proposes a new viewpoint for balancing the base and novel performance across classes by considering this problem as an OOD detection problem. Based on this idea, we propose a competition based scoring function to generate dynamic weight of zero-shot and few-shot classifier. By extensive experiments, even with weak OOD detectors, fused classifier demonstrates better generalization ability compared to the original one. One limitation of our method is that the effectiveness may be less significant when the two classifiers being fused have little complementarity. Future work will focus on exploiting stronger OOD detectors and addressing the above limitation.
Acknowledgments

This research was supported by the National Natural Science Foundation of China under Grants 62306310 and 62076242, and the National Key Research and Development Program of China under Grant No. 2018AAA0100400.

References


Zhu, B.; Niu, Y.; Han, Y.; Wu, Y.; and Zhang, H. 2023. Prompt-aligned Gradient for Prompt Tuning. ICCV.