Bayesian Prompt Learning for Image-Language Model Generalization

Derakhshani, M.M.; Sanchez, E.; Bulat, A.; Turrisi da Costa, V.G.; Snoek, C.G.M.; Tzimiropoulos, G.; Martinez, B.

DOI
10.48550/arXiv.2210.02390
10.1109/ICCV51070.2023.01398

Publication date
2023

Document Version
Final published version

Published in
2023 IEEE/CVF International Conference on Computer Vision

License
Article 25fa Dutch Copyright Act (https://www.openaccess.nl/en/in-the-netherlands/you-share-we-take-care)

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible. 

UvA-DARE is a service provided by the Library of the University of Amsterdam (https://dare.uva.nl)
Bayesian Prompt Learning for Image-Language Model Generalization

Mohammad Mahdi Derakhshani1, Enrique Sanchez4, Adrian Bulat4
Victor Guilherme Turrisi da Costa2, Cees G. M. Snoek1†, Georgios Tzimiropoulos3,4†, Brais Martinez4†
1University of Amsterdam 2University of Trento 3Queen Mary University of London 4Samsung AI Cambridge

Abstract

Foundational image-language models have generated considerable interest due to their efficient adaptation to downstream tasks by prompt learning. Prompt learning treats part of the language model input as trainable while freezing the rest, and optimizes an Empirical Risk Minimization objective. However, Empirical Risk Minimization is known to suffer from distributional shifts which hurt generalizability to prompts unseen during training. By leveraging the regularization ability of Bayesian methods, we frame prompt learning from the Bayesian perspective and formulate it as a variational inference problem. Our approach regularizes the prompt space, reduces overfitting to the seen prompts and improves the prompt generalization on unseen prompts. Our framework is implemented by modeling the input prompt space in a probabilistic manner, as an a priori distribution which makes our proposal compatible with common prompt learning approaches that are unconditional or conditional on the image. We demonstrate empirically on 15 benchmarks that Bayesian prompt learning provides an appropriate coverage of the prompt space, prevents learning spurious features, and exploits transferable invariant features. This results in better generalization of unseen prompts, even across different datasets and domains.

Code available at: https://github.com/saic-fi/Bayesian-Prompt-Learning

1. Introduction

In the continuous quest for better pre-training strategies, models based on image and language supervision have set impressive milestones, with CLIP [38], ALIGN [22] and Flamingo [1] being leading examples. Contrastively trained image-language models consist of image and text encoders that align semantically-related concepts in a joint embedding space. Such models offer impressive zero-shot image classification by using the text encoder to generate classifier weights from arbitrarily newly defined category classes without relying on any visual data. In particular, the class name is used within a handcrafted prompt template and then tokenized and encoded into the shared embedding space to generate new classifier weights. Rather than manually defining prompts, both Lester et al. [28] and Zhou et al. [55] demonstrated prompts can instead be optimized in a data-driven manner through backpropagation. However, as prompt learning typically has access to only a few training examples per prompt, overfitting to the seen prompts in lieu of the unseen prompts is common [55]. In this paper, we strive to mitigate the overfitting behavior of prompt learning so as to improve generalization for unseen prompts.

Others before us have considered the generalization problem in prompt learning as well, e.g., [55, 54], be it they all seek to optimize a so-called Empirical Risk Minimiza-
tion. It is, however, well known that Empirical Risk Minimization based models degrade drastically when training and testing distributions are different [37, 2]. To relax the i.i.d. assumption, Peters et al. [37] suggest exploiting the “invariance principle” for better generalization. Unfortunately, Invariant Risk Minimization methods for deep neural networks as of yet fail to deliver competitive results, as observed in [15, 31, 30]. To alleviate this limitation, Lin et al. [30] propose a Bayesian treatment of Invariant Risk Minimization that alleviates overfitting with deep models by defining a regularization term over the posterior distribution of classifiers, minimizing this term, and pushing the model’s backbone to learn invariant features. We take inspiration from this Bayesian Invariant Risk Minimization [30] and propose the first Bayesian prompt learning approach.

We make three contributions. First, we frame prompt learning from the Bayesian perspective and formulate it as a variational inference problem (see Figure 1). This formulation provides several benefits. First, it naturally injects noise during prompt learning and induces a regularization term that encourages the model to learn informative prompts, scattered in the prompt space, for each downstream task. As a direct result, we regularize the prompt space, reduce overfitting to seen prompts, and improve generalization on unseen prompts. Second, our framework models the input prompt space in a probabilistic manner, as an a priori distribution which makes our proposal compatible with prompt learning approaches that are unconditional [55] or conditional on the image [54]. Third, we empirically demonstrate on 15 benchmarks that Bayesian prompt learning provides an appropriate coverage of the prompt space, prevents learning spurious features, and exploits transferable invariant features, leading to a better generalization of unseen prompts, even across different datasets and domains.

2. Related Work

**Prompt learning in language.** Prompt learning was originally proposed within natural language processing (NLP), by models such as GPT-3 [4]. Early methods constructed prompts by combining words in the language space such that the model would perform better on downstream evaluation [46, 24]. Li and Liang [29] prepend a set of learnable prompts to the different layers of a frozen model and optimize through back-propagation. In parallel, Lester et al. [28] demonstrate that with no intermediate layer prefixes or task-specific output layers, adding prefixes alone to the input of the frozen model is enough to compete with fine-tuning. Alternatively, [16] uses a HyperNetwork to conditionally generate task-specific and layer-specific prompts pre-pended to the values and keys inside the self-attention layers of a frozen model. Inspired by progress from NLP, we propose a prompt learning method intended for image and language models.

