Meta-Learning with Shared Amortized Variational Inference

Ekaterina Iakovleva — Inria

Jakob Verbeek — Facebook

Karteek Alahari — Inria

UGA Université Grenoble Alpes

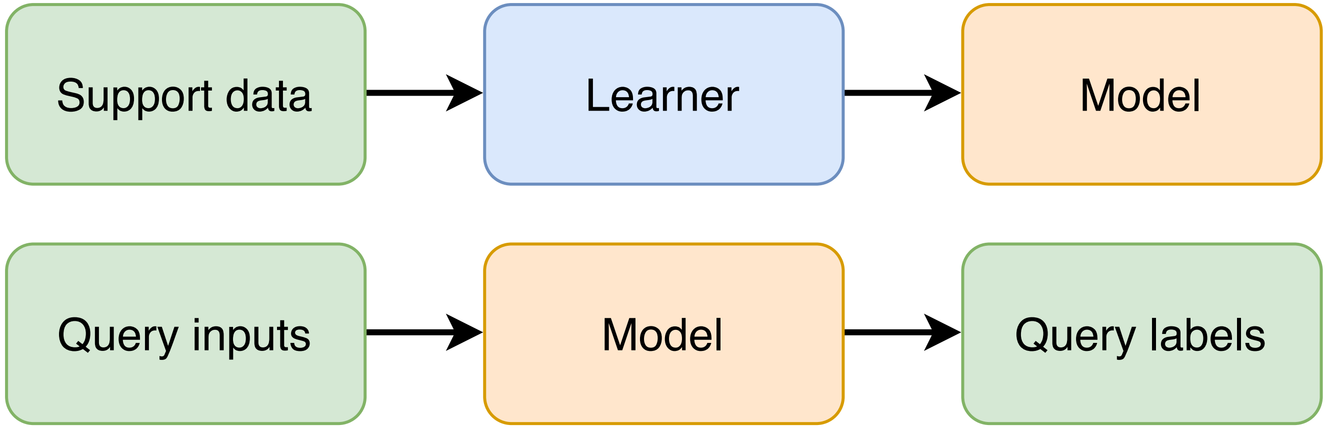
Inria

FACEBOOK AI

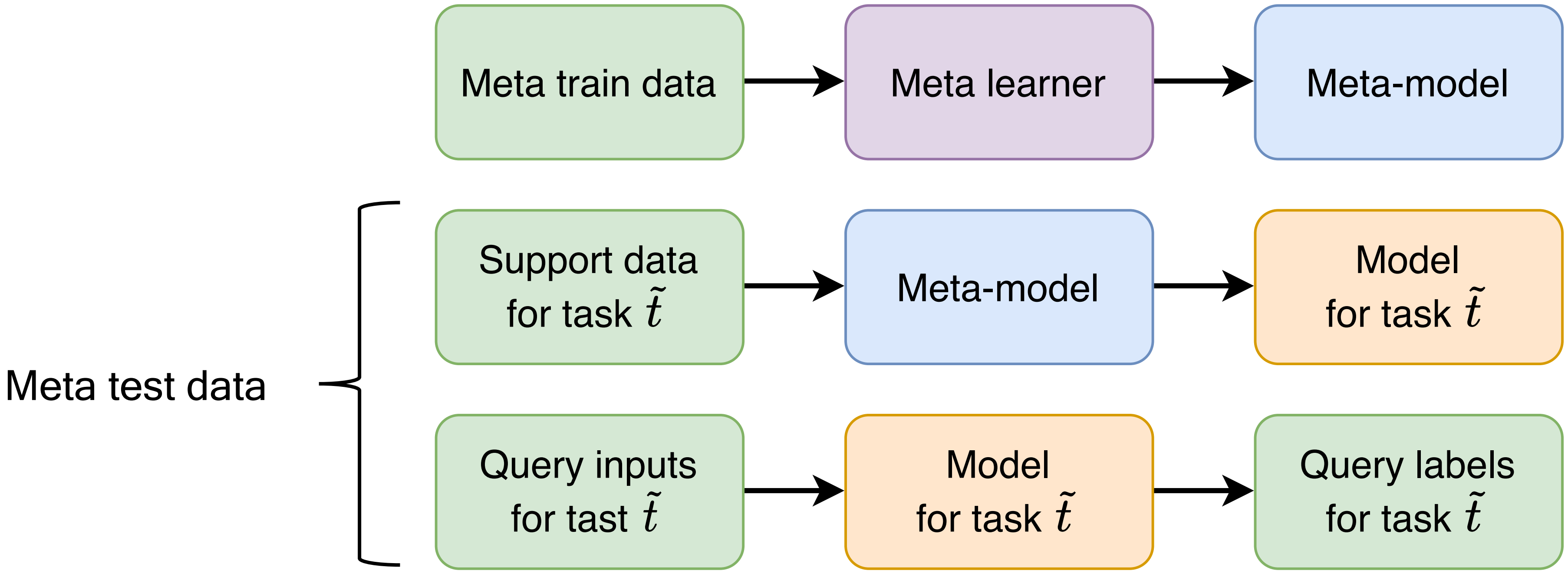
ICML | 2020

Thirty-seventh International Conference on Machine Learning

Standard classification task pipeline



Meta-learning classification task pipeline



Schmidhuber 1999, Ravi & Larochelle ICLR'17

Overview

- This work focuses on the empirical Bayes meta-learning approach.
- We propose a novel scheme for amortized variational inference.
- We demonstrate that earlier work based on Monte-Carlo approximation underestimates model variance.
- We show the advantage of our approach on miniImageNet and FC100.

Meta-learning classification task definition

- K - shot N - way classification task
- Episodic training: each task t is sampled from a distribution over tasks $p(T)$
- Support data $D^t = \{(x_{k,n}^t, y_{k,n}^t)\}_{k,n=1}^{K,N}$
- Query data $\widetilde{D}^t = \{(\tilde{x}_{m,n}^t, \tilde{y}_{m,n}^t)\}_{m,n=1}^{M,N}$

Meta-learning approaches

- **Distance-based classifiers**
  - Learned metric relies on the distance to individual samples or class prototypes.
