Neural Variational Dropout Processes

Insu Jeon[1], Youngin Park[2], Gunhee Kim[1]

The Tenth Internal Conference on Learning Representations (ICLR), April 25, 2022

[1]Seoul National University; [2]Everdoubling LLC., Seoul, South Korea

Background

Meta-Learning

- Humans can generalize well even from only a few learning examples.

Figure 1: Test yourself on one-shot learning. Can you find images similar to the query image?

- But conventional Neural Networks based ML approaches require a large amount of training data to learn a new task.
- **How can we implement an ML model that can generalize well on new tasks even from small learning examples?**
- In *meta-learning*, this goal is addressed by leveraging common (task invariant) structures learned from multiple related tasks.

Bayesian Meta Learning

The meta-learning objective can be formulated as a Bayesian learning framework.

- Given a collection of T tasks, $\{\mathcal{D}^t\}_{t=1}^T = \{(x_i^t, y_i^t)_{i=1}^N\}_{t=1}^T$, the evidence lower-bound (ELBO) over the multi-task dataset can be defined as:

$$\sum_{t=1}^{T} \log p(\mathcal{D}^t) \geq \sum_{t=1}^{T} \{\mathbb{E}_{q(\phi^t|\mathcal{D}^t;\theta)}[\log p(\mathbf{y}^t|\mathbf{x}^t, \phi^t)] - \mathrm{KL}(q(\phi^t|\mathcal{D}^t;\theta)||p(\phi^t))\}$$

- $p(\mathbf{y}^t|\mathbf{x}^t, \phi^t)$ is the likelihood (or NN model) on t-th training data.
- $q(\phi^t|\mathcal{D}^t;\theta)$ is a tractable conditional posterior over the task-specific variable ϕ^t for each task (i.e., latent representation or weight of NN).
- θ is the shared parameter across multiple tasks.
- $p(\phi^t)$ provides a stochastic regularization for the conditional posterior.

The goal here is to find an efficient way to approximate the conditional posterior on new unseen task $q(\phi^*|\mathcal{D}^*;\theta)$ via the learned common structure θ. Essentially, *Bayesian meta-learning* considers the efficient model adaptation as well as task-conditional model uncertainty.

Related Works

Versatile Amortized Inference (VERSA)

VERSA (Gordon et al., 2018) is a model based Bayesian meta-learning approach.

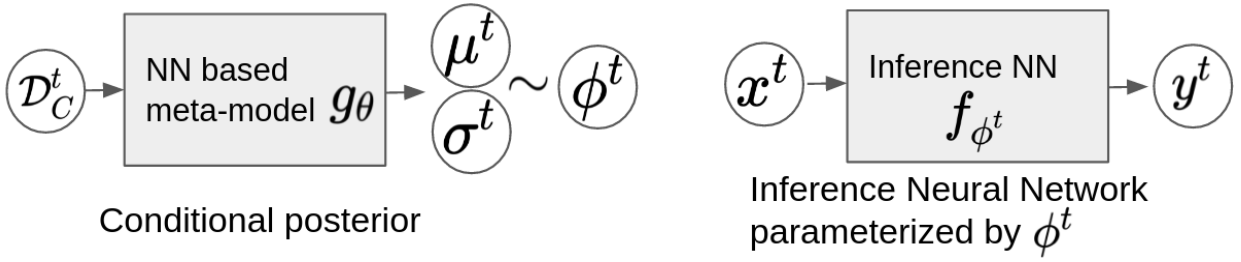

Figure 2: Conditional posterior modeling of VERSA.

- VERSA employs a NN-based meta-model $g_\theta(\cdot)$ to directly predict the Gaussian posterior over the task-specific weight ϕ^t of inference NN.

$$q(\phi^t|\mathcal{D}_C^t;\theta) = \mathcal{N}(\phi^t|(\mu^\mathbf{t},\sigma^\mathbf{t}) = g_\theta(\mathcal{D}_C^t))$$

- In this case, the shared structure θ represents the meta-model's parameters.
- **Pros**: VERSA provides an instant posterior model adaptation as well as uncertainty approximation at test time.
- **Cons**: VERSA did not consider a proper prior for stochastic regularization. The meta-model may hard to scale up due to a large number of NN parameter output.

Neural Processes (NPs)

NPs (Garnelo et al., 2018b) is another model-based meta-learning approach.