**Prompt learning in image and language.** Zhou et al. [55] propose Context Optimization (CoOp), a prompt learner for CLIP, which optimizes prompts in the continuous space through back-propagation. The work demonstrates the benefit of prompt learning over prompt engineering. While CoOp obtains good accuracy on prompts seen during training, it has difficulty generalizing to unseen prompts. It motivated Zhou et al. to introduce Conditional Context Optimization (CoCoOp) [54]. It generates instance-specific prompt residuals through a conditioning mechanism dependent on the image data, which generalizes better. ProGrad by Zhu et al. [56] also strives to bridge the generalization gap by matching the gradient of the prompt to the general knowledge of the CLIP model to prevent forgetting. Alternative directions consist of test-time prompt learning [47], where consistency across multiple views is the supervisory signal, and unsupervised prompt learning [21], where a pseudo-labeling strategy drives the prompt learning. More similar to ours is ProDA by Lu et al. [32], who propose an ensemble of a fixed number of hand-crafted prompts and model the distribution exclusively within the language embedding. Conversely, we prefer to model the input prompt space rather than relying on a fixed number of templates, as it provides us with a mechanism to cover the prompt space by sampling. Moreover, our approach is not limited to unconditional prompt learning from the language embedding, but like Zhou et al. [54], also allows for prompt learning conditioned on an image.

**Prompt learning in vision and language.** While beyond our current scope, it is worth noting that prompt learning has been applied to a wider range of vision problems and scenarios, which highlights its power and flexibility. Among them are important topics such as unsupervised domain adaptation [13], multi-label classification [49], video classification [25], object detection [9, 12] and pixel-level labelling [40]. Finally, prompt learning has also been applied to vision only models [23, 44] providing an efficient and flexible means to adapt pre-trained models.

**Variational inference in computer vision.** Variational inference and, more specifically, variational autoencoder variants have been extensively applied to computer vision tasks as diverse as image generation [39, 43, 42], action recognition [34], instance segmentation [20], few-shot learning [33, 45], domain generalization [10], and continual learning [7]. For example, Zhang et al. [53] focus on the reduction of noise vulnerability and estimation bias in few-shot learning through variational inference. In the same vein, Du et al. [10] propose a variational information bottleneck to better manage prediction uncertainty and unknown domains. Our proposed method also shares the advantages of variational inference in avoiding overfitting in low-shot settings, improving generalization, and encourages the prompt space to be resilient against these chal-
lenges. To the best of our knowledge we are the first to introduce variational inference in prompt learning.

3. Method
3.1. Background

Contrastive Language-Image Pretraining (CLIP) [38] consists of an image encoder \( f(x) \) and text encoder \( g(t) \), each producing a \( d \)-dimensional \((L_2 \text{ normalized})\) embedding from an arbitrary image \( x \in \mathbb{R}^{3 \times H \times W} \), and word embeddings \( t \in \mathbb{R}^{L \times e} \), with \( L \) representing the text length and \( e \) the embedding dimension\(^1\). Both encoders are trained together using a contrastive loss from a large-scale dataset composed of paired images and captions. Once trained, CLIP enables zero-shot classification by generating each of the \( c \) classifier weights \( w_c \) as the \( d \)-dimensional text encoding \( g(t_c) \). Here \( t_c \) results from adding the class-specific word embedding \( e_c \) to a pre-defined prompt \( p \in \mathbb{R}^{L \times 1 \times e} \), i.e., \( p = g(t_c) \) with \( t_c = \{p, e_c\} \). The prompt \( p \) is manually crafted to capture the semantic meaning of the downstream task, e.g., \( t_c = \{\text{An image of a \{class\}}\} \). The probability of image \( x \) being classified as \( y \in \{1...C\} \) is thus defined as

\[
p(y|x) = \frac{e^{f(x)^T w_{y_c}}}{\sum_c e^{f(x)^T w_{c}}}. \]

Context Optimization (CoOp) [55] provides a learned alternative to manually defining prompts. CoOp learns a fixed prompt from a few annotated samples. The prompt is designed as a learnable embedding matrix \( p \in \mathbb{R}^{L \times e} \) which is updated via back-propagating the classification error through the frozen CLIP model. Specifically, for a set of \( N \) annotated meta-training samples \( \{\{x_i, y_i\}\}_{i=1}^{N} \), the prompt \( p \) is obtained by minimizing the cross-entropy loss, as:

\[
p^* = \arg\min_p \mathbb{E}_{x_i, y_i} [-\log p(y_i|x_i, p)]. \quad (1)
\]

Note that this approach, while resembling that of common meta-learning approaches, can still be deployed in a zero-shot scenario provided that for new classes the classification weights will be given by the text encoder. Although this approach generalizes to new tasks with few training iterations, learning a fixed prompt is sensitive to domain shifts between the annotated samples and the unseen prompts.

