DART: Dual-Modal Adaptive Online Prompting and Knowledge Retention for Test-Time Adaptation

Zichen Liu, Hongbo Sun, Yuxin Peng, Jiahuan Zhou*

Wangxuan Institute of Computer Technology, Peking University
lzc20180720@stu.pku.edu.cn, {sunhongbo, pengyuxin, jiahuanzhou}@pku.edu.cn

Abstract

As an up-and-coming area, CLIP-based pre-trained vision-language models can readily facilitate downstream tasks through the zero-shot or few-shot fine-tuning manners. However, they still face critical challenges in test-time generalization due to the shifts between the training and test data distributions, hindering the further improvement of the performance. To address this crucial problem, the latest works have introduced Test-Time Adaptation (TTA) techniques to CLIP which dynamically learn text prompts using only test samples. However, their limited learning capacity due to the overlook of visual modality information, and the underutilization of knowledge in previously seen test samples result in reduced performance. In this paper, we propose a novel Dual-modal Adaptive online prompting and knowledge Retention method called DART to overcome these challenges. To increase the learning capacity, DART captures knowledge from each test sample by learning class-specific text prompts and instance-level image prompts. Additionally, to fully leverage the knowledge from previously seen test samples, DART utilizes dual-modal knowledge retention prompts to adaptively retain the acquired knowledge, thereby enhancing the predictions on subsequent test samples. Extensive experiments on various large-scale benchmarks demonstrate the effectiveness of our proposed DART against state-of-the-art methods.

Introduction

Recently, the emergence of CLIP-based pre-trained vision-language models (Radford et al. 2021; Zhou et al. 2022b,a) has significantly propelled the advancement in computer vision. Through the exploration of appropriately designed text prompts, the pre-trained CLIP can be adapted to various downstream tasks for test-time inference (Luo et al. 2022; Wang et al. 2022a,c; Gal et al. 2022). However, manual crafting of task-specific prompts requires linguistic expertise and is time-consuming. A naive solution, that fine-tuning the entire pre-trained CLIP model on downstream tasks, will inevitably incur a large computational overhead and hinder the generalization ability of CLIP. An alternative approach involves adjusting the text prompts through few-shot fine-tuning, allowing the model to adapt to down-stream tasks (Zhou et al. 2022b,a; Khattak et al. 2023). However, due to the distribution shifts between the training and test data, such methods may encounter severe performance limitations. To address this issue, the latest approaches (Shu et al. 2022; Niu et al. 2023) focus on enhancing the performance of pre-trained models during the test phase by adapting them to fit the distribution of test data, which is known as Test-Time Adaptation (TTA).

Facing the practical challenge of the small batch size of test-time inference, or even one individual sample per batch, various TTA methods have been proposed (Wang et al. 2021; Shu et al. 2022; Niu et al. 2022; Döbler, Marsden, and Yang 2023; Niu et al. 2023). These methods primarily adopt three strategies, batch normalization calibration (Schneider et al. 2020; Wang et al. 2021; Niu et al. 2023).
The recent pre-trained vision-language models including CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021), have presented a surprising ability to learn general visual representations for downstream tasks in a zero-shot manner through proper prompts. Specifically, CLIP aimed to use “a photo of a CLS” as a prompt on the language side for zero-shot image classification. However, its classification performance heavily depends on the elaborately designed text prompts, which requires time-consuming prompt engineering by experts. Therefore, few-shot approaches like CoOp (Zhou et al. 2022b) and CoCoOp (Zhou et al. 2022a) are proposed to regard the text prompt as learnable parameters and employ a small amount of downstream task data for prompt training. Additionally, MaPLe (Khattak et al. 2023) proposes to train cross-modal prompts for the adaptation of CLIP. However, all the above methods have to rely on the collection of training data from downstream tasks. Due to the variations between training and test data, they could not generalize well on unseen test data from shifted data distributions which severely limits their practical effectiveness.