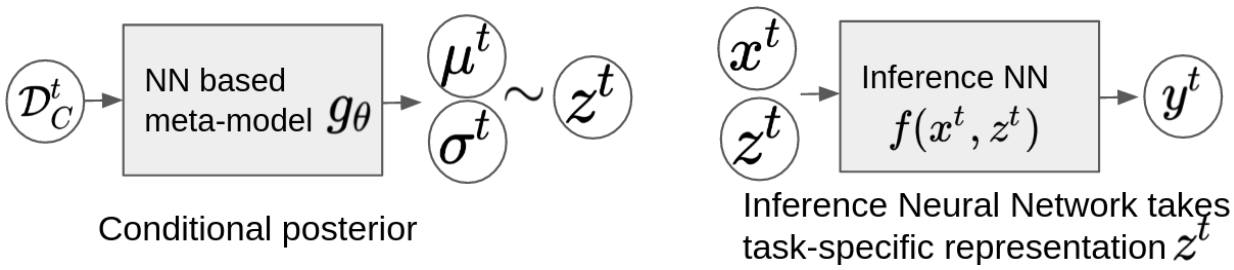

Figure 3: Conditional posterior modeling of NPs

- NPs models the conditional posterior based on a task-specific representation z^t.

$$q(z^t|\mathcal{D}^t;\theta) = \mathcal{N}(z^t|(\mu^t,\sigma^t) = g_\theta(\mathcal{D}^t))$$

- The inference model $p(\mathbf{y}^t|\mathbf{x}^t, z^t)$ takes the task-specific representation z^t as a second input, enabling task-conditional inference.
- **Pros**: We can learn an explicit latent representation that interprets the model uncertainty and distribution over the NN function.
- **Cons**: NPs tend to suffer from under-fitting (Kim et al., 2019) and posterior collapsing behavior (Grover et al. 2019).

Neural Variational Dropout Processes (NVDPs)

Challenges and Motivation

Challenges

- VERSA provides an instant conditional posterior approximation at test time, but has a restriction on scaling up due to the large inference NN's parameter.
- NPs tend to suffer from the under-fitting or posterior-collapsing problem.

Motivation

- Our motivation was to develop a new model-based Bayesian meta-learning approach which can bypass previous approaches' restrictions.

A Conditional Dropout Posterior

- A simple but novel conditional *dropout* posterior based on the task-specific dropout rates $\mathbf{P}^t$ is proposed. The task-specific NN weight ϕ^t can be fully described by the mean and variance of each independent Gaussian distribution via the $\theta_{k,d}$ and $\mathbf{P}^t_{k,d}$:

$$q(\phi^t|\mathcal{D}^t_C;\theta) = \prod_{k=1}^{K}\prod_{d=1}^{D} \mathcal{N}(\phi^t_{k,d}|(1-\mathbf{P}^t_{k,d})\theta_{k,d}, \mathbf{P}^t_{k,d}(1-\mathbf{P}^t_{k,d})\theta^2_{k,d}).$$

- In this case, the shared parameter $\theta_{k,d}$ is just conventional NN's deterministic parameter.
- The key idea here is to employ a NN-based meta-model to predict the task-specific dropout rates $\mathbf{P}^t$ from the small context set $\mathcal{D}^t_C = \{\mathbf{x}^t_i, \mathbf{y}^t_i\}_{i=1}^S$ $(\subseteq \mathcal{D}^t)$ as follows:

$$\mathbf{P}^t_{k,d} = s(\mathbf{a}_k)\cdot s(\mathbf{b}_d)\cdot s(\mathbf{c}), \text{ where } (\mathbf{a},\mathbf{b},\mathbf{c}) = g_\psi(r^t).$$

- The set representation r^t is defined as the mean of features obtained from each data in t-th context set $\mathcal{D}^t_C$ (i.e., $r^t = \sum_{i=1}^S h_\omega(\mathbf{x}^t_i, \mathbf{y}^t_i)/S$, where h_ω is a feature extracting NN parameterized by ω), summarizing the task information.