Conditional Prompt Learning (CoCoOp) [54] attempts to overcome domain shifts by learning an instance-specific continuous prompt that is conditioned on the input image. To ease the training of a conditional prompt generator, CoCoOp defines each conditional token in a residual way, with a task-specific, learnable set of tokens \( p \) and a residual prompt that is conditioned on the input image. Assuming \( p \) to be composed of \( L \) learnable tokens \( p = [p_1, p_2, \ldots, p_L] \), the residual prompt \( r(x) = \pi_{\phi}(f(x)) \in \mathbb{R}^e \) is produced by a small neural network \( \pi_{\phi} \) with as input the image features \( f(x) \). The new prompt is then computed as \( p(x) = [p_1 + r(x), p_2 + r(x), \ldots, p_L + r(x)] \). The training now comprises learning the task-specific prompt \( p \) and the parameters \( \phi \) of the residual network \( \pi_{\phi} \). Defining the context-specific text embedding \( t_c(x) = \{p(x), e_c\} \), and \( p(y|x) \) as:

\[
p(y|x) = \frac{e^{f(x)^T g(t_c(x))}}{\sum_c e^{f(x)^T g(t_c(x))}}, \quad (2)
\]

the learning is formulated as:

\[
p^*, \phi^* = \arg\min_{p, \phi} \mathbb{E}_{x_i, y_i} [-\log p(y_i|x_i, p, \phi)]. \quad (3)
\]

While CoCoOp achieves good results in many downstream tasks, it is still prone to the domain shift problem, considering that \( \pi_{\phi} \) provides a deterministic residual prompt from the domain-specific image features \( f(x) \).

Prompt Distribution Learning (ProDA) [32] learns a distribution of prompts that generalize to a broader set of tasks. It learns a collection of prompts \( P = \{p_k\}_{k=1}^{K} \) that subsequently generate an \textit{a posteriori} distribution of the classifier weights for each of the target classes. For a given mini-batch of \( K \) sampled prompts \( p_k \sim P \), the classifier weights \( w_c \) are sampled from the posterior distribution \( q = \mathcal{N}(\mu_{w_{1,c}}, \Sigma_{w_{1,c}}) \), with mean \( \mu_{w_{1,c}} \) and covariance \( \Sigma_{w_{1,c}} \) computed from the collection \( \{w_{k,c} = g(t_{k,c})\}_{c=1:C, k=1:K} \), with \( t_{k,c} = \{p_k, e_c\} \). The objective is formulated as:

\[
P^* = \arg\min_p \mathbb{E}_{x_i, y_i} [-\log \mathbb{E}_{w_i \sim q} p(y_i|x_i, w_i)]. \quad (4)
\]

Computing \( \mathbb{E}_{w_i} p(y_i|x_i, w_i) \) is intractable and an upper bound to Eq. 4 is derived. During inference, the classifier weights are set to those given by the predictive mean \( w_c = \mu_{w_{1,c}} \), computed across the set of learned prompts \( P \).

3.2. Conditional Bayesian Prompt Learning

We propose to model the input prompt space in a probabilistic manner, as an \textit{a priori} conditional distribution. We define a distribution \( p_{\gamma} \) over the prompts \( p \) that is conditional on the image, i.e., \( p \sim p_{\gamma}(x) \). To this end, we assume that \( p \) can be split into a fixed set of prompts \( p_\gamma \) and an conditional residual prompt \( r \) that act as a latent variable over \( p \). The conditional prompt is then defined as:

\[
p_{\gamma}(x) = [p_1 + r_{\gamma}, p_2 + r_{\gamma}, \ldots, p_L + r_{\gamma}], \quad r_{\gamma} \sim p_{\gamma}(x), \quad (5)
\]

where \( p_{\gamma}(x) \) refers to the real posterior distribution over \( r \) conditioned on the observed features \( x \). Denoting the class-
specific input as \( t_{c,\gamma}(x) = \{p_c(x), e_c\} \), the marginal likelihood \( p(y|x) \) is:

\[
p(y|x) = \int e^{f(x)^T g(t_{c,\gamma}(x))} \sum_{\gamma} e^{f(x)^T g(t_{c,\gamma}(x))} p(\gamma)p_r(x) \, d\gamma. \tag{6}
\]

Solving Eq. 3 with the marginal likelihood from Eq. 6 is intractable, as it requires computing \( p_r(r|x)p_r(x) \). Instead, we resort to deriving a lower bound by introducing a variational posterior distribution \( \pi_\phi(x) \) from which the residual \( r_\gamma \) can be sampled. The variational bound is defined as:

\[
\log p(y|x) \geq \mathbb{E}_{\pi_\phi(r|x)}[\log p(y|x, r)] - D_{KL}[\pi_\phi(r|x) \parallel p_r(r)], \tag{7}
\]

with \( p(y|x, r) \propto e^{f(x)^T g(t_{c,\gamma}(x))} \), where the dependency on \( r \) comes through the definition of \( t_{c,\gamma} \). Following standard variational optimization practices [26, 14], we define \( \pi_\phi \) as a Gaussian distribution conditioned on the input image features \( x \), as \( r(x) \sim \mathcal{N}(\mu(x), \Sigma(x)) \), with \( \mu \) and \( \Sigma \) parameterized by two linear layers followed by two linear heads on top to estimate the \( \mu \) and \( \Sigma \) of the residual distribution. The prior \( p_r(r) \) is defined as \( \mathcal{N}(0, 1) \), and we use the reparameterization trick to generate Monte-Carlo samples from \( \pi_\phi \) to maximize the right side of Eq. 7. The optimization of Eq. 7 comprises learning the prompt embeddings \( \{p_i\}_{i=1}^K \) as well as the parameters of the posterior network \( \pi_\phi \) and the linear layers parameterizing \( \mu \) and \( \Sigma \). Note that this adds little complexity as it requires learning \( p \) and \( \pi_\phi \), given that \( \mu \) and \( \Sigma \) are defined as two linear layers on top of \( \pi_\phi \).