Test-Time Adaptation

In realistic scenarios, the test data always undergo natural variations or corruptions, resulting in data distribution shifting between the training and test phases (Hendrycks and Dietterich 2019; Koh et al. 2021). Therefore, even the large-scale pre-trained models are also difficult to generalize well on test data when domain shifting occurs (Recht et al. 2018). Recently, various Test-Time Adaptation (TTA) approaches are proposed to adaptively adjust the pre-trained models in the test phase to fit the distribution of test data (Schneider et al. 2020; Sun et al. 2020; Wang et al. 2021).

In terms of model parameter optimization, several methods (Schneider et al. 2020; Wang et al. 2021; Shu et al. 2022; Niu et al. 2023) propose to capture domain variations in test data by optimizing the batch normalization layers but are severely limited by the model architecture. Besides, various approaches (Yuan, Xie, and Li 2023; Döbler, Marsden, and Yang 2023) utilize consistency regularization to ensure stable model predictions when the data are perturbed slightly, and the model is optimized by using unlabeled test samples. Typically, these methods employ a teacher-student network architecture, where different augmented samples are fed into the two models whose outputs are constrained to be as close as possible. However, these methods usually update the entire model and can not retain the knowledge of training data well to assist in the prediction of the current data. Based on this, recent works (Wang et al. 2022b; Niu et al. 2022; Shu et al. 2022) propose using anti-forgetting techniques to preserve the knowledge of the training data during TTA, aiding the predictions of test samples. However, they do not effectively explore the historical knowledge from the seen test data, still resulting in limited performance on unseen test samples.

Prompt Learning

Prompts are initially applied in the field of natural language processing (NLP) (Ponti et al. 2020; Brown et al. 2020) by manually designing text prompts to make adaptive adjustments to downstream tasks. Although manually designed prompt templates are intuitive and promising, they require tremendous human effort and specific expertise which are
costly. To address this issue, numerous NLP methods (Li and Liang 2021; Tsimpoukelli et al. 2021) no longer focus on designing human-interpretable natural language prompt templates but treat prompts as learnable parameters, which greatly increase the flexibility and diversity of prompts.

Recently, there are various methods (Jia et al. 2022; Gao et al. 2022; Chen et al. 2022) migrating prompt learning to the field of computer vision. The main idea is concentrated on adopting the pre-trained vision transformer (Dosovitskiy et al. 2020) to downstream tasks by training a small number of prompt parameters. After the emergence of VPT (Jia et al. 2022) which initially migrates prompt learning to the field of image recognition, prompting in vision has been rapidly spread to various tasks such as image recognition (Chen et al. 2022), incremental learning (Wang et al. 2022e,d; Wang, Huang, and Hong 2022). However, the aforementioned methods need to utilize sufficient training data to train prompts before they can be applied to downstream tasks, thus they can hardly adjust prompts for online test data during test-time inference.

Therefore, few works have been proposed recently to focus on the important and promising direction of test-time online prompting. DePT (Gao et al. 2022) proposes a hierarchical self-training model to dynamically train the learnable prompts and classifier of the visual model at test time to cope with variations of test data. Moreover, TPT (Shu et al. 2022) is designed as a TTA method for the pre-trained CLIP model, which tunes the instance-level text prompt by minimizing the entropy loss. However, for a vision-language model like CLIP, TPT only concentrates on the prompt on the text side but lacks the utilization of the important image side, which greatly limits the multi-modality ability of the CLIP model to tackle the test data.

**The Proposed Method**

**Problem Setting and Notations**

In this work, we focus on the test-time adaptation (TTA) scenario, where the distribution of data in the test phase differs from the training phase. The training data are denoted as \( \mathcal{X}_n \), and the test data during the online test-time are denoted as \( \mathcal{X} = \{x_i\}_{i=1}^n \) which exhibits a different distribution against \( \mathcal{X}_n \). For each \( x_i \), we first learn dual-modal instance-level prompts \( \mathcal{P} = \{P^T, P^I\} \) for it, then apply \( \mathcal{P} \) into the pre-trained vision-language model \( \theta \) to obtain the predicted category \( y_i = \theta(x_i; \mathcal{P}), y_i \in \mathcal{Y} \) of input sample \( x_i \).