A Conditional Dropout Posterior

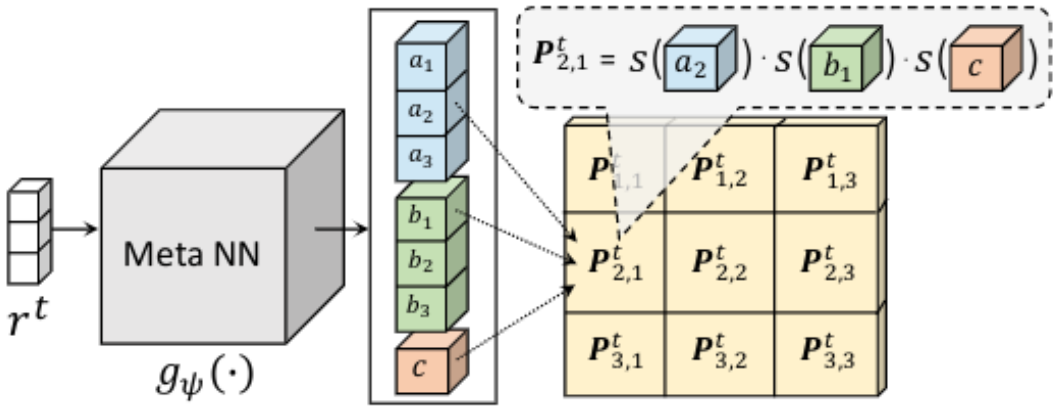

Figure 4: The low-rank product of Bernoulli experts conditional posterior modeling

- **Remark**: the meta-model's role in NVDPs is only to predict the low-rank component of task-specific dropout rates instead of whole task-specific weight ϕ^t.
- This can greatly reduce the complexity of the meta-model which need to predict the high-dimensional task-specific weights with only a few learning examples.
- In addition, the conditional posterior in NVDPs directly modifies the inference NN model through the dropout process, thus can also mitigate the under-fitting and posterior collapsing behavior.

Variational Prior

- A prior distribution in NVDPs is defined as the same dropout posterior model conditioned on the whole task data ($p(\phi^t) \approx q(\phi^t|\mathcal{D}^t)$).).
- Then, the analytical derivation of KL divergence is given as:

$$\mathrm{KL}(q(\phi^t|\mathcal{D}_C^t;\theta)||q(\phi^t|\mathcal{D}^t;\theta))$$

$$= \sum_{k=1}^{K}\sum_{d=1}^{D}\{\frac{\mathbf{P}_{k,d}^t(1-\mathbf{P}_{k,d}^t)+(\widehat{\mathbf{P}}_{k,d}^t-\mathbf{P}_{k,d}^t)^2}{2\widehat{\mathbf{P}}_{k,d}^t(1-\widehat{\mathbf{P}}_{k,d}^t)}+\frac{1}{2}\log\frac{\widehat{\mathbf{P}}_{k,d}^t(1-\widehat{\mathbf{P}}_{k,d}^t)}{\mathbf{P}_{k,d}^t(1-\mathbf{P}_{k,d}^t)}-\frac{1}{2}\}$$

- where $\mathbf{P}_{k,d}$ is obtained from the small context set $\mathcal{D}_C^t$, while $\widehat{\mathbf{P}}_{k,d}$ is from the whole task set $\mathcal{D}^t$.
- Importantly, the analytical derivation of KL does not depend on the deterministic shared parameter θ, this ensures the constant optimization of the ELBO objective w.r.t all independent parameters in the Variational Dropout framework.
- In addition, the denominator and the numerator of the KL divergence term are reversed compared to the conventional approaches such as NPs, which empirically shows better uncertainty generalization.