  - E.g. Prototypical Networks [1], Matching Nets [2].

[1] – Snell et al. NeurIPS’17, [2] – Vinyals et al. NeurIPS’16

Meta-learning approaches

- **Distance-based classifiers**
  - ❖ Learned metric relies on the distance to individual samples or class prototypes.
  - ❖ E.g. Prototypical Networks [1], Matching Nets [2].

- **Optimization-based approaches**
  - ❖ Vanilla SGD approach is replaced by a trainable update mechanism.
  - ❖ E.g. MAML [3], Meta LSTM [4].

[1] – Snell et al. NeurIPS’17, [2] – Vinyals et al. NeurIPS’16, [3] – Finn et al. ICML’17, [4] – Ravi & Larochelle ICLR’17

Meta-learning approaches

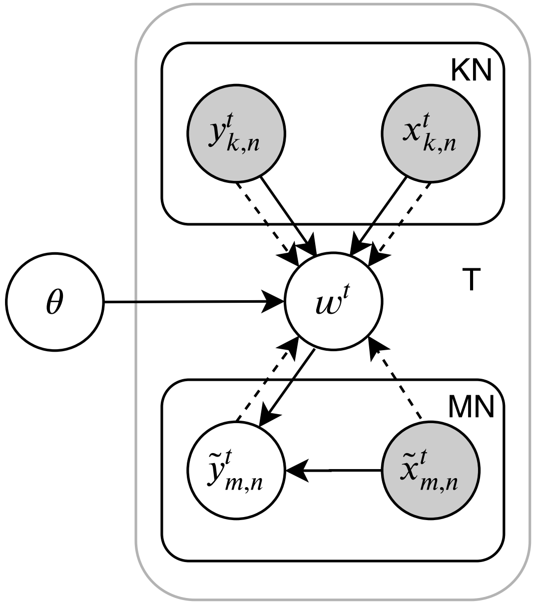
- **Distance-based classifiers**
  - Learned metric relies on the distance to individual samples or class prototypes.
  - E.g. Prototypical Networks [1], Matching Nets [2].

- **Optimization-based approaches**
  - Vanilla SGD approach is replaced by a trainable update mechanism.
  - E.g. MAML [3], Meta LSTM [4].

- **Latent variable models**
  - The model parameters are treated as latent variables.
  - Their variance is explicitly modeled in a Bayesian framework.
  - E.g. Neural Processes [5], VERSA [6].