**Test-Time Adaptation for CLIP**

A pre-trained CLIP model \( \theta = \{E_T, E_I\} \) consists of two encoders, one for the text modality \( E_T \) and the other one for the image modality \( E_I \). These two encoders separately encode text and image inputs into the text-image cross-modality representation space. Generally, the architecture of the text encoder \( E_T \) is a Transformer model (Vaswani et al. 2017), as well as the image encoder is a CNN (He et al. 2016) or a ViT (Dosovitskiy et al. 2020). The CLIP model is trained by a contrastive loss with the goal of maximizing the cosine similarity of the matched text-image pairs and minimizing the cosine similarity of the unmatched pairs.

For a test image \( x \in \mathbb{R}^{C \times H \times W} \), the representation of image \( x \) can be obtained as \( r^I = E_I(x) \). For the labels, a hand-crafted prompt \( P = \text{“a photo of a”} \) combined with all
class names in $\mathcal{Y}$, i.e. $\{[P; c_1], [P; c_2], \ldots, [P; c_N]\}$ are fed into the text encoder $E_T$ to obtain the label representations $\{r^T_1, r^T_2, \ldots, r^T_N\}$, where $r^T_j = E_T(P; c_i)$. Then the prediction probability $P(y = j|x)$ can be obtained as:

$$P(y = j|x) = \frac{\exp(s_j/\tau)}{\sum_{i=1}^N \exp(s_i/\tau)},$$

where $s_j := \cos(r^T_j, r^T_j)$ and $\tau$ is the temperature of the softmax function. Although classification prediction can be obtained via Eq. 1, its performance can be severely limited when test data suffers from significant data distribution shifting against training samples. Therefore, we propose a dual-modal adaptive online prompting and knowledge retention (DART) method for test-time adaptation to improve the online generalization ability of CLIP.

### Dual-Modal Online Prompting in DART

As mentioned above, the text input for the inference of CLIP is $\{[P; c_1], [P; c_2], \ldots, [P; c_N]\}$, which means all classes share a same prompt. The latest work e.g. TPT (Shu et al. 2022) followed this setting and their performance is limited by using a single prompt for different classes.

In order to mitigate the discrepancy between different classes during online test time, we propose to adopt a class-specific text prompt $p^T_j$ for each class $c_i$. Then we initialize a group of class-specific text prompts as below:

$$P^T = \{P^T_1, P^T_2, \ldots, P^T_N\}.$$  

With $P^T$, the text representation of a class $c_i$ can be computed as follows:

$$r^T_i = E_T(P^T_i, c_i).$$

The instance-level class-specific text prompts $P^T$ helps the text encoder $E_T$ to explore the semantic information in the text side related with the current image $x$.

Unlike the existing CLIP-based few-shot or test-time prompt learning methods (Zhou et al. 2022b;a; Shu et al. 2022) only adjusting the text prompts, our DART elaborately integrates visual prompts $p^T$ into the image encoder $E_I$ via an instance-level adjustment for different test samples as below:

$$r^T = E_I(x, P^T),$$

The instance-level visual prompt $P^T$ assists the image encoder $E_I$ in utilizing the intrinsic semantic information.

### Adaptive Knowledge Retention in DART

Through dual-modal online prompting, DART comprehensively captures knowledge from individual test samples. To retain and harness the knowledge unearthed from individual samples, aiding the test of subsequent samples, we first design dual-modal knowledge retention prompts for each category to retain knowledge as below:

$$\hat{P}^T = \{\hat{P}^T_1, \hat{P}^T_2, \ldots, \hat{P}^T_N\};$$

$$\hat{P}^I = \{\hat{P}^I_1, \hat{P}^I_2, \ldots, \hat{P}^I_N\};$$

According to Eq. 1, by using the dual-modal prompts $P^T$, $P^I$ of $t$-th test sample $x_t$, we can obtain its predicted class $j = \hat{y}_t \in \mathcal{Y}$ and similarity with corresponding text prompt $s_j = \cos(r^T_j, r^T_j)$. An intuitive idea is that when the similarity $s_j$ and prediction confidence $P(y = j|x_t)$ are higher, the dual-modal prompts $P^T$, $P^I$ contains more useful knowledge. So a fusing weight $\beta_t$ is calculated as below:

$$\beta_t = 1 - e^{-s_j/h},$$

where $h$ is a temperature hyper-parameter. To adaptively retain knowledge from seen test samples, then we merge $P^T$, $P^I$ with the corresponding class-specific knowledge retention prompts $\hat{P}^T_j$, $\hat{P}^I_j$ using the calculated weight $\beta_t$ to preserve the learned knowledge:

$$\tilde{P}^T_j \leftarrow \hat{P}^T_j \cdot (1 - \beta_t) + P^T \cdot \beta_t,$$

$$\tilde{P}^I_j \leftarrow \hat{P}^I_j \cdot (1 - \beta_t) + P^I \cdot \beta_t,$$

where $\leftarrow$ denotes updating the value of a variable. When the next test sample $x_{t+1}$ is coming, for the test modality, $P^T$ is used for initialization to leverage the knowledge from past test samples:

$$P^T \leftarrow P^T \cdot (1 - w_T) + \tilde{P}^T \cdot w_T,$$

where $w_T$ is a hyper-parameter. For the imaging modality, since only one prompt is used and the category of the image cannot be known in advance, the information from previous samples can only be utilized in the test-time inference phase. Assuming that in the test-time training phase, the class $j$ will get the highest confidence, then the past sample’s knowledge is utilized as follows:

$$P^I \leftarrow P^I \cdot (1 - w_I) + \tilde{P}^I \cdot w_I,$$

where $w_I$ is a hyper-parameter. In summary, our DART can adaptively retain the knowledge learned from high-confidence samples through dual-modal knowledge prompts and utilize them to assist in predicting subsequent unseen samples.

### The Optimization of DART

As shown in Figure 2, the introduced prompts in DART can be readily optimized in an online learning manner where the pre-trained CLIP is frozen. Follow the protocol of TPT (Shu et al. 2022), for each test image $x$, we first augment it to $\{x^a_1, x^a_2, \ldots, x^a_B\}$ where $B$ is the batch size in training. Then, all the $B$ augmented images are fed into the CLIP model to get the prediction probability distribution of $x^a_1$:

$$\{P(p(y = j|x^a_1)), P(p(y = j|x^a_2)), \ldots, P(p(y = j|x^a_B))\},$$

where $P = \{P^T, P^I\}$ is the designed dual-modal instance-level prompts of image $x$. To reduce the noise interference caused by some unsuitable augmentations, we eliminate the predictions with low self-confidence. The self-confidence of one augmented view $x^a_1$ is computed as below:

$$H(x^a_1) = \sum_{j=1}^N P(p(y = j|x^a_1)) \log P(p(y = j|x^a_1)).$$
Then we select the top $\rho$ ratio samples of high self-confidence, where $\rho$ is a pre-defined hyper-parameter. Finally, we optimize the prompts $\mathcal{P}$ by minimizing the entropy of average prediction distribution over the selected confident samples, i.e.,

$$
\arg \min_{\mathcal{P}} - \frac{1}{N} \sum_{j=1}^{N} \mathcal{T}_{\mathcal{P}}(y = j|x) \log \mathcal{T}_{\mathcal{P}}(y = j|x),
$$

where

$$
\mathcal{T}_{\mathcal{P}}(y = j|x) = \frac{1}{\rho \mathcal{B}} \sum_{i=1}^{\rho \mathcal{B}} \mathcal{P}(y = j|x_i^a).
$$