ELBO of NVDPs

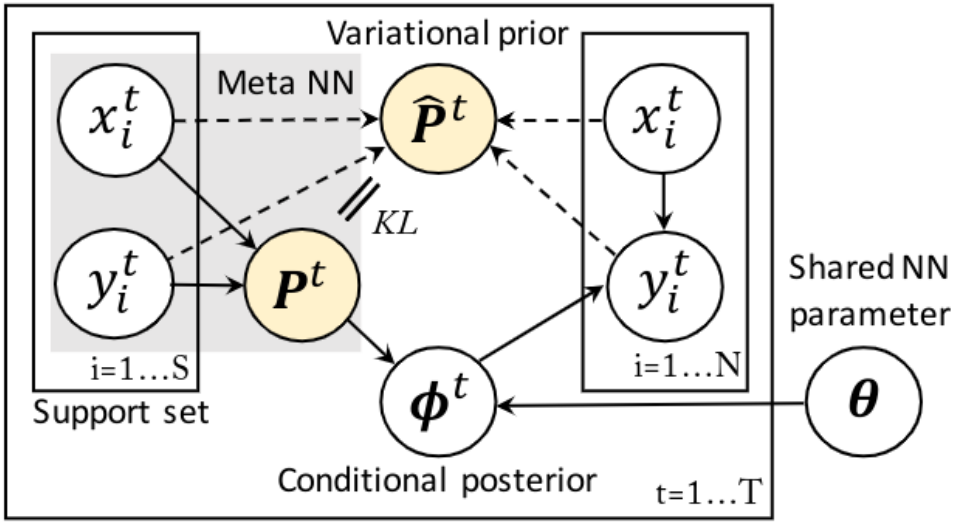

Figure 5: The probabilistic graphical model of NVDPs.

Given the aformentioned development of conditional posterior and prior, we can define the ELBO objective of NVDPs on the multiple tasks as:

$$\sum_{t=1}^{T} \log p(\mathcal{D}^t) \geq \sum_{t=1}^{T} \{\mathbb{E}_{q(\phi^t|\mathcal{D}_C^t;\theta)}[\log p(\mathbf{y}^t|\mathbf{x}^t,\phi^t)] - \mathrm{KL}(q(\phi^t|\mathcal{D}_C^t;\theta)||q(\phi^t|\mathcal{D}^t;\theta))\}.$$

The ELBO of NVDPs can be convinently optimzed by the SGD algorithm. For much detailed stochastic variational inference (SVI) optimzation please refer the manuscript.

Experiment

Tasks and Metrics

Tasks

- We tested NVDPs on various few-shot learning tasks such as 1D regression, image in-painting, and few-shot classification.

Metrics

- In the evaluation of few-shot learning task, the newly observed data $\mathcal{D}^*$ is split into the *context set* $\mathcal{D}_C^* = \{x_i, y_i\}_{i=1}^S$ and the *target set* $\mathcal{D}_T^* = \{x_i, y_i\}_{i=1}^N$ ($\mathcal{D}_C^* \not\subseteq \mathcal{D}_T^*$).

1. The log-likelihood (LL), $\frac{1}{N+S} \sum_{i \in \mathcal{D}_C^* \cup \mathcal{D}_T^*} \mathbb{E}_{q(\phi^*|\mathcal{D}_C^*)}[\log p(y_i|x_i, \phi^*)]$ over the whole task $\mathcal{D}^*$ measures the overall performance of the model.
2. The reconstructive log-likelihood (RLL) on the observed *context set*, $\frac{1}{S} \sum_{i \in \mathcal{D}_C^*} \mathbb{E}_{q(\phi^*|\mathcal{D}_C^*)}[\log p(y_i|x_i, \phi^*)]$, measures how well the model reconstructs the observed data points. A low RLL is a sign of under-fitting.
3. The predictive log-likelihood (PLL) on the unobserved *target set*, $\frac{1}{N} \sum_{i \in \mathcal{D}_T^*} \mathbb{E}_{q(\phi^*|\mathcal{D}_C^*)}[\log p(y_i|x_i, \phi^*)]$, measures the predictive performance outside of the observed dataset. A low PLL is a sign of over-fitting.

