Fine-Tuning with Uncertainty-Aware Priors Makes Vision and Language Foundation Models More Reliable

Tim G. J. Rudner [1] **Xiang Pan** [1] **Yucen Lily Li** [1] **Ravid Shwartz-Ziv** [1] **Andrew Gordon Wilson** [1]

Abstract

Fine-tuning off-the-shelf pre-trained neural networks has become the default starting point for a wide range of challenging prediction tasks—especially in computer vision and natural language processing, where pre-trained models trained on millions or even billions of data points are publicly available and can be fine-tuned with a moderate compute budget. However, while fine-tuned models have been shown to significantly improve predictive performance compared to models trained from scratch, they can exhibit poor calibration and fail to reliably identify challenging distribution shifts. In this paper, we improve uncertainty quantification in fine-tuned models by constructing a data-driven, uncertainty-aware fine-tuning prior that assigns high probability density to parameters that induce predictive functions with high uncertainty on input points that are meaningfully different from the data. We derive a tractable variational objective to perform approximate inference in models with data-driven, uncertainty-aware priors and evaluate models fine-tuned with such priors on different transfer learning tasks. We show that fine-tuning with uncertainty-aware priors significantly improves calibration, selective prediction, and semantic shift detection on computer vision and natural language classification tasks.

1. Introduction

How can we ensure that fine-tuning pre-trained models leads to models with good predictive performance and reliable uncertainty quantification? In this paper, we design uncertainty-aware priors (UAPs) for fine-tuning pre-trained neural networks and show that models fine-tuned with such priors achieve significantly improved predictive uncertainty quantification on image and natural language classification tasks. We illustrate some of our results in Figure 1.

While fine-tuning off-the-shelf pre-trained neural networks has become the default approach for most computer vision and natural language processing tasks, standard fine-tuning methods—such as expected risk minimization (ERM; Vapnik, 1998)—often fall short in providing reliable uncertainty estimates that accurately indicate how confident a model is about its predictions (Tran et al., 2022). Reliable uncertainty quantification, in conjunction with uncertainty-based deferral of predictions to human experts, can be a particularly useful tool in creating scalable oversight for large pre-trained models and help build trust for automated decision-making.

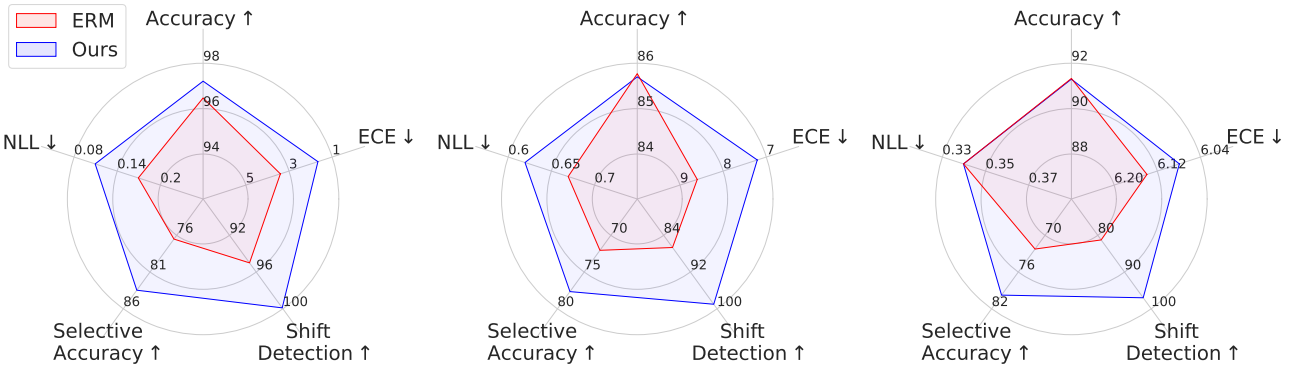
We present a simple addition to standard fine-tuning techniques that enables more reliable uncertainty quantification in fine-tuned models. More specifically, we derive a data-driven, uncertainty-aware prior distribution over neural network parameters designed to lead to improved uncertainty quantification. Prior distributions over model parameters can incorporate relevant information about the parameter values into Bayesian inference. However, specifying meaningful prior distributions over neural network parameters is challenging since it is unclear which parameter values would correspond to specific desired behaviors (e.g., good calibration, high predictive uncertainty under semantic shift, low negative log-likelihood, etc.) (Fortuin et al., 2022; Rudner et al., 2022a).

To tackle this challenge, we construct a *data-driven uncertainty-aware prior distribution* explicitly designed to encourage improved uncertainty quantification. This data-driven prior is specified by constructing a distribution that assigns high probability density to parameter values that induce functions with high predictive entropy on points that are meaningfully different from the training data.

We show how to use uncertainty-aware priors for fine-tuning off-the-shelf pre-trained models using standard mean-field variational inference for neural networks (Graves, 2011; Blundell et al., 2015) and evaluate the fine-tuned models on a wide range of benchmarking tasks—including image

[*]Equal contribution [1]New York University. Correspondence to: Tim G. J. Rudner <tim.rudner@nyu.edu>.

Accepted by the Structured Probabilistic Inference & Generative Modeling workshop of ICML 2024, Vienna, Austria. Copyright 2024 by the author(s).



(a) CIFAR-10 (b) CIFAR-100 (c) Question-Answering NLI

Figure 1. **Fine-tuning with Uncertainty-Aware Priors Significantly Improves Uncertainty Quantification.** The figure showcases the improvement of uncertainty-aware priors (UAPs) over expected risk minimization (ERM: the de facto standard method for fine-tuning) on a representative subset of the datasets considered in our empirical evaluation. All values are estimated from five trials.

classification and text classification—using a diverse set of uncertainty evaluation metrics. Uncertainty-aware priors are probabilistically principled, simple, scalable, and easy to implement in practice. Unlike existing data-driven priors that require separate training procedures (e.g., Shwartz-Ziv et al., 2022), uncertainty-aware priors allow for simple end-to-end training without the need for additional prior pre-training.

To summarize, our main contributions are as follows:

1. We construct a data-driven, uncertainty-aware prior (UAP) distribution for fine-tuning pre-trained neural networks. The prior distribution is explicitly designed to enable reliable uncertainty quantification.
2. We demonstrate how to perform tractable approximate inference in neural networks with data-driven, uncertainty-aware priors using a doubly lowerbounded variational objective.
3. We perform a careful empirical evaluation in which we compare models fine-tuned using the proposed uncertainty-aware prior to ERM fine-tuning and to a state-of-the-art method for Bayesian transfer learning. We find that mean-field variational inference with uncertainty-aware priors significantly outperforms the current state of the art in terms of uncertainty quantification and performs on par or better in terms of predictive accuracy.

2. Background

We consider supervised learning problems with N i.i.d. data realizations $\mathcal{D} = \{x_{\mathcal{D}}^{(n)}, y_{\mathcal{D}}^{(n)}\}_{n=1}^{N} = (x_{\mathcal{D}}, y_{\mathcal{D}})$ of inputs $X \in \mathcal{X}$ and targets $Y \in \mathcal{Y}$ with input space $\mathcal{X} \subseteq \mathbb{R}^D$ and target space $\mathcal{Y} \subseteq \mathbb{R}^K$ for regression and $\mathcal{Y} \subseteq \{0,1\}^K$ for classification tasks with K classes.

2.1. Learning as Probabilistic Inference

For supervised learning tasks, we define a *parametric observation model* $p_{Y|X,\Theta}(y \,|\, x, \theta; f)$, a *prior distribution* $p_\Theta(\theta)$, and a parametric mapping $f(\cdot\,;\theta)$. In Bayesian inference, we wish to find the posterior implied by the prior, the observation model, and the data. The posterior is given by

$$p_{\Theta|Y,X}(\theta \,|\, y_{\mathcal{D}}, x_{\mathcal{D}}) = \frac{p_{Y|X,\Theta}(y_{\mathcal{D}} \,|\, x_{\mathcal{D}}, \theta; f) p_\Theta(\theta)}{p_{Y|X}(y_{\mathcal{D}} \,|\, x_{\mathcal{D}}; f)}. \quad (1)$$

Since neural networks are non-linear in their parameters, exact inference over the stochastic network parameters is analytically intractable.