Experiments

We first introduce the benchmarks used for evaluating our DART and the compared state-of-the-art methods, then the implementation details are demonstrated accordingly. Finally, extensive experiment results and analyses are further presented along with discussions about the ablation study of our proposed method. Moreover, more experimental results and analyses are included in our Supplementary.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Publication</th>
<th>I-A</th>
<th>I-R</th>
<th>I-S</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP</td>
<td>ICM 2021</td>
<td>47.87</td>
<td>73.98</td>
<td>46.09</td>
<td>55.98</td>
</tr>
<tr>
<td></td>
<td>ICM 2021</td>
<td>49.89</td>
<td>77.65</td>
<td>48.24</td>
<td>58.59</td>
</tr>
<tr>
<td>CoOp</td>
<td>ICCV 2022</td>
<td>49.71</td>
<td>77.21</td>
<td>47.99</td>
<td>57.64</td>
</tr>
<tr>
<td>CoCoOp</td>
<td>CVPR 2022</td>
<td>50.63</td>
<td>76.18</td>
<td>48.75</td>
<td>58.52</td>
</tr>
<tr>
<td>MaPLe</td>
<td>CVPR 2023</td>
<td>50.90</td>
<td>76.98</td>
<td>49.15</td>
<td>59.01</td>
</tr>
<tr>
<td>DART</td>
<td>This Paper</td>
<td>60.56</td>
<td>79.56</td>
<td>49.76</td>
<td>63.29</td>
</tr>
</tbody>
</table>

Table 1: The Acc@1 comparison results against CLIP and the latest few-shot fine-tuning methods on three benchmark datasets. The I-A, I-R, and I-S represent ImageNet-A, ImageNet-R, and ImageNet-Sketch respectively.

Datasets

Since the data distribution shifting will inevitably occur in real-world scenarios, the experiments are conducted on three large-scale benchmarks, ImageNet-A (Hendrycks et al. 2021b), ImageNet-R (Hendrycks et al. 2021a), and ImageNet-Sketch (Wang et al. 2019) which are variants of the ImageNet (Deng et al. 2009) dataset to evaluate the performance of different methods for improving the test-time generalization ability of CLIP. These benchmarks have been considered as out-of-distribution data for ImageNet previously (Radford et al. 2021), and we follow the same setting in our experiments.

- **ImageNet-A** (Hendrycks et al. 2021b) is a natural adversarial image dataset that contains natural images misclassified by ResNet-50 (He et al. 2016) in ImageNet (Deng et al. 2009). In total, it contains 7,500 images of 200 categories. These misclassified images in ImageNet-A usually suffer from various distribution shifting which poses critical challenges to the test-time generalization ability of models.

- **ImageNet-R** (Hendrycks et al. 2021a) is a multi-domain (e.g. art, cartoon, painting) image dataset consisting of 30,000 images and 200 categories of ImageNet. All the images in ImageNet-R are collected from 15 different style domains, thus there exist drastic domain gaps between different images.

- **ImageNet-Sketch** (Wang et al. 2019) is a sketch image dataset consisting of 50000 images, 50 images for each of the 1000 ImageNet classes. The images in ImageNet-Sketch are all black and white, and their distribution differs significantly from the training data of CLIP.

Comparison Methods

In the experiments, we compared our proposed DART with state-of-the-art TTA and few-shot fine-tuning methods designed for CLIP. In detail, TPT (Shu et al. 2022) is a test-time prompt-tuning approach focusing on fine-tuning a learnable text prompt for CLIP. Tent (Wang et al. 2021), EATA (Niu et al. 2022), SAR (Niu et al. 2023) and RMT (Dobler, Marsden, and Yang 2023) are general TTA methods. CoOp (Zhou et al. 2022b) and CoCoOp (Zhou et al. 2022a) are few-shot prompt-tuning methods aiming at fine-tuning the text prompts. MaPLe (Khattak et al. 2023) is a few-shot method training cross-modal prompts on each dataset. Following the same protocol in (Zhou et al. 2022b,a), 16-shot extra training images of each category are provided for fine-tuning. For test-time inference, once the text prompts are learned, the aforementioned methods are directly used in the same way as CLIP does. In addition, since the pre-trained CLIP can be directly applied to downstream classification tasks in a zero-shot manner, we also regard it as a baseline method. Two different text prompt settings of CLIP are evaluated, one is the default “a photo of a” and the other one is the ensemble of 80-hand-crafted prompts from (Radford et al. 2021).