GP dataset

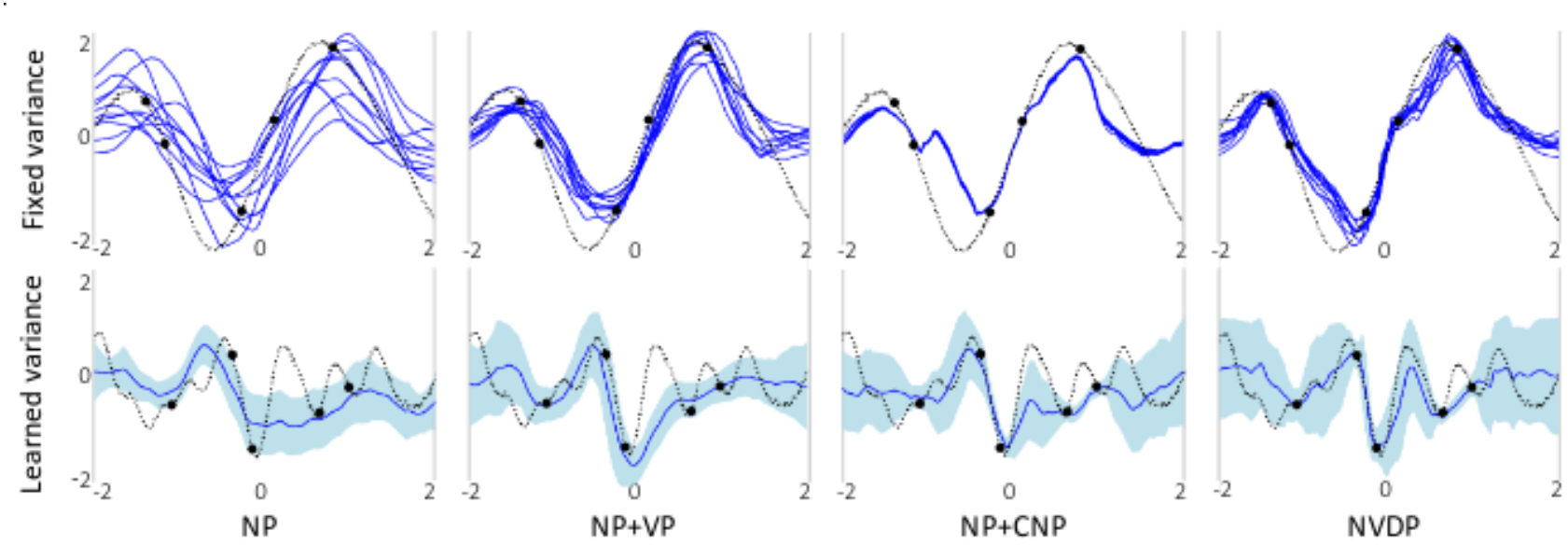

Figure 6: The 1D few-shot regression results of the models on GP dataset in fixed variance and learned variance settings. The black (dash-line) represent the true unknown task function. Black dots are a few context points ($S = 5$) given to the posteriors. The blue lines (and light blue area in learned variance settings) are mean values (and variance) predicted from the sampled NNs.

GP dataset

GP Dataset		NP	NP+VP	NP+CNP	NP+CNP+VP	NVDP
Fixed	LL	$-0.98(\pm0.08)$	$-0.96(\pm0.06)$	$-0.95(\pm0.05)$	$-0.94(\pm0.05)$	$\mathbf{-0.94(\pm0.04)}$
Variance	RLL	$-0.97(\pm0.06)$	$-0.95(\pm0.04)$	$-0.93(\pm0.02)$	$-0.93(\pm0.02)$	$\mathbf{-0.93(\pm0.01)}$
	PLL	$-0.98(\pm0.08)$	$-0.97(\pm0.06)$	$-0.95(\pm0.05)$	$-0.94(\pm0.05)$	$\mathbf{-0.94(\pm0.05)}$
Learned	LL	$0.19(\pm1.87)$	$0.48(\pm0.77)$	$0.70(\pm0.89)$	$0.70(\pm0.73)$	$\mathbf{0.83(\pm0.61)}$
Variance	RLL	$0.72(\pm0.57)$	$0.71(\pm0.55)$	$1.03(\pm0.39)$	$1.00(\pm0.41)$	$\mathbf{1.10(\pm0.34)}$
	PLL	$0.16(\pm1.90)$	$0.46(\pm0.78)$	$0.66(\pm0.92)$	$0.68(\pm0.75)$	$\mathbf{0.81(\pm0.62)}$

Figure 7: The validation result of the 1D regression models on the GP dataset. The higher scores are the better. The models of NP, NP with variational prior (NP+VP), NP with deterministic path (NP+CNP), NP+CNP+VP and NVDP (ours) are compared.