**Inference.** At test time, \( K \) residuals are sampled from the conditional distribution \( \pi_\phi(x) \), which are used to generate \( K \) different prompts per class \( p_k = [p_1 + r_k, p_2 + r_k, \ldots, p_L + r_k] \). Each prompt is prepended to the class-specific embedding to generate a series of \( K \) separate classifier weights \( w_{k,e} \). We then compute \( p(y=c|x) = (1/K) \sum_k p(y=c|x, w_{k,e}) \) and select \( \hat{c} = \arg\max_c p(y=c|x) \) as the predicted class. It is worth noting that because the posterior distribution is generated by the text encoder, it is not expected that for \( K \to \infty \), \((1/K) \sum_k p(y=c|x, w_{k,e}) \to p(y=c|x, g(\{\mu(x), e_c\}))\), meaning that sampling at inference time remains relevant.

### 3.3. Unconditional Bayesian Prompt Learning

Notably, our framework can also be reformulated as an unconditional case by simply removing the dependency of the input image from the latent distribution. In such a scenario, we keep a fixed set of prompt embeddings and learn a global latent distribution \( p_r \) over the residual prompts \( r \), as \( r \sim \mathcal{N}(\mu, \Sigma) \), where \( \mu \) and \( \Sigma \) are parameterized by two learnable vectors. To this end, we assume that \( p \) can be split into a fixed set of prompts \( p \) and a residual prompts \( r \) that act as a latent variable over \( p \). For the unconditional case, the prompt is defined as:

\[
p_\gamma = [p_1 + r_\gamma, p_2 + r_\gamma, \ldots, p_L + r_\gamma], \quad r_\gamma \sim p_\gamma. \tag{8}
\]

In this case, \( p_\gamma \) is a general distribution learned during training with no dependency on the input sample \( x \).

Having defined both unconditional Bayesian prompt learning and conditional Bayesian prompt learning we are now ready to evaluate their generalization ability.

### 4. Experiments and Results

#### 4.1. Experimental Setup

**Three tasks and fifteen datasets.** We consider three tasks: *unseen prompts generalization*, *cross-dataset prompts generalization*, and *cross-domain domain generalization*. For the first two tasks, we rely on the same 11 image recognition datasets as Zhou et al. [55, 54]. These include image classification (ImageNet [6] and Caltech101 [11]), fine-grained classification (OxfordPets [36], StanfordCars [27], Flowers102 [35], Food101 [3] and FGVC-Aircraft [33]), scene recognition (SUN397 [52]), action recognition (UCF101 [48]), texture classification (DTD [5]), and satellite imagery recognition (EuroSAT [17]). For cross-domain prompts generalization, we train on ImageNet and report on ImageNetV2 [41], ImageNet-Sketch [51], ImageNet-A [19], and ImageNet-R [18].

**Evaluation metrics.** For all three tasks we report average accuracy and standard deviation.

**Implementation details.** Our conditional variational prompt learning contains three sub-networks: an image encoder \( f(x) \), a text encoder \( g(t) \), and a posterior network \( \pi_\phi \). The image encoder \( f(x) \) is a ViT-B/16 [8] and the text encoder \( g(t) \) a transformer [50], which are both initialized with CLIP’s pre-trained weights and kept frozen during training, as in [55, 54]. The posterior network \( \pi_\phi \) consists of two linear layers followed by an ELU activation function as trunk and two linear heads on top to estimate the \( \mu \) and \( \Sigma \) of the residual distribution. For each task and dataset, we optimize the number of samples \( K \) and epochs. Other hyper-parameters as well as the training pipeline in terms of few-shot task definitions are identical to [55, 54] (see Table 9 and 10 in the appendix).

#### 4.2. Comparisons

We first compare against CoOp [55], CoCoOp [54], and ProDA [32] in terms of the generalization of learned prompts on unseen classes, datasets, or domains. For CoOp and CoCoOp, all results are adopted from [54], and we report results for ProDA using our re-implementation.

