Make Prompts Adaptable: Bayesian Modeling for Vision-Language Prompt Learning with Data-Dependent Prior

Youngjae Cho[1], **HeeSun Bae**[1], **Seungjae Shin** [1],**Yeo Dong Youn** [2], **Weonyoung Joo** [3], **Il-Chul Moon** [1]
[1] KAIST
[2] Seoul National University
[3] Ewha Womans University
{genius.cho, cat2507, tmdwo0910}@kaist.ac.kr, dbsduehd95@snu.ac.kr, weonyoungjoo@ewha.ac.kr, icmoon@kaist.ac.kr

Abstract

Recent Vision-Language Pretrained (VLP) models have become the backbone for many downstream tasks, but they are utilized as frozen model without learning. Prompt learning is a method to improve the pre-trained VLP model by adding a learnable context vector to the inputs of the text encoder. In a few-shot learning scenario of the downstream task, MLE training can lead the context vector to over-fit dominant image features in the training data. This overfitting can potentially harm the generalization ability, especially in the presence of a distribution shift between the training and test dataset. This paper presents a Bayesian-based framework of prompt learning, which could alleviate the overfitting issues on few-shot learning application and increase the adaptability of prompts on unseen instances. Specifically, modeling data-dependent prior enhances the adaptability of text features for both seen and unseen image features without the trade-off of performance between them. Based on the Bayesian framework, we utilize the Wasserstein Gradient Flow in the estimation of our target posterior distribution, which enables our prompt to be flexible in capturing the complex modes of image features. We demonstrate the effectiveness of our method on benchmark datasets for several experiments by showing statistically significant improvements on performance compared to existing methods. The code is available at https://github.com/youngjae-cho/APP.

Introduction

Recently, Vision-Language Pretrained models (VLP) (Radford et al. 2021; Jia et al. 2021) have been used as backbones for various downstream tasks (Shen et al. 2022; Ruan, Dubois, and Maddison 2022), and the pre-trained models have shown successful adaptation. Since these pre-trained models are used as-is in downstream tasks, *prompt learning* adds a context vector to the input of pre-trained model, so the context vector becomes the learnable parameter to improve the representation from the pre-trained model (Zhou et al. 2022b) for the downstream task. For instance, a text input is concatenated to a context vector, and the new text input could be fed to the text encoder. The learning of context vector comes from the back-propagation after the feed-forward of the concatenated text input. Since there is only a single context vector without being conditioned by inputs, the inferred value of context vector becomes a static single context defined for the given downstream task.

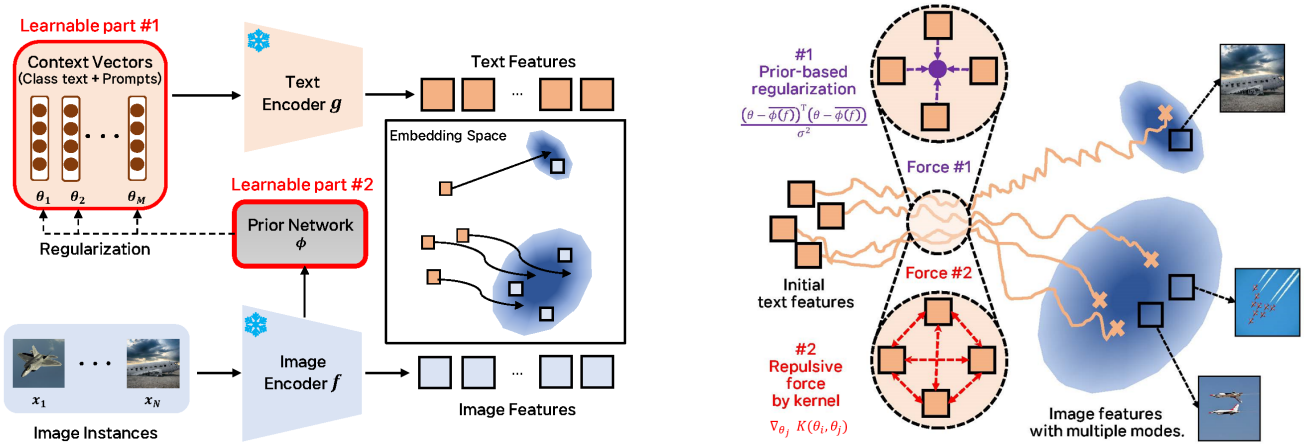
Whereas improving parameter-frozen VLP models by additional input context vector is a feasible solution, it can potentially overfit to a dense area of image features in few-shot learning. Since text features are hard to capture the multi-modes of image features in MLE training, it could fail to infer the minor area of image features, eventually degrading the performance. In addition, MLE training undermines the generalization capability of VLP models especially when there is a distribution shift between the training and testing (Zhou et al. 2022a). Although several input-conditioned prompt learning (Zhou et al. 2022a; Derakhshani et al. 2023) tried to generalize unseen datasets, it inevitably undermines the performance of seen datasets.