[1] – Snell et al. NeurIPS'17, [2] – Vinyals et al. NeurIPS'16, [3] – Finn et al. ICML'17, [4] – Ravi & Larochelle ICLR'17, [5] – Garnelo et al. ICML'18, [6] – Gordon et al. ICLR'19

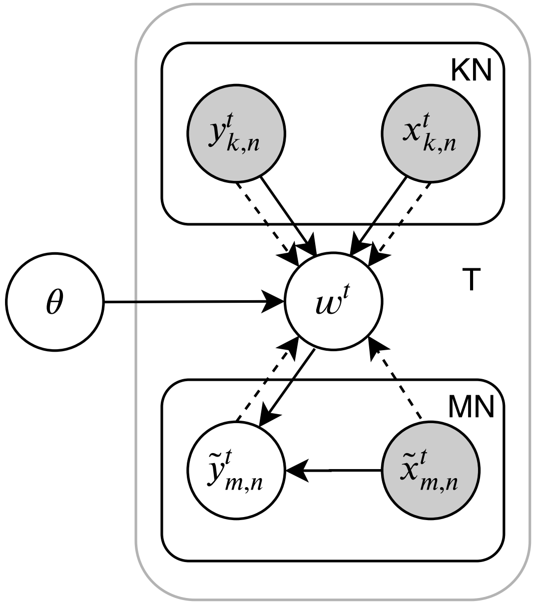
Multi-task generative model



The multi-task graphical model includes:

- task-agnostic parameters θ
- task-specific latent parameters $\{w^t\}_{t=1}^T$

Multi-task generative model



The multi-task graphical model includes:

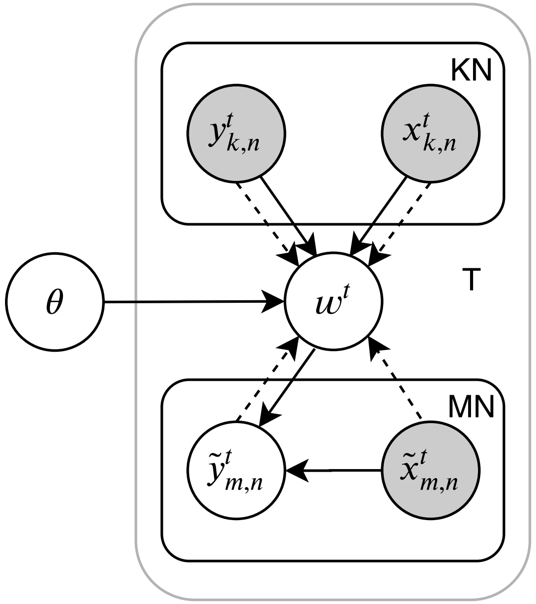
- task-agnostic parameters θ
- task-specific latent parameters $\{w^t\}_{t=1}^T$

Marginal likelihood of the query labels $\tilde{Y} = \{\tilde{Y}^t\}_{t=1}^T$ given query samples $\tilde{X} = \{\tilde{X}^t\}_{t=1}^T$ and the support sets $D = \{D^t\}_{t=1}^T = \{(X^t, Y^t)\}_t^T$

$$p\left(\tilde{Y} | \tilde{X}, D, \theta\right) = \prod_{t=1}^{T} \int p\left(\tilde{Y} | \tilde{X}, w^t\right) p_\phi(w^t | D^t, \theta)\, dw^t$$

Intractable integral requires approximation for training and prediction.

Multi-task generative model



The multi-task graphical model includes:

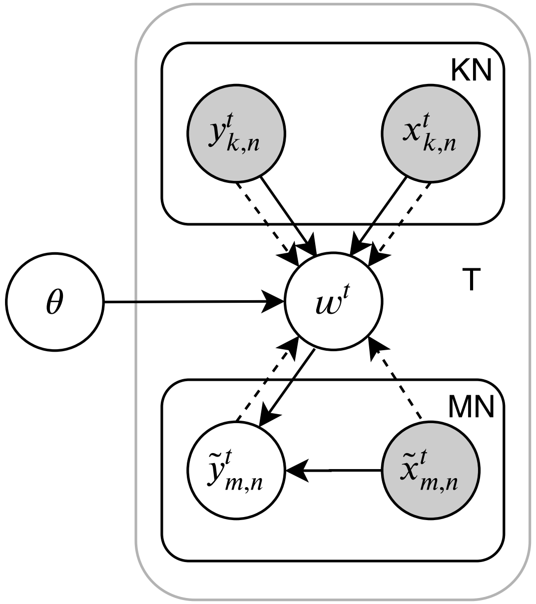
- task-agnostic parameters θ
- task-specific latent parameters $\{w^t\}_{t=1}^T$

Marginal likelihood of the query labels $\tilde{Y} = \{\tilde{Y}^t\}_{t=1}^T$ given query samples $\tilde{X} = \{\tilde{X}^t\}_{t=1}^T$ and the support sets $D = \{D^t\}_{t=1}^T = \{(X^t, Y^t)\}_t^T$

$$p(\tilde{Y}|\tilde{X}, D, \theta) = \prod_{t=1}^{T} \int p(\tilde{Y}|\tilde{X}, w^t)\, p_\phi(w^t|D^t, \theta)\, dw^t$$

Intractable integral requires approximation for training and prediction.