**Task 1: unseen prompts generalization** We report the unseen prompts generalization of our method on 11 datasets for three different random seeds. Each dataset is divided into two disjoint subsets: seen classes and unseen classes. We train our method on the seen classes and evaluate it on the unseen classes. For a fair comparison, we follow [55, 54] in terms of dataset split and number of shots. From
Table 1: **Task I: unseen prompts generalization** comparison between conditional Bayesian prompt learning and alternatives. Our model provides better generalization on unseen prompts compared to CoOp, CoCoOp and ProDA.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Caltech101</td>
<td>89.81</td>
<td>93.81</td>
<td>93.23</td>
<td><strong>94.93 ± 0.1</strong></td>
</tr>
<tr>
<td>DTD</td>
<td>41.18</td>
<td>56.00</td>
<td>56.48</td>
<td><strong>60.80 ± 0.5</strong></td>
</tr>
<tr>
<td>EuroSAT</td>
<td>54.74</td>
<td>60.04</td>
<td>60.00</td>
<td><strong>75.30 ± 0.7</strong></td>
</tr>
<tr>
<td>FGVC Aircraft</td>
<td>22.30</td>
<td>23.71</td>
<td>24.13</td>
<td><strong>35.00 ± 0.5</strong></td>
</tr>
<tr>
<td>Flowers102</td>
<td>59.67</td>
<td>71.75</td>
<td>68.68</td>
<td><strong>70.40 ± 0.1</strong></td>
</tr>
<tr>
<td>Food101</td>
<td>82.26</td>
<td>91.29</td>
<td>88.57</td>
<td><strong>92.13 ± 0.1</strong></td>
</tr>
<tr>
<td>ImageNet</td>
<td>67.88</td>
<td>70.43</td>
<td>70.23</td>
<td><strong>70.93 ± 0.1</strong></td>
</tr>
<tr>
<td>OxfordPets</td>
<td>95.29</td>
<td>97.69</td>
<td>97.83</td>
<td><strong>98.00 ± 0.1</strong></td>
</tr>
<tr>
<td>StanfordCars</td>
<td>60.40</td>
<td><strong>73.59</strong></td>
<td>71.20</td>
<td><strong>73.23 ± 0.2</strong></td>
</tr>
<tr>
<td>SUN397</td>
<td>65.89</td>
<td>76.86</td>
<td>76.93</td>
<td><strong>77.87 ± 0.5</strong></td>
</tr>
<tr>
<td>UCF101</td>
<td>56.05</td>
<td>73.45</td>
<td>71.97</td>
<td><strong>75.77 ± 0.1</strong></td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>63.22</td>
<td>71.69</td>
<td>72.30</td>
<td><strong>74.94 ± 0.2</strong></td>
</tr>
</tbody>
</table>

Table 1, it can be seen that our best-performing method, conditional Bayesian prompt learning, outperforms CoOp and CoCoOp in terms of unseen prompts generalization by 11.72% and 3.25%, respectively. Our proposal also demonstrates minimal average variance across all datasets. This is achieved by regularizing the optimization by virtue of the variational formulation. Moreover, our model also performs better than ProDA [32], its probabilistic counterpart, by 2.64%. This mainly happens because our proposed method learns the prompt distribution directly in the prompt space and allows us to sample more informative prompts for the downstream tasks.

**Task II: cross-dataset prompts generalization** For the next task, cross-dataset prompts generalization, the model is trained on a source dataset (ImageNet) and then assessed on 10 distinct target datasets. This experiment tries to determine how effectively our method generalizes beyond the scope of a single dataset. As reported in Table 2, our conditional Bayesian prompt learning outperforms CoOp and CoCoOp on the target dataset by 2.07% and 0.36%. This highlights that our method encourages the model to exploit transferable invariant features beneficial for datasets with a non-overlapping label space. Furthermore, our method improves accuracy for 7 out of 10 target datasets. Unlike CoCoOp, which performs better in ImageNet-like datasets such as Caltech101 and StanfordCars, our Bayesian method exhibits improvement on dissimilar datasets (e.g., FGVC-Aircraft, DTD, and EuroSAT), demonstrating its capacity to capture the unique characteristics of each dataset.

**Task III: cross-domain prompts generalization** Lastly, we examine our conditional Bayesian prompt learning through the lens of distribution shift and robustness. We train our model on the source dataset ImageNet for three different random seeds, and assess it on ImageNetV2, ImageNet-Sketch, ImageNet-A, and ImageNet-R. Prior works such as CoOp [55] and CoCoOp [54] demonstrate empirically that learning a soft-prompt improves the model’s resilience against distribution shift and adversarial attack. Following their experiments, we are also interested in determining if treating prompts in a Bayesian manner maintains or improves performance. As reported in Table 2 compared with CoOp, our method improves the performance on ImageNet-Sketch, ImageNet-A, and ImageNet-R by 1.21%, 1.62%, and 1.79%. Compared with CoCoOp, our proposed method consistently enhances the model accuracy on ImageNetV2, ImageNet-Sketch, ImageNet-A, and ImageNet-R by 0.16%, 0.45%, 0.7%, and 0.82%. This is because our proposed method prevents learning spurious features, e.g., high-frequency features, and instead learns invariant ones by virtue of the Bayesian formulation.

**In-domain performance** Different from other prompt works, we focus on prompt generalization when we have distribution shift in the input space, the label space, or both.

Table 2: **Task II: cross-dataset prompts generalization.** Our proposed model is evaluated on 11 datasets with different label spaces. As shown, conditional Bayesian prompt learning performs better than non-Bayesian alternatives on 7 out of 10 datasets. **Task III: cross-domain prompts generalization.** Our model is evaluated on four datasets sharing the same label space as the training data. Our method outperforms alternatives on 3 out of 4 datasets.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Caltech101</td>
<td>93.70</td>
<td><strong>94.43</strong></td>
<td>93.67 ± 0.2</td>
</tr>
<tr>
<td>DTD</td>
<td>41.92</td>
<td>45.73</td>
<td><strong>46.10 ± 0.1</strong></td>
</tr>
<tr>
<td>EuroSAT</td>
<td><strong>46.39</strong></td>
<td>45.37</td>
<td><strong>45.87 ± 0.7</strong></td>
</tr>
<tr>
<td>FGVC Aircraft</td>
<td>18.47</td>
<td>22.94</td>
<td><strong>24.93 ± 0.2</strong></td>
</tr>
<tr>
<td>Flowers102</td>
<td>68.71</td>
<td>70.20</td>
<td><strong>70.90 ± 0.1</strong></td>
</tr>
<tr>
<td>Food101</td>
<td>85.30</td>
<td>86.06</td>
<td><strong>86.30 ± 0.1</strong></td>
</tr>
<tr>
<td>OxfordPets</td>
<td>89.14</td>
<td>90.14</td>
<td><strong>90.63 ± 0.1</strong></td>
</tr>
<tr>
<td>StanfordCars</td>
<td>64.51</td>
<td><strong>65.50</strong></td>
<td>65.00 ± 0.1</td>
</tr>
<tr>
<td>SUN397</td>
<td>64.15</td>
<td>67.36</td>
<td><strong>67.47 ± 0.1</strong></td>
</tr>
<tr>
<td>UCF101</td>
<td>66.55</td>
<td>68.21</td>
<td><strong>68.67 ± 0.2</strong></td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>63.88</td>
<td>65.59</td>
<td><strong>65.95 ± 0.2</strong></td>
</tr>
</tbody>
</table>