To alleviate the impact of uncertainty arising from a few-shot learning scenario, our paper proposes Adaptive Particle-based Prompt Learning (APP), which utilizes a Bayesian inference for prompt learning with a data-dependent prior as shown in Figure 1. Through regularization using this data-dependent prior, the context vector is directed toward capturing the diverse modes in image features among the seen data instances. Additionally, we approximate the posterior distribution via Wasserstein Gradient Flow to enhance the flexibility of our text features to infer the complex image features. Furthermore, we extend the modeling of the data-dependent prior to unseen test instances to adapt the distribution shift. This adaptation of the context vector to the unseen instances enhances the model's resilience in the face of distribution shifts, providing robustness to these variations.

We summarize our contributions in two aspects.

1. Enhancing Flexibility of Prompt: By approximating prompt posterior with Wasserstein Gradient Flow, our context vectors can be more flexibly utilized to infer the complex image feature spaces.

2. Enhancing Adaptability of Prompt: By modeling data-dependent prior based on the image feature information, text features capture the multi-modes of seen image features and adapt to unseen image features, which leads to the improved performances of seen and unseen datasets without trade-off.



(a) Structure view of APP

(b) Learning dynamics of APP

Figure 1: Structure (left) and learning dynamics (right) of APP. Multiple context vectors are particles of approximated distribution and image conditioned prior can guide the context vector to capture the multi modes.

Preliminary

Formulation of Prompt Learning

Deterministic Prompt Learning The goal of prompt learning, i.e. CoOp (Zhou et al. 2022b), is to facilitate adaptation of a given Vision-Language Pretrained model for a target task through learning arbitrary context vectors. In this setting, the pre-trained VLP model is frozen under the prompt learning task, and the additional learnable input is added to text inputs. Finally, the learning gradient is obtained from adapting the target task, a.k.a. downstream task. When we define (X, Y) as a pair of image and its label (i.e. text phrase), Eq. 1 shows a log-likelihood of prompt learning in the image classification.

$$\mathcal{L}_{CE}(\theta, X, Y) = -\log p(Y|X, \theta)$$
$$= -\sum_{i=1}^{N}\{\log(\frac{\exp(\text{sim}(g(\theta, y_i), f(x_i))/\tau)}{\sum_{k=1}^{C}\exp(\text{sim}(g(\theta, y_k), f(x_i))/\tau)})\} \quad (1)$$

Here, f and g are image and text encoders from VLP models, respectively; and they are frozen in prompt learning. Therefore, the only learnable part is θ, which is a context vector. $\theta \in \mathbb{R}^d$ is learned as a unique vector for each downstream task without discriminating the data instances. Since g is often implemented as a transformer to take sequential inputs of any length, g does not need to be modified to accept y and θ. Additionally, $\text{sim}(\cdot)$ represents cosine similarity, and τ is the annealing temperature.

From Eq. 1, we define the prompt as $\{\theta, y_i\}$, which is the concatenation of the context vector and the label. By minimizing Eq. 1, θ is learned to maximize the alignment of the space between the image feature $f(x)$ and the text feature $g(\theta, y_i)$. Therefore, the learnable part of prompt learning contributes from the input space side, rather than the frozen VLP model parameters. There were some follow-up researches on CoOp. For example, CoCoOp (Zhou et al. 2022a) extended CoOp by learning the image-conditional context vector as $\theta + \phi(f(x_i))$, where ϕ is a neural network to map image feature to the prompt space.

Probabilistic Prompt Learning Given a few instance of training dataset, the point estimate of text feature given prompt is hard to capture unseen image feature. ProDA (Lu et al. 2022) is the first probabilistic model, where the text feature-given prompt is approximated as a Gaussian distribution with a regularizer to enhance the diversity of text feature. PLOT (Chen et al. 2023) formulates the prompt learning as optimal transport, where image and text features are defined as a discrete distribution by Dirac measure. The text features, given as multiple prompts, are assigned to the locality of image features to learn diverse semantics.

Bayesian Probabilistic Prompt Learning Since MLE training can induce overfitting with a few training dataset, Bayesian inference is needed to mitigate the high data variance from such a limited dataset. BPL (Derakhshani et al. 2023) is the first prompt learning model from the view of Bayesian inference, which uses variational inference to approximate the posterior distribution with a parameterized Gaussian distribution. The objective function of BPL is defined as follows:

$$\mathbb{E}_{q(r|X)}[\log(p(Y|X, \theta, r))] - D_{KL}(q(r|X)||p(r)) \quad (2)$$

where r is a random variable conditioned on X, and r is added to a deterministic θ to turn it into a random variable, which is a reparameterization trick. In detail, the distribution of r is given by $q(r|X)$, which is a Gaussian distribution parameterized by $m(f(X))$ and $\Sigma(f(X))$, where $m(f(X))$ and $\Sigma(f(X))$ are functions of the image feature $f(X)$. The prior distribution over r is $p(r)$, which is a standard Gaussian distribution $N(0, I)$.