Variational Inference. Variational inference is an approximate inference method that seeks to avoid the intractability of exact inference by framing posterior inference as a variational optimization problem,

$$\min_{q_{\boldsymbol{\Theta}} \in \mathcal{Q}_{\boldsymbol{\Theta}}} \mathbb{D}_{\mathrm{KL}}(q_{\boldsymbol{\Theta}} \,\|\, p_{\boldsymbol{\theta}|\mathcal{D}}) \iff \max_{q_{\boldsymbol{\Theta}} \in \mathcal{Q}_{\boldsymbol{\Theta}}} \mathcal{F}(q_{\boldsymbol{\Theta}}),$$

where $\mathcal{F}(q_{\boldsymbol{\Theta}})$ is the variational objective

$$\mathcal{F}(q_\Theta) = \mathbb{E}_{q_\Theta}[\log p_{Y|X,\Theta}(y_{\mathcal{D}} \,|\, x_{\mathcal{D}}, \theta; f)] - \mathbb{D}_{\mathrm{KL}}(q_\Theta \,\|\, p_\Theta)$$

and $\mathcal{Q}_{\boldsymbol{\Theta}}$ is a variational family of distributions (Wainwright and Jordan, 2008). A commonly chosen variational family is that of mean-field Gaussian distribution, where $q_\Theta(\theta) = \mathcal{N}(\theta; \mu, \Sigma)$ with diagonal Σ.

Priors over Parameters as Regularizers. In optimization-based inference methods such as variational inference or stochastic gradient Langevin dynamics (SGLD), the prior $\log p_\Theta(\theta)$ effectively acts as a regularizer. For example, in SGLD training, selecting a prior $p_\Theta(\theta) = \mathcal{N}(\theta; \mathbf{0}, \tau_0^{-1} I)$ results in standard L_2-norm regularization (weight decay) while $p_\Theta(\theta) = \mathrm{Laplace}(\theta; \mathbf{0}, \tau_0^{-1} I)$ leads to sparsity-inducing L_1-norm regularization (LASSO) (Bishop, 2006; Murphy, 2013).

3. Related Work

Informative Priors in Bayesian Deep Learning. Informative priors have the potential to greatly improve model performance. However, certain challenges, such as cold posterior effects and performance concerns compared to traditional neural networks, have been reported in the literature (Wenzel et al., 2020). To address these shortfalls, recent studies have focused on exploring informative priors, including heavy-tailed priors (Fortuin et al., 2022; Izmailov et al., 2021b), noise-contrastive priors (Hafner et al., 2020), and input-dependent priors for domain generalization (Izmailov et al., 2021a; Rudner et al., 2023). Data-driven priors, derived from relevant context data, have been shown to improve group robustness (Rudner et al., 2024), clinical decision-making (Lopez et al., 2023), and out-of-distribution molecular and protein design (Klarner et al., 2024). Furthermore, function-space priors (Sun et al., 2019; Rudner et al., 2022a;b; Klarner et al., 2023; Qiu et al., 2023), sparsity-promoting weight-space priors (Carvalho et al., 2009; Ghosh et al., 2018), structured configuration priors (Louizos and Welling, 2016), and priors driven by learning algorithms (Khan et al., 2019; Immer et al., 2021) have been developed. Several studies have also explored formulating prior distributions in reference to related tasks. For example, Bayesian meta-learning (Finn et al., 2018; Rothfuss et al., 2021; Pavasovic et al., 2023) can be viewed as a way to learn priors from data, and empirical prior learning has been investigated in Robbins (1992). Fortuin et al. (2022) investigated learning empirical weight distributions using stochastic gradient descent, Wu et al. (2019) applied a moment-matching technique, and Krishnan et al. (2020) proposed to learn the mean of a Gaussian prior distribution from a relevant dataset. To improve transfer learning, Shwartz-Ziv et al. (2022) proposed a method to learn priors from large datasets using the Stochastic Weight Averaging-Gaussian (SWAG) framework (Maddox et al., 2019) which resulted in state-of-the-art predictive performance on computer vision transfer learning tasks (Harvey et al., 2024).

Transfer Learning. In transfer learning, representations learned from one task are adapted to enhance another task. This approach is foundational in deep learning, where large-scale neural networks are pre-trained on extensive datasets and then fine-tuned for specific tasks (Bommasani et al., 2021; Tran et al., 2022). In computer vision, models with vast pre-trained feature extractors are the default starting point for any image classification task (Dai et al., 2021; Dosovitskiy et al., 2021). Similarly, segmentation and object detection models use ImageNet pre-trained CNN or Transformer backbones combined with other modules (Chen et al., 2018; Ren et al., 2015; Dosovitskiy et al., 2021). In NLP, text classification is based on fine-tuned transformer models (Devlin et al., 2019; Howard and Ruder, 2018).

Knowledge Transfer with Bayesian Principles. Leveraging auxiliary data through Bayesian modeling can lead to improved generalization and robustness. Xuan et al. (2021) explored transfer with probabilistic graphical models, and Karbalayghareh et al. (2018) presented a theoretical framework for optimal Bayes classifiers informed by several domains. Several studies have used Bayesian tools to exploit auxiliary data across domains. For example, Chandra and Kapoor (2020) proposed to learn from multiple domains using round-robin task sampling with a single-layer neural network. Bayesian continual learning methods update the posterior to accommodate new tasks without neglecting previous ones (Ebrahimi et al., 2019; Kapoor et al., 2021; Nguyen et al., 2018; Tseran et al., 2018; Rudner et al., 2022b) or formulate kernels based on previously-trained neural networks for Gaussian process inference (Maddox et al., 2021; Pan et al., 2020; Titsias et al., 2020). Semi-supervised algorithms can blend unlabeled data in Bayesian neural network training pipelines, as shown in various studies (Do et al., 2021; Jean et al., 2018; Ravichandran et al., 2021; Shwartz Ziv and LeCun, 2024).

4. Uncertainty-Aware Fine-Tuning

In this section, we consider a family of data-driven priors, build on this family of priors to construct uncertainty-aware priors (UAPs) for fine-tuning pre-trained neural network models, and finally, we show to perform tractable variational inference in neural networks with such priors. Crucially, the proposed uncertainty-aware prior does not require any further pre-training and allows for straightforward end-to-end fine-tuning with existing pre-trained models.

4.1. A Family of Data-Driven Priors

Consider again a parametric observation model $p_{Y|X,\Theta}(y \,|\, x, \theta; f)$, and let the mapping f be defined by $f(\cdot\,;\theta) \doteq h(\cdot\,;\theta_h)\theta_L$, where $h(\cdot\,;\theta_h)$ is the post-activation output of the penultimate layer, Θ_L is the set of stochastic final-layer parameters, Θ_h is the set of stochastic non-final-layer parameters, and $\Theta \doteq \{\Theta_h, \Theta_L\}$ is the full set of stochastic parameters. We assume access to pre-trained feature parameters, θ_h^*, and context data that encodes useful information about the downstream tasks. We denote a batch of context inputs with corresponding context labels by $x_\mathrm{c} = \{x_1, ..., x_M\}$ and $y_\mathrm{c} = \{y_1, ..., y_M\}$, respectively, and let $p_{X_\mathrm{c}, Y_\mathrm{c}}$ be a joint distribution over context batches.

To construct a family of data-driven priors, we begin by specifying a *context inference problem*. We consider a Bernoulli random variable $\check{Z}$ denoting whether a given set of neural network parameters induces predictions that exhibit some desired property (e.g., high uncertainty on a set of evaluation points). Furthermore, we define a *context observation model*

$\check{p}_{\check{Z}\,|\,\Theta}(\check{z}\,|\,\theta; f, p_{X_c,Y_c})$—which denotes the likelihood of observing a yet-to-be-specified outcome $\check{z}$ under $\check{p}_{\check{Z}\,|\,\Theta}$ given θ and p_{X_c,Y_c}—and specify a *base prior* over the model parameters, $p_\Theta(\theta)$. For notational simplicity, we will drop the subscripts going forward except when needed for clarity. With this setup, we can now define the context inference problem as finding the conditional distribution over neural network parameters that we *would* obtain if we conditioned on the desired property being satisfied. This conditional distribution will serve as our data-driven prior, and by Bayes' Theorem, we can express it as

$$p(\theta\,|\,\check{z}; p_{X_c,Y_c}) = \frac{\check{p}(\check{z}\,|\,\theta; p_{X_c,Y_c})p(\theta)}{\check{p}(\check{z}; p_{X_c,Y_c})}. \tag{2}$$