Table 3: **In-domain performance.** CoOp provides the best in-domain performance, but suffers from distribution shifts. Our proposal provides the best trade-off.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Task I</td>
<td>82.66</td>
<td>80.47</td>
<td>81.56</td>
<td><strong>80.10 ± 0.1</strong></td>
</tr>
<tr>
<td>Task II (III)</td>
<td><strong>71.51</strong></td>
<td>71.02</td>
<td>-</td>
<td>70.70 ± 0.2</td>
</tr>
</tbody>
</table>

Authorized licensed use limited to: Universiteit van Amsterdam. Downloaded on June 17,2024 at 14:51:38 UTC from IEEE Xplore. Restrictions apply.
Table 4: Effect of variational formulation. Formulating prompt learning as variational inference improves model generalization on unseen prompts compared to a non-Bayesian baseline [55], for both the unconditional and conditional formulation.

<table>
<thead>
<tr>
<th></th>
<th>DTD</th>
<th>EuroSAT</th>
<th>FGVC</th>
<th>Flowers102</th>
<th>UCF101</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>41.18</td>
<td>54.74</td>
<td>22.30</td>
<td>59.67</td>
<td>56.05</td>
</tr>
<tr>
<td>Unconditional</td>
<td>58.70</td>
<td>71.63</td>
<td>33.80</td>
<td>75.90</td>
<td>74.63</td>
</tr>
<tr>
<td>Conditional</td>
<td>60.80</td>
<td>75.30</td>
<td>35.00</td>
<td>70.40</td>
<td>75.77</td>
</tr>
</tbody>
</table>

Table 5: Benefit of the posterior distribution. The conditional posterior distribution \(\mathcal{N}(\mu(x), \Sigma(x))\) outperforms the two distributions \(\mathcal{U}(0, 1)\) and \(\mathcal{N}(0, 1)\) by a large margin for all datasets, indicating that the conditional variant captures more informative knowledge regarding the underlying distribution of prompts in the prompt space.

<table>
<thead>
<tr>
<th></th>
<th>DTD</th>
<th>EuroSAT</th>
<th>FGVC</th>
<th>Flowers102</th>
<th>UCF101</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\mathcal{U}(0, 1))</td>
<td>33.20</td>
<td>54.20</td>
<td>10.50</td>
<td>45.30</td>
<td>55.70</td>
</tr>
<tr>
<td>(\mathcal{N}(0, 1))</td>
<td>26.60</td>
<td>50.00</td>
<td>07.70</td>
<td>36.10</td>
<td>48.80</td>
</tr>
<tr>
<td>(\mathcal{N}(\mu(x), \Sigma(x)))</td>
<td>56.40</td>
<td>64.50</td>
<td>33.00</td>
<td>72.30</td>
<td>75.60</td>
</tr>
<tr>
<td>(\mathcal{N}(\mu(x), 0))</td>
<td>59.80</td>
<td>59.90</td>
<td>34.10</td>
<td>73.50</td>
<td>76.50</td>
</tr>
</tbody>
</table>

For reference, we provide the in-domain average performance in Table 3. As expected, our generalization ability comes with reduced overfitting on the in-domain setting, leading to CoOp exhibiting better in-domain performance. However, when comparing average results in Tables 1, 2 and 3, our proposal brings the best trade-off and we maintain performance on the in-domain setting while delivering a consistent performance gain on the out-of-domain setting.

4.3. Ablations

Effect of variational formulation. When prompt learning is formulated as a variational inference problem, it improves model generalization by regularizing the prompt space. To demonstrate this, we first report the model performances on unseen prompts for DTD, EuroSAT, FGVC, Aircraft, Flowers102, and UCF101 in Table 4. Regardless of whether a variational formulation is unconditional or conditional, it consistently increases unseen prompt generalization compared with the CoOp baseline [55]. This is achieved by sampling more informative prompts and better coverage of the prompt space (which we detail in a following ablation). Moreover, conditional variational prompt learning enhances the model performance on 4 out of 5 datasets compared to its unconditional counterpart, indicating its effectiveness in capturing the prompt distribution. This happens as the conditional formulation provides us with the potential to successfully capture fine-grained prompts by conditioning on the input image.