Implementation Details

The pre-trained CLIP model with ViT-B/16 is used as our backbone (Radford et al. 2021). For each test image, we initialize all the text prompts in our DART as “a photo of a”. The image prompts are initialized with a uniform distribution of $(-1, 1)$ following the previous visual prompting methods (Wang et al. 2022e,d). The length of image prompts is set to 2, and they are added to the second layer of the CLIP image encoder. The hyper-parameters $h$, $w_T$, and $w_I$ of dual-modal knowledge retention prompts are set to 5000, 0.1, and 0.1 respectively. For the learning of DART, we use randomly resized crops to augment the single test sample to obtain a batch of $B = 64$ images, and the confidence threshold $\rho$ follows the same setting in (Shu et al. 2022). An Adam optimizer with a learning rate of 0.003 is used to optimize the prompts $\mathcal{P}$. All experiments are implemented on a single NVIDIA 4090 GPU.

Comparison with State-of-the-arts

The overall comparison results against the state-of-the-art TTA and few-shot fine-tuning methods on ImageNet-A, ImageNet-R, and ImageNet-Sketch are reported in Table 1.
Table 2: The comparison results against state-of-the-art TTA methods on three benchmark datasets. ViT represents the ViT-B/16 model pre-trained on ImageNet, and CLIP represents the pre-trained CLIP model with ViT-B/16 architecture.

Table 3: Ablation study about the different components of DART. $P^T$ and $P^I$ represent the class-specific text prompts and image prompts respectively. $\tilde{P}^T$ and $\tilde{P}^I$ represent knowledge retention text prompts and knowledge retention image prompts respectively. ✓ and ✗ represent without or with this component. When none of the components is used, the model degenerates to the baseline CLIP.

Table 4: Ablation study about the influence of different learnable text prompts.

As for the TTA methods including Tent, EATA, SAR, and RMT, they originally employed the pre-trained ViT on ImageNet as the backbone. Considering the distinct training data distributions and generalization capability between the pre-trained ViT and CLIP, to ensure a fair and equitable comparison, we conduct additional experiments with all comparison methods and DART, utilizing the pre-trained CLIP model as the same backbone. As demonstrated in Table 2, DART outperforms the second-best player TPT by 3.37% at average Acc@1 on all three datasets. Specifically, our proposed DART significantly outperforms TPT by 5.79% at Acc@1 on ImageNet-A. Since ImageNet-A consists of natural images misclassified by ResNet-50, this result verifies that our DART can well handle natural distribution shifting in the test phase. The same conclusion can also be confirmed by the experimental results on ImageNet-R and ImageNet-Sketch. Notably, when SAR employs ViT as the backbone, it achieves comparable performance as our DART at Acc@1 on ImageNet-Sketch. However, when using the same CLIP backbone, DART significantly outperforms SAR by 2.17%. This is credited to DART’s dual-modal online prompting and knowledge retention prompts, which effectively tap into and utilize information from samples during test time, even in cases of significant style variations in the samples (e.g. cartoon).

**Ablation Studies and Analyses**

The Influence of Different Components in DART. To verify the effectiveness of the proposed dual-modal adaptive online prompting and knowledge retention components in our DART, an ablation experiment is conducted on ImageNet-A. As demonstrated in Table 3, utilizing either the instance-level text prompts $P^T$ or the image prompts $P^I$ consistently enhance the robustness against the distribution shifting compared to the naive zero-shot CLIP. Moreover, with the integration of the text knowledge retention prompts $\tilde{P}^T$ or image knowledge retention prompts $\tilde{P}^I$, performance is further elevated beyond the utilization of single-modal prompts $P^T$, $P^I$ alone. This improvement can be attributed to the inherent ability of knowledge retention prompts to adaptively retain knowledge. Notably, employing the whole dual-modal adaptive online prompting and knowledge retention components yields the best results, as it efficiently captures information from each individual test sample and retains the knowledge from previously seen test samples, thereby facilitating the performance of the pre-trained CLIP.