To define a family of data-driven priors that place high probability density on neural network parameter values that induce predictive functions with reliable uncertainty estimates, we specify a Bernoulli context observation model $\check{p}_{\check{Z}\,|\,\Theta}$ in which $\check{Z} = 1$ denotes the outcome of 'achieving reliable uncertainty quantification' and $\check{p}(\check{z} = 1\,|\,\theta; p_{X_c,Y_c})$ denotes the likelihood of $\check{z} = 1$ given θ and p_{X_c,Y_c}. More specifically, we define

$$\begin{aligned} \check{p}(\check{z} = 1\,|\,\theta; p_{X_c,Y_c}) &= \exp(-\mathbb{E}_{p_{X_c,Y_c}}[c(X_c, Y_c, \theta)]) \\ \check{p}(\check{z} = 0\,|\,\theta; p_{X_c,Y_c}) &= 1 - \check{p}(\check{z} = 1\,|\,\theta; p_{X_c,Y_c}), \end{aligned} \tag{3}$$

where $c : \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^P \to \mathbb{R}_{\geq}$ is a *cost function*. By specifying the outcome $\check{z} = 1$ along with a distribution p_{X_c,Y_c} we obtain a conditional distribution $\check{p}(\theta\,|\,\check{z}; f, p_{X_c,Y_c})$—the distribution over neural network parameters that we *would* infer if we observed outcome $\check{z} = 1$ under the base prior and the Bernoulli context observation model defined in Equation (3). Naturally, the quality (i.e., the usefulness) of this conditional distribution is determined by the quality of the context observation model $\check{p}_{\check{Z}\,|\,\Theta}$, the data, and the prior. As a result, the primary challenge in designing effective uncertainty-aware priors lies in constructing a context observation model—via a cost function c and a context distribution p_{X_c,Y_c}—that is as well-specified as possible. The better specified the context observation model, the more useful the data-driven prior.

4.2. Defining Data-Driven, Uncertainty-Aware Priors for Fine-Tuning Pre-trained Models

In this section, we present a specific instantiation of an uncertainty-aware prior for fine-tuning foundation models. To define a data-driven prior $\check{p}(\theta\,|\,\check{z}; p_{X_c,Y_c})$ that incorporates useful information from the pre-trained parameters θ_h^* and assigns high probability density to parameter values θ that induce models with reliable uncertainty quantification, we need to specify a suitable context likelihood and suitable layer-specific base priors $p(\theta_h)$ and $p(\theta_L)$. For the base priors, we let $p(\theta_h) = \mathcal{N}(\theta_h; \theta_h^*, \tau_h^{-1} I)$, which assigns high probability to parameters θ_h that are close to the pre-trained parameters θ_h^*, and $p(\theta_L) = \mathcal{N}(\theta_L; \mathbf{0}, \tau_L^{-1} I)$.

To define a context observation model that induces a data-driven prior with desirable properties, we specify a cost function c of the form

$$c(x_c, y_c, \theta) \doteq \tau \sum_{k=1}^{K} D^2_{\mathcal{M}}([f(x_c;\theta)]_k, m(x_c, y_c)_k, C(x_c)), \tag{4}$$

where K is the number of output dimensions, p_{X_c,Y_c} is a joint distribution over context batches x_c and y_c (each of size M),

$$D^2_{\mathcal{M}}([f(x_c;\theta)]_k, m(x_c, y_c)_k, C(x_c)) \doteq \mathbf{v}_k^\top C(x_c)^{-1} \mathbf{v}_k \tag{5}$$

with $\mathbf{v}_k \doteq [f(x_c;\theta)]_k - m(x_c, y_c)_k$ is the squared Mahalanobis distance between model predictions $[f(x_c;\theta)]_k$ and an input-dependent distribution with mean $m(x_c, y_c)_k$ and M-by-M covariance matrix $C(x_c)$. To obtain a data-driven prior that assigns high probability density to parameters θ that induce models with reliable uncertainty estimates, we specify a data-dependent mean function, $m(x_c, y_c)_k \doteq [y_c]_k$, and a covariance function

$$C(\cdot) \doteq s_1 h(\cdot; \theta_h^*) h(\cdot; \theta_h^*)^\top + s_2 I, \tag{6}$$

parameterized by pre-trained model parameters θ_h^* and fixed scaling parameters τ, s_1, and s_2, that reflects structure in the pre-trained model representations $h(\cdot; \theta_h^*)$. Finally, we define the context distribution as $p(x_c, y_c) = p(y_c\,|\,x_c)p(x_c)$, where $p(y_c\,|\,x_c) \doteq \delta(\{\mathbf{0}, ..., \mathbf{0}\} - y_c)$ and $p(x_c)$ is an empirical distribution constructed from a larger set of (domain- and task-specific) context inputs.[1]

Under this cost function and context distribution, the data-driven prior defined in Equation (3), by design, assigns high probability density to parameters θ that induce predictions $f(x_c;\theta)$ that have high predictive uncertainty on the context inputs. If the distribution over context inputs, p_{X_c}, is specified to place high probability density on context batches which contain input points that are meaningfully distinct from the training inputs, then the data-driven prior favors models that exhibit high predictive uncertainty on such meaningfully distinct inputs.

4.3. Variational Inference with Uncertainty-Aware Priors

In this section, we show how to perform variational inference with uncertainty-aware priors. We start by specifying a probabilistic model with an uncertainty-aware prior,

$$p(y_\mathcal{D}, \theta\,|\,x_\mathcal{D}, \check{z}; p_{X_c,Y_c}) = \underbrace{p(y_\mathcal{D}\,|\,x_\mathcal{D}, \theta; f)}_{\text{Likelihood}}\ \underbrace{p(\theta\,|\,\check{z}; p_{X_c,Y_c}))}_{\text{Uncertainty-aware prior}}.$$

To perform variational inference in this model and approximate the posterior distribution over the parameters of interest, we begin by defining a variational distribution,

$$q(\theta) \doteq q(\theta_h)\, q(\theta_L),$$

[1] Defining $p(y_c\,|\,x_c) \doteq \delta(\{\mathbf{0}, ..., \mathbf{0}\} - y_c)$ implies that under p_{X_c,Y_c}, all context batch samples have $y_c = \mathbf{0}$, and therefore, we effectively have $m(x_c, y_c)_k \doteq \mathbf{0}$ for all context batch samples.

where $q(\theta_L) = \mathcal{N}(\theta_L; \mu_L, \Sigma_L)$ and $q(\theta_h) = \mathcal{N}(\theta_h; \mu_h, \Sigma_h)$ with learnable variational parameters $\mu \doteq \{\mu_h, \mu_L\}$ and $\Sigma \doteq \{\Sigma_h, \Sigma_L\}$, and frame the inference problem of finding the posterior $p(\theta \mid x_{\mathcal{D}}, y_{\mathcal{D}}, \check{z})$ variationally as

$$\min_{q_\Theta \in \mathcal{Q}} \mathbb{D}_{\mathrm{KL}}(q_\Theta \parallel p_{\Theta \mid X_{\mathcal{D}}, Y_{\mathcal{D}}, \check{Z}}),$$

where $\mathcal{Q}$ is a mean-field Gaussian variational family. This variational problem can equivalently be expressed as maximizing the variational objective

$$\bar{\mathcal{F}}(q_\Theta) \doteq \mathbb{E}_{q_\Theta}[\log p(y_{\mathcal{D}} \mid x_{\mathcal{D}}, \Theta; f)] - \mathbb{D}_{\mathrm{KL}}(q_\Theta \parallel p_{\Theta \mid \check{Z}}).$$

Unfortunately, this variational objective is intractable since the data-driven prior $\check{p}(\theta \mid \check{z}; p_{X_c, Y_c})$ defined in Equation (2)—which is required to compute $\mathbb{D}_{\mathrm{KL}}(q_\Theta \parallel p_{\Theta \mid \check{Z}})$—is not in general tractable.