Benefit of the posterior distribution \(q_\phi\). We ablate the effectiveness of the variational posterior distribution. To do so, we consider sampling one residual prompt from the uniform distribution \(\mathcal{U}(0, 1)\), normal distribution \(\mathcal{N}(0, 1)\), normal distribution \(\mathcal{N}(\mu(x), \Sigma(x))\), and normal distribution \(\mathcal{N}(\mu(x), 0)\) and report the unseen prompts accuracy for one random seed in Table 5. We find that drawing one sample from \(\mathcal{N}(\mu(x), \Sigma(x))\) yields superior results compared to drawing one sample from uniform distribution \(\mathcal{U}(0, 1)\) and normal distribution \(\mathcal{N}(0, 1)\), demonstrating the informativeness of the prompt samples generated by our conditional setting. In addition, except for the EuroSAT dataset, a sample from the normal distribution \(\mathcal{N}(\mu(x), 0)\) obtains the best performance in comparison with alternatives, indicating the mean of the normal distribution \(\mu(x)\) is the most effective sample.

Influence of Monte Carlo sampling. When approximating the log-likelihood of input data, the number of Monte Carlo samples is an important hyperparameter. Generally, a large number of samples should lead to a better approximation and unseen prompt generalization. We ablate this hyperparameter on unseen prompts by varying the number of samples from the normal distribution \(\mathcal{N}(\mu(x), \Sigma(x))\) to understand the informativeness of the learned variational distribution at inference time in Table 6. Increasing the Monte Carlo samples from 1 to 20 consistently improves the unseen prompts accuracy, afterwards, the accuracy saturates. Table 7 further shows the impact of the time complexity as the number of Monte Carlo samples increases on one NVIDIA RTX 2080ti GPU. As expected, it shows the increase in sample size is accompanied by an increase in model inference time. We obtain a good trade-off in terms of accuracy and inference cost for 20 samples. For bet-
ter accuracy we recommend evaluating on a larger number of Monte Carlo samples as they capture the prompt space more appropriately. We optimize this hyper-parameter using cross-validation for all datasets and use them in the remaining experiments (see Table 10 in the appendix.)

Effect of prompt initialization. Next we ablate the effectiveness of the prompt initialization on the unseen prompts accuracy for one random seed. We consider two variants. In the first we initialize the context tokens randomly using a normal distribution and in the second we initialize the context tokens with “An image of a {class}.” Comparing the two variants in Table 8 demonstrates that an appropriately initialized prompt consistently outperforms a randomly initialized prompt. This is mainly because the prompt samples distributed in the vicinity of a guided prompt, e.g., “An image of a {class}”, are more informative than a random prompt. It can be viewed as informative prior knowledge about the prompt distribution, which facilitates the optimization towards a better prompt sub-space.

Benefit of variational ensemble. Variational prompt learning can be considered as an efficient ensemble approach as it generates several samples from the prompt space and uses them in conjunction. To demonstrate the value of our proposal over a naïve ensemble of prompts we compare against a modified version of CoCoOp [54]. We implement this version by initializing several learnable prompts per class using $M$ hand-crafted templates provided in the CLIP codebase and fine-tune them together. The final text feature representation is computed as a simple average of the ensemble of prompts for each class (as in Section 3.1.4 of [38]). For a fair comparison, the number of ensembles, $M$, equals the number of Monte Carlo samples $K$ per dataset (see Table 10 in the appendix). The results in Table 9 demonstrate that adding a naïve ensemble of prompts on top of CoCoOp does not necessarily lead to an improvement. For instance on DTD and UCF101, CoCoOp performs better by itself, while on Flowers102, EuroSAT, and FGVC Aircraft, the naïve ensemble obtains a higher unseen prompts accuracy. When we compare the naïve ensemble with our proposal, we note our method performs better in terms of unseen prompts accuracy for 4 out of 5 datasets. The reason is that, in naïve ensemble, these $M$ learnable prompts likely converge into contradictory prompts, that deteriorate the final performance, due to the lack of a proper regularization term in its objective function. Moreover, the

Table 8: Effect of prompt initialization. Initializing the context tokens with an appropriate prompt “An image of a {class}” (denoted as Guided) improves the performance compared to random initialization.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>DTD</th>
<th>EuroSAT</th>
<th>FGVC</th>
<th>Flowers102</th>
<th>UCF101</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>74.10</td>
<td>75.10</td>
<td>34.30</td>
<td>67.23</td>
<td>81.30</td>
</tr>
<tr>
<td>Guided</td>
<td>74.80</td>
<td>77.90</td>
<td>35.50</td>
<td>70.05</td>
<td>82.50</td>
</tr>
</tbody>
</table>

Table 9: Benefit of variational ensemble. Our conditional Bayesian prompt learning outperforms CoCoOp and its naïve ensemble for 4 out of 5 datasets regarding unseen prompts accuracy, demonstrating that our method is effective in constructing informative prompt samples for ensemble learning in a data-driven manner.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>DTD</th>
<th>EuroSAT</th>
<th>FGVC</th>
<th>Flowers102</th>
<th>UCF101</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoCoOp</td>
<td>56.00</td>
<td>60.04</td>
<td>23.71</td>
<td>71.75</td>
<td>73.45</td>
</tr>
<tr>
<td>Naïve ensemble</td>
<td>52.17</td>
<td>66.10</td>
<td>32.63</td>
<td>72.17</td>
<td>73.90</td>
</tr>
<tr>
<td>This paper</td>
<td>60.80</td>
<td>75.30</td>
<td>35.00</td>
<td>70.40</td>
<td>75.77</td>
</tr>
</tbody>
</table>

KL divergence in Eq. 7 serves to regulate and expand the norm of the prompt space. This allows our proposal to freely explore the prompt space, leading to complementary prompt samples. We believe the prompt samples are not contradictory as they are jointly optimized to cover inherent variability in the model predictions and maximize model performance.