Wasserstein Gradient Flow

Bayesian inference is a solution to mitigate the uncertainty from modeling the posterior distribution of parameters. Of-

ten, the hurdle of the Bayesian inference is the inference of the posterior distribution, which could be difficult, i.e. modeling either prior or likelihood to be flexible without being conjugate to each other. For improving the posterior distribution to be flexible and multi-modal, we need an inference tool for this complex posterior distribution.

For instance, the JKO scheme (Jordan, Kinderlehrer, and Otto 1998) interprets variational inference as gradient flow, which minimizes the KL divergence between the variational distribution q and the true posterior distribution $\pi \propto \exp(-V(\theta))$ in Wasserstein Space, where $V(\theta)$ is an energy function of posterior distribution. The learning objective of this variational posterior inference $F(q)$ becomes the KL Divergence as follows.

$$F(q) \coloneqq D_{KL}(q||\pi) \approx \mathbb{E}_q[V(\theta) + \log q] \tag{3}$$

To compute the steepest gradient of $F(q)$, we define the Wasserstein Gradient Flow (WGF) as follows:

Definition 1. *Suppose we have a Wasserstein space $\mathcal{W}_2 = (\mathcal{P}_2(\mathbb{R}^d), W_2)$, $\mathcal{P}_2(\mathbb{R}^d) = \{\mu \in \mathcal{P}_2, \int ||\theta||^2 d\mu(\theta) < \infty\}$, $W_2(\mu_1, \mu_2) = min_{\omega \in \Pi(\mu_1,\mu_2)} \int ||\theta - \theta'||^2 d\omega(\theta, \theta')$.*

A curve of μ_t is a Wasserstein Gradient Flow for functional F, if it satisfies Eq. 4.

$$\partial_t \mu_t = \nabla \cdot (\mu_t \nabla_\theta \frac{\delta F(\mu_t)}{\delta \mu}) = \nabla \cdot (\mu_t \nabla_{W_2} F(\mu_t)) \tag{4}$$

WGF can be discretized as Stochastic Gradient Langevin Dynamics (SGLD) (Welling and Teh 2011; Chen et al. 2018) as Eq. 5, where each particle θ follows the true posterior distribution π with Gaussian perturbation.

$$\theta^i_{t+1} = \theta^i_t - h(\nabla_{\theta^i_t} V(\theta^i_t)) + \sqrt{2h}\epsilon, \epsilon \sim N(0, I) \tag{5}$$

Whereas the Gaussian noise can assure the diversity of parameters, the learning can be unstable, when the learning rate is high.

Hence, this paper relies on Stein Variational Gradient Descent (SVGD) (Liu and Wang 2016), which is a version of Wasserstein Gradient Flow with Reproducing Kernel Hilbert Space (RKHS) (Chen et al. 2018). In SVGD, interaction between particles θ_i guarantees their convergence to true posterior distribution with forcing diversity.

$$\theta^i_{t+1} = \theta^i_t - \frac{h}{M} \sum_{j=1}^{M} K(\theta^i_t, \theta^j_t) \nabla_{\theta^j_t} V(\theta^j_t) - \nabla_{\theta^j_t} K(\theta^i_t, \theta^j_t) \tag{6}$$

By using SVGD to approximate true posterior distribution, context vectors θ^j can be optimized to follow the true posterior distribution, effectively capturing a representation space of the image features.

Data-Dependent Prior In a Bayesian neural network, the prior is commonly chosen as the Standard normal distribution, i.e. BPL, which is data-independently initialized with zero means. Since such distribution does not include any information on data, it only regularizes the context vector in the neighbor of zero-mean. In many domains (Li et al. 2020; gil Lee et al. 2022), the data-dependent prior is utilized to improve the prior knowledge more informative.

This paper utilizes the data-dependent prior for prompt learning, which is not restricted to the standard Gaussian distribution. Specifically, this paper derives the prior distribution to be dependent on image features, which can have multiple modes in their distributions.

Method

This section introduces adaptive particle-based prompt learning (APP) by enumerating the model formulation and by explaining its inference method.

Formulation of Prompt Posterior Distribution

Following the CoOp formulation (Zhou et al. 2022b), we additionally reformulate the posterior distribution of context vector θ as Eq. 7.

$$\begin{aligned} \pi(\theta) = p(\theta|X, Y) &= \frac{p(Y|X, \theta)p(\theta|X)p(X)}{p(X, Y)} \\ &\propto p(Y|X, \theta)p(\theta|X) \end{aligned} \tag{7}$$

$\pi(\theta)$ is true posterior distribution, which is factorized with likelihood $p(Y|X, \theta)$ and conditioned prior $p(\theta|X)$. For likelihood Eq. 1, we follow the formulation of CoOp, Eq. 1, and we propose the image feature conditioned prior Eq. 8, where mean is parametrized as ϕ and $\overline{\phi(f)} \coloneqq \frac{1}{N}\sum_{i=1}^{N} \phi(f(x_i))$. We set the standard deviation of prior σ as a hyper-parameter.