To overcome this intractability, we take advantage of the properties of the KL divergence and note that we can express $\mathbb{D}_{\mathrm{KL}}(q_\Theta \parallel p_{\Theta \mid \check{Z}})$ as

$$\begin{aligned}\mathbb{D}_{\mathrm{KL}}(q_\Theta \parallel p_{\Theta \mid \check{Z}}) = \mathbb{E}_{q_{\Theta_h} q_{\Theta_L}} &\left[\log \frac{q(\Theta_h)\, q(\Theta_L)}{p(\Theta_h)\, p(\Theta_L)}\right] \\ &- \mathbb{E}_{q_{\Theta_h} q_{\Theta_L}}[\log \check{p}(\check{z} \mid \Theta; p_{X_c, Y_c})] \\ &+ \log \check{p}(\check{z}; p_{X_c, Y_c}),\end{aligned}$$

where the intractable log-marginal likelihood $\log \check{p}(\check{z}; p_{X_c, Y_c})$ was factored out as an additive constant independent of any learnable parameters. Using this result, we can obtain a tractable lower bound

$$\begin{aligned}\mathbb{D}_{\mathrm{KL}}(q_\Theta \parallel p_{\Theta \mid \check{Z}}) \geq &-\mathbb{E}_{q_{\Theta_h} q_{\Theta_L}}[\log \check{p}(\check{z} \mid \Theta; p_{X_c, Y_c})] \\ &+ \mathbb{D}_{\mathrm{KL}}(q_{\Theta_h} \parallel p_{\Theta_h}) \\ &+ \mathbb{D}_{\mathrm{KL}}(q_{\Theta_L} \parallel p_{\Theta_L}),\end{aligned} \tag{7}$$

where each KL divergence term can be computed analytically, and we can obtain an unbiased estimator of the negative log-likelihood using simple Monte Carlo estimation.

Variational Objective. Since q_{Θ_h} and q_{Θ_L} are both mean-field Gaussian distributions, we can obtain a doubly lower-bounded variational objective

$$\begin{aligned}\mathcal{F}(\mu, \Sigma) \doteq &\underbrace{\mathbb{E}_{q_\Theta}[\log p(y_{\mathcal{D}} \mid x_{\mathcal{D}}, \Theta; f)]}_{\text{Expected log-likelihood}} \\ &- \underbrace{\mathbb{D}_{\mathrm{KL}}(q_{\Theta_L} \parallel p_{\Theta_L})}_{\text{Pre-training regularization}} - \underbrace{\mathbb{D}_{\mathrm{KL}}(q_{\Theta_L} \parallel p_{\Theta_L})}_{\text{Final-layer regularization}} \\ &- \underbrace{\mathbb{E}_{q_\Theta}[\mathbb{E}_{p_{X_c, Y_c}}[c(X_c, Y_c, \Theta)]]}_{\text{Uncertainty regularization}},\end{aligned} \tag{8}$$

where the cost function and context distribution are as defined above. We can estimate all expectations in the objective using simple Monte Carlo estimation, giving the final variational objective

$$\begin{aligned}\hat{\mathcal{F}}(\mu, \Sigma) \doteq &\frac{1}{J} \sum\nolimits_{j=1}^{J} \log p(y_{\mathcal{D}} \mid x_{\mathcal{D}}, \theta^{(j)}; f) - \mathbb{D}_{\mathrm{KL}}(q_\Theta \parallel p_\Theta) \\ &- \frac{1}{JJ'} \sum\nolimits_{j=1}^{J} \sum\nolimits_{j'=1}^{J'} c(x_c^{(j')}, y_c^{(j')}, \theta^{(j)}),\end{aligned} \tag{9}$$

with $\theta^{(j)} \sim q_\Theta$, $x_c^{(j')} \sim p_{X_c}$, and $y_c^{(j')} \sim p_{Y_c \mid X_c}$ for $j = 1, ..., J$ and $j' = 1, ..., J'$. This objective can be maximized with stochastic variational inference (Hoffman et al., 2013).

When maximizing this variational objective, the uncertainty-aware prior explicitly encourages the predictive distribution induced by the learned variational distribution q_Θ to have high uncertainty on the context inputs while taking into account the covariance under the pre-trained features $h(\cdot; \theta_h^*)$ and pulling μ_h towards the pre-trained parameters θ_h^*

4.4. Practical Considerations

Computational Complexity. Computing the uncertainty regularizer requires inverting an M-by-M covariance matrix, which limits the size of the context batches. However, we find that in practice, a small context batch size M (roughly 25% of the mini-batch size) is sufficient, and the main computational expense is the forward pass needed to compute the context likelihood—not the matrix inversion. As the context distribution p_{X_c, Y_c} is defined over context batches (i.e., as defined above, each sample $x_c^{(j')}$ and $y_c^{(j')}$ is a randomly sampled batch of size M), we find that setting $J' = 1$ for each stochastic gradient descent step is sufficient in practice, which significantly reduces the computational cost of the nested Monte Carlo estimation in Equation (9).

Defining a Suitable Context Distribution. In order for a data-driven prior to favor models that generate reliable uncertainty estimates, the context input distribution must be chosen carefully. We find that distributions over context inputs that retain domain-specific structure work particularly well. For example, when performing image classification, specifying a distribution over images that are meaningfully distinct from the training images is more effective than white noise, and when performing text classification, specifying a distribution over sentences that are meaningfully distinct from the training data but make sense grammatically is more effective than random sequences of words.

5. Empirical Evaluation

In this section, we evaluate the usefulness of uncertainty-aware priors for fine-tuning pre-trained models using uncertainty-based evaluation metrics, including the negative log-likelihood, selective prediction accuracy, calibration, and semantic shift detection AUROC. We find that using uncertainty-aware priors leads to significant empirical improvements in uncertainty quantification on a diverse set of computer vision and natural language classification tasks.

5.1. Experiment Setup

Datasets. We evaluate our methods on the CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), and Flowers102 (Nilsback and Zisserman, 2008) computer vision datasets, and on the MultiNLI and QNLI natural language datasets (Wang et al., 2018).

Semantic Shift Data. For CIFAR-10 and CIFAR-100, we use the SVHN dataset as an example of semantic shift. For Flowers, we use the Plantae subset from the iNaturalist dataset (Van Horn et al., 2018) as an example of semantic shift. For QNLI, we use the MultiNLI dataset and for MultiNLI, we use the Emotions dataset (Saravia et al., 2018) as an example of semantic shift.

Context Distributions. For all image classification tasks, we sample uniformly from the ImageNet dataset (Deng et al., 2009) to construct the context distribution. For MultiNLI and QNLI, we use sample uniformly from the MathQA dataset (Amini et al., 2019) as the context.

Models. We use a pre-trained ResNet-50 (He et al., 2016) for image and a pre-trained BERT-base (uncased) model (Devlin et al., 2019) for text classification tasks.

Baselines. The default—and most commonly used—method for fine-tuning pre-trained neural networks is Expected Risk Minimization (ERM) (Chen et al., 2020b; Bardes et al., 2022). In addition to ERM, we also compare our method to the *Pretrain Your Loss* (PTYL) method by Shwartz-Ziv et al. (2022), the state of the art for Bayesian transfer learning from pre-trained models. For implementation details, see Appendix B.

Training Details. For full training details, see Appendix B.

5.2. Accuracy and Negative Log-Likelihood

While having reliable uncertainty estimates is a crucial component of trustworthy models, it is important that efforts to improve uncertainty quantification do not compromise model accuracy. For this reason, we begin by evaluating predictive accuracies and negative log-likelihoods obtained with different methods.

Results. We report the accuracy and negative log-likelihood (NLL) in Tables 1 and 2 and Figure 2. We find that training with uncertainty-aware priors results in similar or improved performance accuracy and a consistent improvement in the negative log-likelihood, reflecting improved uncertainty quantification without deterioration in generalization.

5.3. Selective Prediction

Selective prediction modifies the standard prediction pipeline by introducing a rejection class (El-Yaniv et al., 2010). In selective prediction pipelines, input points for which a model abstains from making a prediction can be deferred to a human expert for review, enabling collaborative human-machine decision-making. This way, selective prediction can provide a responsible approach to leveraging machine learning systems in safety-critical domains while maintaining human oversight. Given classifier f where $f(x) = \arg\max_{k \in \mathcal{Y}} f(x|k)$, the selective prediction model introduces a selection function s which determines whether a prediction should be made. This selection function can be based on the outputs of f, such as $s(x) = \max_{k \in \mathcal{Y}} f(x|k)$ The selective prediction model f_s is then defined as

$$f(x; \tau) = \begin{cases} f(x) & \text{if } s(x) \geq \tau \\ \perp & \text{otherwise } s(x) < \tau \end{cases} \tag{10}$$

where τ represents the rejection threshold.