Image completion tasks

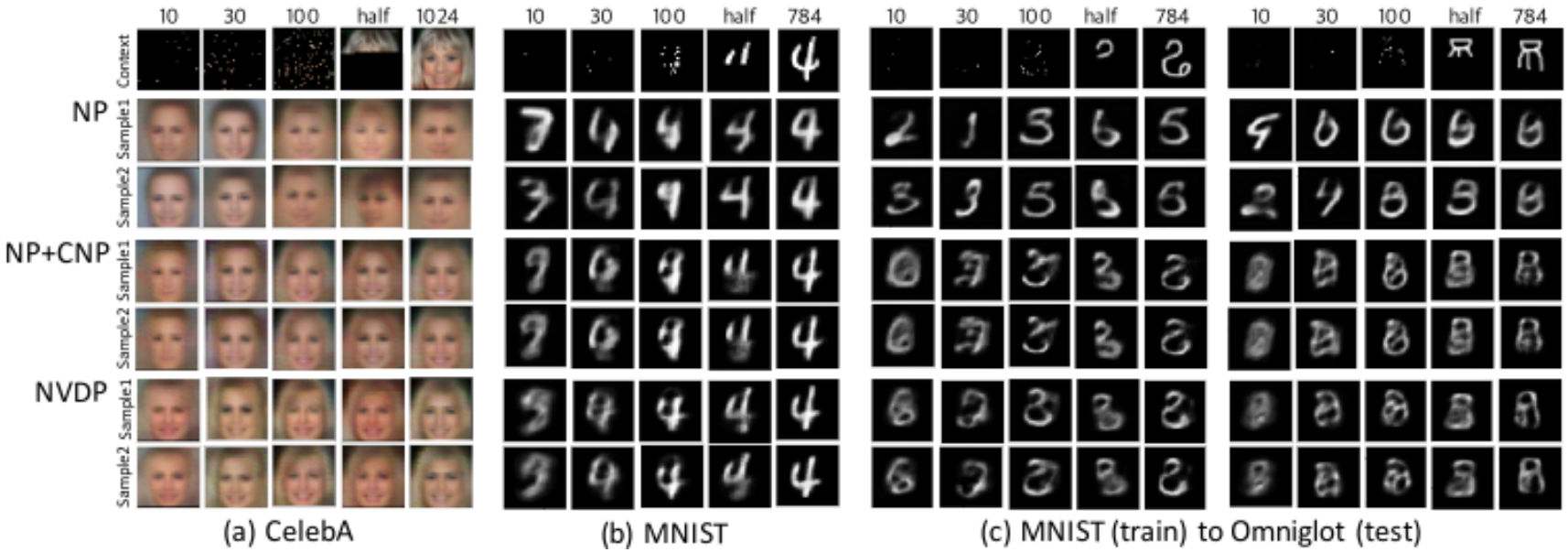

Figure 8: The results from the 2D image completion tasks on CelebA, MNIST, and Omniglot dataset. Given the observed context points (10, 30, 100, half, and full pixels), the mean values of two independently sampled functions are presented.

Image completion tasks

Image Dataset		NP	NP+VP	NP+CNP	NP+CNP+VP	NVDP
MNIST	LL	$0.54(\pm0.51)$	$0.76(\pm0.14)$	$0.83(\pm0.21)$	$0.86(\pm0.16)$	$\mathbf{0.90(\pm0.16)}$
	RLL	$0.94(\pm0.18)$	$0.90(\pm0.10)$	$1.12(\pm0.09)$	$1.12(\pm0.09)$	$\mathbf{1.15(\pm0.05)}$
	PLL	$0.51(\pm0.51)$	$0.75(\pm0.14)$	$0.80(\pm0.20)$	$0.83(\pm0.15)$	$\mathbf{0.88(\pm0.15)}$
MNIST (train)	LL	$0.35(\pm0.29)$	$0.56(\pm0.10)$	$0.64(\pm0.17)$	$0.68(\pm0.13)$	$\mathbf{0.70(\pm0.13)}$
to	RLL	$0.72(\pm0.19)$	$0.73(\pm0.12)$	$0.95(\pm0.09)$	$0.98(\pm0.09)$	$\mathbf{0.99(\pm0.07)}$
Omniglot (test)	PLL	$0.32(\pm0.28)$	$0.54(\pm0.11)$	$0.60(\pm0.16)$	$0.64(\pm0.13)$	$\mathbf{0.66(\pm0.12)}$
CelebA	LL	$0.51(\pm0.27)$	$0.60(\pm0.12)$	$0.76(\pm0.15)$	$0.77(\pm0.15)$	$\mathbf{0.83(\pm0.15)}$
	RLL	$0.63(\pm0.11)$	$0.68(\pm0.06)$	$0.91(\pm0.05)$	$0.92(\pm0.05)$	$\mathbf{0.99(\pm0.04)}$
	PLL	$0.50(\pm0.27)$	$0.59(\pm0.12)$	$0.75(\pm0.15)$	$0.76(\pm0.15)$	$\mathbf{0.82(\pm0.15)}$

Figure 9: The summary of 2D image completion tasks on the MNIST, CelebA, and Omniglot dataset.