$$\log p(\theta|X) \propto -\frac{(\theta - \overline{\phi(f)})^T(\theta - \overline{\phi(f)})}{\sigma^2} \tag{8}$$

Bayesian Adaptation of Prompt to Test data In few-shot learning, there is uncertainty on whether the test data will follow the distribution of training data or not. If there is a difference between the two data distributions, such a difference will harm the generalization ability of VLP model. To model the uncertainty from the mismatch between training and test datasets in the few-shot learning framework, we reformulate the posterior distribution to consider the uncertainty of a test image x'. Since training and test datasets are i.i.d sample; and because the prior is assumed to follow the Gaussian distribution; we derive the posterior distribution as Eq.9.

$$\begin{aligned} \pi(\theta) = p(\theta|X, Y, x') &= \frac{p(Y|X, \theta)p(\theta|X)p(\theta|x')p(X)}{p(X, Y)} \\ &\propto \underbrace{p(Y|X, \theta)p(\theta|X)}_{Training}\underbrace{p(\theta|x')}_{Testing} \end{aligned} \tag{9}$$

After approximating the Eq.7, we adapt the context vector θ with test data-dependent prior.

Variational Inference for Prompt Posterior

Since Eq. 7 is not tractable, we approximate the posterior distribution of π, using particle-based variational inference. Suppose that q is a probabilistic measure of variational distribution, which generates the context vector θ, in Wasserstein space. Eq. 11 defines the optimization problem to approximate the model posterior distribution by the variational distribution.

$$V(\theta) \coloneqq -\log p(Y|X, \theta) - \log p(\theta|X) \tag{10}$$

$$F(q) \coloneqq D_{KL}(q||\pi) \approx \mathbb{E}_q[V(\theta) + \log q] \tag{11}$$

To define the steepest direction of Eq. 11, we follow Wasserstein Gradient Flow Eq.4.

$$\begin{aligned}\partial_t \theta_t &= -\mathcal{K}_q \nabla_\theta (\frac{\delta F(q)}{\delta q}) \\ &= -[\int K(\theta, \theta') \nabla_{\theta'} V(\theta') dq - \int \nabla_{\theta'} K(\theta, \theta') dq]\end{aligned} \tag{12}$$

By discretizing Eq. 12, we derive the Stein Variational Grad (Liu and Wang 2016), where each context vector θ^j can be optimized as Eq. 6.

For Eq. 6, the first term can be interpreted as a smoothing gradient between context vectors θ^j and assure the convergence toward the true posterior distribution. The second term can be interpreted as the repulsive force between context vectors θ^j and guide the text features can cover the multi modes sparsely.

Parameter Training of Data-Dependent Prior

Since the prior has the parametrized mean ϕ, we pre-train the ϕ, which can map the image feature on the prompt space. To preserve the image feature information within our prior distribution, we propose to maximize the mutual information $I\left(\phi\left(f\left(x\right)\right); f(x)\right)$. In other words, this mutual information encourages the prior to capture the dependencies between the image features and the parameter of the prompt distribution. Due to the data processing inequality (Beaudry and Renner 2012), we derive inequality as follows:

$$I(\phi(f(X)); f(X)) \geq I(g(\phi(f(X)), \cdot); f(X)) \tag{13}$$

where ϕ can be learned to maximize the mutual information.

Proposition 1. *Suppose that the Markov chain assumption holds as* $f(X) \rightarrow \phi(f(X)) \rightarrow g(\phi(f(X)), \cdot)$*, then the lower bound of the mutual information,* $I(f(X); \phi(f(X)))$*, is derived as follows:*
$I(f(X); \phi(f(X))) \geq I(f(X); g(\phi(f(X)), \cdot)) \geq \log C - \mathcal{L}_{CE}(\phi(f(X)), X, Y)$

Based on Proposition 1, we can maximize the mutual information by minimizing the cross entropy. The full training scenario is reported in Algorithm 1.

Adaptation θ with Test Data-Dependent Prior

Following the training of the posterior distribution as described in Eq. 7, we extend our approach to accommodate an unseen data instance, x', within the posterior distribution. This involves updating the context vector $\theta \sim q$ through a linear combination with the prior mean $\phi(f(x'))$. For the sake of simplicity, we perform a weighted average of the text features, which can be outlined as follows. The adaptation scenario is reported in Algorithm 2.

$$g(\theta^*, y) = \alpha g(\theta, y) + (1 - \alpha) g(\phi(f(x')), y) \tag{14}$$