Predictive uncertainty is a natural choice for a selection function: If a model's predictive uncertainty is above a certain threshold, the selection function will decline to make a prediction. Using predictive uncertain as a selection function, selective prediction allows us to assess both a model's predictive accuracy and the quality of its uncertainty estimates. If a model is successful at rejecting data points for which it would have made an incorrect prediction based on its level of predictive uncertainty, the accuracy of the remaining, non-rejected points will increase as more and more points are rejected. To evaluate the selective prediction model, we compute its predictive accuracy over a range of thresholds τ and compute the area under the selective prediction accuracy curve. Successful selective prediction models are able to obtain high accuracy across many thresholds.

Results. In Table 3 and Figure 3, we show the selective prediction results on data with a mixture of in-domain and out-of-distribution samples. The quality of the uncertainty estimates determines the accuracy of the predicted inputs, where in-domain samples are compared to their true label, and any out-of-distribution samples are marked as incorrect. We find that our method consistently outperforms the baseline methods across all downstream tasks, both for images and language classification tasks.

Method	CIFAR-10	CIFAR-100	Flowers
ERM	96.45±0.08	**85.76**±0.20	89.64±0.24
PTYL	**97.35**±0.34	**85.82±0.23**	89.73±0.51
Ours	**97.19**±0.08	**85.69**±0.13	**90.35**±0.18

(a) **Computer Vision Tasks.**

Method	QNLI	MultiNLI
ERM	91.31±0.10	84.65±0.13
PTYL	91.29±0.11	84.60±0.19
Ours	91.28±0.13	84.64±0.06

(b) **Language Tasks.**

Table 1. **Predictive Accuracy (in %).** Our method achieves competitive accuracy across modalities and datasets. The mean and standard error are estimated from five trials.

Method	CIFAR-10	CIFAR-100	Flowers
ERM	0.17±0.01	0.67±0.02	0.48±0.02
PTYL	**0.11**±0.01	0.68±0.01	0.47±0.01
Ours	**0.11**±0.01	**0.62**±0.01	**0.45**±0.02

Table 2. **Negative Log Likelihood.** Lower values are better. Our method consistently achieves the best negative log-likelihood compared to other methods. The mean and standard error are estimated from five trials.

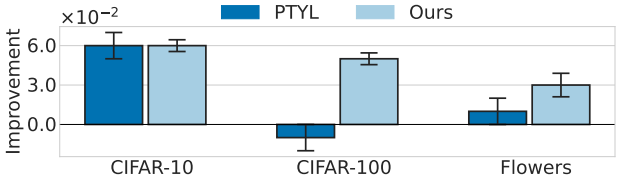


Figure 2. **Improvement in Negative Log Likelihood.** We plot the difference in the mean negative log-likelihood between ERM and other methods. The mean and standard error are estimated from five trials.

Method	CIFAR-10	CIFAR-100	Flowers
ERM	76.46±0.27	71.99±1.01	78.87±0.62
PTYL	82.48±0.92	72.10±1.69	77.88±0.34
Ours	**83.42**±0.03	**77.61**±0.10	**79.68**±0.08

Table 3. **Selective Prediction Accuracy (in %).** Higher values are better. Our method consistently achieves the best compared to other methods. The mean and standard error are estimated from five trials.

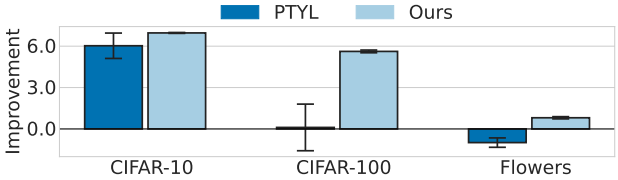


Figure 3. **Improvement in Selective Prediction Accuracy (in %).** We plot the difference in the mean selective accuracy between ERM and other methods. The mean and standard error are estimated from five trials.

5.4. Model Calibration

Calibration expresses how closely the confidence of a model's predictions is aligned with its accuracy, and well-calibrated models allow users to reliably estimate the accuracy of a model's prediction from its confidence. This is especially important in safety-critical applications, where it is crucial to identify inaccurate predictions. Well-calibrated models are also more trustworthy, as they provide users with a clearer understanding of when to rely on the model.

One notion of miscalibration is the Expected Calibration Error (ECE; Naeini et al., 2015), which computes the gap between model accuracy and confidence. The ECE estimator is defined as

$$\mathrm{ECE} = \sum_{m=1}^{M'} \frac{|B_m|}{n} |\mathrm{Accuracy}(B_m) - \mathrm{Confidence}(B_m)|,$$

where n is the number of samples, M' is the number of bins, $\mathrm{Accuracy}(B_m)$ is the accuracy of samples within bin B_m, and $\mathrm{Confidence}(B_m)$ is the average maximum probability outputs of the classifier of all samples within the bin. A perfectly calibrated model has an ECE of zero since its accuracy is perfectly aligned with its confidence, and a more miscalibrated model has a higher ECE.

Results. In Table 4 and Figure 4, we see that our method significantly improves model calibration on all downstream image tasks. We note that this improvement in performance is dependent on the uncertainty-aware fine-tuning prior, as the prior used by PTYL does not see similar benefits for calibration. For language tasks such as QNLI and MultiNLI, we find our method leads to slight improvements in calibration compared to the ERM baseline (see Appendix A).

5.5. Semantic Shift in Vision and Language Data

In real-world prediction tasks, models may be tested on data that is meaningfully different from the training data. *Semantic shift* is a type of distribution shift where the test data contains labels that are semantically different from any labels seen during training, that is, $p_{\text{train}}(x, y) \neq p_{\text{test}}(x, y)$. This poses a significant challenge to model performance, as a model will, by definition, not be able to output the correct prediction for semantically different inputs. It is, therefore, important to detect instances of semantic shift so that users can trust a model's prediction.

Semantic shift in natural language can be particularly subtle, and semantic shifts may occur unexpectedly at deployment, which can lead to catastrophic failures. To assess the reliability of semantic shift detection in language data, we evaluate our method on two different language datasets, MultiNLI and QNLI.

Uncertainty estimates can be used to identify semantic shifts in the data, and reliable models should have high uncertainty for input points whose true labels are semantically different from the training labels. To evaluate the model's performance on detecting semantic shift, we compute the predictive entropy $\mathcal{H}(p(y \mid x))$. We design test sets that consist of a mixture of in-distribution and semantically shifted test samples, and we construct a binary classifier that only uses the uncertainty score to separate the two groups. We then compute the area under the ROC curve (AUROC) of this classifier as an evaluation metric for model uncertainty performance. Models that are able to successfully detect a semantic shift achieve higher AUROC scores.

Method	CIFAR-10	CIFAR-100	Flowers
ERM	3.42±0.13	8.61±0.26	11.00±0.47
PTYL	3.10±0.27	8.65±0.56	10.99±0.64
Ours	**1.68**±0.08	**7.22**±0.20	**10.07**±0.18

Table 4. **Expected Calibration Error (ECE).** Lower values are better. Our method consistently achieves the best ECE compared to other methods. The mean and standard error are estimated from five trials.

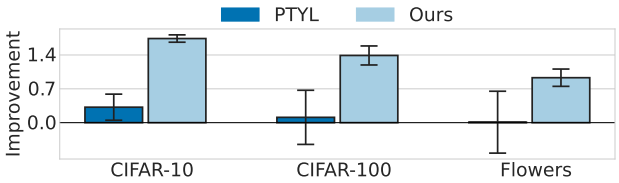


Figure 4. **Improvement in ECE.** Lower values are better. We plot the difference in the mean ECE between ERM and other methods. The mean and standard error are estimated from five trials.

Results. As can be seen in Table 5 and Figure 5, we find that our method improves semantic shift detection by a significant margin across all computer vision datasets, and we are able to achieve an AUROC of above 96% for all tasks. This indicates that uncertainty-aware fine-tuning priors enable models to successfully separate in-domain data from semantically shifted data using only uncertainty values at a much higher rate than existing methods. Additionally, as can be seen in Table 6, uncertainty-aware fine-tuning priors significantly improve models' ability to detect semantic shifts in natural language, as evidenced by both the selective prediction accuracy and the semantic shift detection AUROC relative to the ERM baseline and PTYL. For semantic shift examples, see Appendix B.