Multi-task generative model



The multi-task graphical model includes:

- task-agnostic parameters θ
- task-specific latent parameters $\{w^t\}_{t=1}^T$

Marginal likelihood of the query labels $\tilde{Y} = \{\tilde{Y}^t\}_{t=1}^T$ given query samples $\tilde{X} = \{\tilde{X}^t\}_{t=1}^T$ and the support sets $D = \{D^t\}_{t=1}^T = \{(X^t, Y^t)\}_t^T$

$$p(\tilde{Y}|\tilde{X}, D, \theta) = \prod_{t=1}^{T} \int p(\tilde{Y}|\tilde{X}, w^t)\, p_\phi(w^t|D^t, \theta)\, dw^t$$

Intractable integral requires approximation for training and prediction.

Monte Carlo approximation

- Monte Carlo approximation of the marginal log-likelihood using $w_l^t \sim p_\phi(w^t|D^t, \theta)$:

$$\log p\left(\tilde{Y}^t \middle| \tilde{X}^t, D^t, \theta\right) \approx \frac{1}{TM} \sum_{t=1}^{T} \sum_{m=1}^{M} \log \frac{1}{L} \sum_{l=1}^{L} p(\tilde{y}_m^t | \tilde{x}_m^t, w_l^t).$$

- This objective function has been used in VERSA [1].

- Our experiments show that this approach learns degenerate prior $p_\phi(w^t|D^t, \theta)$.

[1] – Gordon et al. ICLR'19

Amortized variational inference

- Variational evidence lower bound (ELBO) with the amortized approximate posterior [1] parameterized by ψ:

$$\log p\left(\tilde{Y}^t|\tilde{X}^t, D^t, \theta\right) \geq \mathbb{E}_{q_\psi}\left[\log p\left(\tilde{Y}^t\middle|\tilde{X}^t, w^t\right)\right] - \mathcal{D}_{KL}\left(q_\psi\left(w^t\middle|\tilde{Y}^t, \tilde{X}^t, D^t, \theta\right)||p_\phi(w^t|D^t, \theta)\right)$$

[1] – Kingma & Welling ICLR'14

Amortized variational inference

- Variational evidence lower bound (ELBO) with the amortized approximate posterior [1] parameterized by ψ:

$$\log p(\tilde{Y}^t|\tilde{X}^t, D^t, \theta) \geq \underbrace{\mathbb{E}_{q_\psi}\left[\log p(\tilde{Y}^t|\tilde{X}^t, w^t)\right]}_{\text{Reconstruction loss}} - \mathcal{D}_{KL}\left(q_\psi(w^t|\tilde{Y}^t, \tilde{X}^t, D^t, \theta)||p_\phi(w^t|D^t, \theta)\right)$$

[1] – Kingma & Welling ICLR'14

Amortized variational inference

- Variational evidence lower bound (ELBO) with the amortized approximate posterior [1] parameterized by ψ:

$$\log p(\tilde{Y}^t | \tilde{X}^t, D^t, \theta) \geq \mathbb{E}_{q_\psi}[\log p(\tilde{Y}^t | \tilde{X}^t, w^t)] - \underbrace{\mathcal{D}_{KL}\left(q_\psi(w^t | \tilde{Y}^t, \tilde{X}^t, D^t, \theta) || p_\phi(w^t | D^t, \theta)\right)}_{\text{Regularization}}$$

[1] – Kingma & Welling ICLR'14

Amortized variational inference

- Variational evidence lower bound (ELBO) with the amortized approximate posterior [1] parameterized by ψ:

$$\log p(\tilde{Y}^t|\tilde{X}^t, D^t, \theta) \geq \mathbb{E}_{q_\psi}[\log p(\tilde{Y}^t|\tilde{X}^t, w^t)] - \beta \mathcal{D}_{KL}\left(q_\psi(w^t|\tilde{Y}^t, \tilde{X}^t, D^t, \theta)||p_\phi(w^t|D^t, \theta)\right)$$

- We use regularization coefficient β [2] to weight KL term.