Algorithm 1: Training Scenario of APP

1: **Input**: Dataset $\mathcal{D} = \{X, Y\}$, Context vector θ^i, Prior Network ϕ
2: **while** not converged **do**
3: Compute $\mathcal{L}_{CE}(\phi(f(X)), X, Y)$
4: Update $\phi_{t+1} = \phi_t - h\nabla_\phi \mathcal{L}_{CE}$
5: **end while**
6: **while** not converged **do**
7: Compute $V(\theta) = -\log p(Y|X, \theta) - \log p(\theta|X)$
8: Update $\theta^i_{t+1} = \theta^i_t - \frac{h}{M}\sum_{j=1}^{M}[K(\theta^i_t, \theta^j_t)\nabla_{\theta^j_t} V(\theta^j_t) - \nabla_{\theta^j_t} K(\theta^i_t, \theta^j_t)]\,, \forall i \in [1, ..., M]$
9: **end while**

Algorithm 2: Test Scenario of APP

1: **Input**: Test image instance x', Context vector θ^i, Prior Network ϕ
2: Training as Algorithm 1
3: Compute $g(\theta^*, y_j)$ as Eq.14, $\forall j \in [1, .., K]$
4: $y' = \text{argmax}_{y_j} \text{sim}(g(\theta^*, y_j), f(x'))$

Results

Experiment Settings

We conduct three distinct experiments; Few-shot classification, domain generalization of ImageNet, and base-to-new generalization following PLOT (Chen et al. 2023). For Few-shot classification, we conduct 11 image datasets, including Caltech101 (Fei-Fei, Fergus, and Perona 2004), DTD (Cimpoi et al. 2014), EuroSAT (Helber et al. 2019), FGVCAircraft (Maji et al. 2013), Oxford 102 Flower (Nilsback and Zisserman 2008), OxfordPets (Parkhi et al. 2012), Food101 (Bossard, Guillaumin, and Van Gool 2014), StanfordCars (Krause et al. 2013), Sun397 (Xiao et al. 2010), UCF101 (Soomro, Zamir, and Shah 2012), and ImageNet (Deng et al. 2009). We followed the training setting of PLOT (Chen et al. 2023), where the training shots are chosen in 1, 2, 4, 8, 16 shots, and we train 50, 100, 100, 200, and 200 epochs for each shot. For ImageNet, we train the prompts in 50 epochs for all shots. Before the training context vector θ, we train the prior mean ϕ in 10, 20, 20, 40, and 40 epochs for each shot. For domain generalization of ImageNet, we train prompts about ImageNet as a source dataset and report the accuracy of the source dataset and target datasets, including ImageNetV2 (Recht et al. 2019), ImageNet-A (Hendrycks et al. 2021b), ImageNet-R (Hendrycks et al. 2021a), and ImageNet-Sketch (Wang et al. 2019). For base-to-new generalization, we train prompts using 16 shots for each of 11 datasets for the base class and report the performance of base and new classes.

As a common setting, we conduct three replicated experiments to report the performances, and we use CLIP (Jia et al. 2021) as the backbone network, where ResNet50 (He et al. 2016) is chosen as the image encoder. We report more details of the setting in the Appendix.