Few-shot Classification tasks

Method	*Omniglot* dataset 5-way accuracy (%)		*Omniglot* dataset 20-way accuracy (%)		*MiniImageNet* dataset 5-way accuracy (%)	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Matching Nets	98.1	98.9	93.8	98.5	46.6	60.0
Prototypical Nets	97.4	99.3	95.4	98.7	46.61 ± 0.78	65.77 ± 0.70
CNP	95.3	98.5	89.9	96.8	48.05 ± 2.85	62.71 ± 0.58
Meta-SGD (MSGD)	-	-	96.16 ± 0.14	98.54 ± 0.07	48.30 ± 0.64	65.55 ± 0.56
MSGD + Meta-dropout	-	-	97.02 ± 0.13	$\mathbf{99.05 \pm 0.05}$	50.87 ± 0.63	65.55 ± 0.57
MAML	98.70 ± 0.40	$\mathbf{99.90 \pm 0.10}$	95.80 ± 0.63	98.90 ± 0.20	48.70 ± 1.84	63.11 ± 0.92
VERSA	99.70 ± 0.20	99.75 ± 0.13	97.66 ± 0.29	98.77 ± 0.18	53.40 ± 1.82	67.37 ± 0.86
NVDP	$\mathbf{99.70 \pm 0.12}$	99.86 ± 0.28	$\mathbf{97.98 \pm 0.22}$	98.99 ± 0.22	$\mathbf{54.06 \pm 1.86}$	$\mathbf{68.12 \pm 1.04}$

Figure 10: Few-shot classification results on Omniglot and MiniImageNet dataset. The baselines are Matching Nets, Prototypical Nets, MAML, Meta-SGD, Meta-dropout, CNP, VERSA, and NVDP (our). Each value corresponds to the classification accuracy (%) (and *std*) on validation set.

Experiment on ranodom trigonometry function

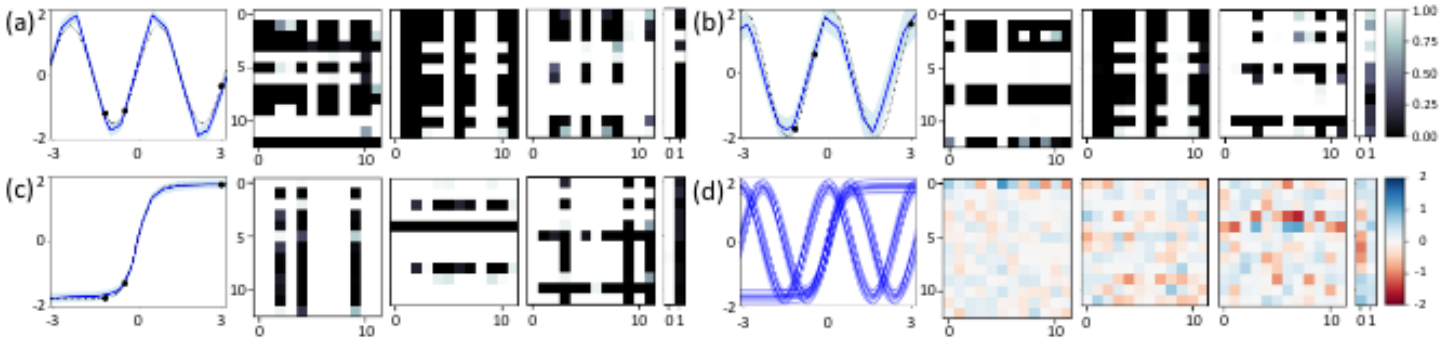

Figure 11: (a) sine, (b) cosine, (c) tanh function with the probabilities of using parameters (1-$\mathbf{P}^t$) predicted with small NVDPs conditioned on 4-shot context points (black dots), and (d) the trigonometry dataset (left) and the deterministic shared NN parameters θ (right). Dropout can modulate the NN parameters.

Conclusion

- NVDPs extend the *dropout* posterior of variational dropout (VD) as a conditional posterior in the context of the Bayesian meta-learning framework.
- NVDPs also introduce a new concept of *variational prior* that can be universally applied to other Bayesian learning approaches.
- NVDPs can greatly resolve some limitations of previous model-based Bayesian learning approaches such as under-fitting and posterior collapsing and achieved state-of-the-art performances on various few-shot learning tasks.

Thank you!

insuj3on@gmail.com