Variational distribution interpretation. Next, we provide an intuition on how prompt samples are distributed within the variational distribution. For an input image $x$ and its corresponding target $y$, $K$ residuals are sampled from the conditional distribution $p_{\phi}(x)$, which are used to generate $K$ different prompts per class $p_k = [p_1 + r_k, p_2 + r_k, \ldots, p_L + r_k]$. We also construct a new prompt based on the mean of the variational distribution $\mu(x)$ dubbed as the mean prompt $p_{\mu(x)} = [p_1 + \mu(x), p_2 + \mu(x), \ldots, p_L + \mu(x)]$. These $K$ prompts are sorted based on their distance to the mean of the variational distribution $\mu(x)$, where for any $i$ and $j$ ($i \leq j$), $d(\mu(x), r_i) \leq d(\mu(x), r_j)$. All these $K$ prompts and the mean prompt are prepended to the class-specific embedding to generate a series of $K$ separate text encodings $w_{\mu(x), c}, w_{1,c}, w_{2,c}, \ldots, w_{K,c}$. For each class $c$, we compute the cosine similarity of the image encoding $f(x)$ and $w_{\mu(x), c}$, $w_{1,c}, w_{2,c}, \ldots, w_{K,c}$, and visualize them for $K=20$ prompts in Figure 2. The mean prompt $p_{\mu(x)}$ has the largest cosine similarity between its corresponding text and image encoding among other prompts, indicating it is the most informative sample. As we move further away from the mean prompt, the cosine similarity score decreases. Thus, the further a prompt is from the mean prompt, the less informative it becomes.

Next we demonstrate that prompts distributed within the variational distribution positively complement each other and maximize cosine similarity. Given the text encodings $w_{\mu(x), y}, w_{1,y}, w_{2,y}, \ldots, w_{M,y}$ we further define a new weight as $w_{j,y}^f = w_{\mu(x), y} + \sum_{j=1}^{J} w_{j,y}$, where $w_{j,y}^f$ is the cumulative sum of text encodings regarding prompt $p_1$ to $p_J$ and mean prompt $p_{\mu(x)}$. The cosine similarity between the image encoding $f(x)$ and all $\{w_{j,y}^f\}_{y=1}^{M}$ are computed and visualized in Figure 2. As shown, summing up, or ensembling, text encodings together increases the cosine similarity, where the maximum similarity is obtained when all clas-
Figure 2: **Variational distribution interpretation** on the EuroSAT dataset. The text encoding of the mean prompt \( p_{\mu(x)} \) (■) is the most similar to the image encoding. As we move further away from the mean prompt, the cosine similarity scores between the text encoding and image encoding decrease further ( ). When we ensemble the text encoding of different prompts the cosine similarity increases ( ■), where the maximum similarity is obtained when all text encodings are combined.

Figure 3: **Factor of variation analysis** on Flowers102 for two different classes. Left: we plot prompt samples across five clusters. Right: we show the top-3 most representative samples within each cluster (e.g., 3 closest images to the centroids). There is a region where the contours for five different clusters intersect, indicating shared knowledge related to its corresponding class, while they diverge slightly, indicating a particular variation factor.

**Factor of variation analysis** Here we analyze whether prompts that are sampled at the distributional modes correlate with features that characterize any known factors of variation (e.g., sub-domains) in a dataset. We perform this experiment on our conditional Bayesian prompt learning model. First, we randomly select two classes \( c_1 \) and \( c_2 \) from the Flowers102 dataset (Passion Flower, and Cyclamen, for reference). Then, for each class, we compute image features by forwarding the respective input images \( x \) into the image encoder of CLIP. Next, we apply \( k \)-means to group the image features into \( Q=5 \) cluster centroids. Note that each cluster centroid is assumed to capture some factors of visual variation within each class. Afterward, we treat the cluster assignments as pseudo-labels of those images. Note that this is done separately for each class. Additionally, for an input image \( x \), we generate \( M=21 \) different prompts. These prompts are fed into the clip’s text encoder, which generates \( M \) different weights \( w_1, w_2, \cdots, w_M \) per image instance. We now have a set of \( M \) weights for \( Q \) pseudo-labels across two classes, which are visualized in each row of Figure 3. From this figure, it can be seen that there is a high intersection between the distribution of all sub-classes, representing shared knowledge. However, some particular visual differences are expressed in regions without intersections in the distribution. Hence, we believe there is a correlation between the modes of the variational distribution and visual particularities of subsets of the same class.

5. Conclusion

In this paper, we formulate prompt learning for image-language models as a variational inference problem. The Bayesian perspective allows regularizing the prompt space, reducing overfitting to seen prompts, and improving unseen prompt generalization. We model the input prompt space as an a priori distribution, making our framework compatible with unconditional or conditional prompt learning approaches. Using 15 benchmarks, we show that Bayesian prompt learning provides an appropriate coverage of the prompt space, prevents learning spurious features, and exploits transferable invariant features, leading to better generalization of unseen prompts, even across datasets and domains. We believe prompt distributions can be estimated...
also using other probabilistic methods, like normalizing flows, energy-based models and diffusion models, which we will leave open for future investigation.

References


[22] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In International Conference on Machine Learning, 2021.


[34] Nnaznin Mehrasa, Akash Abdu Jyothi, Thibaut Durand, Jiawei He, Leonid Sigal, and Greg Mori. A variational autoencoder model for stochastic point processes. In IEEE Conference on Computer Vision and Pattern Recognition, 2019.