Method	Selective Acc. (↑)	Det. AUROC (↑)
ERM	72.18±0.42	81.15±0.42
PTYL	73.17±0.23	84.27±0.35
Ours	**79.71**±0.10	**96.90**±0.34

(a) **Multi-Genre NLI (MultiNLI).**

Method	Selective Acc. (↑)	Det. AUROC (↑)
ERM	75.41±0.13	94.28±0.59
PTYL	75.88±0.15	94.74±0.41
Ours	**76.34**±0.13	**98.17**±0.72

(b) **Question-Answering NLI (QNLI).**

Table 6. **Semantic Shift Detection Selective Prediction.** Our method significantly outperforms ERM at selective prediction with in-domain and semantically shifted data (Selective Acc.) and at uncertainty-based semantic shift detection (Detection AUROC). The mean and standard error are estimated from five trials.

Method	CIFAR-10	CIFAR-100	Flowers
ERM	94.96±0.72	86.57±1.21	95.68±1.85
PTYL	96.94±1.92	88.23±2.98	92.30±0.57
Ours	**99.86**±0.04	**98.93**±0.22	**97.81**±0.10

Table 5. **Semantic Shift Detection AUROC (in %).** Higher values are better. Our method consistently achieves the best AUROC compared to other methods. The mean and standard error are estimated from five trials.

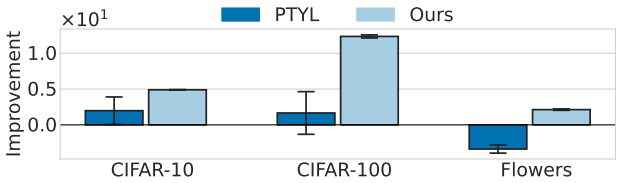


Figure 5. **Improvement in Semantic Shift Detection (in %).** We plot the difference in the mean detection AUROC between ERM and other methods. The mean and standard error are estimated from five trials.

6. Discussion and Limitations

Reliable uncertainty quantification is a key ingredient for creating trust in machine learning systems. We showed that explicitly incorporating uncertainty-aware priors into fine-tuning routines for pre-trained models consistently and, in many cases, significantly improves uncertainty quantification across modalities and datasets. Intriguingly, the extent to which fine-tuning pre-trained models with data-driven, uncertainty-aware priors improves uncertainty quantification depends heavily on the design of the context distribution, which governs the input points on which the prior encourages the model to have high uncertainty. In our empirical evaluation, we used fairly naive context distributions: For computer vision tasks, we used ImageNet, and for language tasks, we used the MathQA dataset to define domain-related context distributions. We hypothesize that using more carefully tailored, domain-specific context distributions will further improve performance. Finally, we emphasize that uncertainty-aware priors are complementary to other efforts for improving model performance, such as more sophisticated pre-training techniques or alternative architectures, can be used with any Bayesian inference method that only requires access to an unnormalized prior density (like SGLD, SG-HMC, or the Laplace approximation), and we recommend uncertainty-aware priors as a simple, scalable, and probabilistically principled plug-and-play addition to standard fine-tuning routines.

References

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1245.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Christopher M. Bishop. Pattern recognition and machine learning (information science and statistics). 2006.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622, Lille, France, 07–09 Jul 2015. PMLR.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. Handling sparsity via the horseshoe. In David van Dyk and Max Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 73–80, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR.

Rohitash Chandra and Arpit Kapoor. Bayesian neural multi-source transfer learning. *Neurocomputing*, 378:54–64, 2020.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*, pages 1683–1691. PMLR, 2014.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020b.

Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. CoAtNet: Marrying convolution and attention for all data sizes. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3965–3977. Curran Associates, Inc., 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

Kien Do, Truyen Tran, and Svetha Venkatesh. Semi-supervised learning with variational Bayesian inference and maximum uncertainty regularization. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 7236–7244. AAAI Press, 2021. doi: 10.1609/AAAI.V35I8.16889.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Sayna Ebrahimi, Mohamed Elhoseiny, Trevor Darrell, and Marcus Rohrbach. Uncertainty-guided continual learning in bayesian neural networks - extended abstract. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2019.

Ran El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.

Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. *Advances in neural information processing systems*, 31, 2018.

Vincent Fortuin, Adrià Garriga-Alonso, Sebastian W. Ober, Florian Wenzel, Gunnar Ratsch, Richard E Turner, Mark van der Wilk, and Laurence Aitchison. Bayesian neural network priors revisited. In *International Conference on Learning Representations*, 2022.

Soumya Ghosh, Jiayu Yao, and Finale Doshi-Velez. Structured variational learning of Bayesian neural networks with horseshoe priors. In *International Conference on Machine Learning*, pages 1744–1753. PMLR, 2018.

Alex Graves. Practical variational inference for neural networks. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, page 2348–2356, Red Hook, NY, USA, 2011. Curran Associates Inc. ISBN 9781618395993.

Danijar Hafner, Dustin Tran, Timothy Lillicrap, Alex Irpan, and James Davidson. Noise contrastive priors for functional uncertainty. In *Uncertainty in Artificial Intelligence*, pages 905–914. PMLR, 2020.

Ethan Harvey, Mikhail Petrov, and Michael C Hughes. Transfer learning with informative priors: Simple baselines better than previously reported. In *Transactions on Machine Learning Research*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, May 2013. ISSN 1532-4435.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1031.

Alexander Immer, Matthias Bauer, Vincent Fortuin, Gunnar Rätsch, and Khan Mohammad Emtiyaz. Scalable marginal likelihood estimation for model selection in deep learning. In *International Conference on Machine Learning*, pages 4563–4573. PMLR, 2021.

Pavel Izmailov, Patrick Nicholson, Sanae Lotfi, and Andrew Gordon Wilson. Dangers of bayesian model averaging under covariate shift. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021a.

Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Gordon Wilson. What are Bayesian neural network posteriors really like? In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4629–4640. PMLR, 18–24 Jul 2021b.

Neal Jean, Sang Michael Xie, and Stefano Ermon. Semi-supervised deep kernel learning: Regression with unlabeled data by minimizing predictive variance. *Advances in Neural Information Processing Systems*, 31, 2018.

Sanyam Kapoor, Theofanis Karaletsos, and Thang D Bui. Variational auto-regressive gaussian processes for continual learning. In *International Conference on Machine Learning*, pages 5290–5300. PMLR, 2021.

Alireza Karbalayghareh, Xiaoning Qian, and Edward R Dougherty. Optimal Bayesian transfer learning. *IEEE Transactions on Signal Processing*, 66(14):3724–3739, 2018.

Mohammad Emtiyaz E Khan, Alexander Immer, Ehsan Abedi, and Maciej Korzepa. Approximate inference turns deep networks into Gaussian processes. In *Advances in Neural Information Processing Systems 32*, pages 3094–3104. Curran Associates, Inc., 2019.

Leo Klarner, Tim G. J. Rudner, Michael Reutlinger, Torsten Schindler, Garrett M. Morris, Charlotte Deane, and Yee Whye Teh. Drug Discovery under Covariate Shift with Domain-Informed Prior Distributions over Functions. In *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2023.

Leo Klarner, Tim G. J. Rudner, and Yee Whye Teh. Garrett M. Morris, Charlotte Deane. Domain-aware guidance for out-of-distribution molecular and protein design. In *Proceedings of the 41th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2024.

Ranganath Krishnan, Mahesh Subedar, and Omesh Tickoo. Specifying weight priors in Bayesian deep neural networks with empirical Bayes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4477–4484, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Julian Lechuga Lopez, Tim G. J. Rudner, and Farah Shamout. Informative priors improve the reliability of multimodal clinical data classification. In *Machine Learning for Health Symposium Findings*, 2023.

Christos Louizos and Max Welling. Structured and efficient variational deep learning with matrix gaussian posteriors. In *International conference on machine learning*, pages 1708–1716. PMLR, 2016.

Wesley Maddox, Shuai Tang, Pablo Moreno, Andrew Gordon Wilson, and Andreas Damianou. Fast adaptation with linearized neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 2737–2745. PMLR, 2021.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for Bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems*, pages 13153–13164, 2019.

Kevin P. Murphy. *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.], 2013. ISBN 9780262018029 0262018020.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Cuong V Nguyen, Yingzhen Li, Thang D. Bui, and Richard E. Turner. Variational continual learning. In *International Conference on Learning Representations*, 2018.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008.

Pingbo Pan, Siddharth Swaroop, Alexander Immer, Runa Eschenhagen, Richard Turner, and Mohammad Emtiyaz E Khan. Continual deep learning by functional regularisation of memorable past. *Advances in Neural Information Processing Systems*, 33:4453–4464, 2020.

Krunoslav Lehman Pavasovic, Jonas Rothfuss, and Andreas Krause. MARS: Meta-learning as score matching in the function space. In *The Eleventh International Conference on Learning Representations*, 2023.