---

**Learnable Text Prompt**

<table>
<thead>
<tr>
<th>Prompt</th>
<th>ImageNet-A</th>
<th>Learnable Text Prompt</th>
<th>Acc@1</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓  ✓  ✗</td>
<td>47.87/79.09</td>
<td>(“a photo of a [CLS]”)</td>
<td>59.53</td>
</tr>
<tr>
<td>✓  ✓  ✓</td>
<td>57.12/81.59</td>
<td>(“a photo of a”) in DART</td>
<td>60.56</td>
</tr>
</tbody>
</table>
The Influence of Different Learnable Text Prompts in DART. Different from TPT which learns a unified text prompt for all categories, our proposed DART proposes to utilize class-specific text prompts initialized as “a photo of a” for different categories during the test time. From the results on ImageNet-A in Table 4, our proposed class-specific text prompts outperformed the unified text prompt by 3.29% at Acc@1. The reason is that our class-specific text prompts are more flexible to capture and highlight the class-specific information of the test sample. Moreover, an extra experiment that simultaneously updates our class-specific text prompts and the CLS token at test time reports inferior performance than ours. This is mainly because the important semantic information contained by the CLS token may be hindered during online learning.

The Generalization across Different Backbones. To validate the generalization ability of DART, across different backbones, we conduct experiments using pre-trained CLIP models with various backbones. As shown in Figure 3, when employing the ViT-B/32, DART exhibits significant improvements at Acc@1 compared to CLIP and TPT, with increments of 11.81% and 6.79% respectively. For the higher-parameter ViT-L/14, DART demonstrates enhancements of 8.2% and 2.14% at Acc@1 compared to CLIP and TPT respectively. Furthermore, across different backbones, DART consistently exhibits further enhancements at Acc@5 over CLIP and TPT. This can be credited to DART’s dual-modal online prompting, which introduces additional learnable parameters as the backbone is frozen. This enables DART to adapt the pre-trained CLIP model across both modalities. Moreover, the dual-modal knowledge retention prompts effectively preserve and leverage knowledge learned from seen test samples, resulting in superior performance.

The Influence of Different Hyper-parameters in DART. There are several hyper-parameters in our DART. We initially conduct experiments to investigate the influence of the hyper-parameter $h$ in Eq. 7, which is responsible for generating the adaptive weight $\beta_t$. As illustrated in Figure 4, DART demonstrates insensitivity to variations in $h$. Notably, DART shows remarkable performance when $h$ falls within the range of 4500 to 6000. This is attributed to a favorable balance achieved between the retention of newly acquired knowledge and historical knowledge. Subsequently, we explore the impact of the fusion coefficients $w_T$ and $w_I$ employed in Eq. 10 and Eq. 11. These coefficients exhibit a similar trend, with their optimal performance observed within the range of 0.05 to 0.15. This outcome demonstrates that while knowledge retained through retention prompts provides auxiliary support, newly captured knowledge remains more crucial and tailored for accurate predictions on the current samples. Then we conduct experiments to investigate which layer of the image encoder the proposed image prompts $p^I$ should be added. Considering the limited learning condition (only one test sample available) during online test time, we propose to add the prompts to only one layer. As presented in Figure 4, adding the prompts to the second layer of the image encoder performs the best.

Conclusion

In conclusion, we propose a novel dual-modal adaptive online prompting and knowledge retention method, named DART, for test-time adaptation of CLIP-based pre-trained vision-language models. Specifically, our approach involves learning class-specific text prompts and instance-level image prompts for each test sample, effectively capturing the knowledge within an individual test sample to enhance the model’s prediction accuracy. Moreover, we design text and image knowledge retention prompts to adaptively retain and utilize the knowledge from previously seen test samples, to facilitate the predictions of subsequent test samples. This enables our DART to adapt to new test instances and improve overall performance during test-time adaptation. Extensive experiments on various large-scale benchmarks demonstrate the effectiveness of DART against state-of-the-art approaches. Our work investigates a promising direction, addressing the challenging problem of training-test data distribution shifting in pre-trained vision-language models using dual-modal prompting and knowledge retention.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62376011, 61925201, 62132001).
References