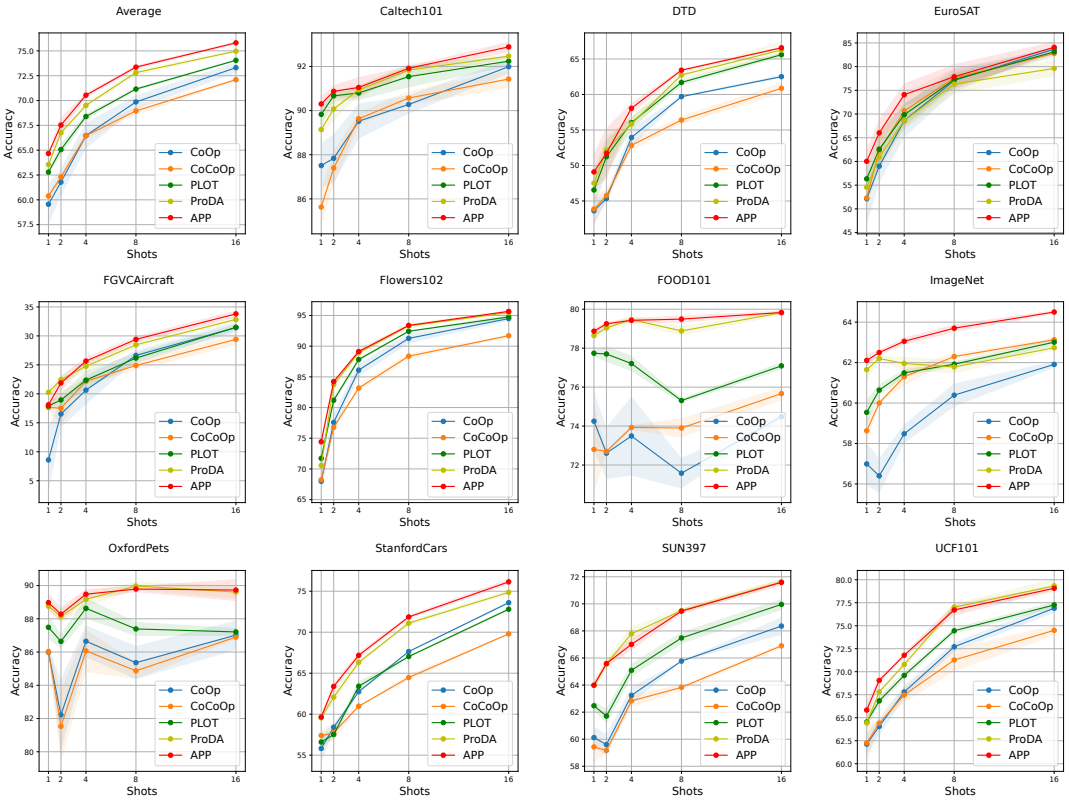


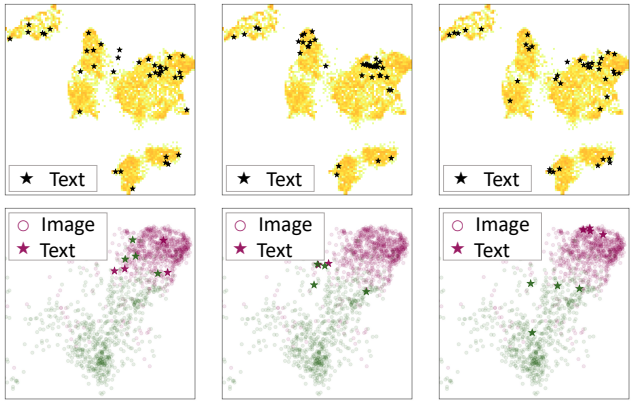
Figure 2: Result of Few-shot Classification. We conduct three-replicated experiments.

Baselines We compare the performance of our method, APP with CoOp (Zhou et al. 2022b), CoCoOP (Zhou et al. 2022a), PLOT (Chen et al. 2023), and ProDA (Lu et al. 2022). We do not include BPL (Derakhshani et al. 2023) as our baselines due to reasons in the Appendix. We initialize the four context vectors for PLOT, ProDA, and APP randomly.

Few-Shot Classification

Quantitative Analysis Figure 2 indicates that our method outperforms baselines on all benchmark datasets on average, where our performance is superior to 45 out of 55 experiment cases (11 datasets $\times$ 5 shots). The advantage of APP stands out in Caltech101, DTD, EuroSAT, and ImageNet, which consist of more diverse images. Since the data-dependent prior and the repulsive force of Eq. 6 enable text features to infer the multi-modes of image features, our context vectors are learned to capture the diverse semantics of image features.

Qualitative Analysis To demonstrate the efficacy of our method in capturing intricate image features, we provide visualizations of both image features ($f(x_i)$) and text features ($g(\theta_j, y_i)$). In Figure 3, we present Umap (McInnes, Healy, and Melville 2018) representations for EuroSAT test dataset. The upper figures depict image and text representations of all classes. A richer yellow hue indicates a denser allocation of image features. Our method, APP, well captures and comprehensively spans various modes within image features. The lower figures spotlight two arbitrarily selected classes, revealing how APP's text features harmoniously align with the image features of each class. This alignment is particu-



(a) PLOT (b) ProDA (c) APP

Figure 3: Umap visualization about image features and text features for EuroSAT. Upper Histograms correspond to image features and $\star$ means text features of all classes. Lower Image and text features of arbitrary two classes. The color coding corresponds to each class.

Method	Base	New	**H**
CoCoOp	71.7	53.6	61.4
PLOT	82.2	60.5	69.7
ProDA	82.4	63.6	71.8
APP	**83.0**	**65.8**	**73.4**

(a) Average

Method	Base	New	**H**
CoCoOp	**95.2**	87.4	91.2
PLOT	94.7	88.1	91.3
ProDA	**95.2**	86.8	90.8
APP	**95.2**	**91.0**	**93.0**

(b) Caltech101

Method	Base	New	**H**
CoCoOp	74.6	38.9	51.1
PLOT	78.1	42.7	55.2
ProDA	78.0	47.0	58.6
APP	**78.4**	**48.9**	**60.2**

(c) DTD

Method	Base	New	**H**
CoCoOp	91.4	35.6	51.3
PLOT	92.9	39.3	55.2
ProDA	89.6	39.0	54.4
APP	**93.6**	**47.6**	**63.1**

(d) EuroSAT

Method	Base	New	**H**
CoCoOp	29.1	14.1	19.0
PLOT	43.3	20.4	27.8
ProDA	44.3	24.1	31.2
APP	**44.9**	**26.0**	**33.0**

(e) FGVC-Aircraft

Method	Base	New	**H**
CoCoOp	80.7	78.8	79.7
PLOT	83.4	84.2	83.8
ProDA	84.5	**86.2**	**85.4**
APP	**84.6**	86.1	**85.4**

(f) Food101

Method	Base	New	**H**
CoCoOp	68.3	60.5	64.1
PLOT	68.3	58.4	62.9
ProDA	68.8	63.0	65.7
APP	**69.9**	**63.2**	**66.4**