Shikai Qiu, Tim G. J. Rudner, Sanyam Kapoor, and Andrew Gordon Wilson. Should we learn most likely functions or parameters? In *Advances in Neural Information Processing Systems 36*, 2023.

Naresh Balaji Ravichandran, Anders Lansner, and Pawel Herman. Semi-supervised learning with Bayesian confidence propagation neural network. *arXiv preprint arXiv:2106.15546*, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

Herbert E. Robbins. *An Empirical Bayes Approach to Statistics*, pages 388–394. Springer New York, New York, NY, 1992. ISBN 978-1-4612-0919-5. doi: 10.1007/978-1-4612-0919-5_26.

Jonas Rothfuss, Vincent Fortuin, Martin Josifoski, and Andreas Krause. PACOH: Bayes-optimal meta-learning with PAC-guarantees. In *International Conference on Machine Learning*, pages 9116–9126. PMLR, 2021.

Tim G. J. Rudner, Zonghao Chen, Yee Whye Teh, and Yarin Gal. Tractable function-space variational inference in Bayesian neural networks. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022a.

Tim G. J. Rudner, Freddie Bickford Smith, Qixuan Feng, Yee Whye Teh, and Yarin Gal. Continual Learning via Sequential Function-Space Variational Inference. In *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2022b.

Tim G. J. Rudner, Sanyam Kapoor, Shikai Qiu, and Andrew Gordon Wilson. Function-Space Regularization in Neural Networks: A Probabilistic Perspective. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Tim G. J. Rudner, Ya Shi Zhang, Andrew Gordon Wilson, and Julia Kempe. Mind the GAP: Improving robustness to subpopulation shifts with group-aware priors. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, 2024.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels,

Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1404.

Ravid Shwartz Ziv and Yann LeCun. To compress or not to compress—self-supervised learning and information theory: A review. *Entropy*, 26(3), 2024. ISSN 1099-4300. doi: 10.3390/e26030252.

Ravid Shwartz-Ziv, Micah Goldblum, Hossein Souri, Sanyam Kapoor, Chen Zhu, Yann LeCun, and Andrew G Wilson. Pre-train your loss: Easy Bayesian transfer learning with informative priors. *Advances in Neural Information Processing Systems*, 35:27706–27715, 2022.

Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. Functional variational Bayesian neural networks. In *International Conference on Learning Representations*, 2019.

Michalis K. Titsias, Jonathan Schwarz, Alexander G. de G. Matthews, Razvan Pascanu, and Yee Whye Teh. Functional regularisation for continual learning with Gaussian processes. In *International Conference on Learning Representations*, 2020.

Dustin Tran, Jeremiah Liu, Michael W. Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda Mariet, Huiyi Hu, Neil Band, Tim G. J. Rudner, Karan Singhal, Zachary Nado, Joost van Amersfoort andAndreas Kirsch, Rodolphe Jenatton, Nithum Thain, Honglin Yuan, Kelly Buchanan, Kevin Murphy, D. Sculley, Yarin Gal, Zoubin Ghahramani, Jasper Snoek, and Balaji Lakshminarayanan. Plex: Towards Reliability Using Pretrained Large Model Extensions. In *ICML 2022 Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward*, 2022.

Hanna Tseran, Mohammad Emtiyaz Khan, Tatsuya Harada, and Thang D Bui. Natural variational continual learning. In *Continual Learning Workshop@ NeurIPS*, volume 2, 2018.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Vladimir Vapnik. *Statistical learning theory*. Wiley, 1998. ISBN 978-0-471-03003-4.

Martin J Wainwright and Michael I Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., Hanover, MA, USA, 2008. ISBN 1601981848.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446.

Florian Wenzel, Kevin Roth, Bastiaan Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the Bayes posterior in deep neural networks really? In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10248–10259. PMLR, 13–18 Jul 2020.

Anqi Wu, Sebastian Nowozin, Edward Meeds, Richard E. Turner, Jose Miguel Hernandez-Lobato, and Alexander L. Gaunt. Deterministic variational inference for robust bayesian neural networks. In *International Conference on Learning Representations*, 2019.

Junyu Xuan, Jie Lu, and Guangquan Zhang. Bayesian transfer learning: An overview of probabilistic graphical models for transfer learning. *arXiv preprint arXiv:2109.13233*, 2021.

Appendix

A. Further Empirical Results

A.1. Tabular Results

Dataset	Method	Accuracy(↑)	Selective Acc.(↑)	NLL(↓)	ECE(↓)	Detection AUROC(↑)
CIFAR-10	ERM	96.45±0.08	76.46±0.27	0.17±0.01	3.42±0.13	94.96±0.72
	PTYL	**97.35**±0.34	82.48±0.92	**0.11**±0.01	3.10±0.27	96.94±1.92
	Ours	**97.19**±0.08	**83.42**±0.03	**0.11**±0.01	**1.68**±0.08	**99.86**±0.04
CIFAR-100	ERM	**85.76**±0.20	71.99±1.01	0.67±0.02	8.61±0.26	86.57±1.21
	PTYL	**85.82**±0.23	72.10±1.69	0.68±0.01	8.65±0.56	88.23±2.98
	Ours	**85.69**±0.13	**77.61**±0.10	**0.62**±0.01	**7.22**±0.20	**98.93**±0.22
Flowers	ERM	89.64±0.24	78.87±0.62	0.48±0.02	11.00±0.47	95.68±1.85
	PTYL	89.73±0.51	77.88±0.34	0.47±0.01	10.99±0.64	92.30±0.57
	Ours	**90.35**±0.18	**79.68**±0.08	**0.45**±0.02	**10.07**±0.18	**97.81**±0.10
MultiNLI	ERM	**91.31**±0.10	72.18±0.42	**0.34**±0.01	**6.14**±0.11	81.15±0.42
	PTYL	**91.29**±0.11	73.17±0.23	**0.35**±0.01	**6.15**±0.09	84.27±0.35
	Ours	**91.28**±0.13	**79.71**±0.10	**0.34**±0.01	**6.08**±0.06	**96.90**±0.75
QNLI	ERM	**84.65**±0.13	75.41±0.13	**0.51**±0.01	**8.44**±0.20	94.28±0.59
	PTYL	**84.60**±0.19	75.88±0.15	**0.52**±0.02	**8.48**±0.31	94.74±0.41
	Ours	**84.64**±0.06	**76.34**±0.13	**0.51**±0.01	**8.37**±0.20	**98.17**±0.72

Table 7. **Comparison of Predictive Performance Across Datasets.** Our method consistently outperforms existing methods on all uncertainty metrics across downstream image and language tasks, and remains competitive on accuracy and negative log-likelihood. We achieved improved results for Expected Calibration Error (ECE), Selective Prediction Accuracy (Selective Acc.), and Semantic Shift Detection AUROC on all five datasets. We show the mean and standard error over five trials.

A.2. Ablation Study

To better understand the difference between using our uncertainty-aware fine-tuning prior compared to standard Gaussian prior, we show how KL divergence between the variational distributions increases between our empirical prior compared to a standard Gaussian prior as we fine-tune in Figure 6. We have the mean and the $\log$ of the variance of the variational distribution q_{ours} and the standard Gaussian prior q_{standard}. We compute the KL divergence between the two distributions as follows:

$$D_{\text{KL}}(q_{\text{ours}}||q_{\text{standard}}) = \frac{1}{2}\left(\log\frac{\sigma_{\text{standard}}}{\sigma_{\text{ours}}} + \frac{\sigma^2_{\text{ours}} + (\mu_{\text{ours}} - \mu_{\text{standard}})^2}{\sigma^2_{\text{standard}}} - 1\right)$$

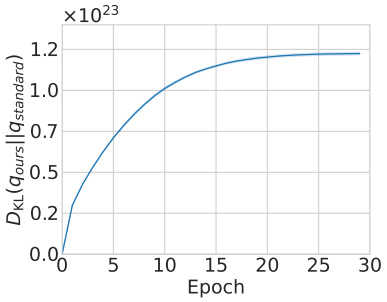


Figure 6. **Comparison of Learned Variational Distributions Under Different Priors.** The plot shows the KL divergence from the variational distribution learned with an uncertainty-aware prior to the variational distribution learned with an uninformative Gaussian prior. The variational distribution learned under the uncertainty-aware prior differs significantly from the variational distribution learned with an uninformative Gaussian prior. The plot shows the mean and the standard error of the KL divergence on CIFAR-10 with ResNet18 over five trials. We fixed the parameter initialization and the stochasticity in the data loader to ensure comparability.

A.3. Improvement Over Expected Risk Minimization

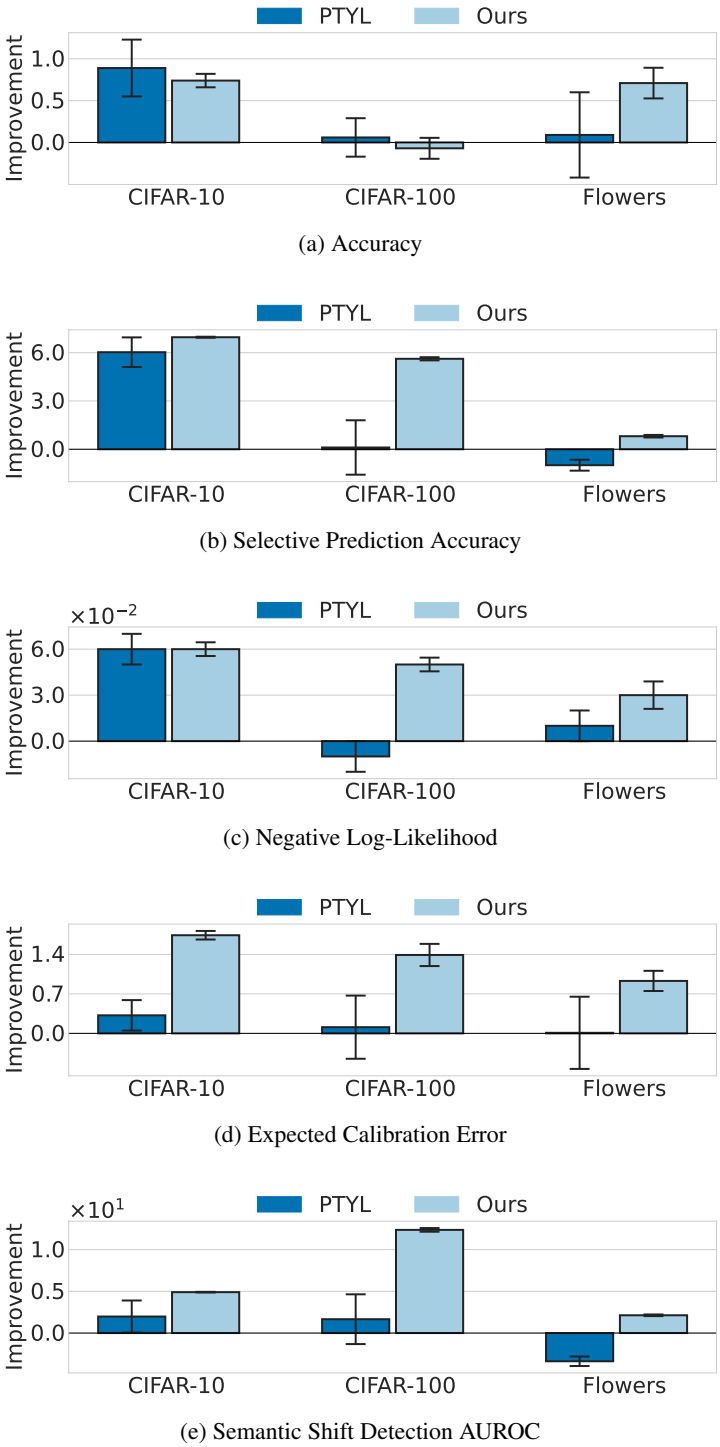


Figure 7. **On downstream image tasks, our method improves on all uncertainty metrics compared to the ERM baseline while maintaining similar or improved levels of accuracy.** For each metric, we plot the improvement over ERM, where positive values indicate the metric has changed in the preferred direction (e.g., increased accuracy or decreased ECE). Means and standard errors are calculated over five trials.

B. Experiment Details

B.1. Training Details

Image Tasks. We use the SGD optimizer with momentum of 0.9 and learning rate of 0.005 with a batch size of 128 and context batch size of 32 for methods with context dataset. We train all models for 50 epochs with a cosine annealing learning rate scheduler and use the paramters at the last epoch to evaluate the models.

Language Tasks. We use the AdamW optimizer with a learning rate of 0.0001 with a batch size of 32 and a context batch size of 8 for methods with context datasets. We train our models for three epochs with a linear learning rate scheduler and use the parameters at the last epoch to evaluate our models.

Shared Details. We use the Monte Carlo sampling with 1 sample during training and 10 samples during evaluation for all non-deterministic methods.

Sweep Protocol. We keep the same training setting for ERM training from SoTA results from (Shwartz-Ziv et al., 2022). For our method, we sweep the cov-scale from 1e-5 to 1e-3, we report the best hyperparameter as our result. We use the converge point to evaluate the models.

Computational Resources. All the experiments can be run on a single NVIDIA RTX8000 GPU, A40 GPU, or A100 GPU, with 50GB of RAM and 16 core CPU 3.4GHZ (Intel Cascade Lake Platinum 8268 chips). The training time of CIFAR-10 is around six hours for image tasks. The average running time is 294 minutes for ERM and 312 minutes for our method with a ResNet-18 architecture, and 294 minutes for ERM and 432 minutes for our method with a ResNet-50 architecture.

B.2. Bayesian Transfer Learning with "Pretrain Your Loss" (Shwartz-Ziv et al., 2022)

Bayesian transfer learning method with pre-trained priors as presented in Shwartz-Ziv et al. (2022) serves as our baseline for comparison in the subsequent experiments. This approach enables the transfer of knowledge acquired through pre-training to downstream tasks by following a three-step pipeline. First, we fit a probability distribution with a closed-form density to the posterior distribution over feature extractor parameters using a pre-trained checkpoint. Second, we adapt this distribution to reflect the discrepancies between the pre-training and downstream tasks. Finally, we employ this re-scaled prior within a Bayesian inference algorithm, accompanied by a zero-mean isotropic Gaussian prior for added parameters, such as classification heads, to effectively learn on the downstream task.

In our experiments, following the original setup in Shwartz-Ziv et al. (2022), we utilize a prior learned over the parameters of a pre-trained SimCLR ResNet-50 feature extractor trained on ImageNet (Deng et al., 2009; Chen et al., 2020a; He et al., 2016). For Bayesian inference, we adopt the SWA-Gaussian (SWAG) method (Maddox et al., 2019), known for its strong performance. SWAG involves an initial phase of exploring a basin in the loss function using SGD with a cyclic learning rate. Subsequently, a Gaussian distribution is fitted to the SGD iterates, with the covariance matrix composed of diagonal and low-rank components, including ten components. After obtaining a closed-form distribution using SWAG, we exclude the head from the feature extractor, focusing solely on the distribution's parameters related to the feature extractor. Any new layers added for downstream tasks receive a non-learned prior over their parameters.

To perform approximate Bayesian inference, we employ stochastic gradient Hamiltonian Monte Carlo (SGHMC) (Chen et al., 2014).

B.3. Dataset Details

Training Examples	Context Point Examples	Semantic Shift Examples
CIFAR-10 **Label:** Dog	**ImageNet** **Label:** Maltese (Dog)	**SVHN** **Label:** 19
CIFAR-100 **Label:** Sunflower	**ImageNet** **Label:** Maltese (Dog)	**SVHN** **Label:** 19
Flowers102 **Label:** Common Dandelion	**ImageNet** **Label:** Maltese (Dog)	**iNaturalist (Plantae)** **Label:** Plantae
MultiNLI **Premise**: He started slowly back to the bunkhouse. **Hypothesis**: He returned slowly to the bunkhouse. **Label:** Neutral	**MathQA** **Problem**: the banker ' s gain of a certain sum due 3 years hence at 10 % per annum is rs . 36 . what is the present worth ?	**Emotions** I feel so cold
QNLI **Question**: What was the port known as prior to the Swedish occupation of St. Barts? **Sentence**: Earlier to their occupation, the port was known as "Carénage". **Label:** Not Entailment	**MathQA** **Problem**: the banker ' s gain of a certain sum due 3 years hence at 10 % per annum is rs . 36 . what is the present worth ?	**MultiNLI** **Premise**: He started slowly back to the bunkhouse. **Hypothesis**: He returned slowly to the bunkhouse. **Label:** Neutral

Table 8. Representative training, context input, and semantic shift examples.