(g) ImageNet

Method	Base	New	**H**
CoCoOp	94.7	58.6	72.4
PLOT	**97.4**	54.2	69.6
ProDA	97.0	58.5	73.0
APP	96.8	**61.0**	**74.8**

(h) Flower102

Method	Base	New	**H**
CoCoOp	89.4	**91.0**	90.2
PLOT	95.9	87.6	91.5
ProDA	96.4	88.6	**92.4**
APP	**96.8**	88.3	**92.4**

(i) Oxford Pets

Method	Base	New	**H**
CoCoOp	68.7	51.6	58.9
PLOT	84.2	62.6	71.8
ProDA	84.5	68.1	75.5
APP	**85.9**	**69.5**	**76.8**

(j) Stanford Cars

Method	Base	New	**H**
CoCoOp	73.3	64.0	68.4
PLOT	79.8	65.3	71.8
ProDA	**80.9**	70.8	75.5
APP	80.6	**73.3**	**76.8**

(k) Sun397

Method	Base	New	**H**
CoCoOp	79.2	47.0	59.0
PLOT	86.5	62.7	72.7
ProDA	**86.9**	67.9	76.2
APP	86.2	**69.2**	**76.8**

(l) UCF101

Table 1: Test accuracies (%) of the unseen classes generalization settings. H means the harmonic mean between the base accuracy and the new accuracy. Bold means the best accuracy of each column.

larly pronounced in comparison to other methods.

Ablation Study We show two ablation studies, considering our method and the number of prompts.

To identify the key enabler, we conduct additional ablation studies for APP by experimenting on 1) data-dependent prior and 2) Stein Variational Gradient Descent (SVGD). For all cases, four context vectors are initialized, and we select MLE training as the baseline optimized by SGD. Table 2 shows the posterior approximation with data-dependent prior improved performances generally than MLE Training. SVGD shows a more robust performance in a few data instances than SGD. The posterior approximation by SVGD also shows consistently better performance, outperforming in Caltech101 and EuroSAT dataset.

Dataset	**Methods**	**Number of shots**		
		1	**2**	**4**
Caltech101	SGD	89.87	90.29	91.00
	SVGD	89.90	90.56	91.02
	SGD+Prior	90.26	90.70	91.01
	APP	**90.30**	**90.87**	**91.05**
EuroSAT	SGD	56.43	63.93	72.63
	SVGD	58.93	64.36	73.64
	SGD+Prior	59.96	63.99	72.63
	APP	**60.04**	**66.02**	**74.08**
Food101	SGD	78.53	78.88	78.88
	SVGD	78.51	78.86	78.87
	SGD+Prior	78.81	79.15	79.32
	APP	**78.87**	**79.25**	**79.43**

Table 2: An ablation study about our method, APP. Experiments are replicated over three times.

Generalization Experiment

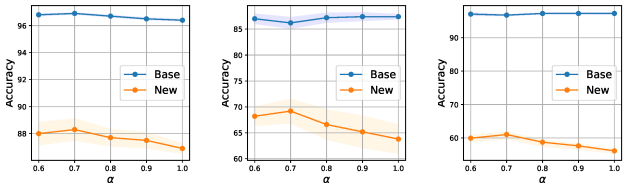
It is well known that VLP model is rather robust for domain shift (Radford et al. 2021), yet this good property can be corrupted when the model parameter is fine-tuned on the downstream task (Wortsman et al. 2022). Therefore, if this robustness regarding domain shift from VLP models could be sustained with prompt learning, it implies that this technique can be utilized more generally. For comparing the robustness, we conduct two experiments: 1) Unseen classes generalization setting in 11 datasets. and 2) Domain generalization setting in ImageNet.

Dataset	M	Number of shots		
		1	2	4
Caltech101	2	89.19	90.08	90.67
	4	**90.30**	**90.87**	91.05
	8	90.18	90.33	**91.38**
EuroSAT	2	52.06	60.65	71.19
	4	**60.04**	**66.02**	74.08
	8	58.60	64.22	**74.32**
Food101	2	78.27	78.74	78.93
	4	78.87	79.25	79.43
	8	**78.90**	**79.29**	**79.45**

Table 3: An ablation study with regard to the number of prompts (M). Experiments are replicated over three times.

Unseen classes Generalization in 11 Datasets. Following Zhou et al. (2022a), we report the robustness over unseen classes in 11 datasets. Table 1 shows the test accuracies with regard to both seen (base) classes and unseen (new) classes. Note that APP is robust to unseen (new) class data, maintaining the performance of seen class data, while other baselines have a performance trade-off between seen and unseen classes.

Sensitive Analysis of α We additionally investigate the impact of test data-dependent prior, $p(\theta|x')$, which adapts our posterior distribution to unseen instance. Figure 4 shows that adaptation of test data is beneficial for both seen and unseen performances. Additionally, balancing between posterior of seen data and prior of unseen data holds significance in achieving effective generalization for both scenarios.



(a) Oxford Pets (b) UCF101 (c) Flower102

Figure 4: Sensitivity analysis on α, effect of test data-dependent prior.

Domain Generalization on ImageNet Despite the tendency for a potential performance trade-off between the source and target datasets, table 4 demonstrates APP attains the improved performance achieved on both the source and target datasets, highlighting the robustness of APP in dealing with distribution shifts.

Conclusion

We propose the Bayesian framework for prompt learning to consider the uncertainty from few-shot learning scenario, where the image features are possible to be multi-modal and a distribution shift exists between the train and test dataset. We enhance flexibility via Wasserstein Gradient Flow. Furthermore, we propose a novel data-dependent prior distribution that is conditioned on averaged image features. This approach is designed to capture minor modes of image features and facilitate adaptation to previously unseen distributions. We demonstrate substantial performance improvements in various scenarios, including few-shot classifications, domain generalizations, and unseen class generalizations.

	Dataset	Methods	Acc (%)
Source	ImageNet	CoCoOp	63.13
		PLOT	63.14
		ProDA	62.73
		APP	**64.50**
Target	ImageNetV2	CoCoOp	55.23
		PLOT	54.23
		ProDA	54.97
		APP	**57.10**
	ImageNet-Sketch	CoCoOp	34.07
		PLOT	33.93
		ProDA	34.60
		APP	**35.70**
	ImageNet-R	CoCoOp	56.03
		PLOT	56.86
		ProDA	58.57
		APP	**58.70**
	ImageNet-A	CoCoOp	22.37
		PLOT	22.63
		ProDA	23.47
		APP	**23.80**

Table 4: Result of domain generalization in ImageNet. Acc represents the accuracy. Bold means the best accuracy.

Acknowledgments

This research was supported by AI Technology Development for Commonsense Extraction, Reasoning, and Inference from Heterogeneous Data (IITP) funded by the Ministry of Science and ICT(2022-0-00077).

References

Beaudry, N. J.; and Renner, R. 2012. An intuitive proof of the data processing inequality. arXiv:1107.0740.

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, 446–461. Springer.

Chen, C.; Zhang, R.; Wang, W.; Li, B.; and Chen, L. 2018. A Unified Particle-Optimization Framework for Scalable Bayesian Sampling. arXiv:1805.11659.

Chen, G.; Yao, W.; Song, X.; Li, X.; Rao, Y.; and Zhang, K. 2023. PLOT: Prompt Learning with Optimal Transport for Vision-Language Models. In *The Eleventh International Conference on Learning Representations*.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3606–3613.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Derakhshani, M. M.; Sanchez, E.; Bulat, A.; da Costa, V. G. T.; Snoek, C. G. M.; Tzimiropoulos, G.; and Martinez, B. 2023. Bayesian Prompt Learning for Image-Language Model Generalization. arXiv:2210.02390.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, 178–178. IEEE.

gil Lee, S.; Kim, H.; Shin, C.; Tan, X.; Liu, C.; Meng, Q.; Qin, T.; Chen, W.; Yoon, S.; and Liu, T.-Y. 2022. PriorGrad: Improving Conditional Denoising Diffusion Models with Data-Dependent Adaptive Prior. In *International Conference on Learning Representations*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. 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.

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; et al. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8340–8349.

Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; and Song, D. 2021b. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15262–15271.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 4904–4916. PMLR.

Jordan, R.; Kinderlehrer, D.; and Otto, F. 1998. The Variational Formulation of the Fokker–Planck Equation. *SIAM Journal on Mathematical Analysis*, 29(1): 1–17.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, 554–561.

Li, Z.; Wang, R.; Chen, K.; Utiyama, M.; Sumita, E.; Zhang, Z.; and Zhao, H. 2020. Data-dependent Gaussian Prior Objective for Language Generation. In *International Conference on Learning Representations*.

Liu, Q.; and Wang, D. 2016. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. In Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Lu, Y.; Liu, J.; Zhang, Y.; Liu, Y.; and Tian, X. 2022. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5206–5215.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

McInnes, L.; Healy, J.; and Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, 722–729. IEEE.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. V. 2012. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 3498–3505.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, 5389–5400. PMLR.

Ruan, Y.; Dubois, Y.; and Maddison, C. J. 2022. Optimal Representations for Covariate Shift. In *International Conference on Learning Representations*.

Shen, S.; Li, L. H.; Tan, H.; Bansal, M.; Rohrbach, A.; Chang, K.-W.; Yao, Z.; and Keutzer, K. 2022. How Much Can CLIP Benefit Vision-and-Language Tasks? In *International Conference on Learning Representations*.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Wang, H.; Ge, S.; Lipton, Z.; and Xing, E. P. 2019. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32.

Welling, M.; and Teh, Y. W. 2011. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, 681–688.

Wortsman, M.; Ilharco, G.; Kim, J. W.; Li, M.; Kornblith, S.; Roelofs, R.; Lopes, R. G.; Hajishirzi, H.; Farhadi, A.; Namkoong, H.; et al. 2022. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7959–7971.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, 3485–3492. IEEE.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.