[1] – Kingma & Welling ICLR'14, [2] – Higgins et al. ICLR'17

Amortized variational inference

- Variational evidence lower bound (ELBO) with the amortized approximate posterior [1] parameterized by ψ:

$$\log p(\tilde{Y}^t|\tilde{X}^t, D^t, \theta) \geq \mathbb{E}_{q_\psi}[\log p(\tilde{Y}^t|\tilde{X}^t, w^t)] - \beta \mathcal{D}_{KL}\left(q_\psi(w^t|\tilde{Y}^t, \tilde{X}^t, D^t, \theta)||p_\phi(w^t|D^t, \theta)\right)$$

- We use regularization coefficient β [2] to weight KL term.

- Predictions are made via Monte Carlo sampling from the learned prior:

$$p(\tilde{y}_m^t|\tilde{x}_m^t, D^t, \theta) \approx \frac{1}{L}\sum_{l=1}^{L} p(\tilde{y}_m^t|\tilde{x}_m^t, w_l^t), \qquad \text{where} \quad w_l^t \sim p_\phi(w^t|D^t, \theta).$$

[1] – Kingma & Welling ICLR'14, [2] – Higgins et al. ICLR'17

Shared amortized variational inference: SAMOVAR

- Both prior and posterior are conditioned on labeled sets.

$$\log p\left(\tilde{Y}^t|\tilde{X}^t, D^t, \theta\right) \geq \mathbb{E}_{q_\phi}\left[\log p\left(\tilde{Y}^t\middle|\tilde{X}^t, w^t\right)\right] - \beta \mathcal{D}_{KL}\left(q_\psi\left(w^t\middle|\tilde{Y}^t, \tilde{X}^t, D^t, \theta\right)||p_\phi(w^t|D^t, \theta)\right)$$

Shared amortized variational inference: SAMOVAR

- Both prior and posterior are conditioned on labeled sets.

- The inference network can be shared between prior and posterior.

$$\log p(\tilde{Y}^t|\tilde{X}^t, D^t, \theta) \geq \mathbb{E}_{q_\phi}[\log p(\tilde{Y}^t|\tilde{X}^t, w^t)] - \beta \mathcal{D}_{KL}\left(q_\phi(w^t|\tilde{Y}^t, \tilde{X}^t, D^t, \theta)||p_\phi(w^t|D^t, \theta)\right)$$

Shared amortized variational inference: SAMOVAR

- Both prior and posterior are conditioned on labeled sets.

- The inference network can be shared between prior and posterior.

$$\log p\left(\tilde{Y}^t|\tilde{X}^t, D^t, \theta\right) \geq \mathbb{E}_{q_\phi}\left[\log p\left(\tilde{Y}^t\middle|\tilde{X}^t, w^t\right)\right] - \beta \mathcal{D}_{KL}\left(q_\phi\left(w^t\middle|\tilde{Y}^t, \tilde{X}^t, D^t, \theta\right) || p_\phi\left(w^t|D^t, \theta\right)\right)$$

- Sharing reduces memory footprint, and encourages learning non-degenerate prior.

SAMOVAR design based on VERSA

- **Task-agnostic feature extractor** f_θ produces embeddings of the input images x.

- **Task-specific linear classifier** w^t predicts labels for query samples $\tilde{x}$:

$$p(\tilde{y}_m^t | \tilde{x}_m^t, w^t) = \mathrm{softmax}\big(w^t f_\theta(\tilde{x}_m^t)\big).$$

- **Shared amortized inference network** g_ϕ returns the parameters $\{\mu_n^t, \sigma_n^t\}$ of a Gaussian over weight vector w_n^t for each class n:

$$p(w_n^t | D^t, \theta) = \mathcal{N}(\mu_n^t, \mathrm{diag}(\sigma_n^t)), \qquad \text{where } \begin{pmatrix} \mu_n^t \\ \sigma_n^t \end{pmatrix} = g_\phi\left(\frac{1}{K}\sum_{k=1}^{K} f_\theta(x_{k,n}^t)\right)$$

VERSA – Gordon et al. ICLR’19

Improved architectural design based on TADAM

- **Scaled cosine similarity (-SC).**

The linear classifier is replaced with the cosine similarity classifier scaled with α.

- **Task encoding network (-TEN).**

TEN provides task-conditioned batch norm parameters for feature maps in f_θ.

- **Auxiliary co-training (-AT).**

f_θ is shared with an auxiliary classification task across all meta-train classes.

TADAM – Oreshkin et al. NeurIPS’18

Experiments: synthetic task

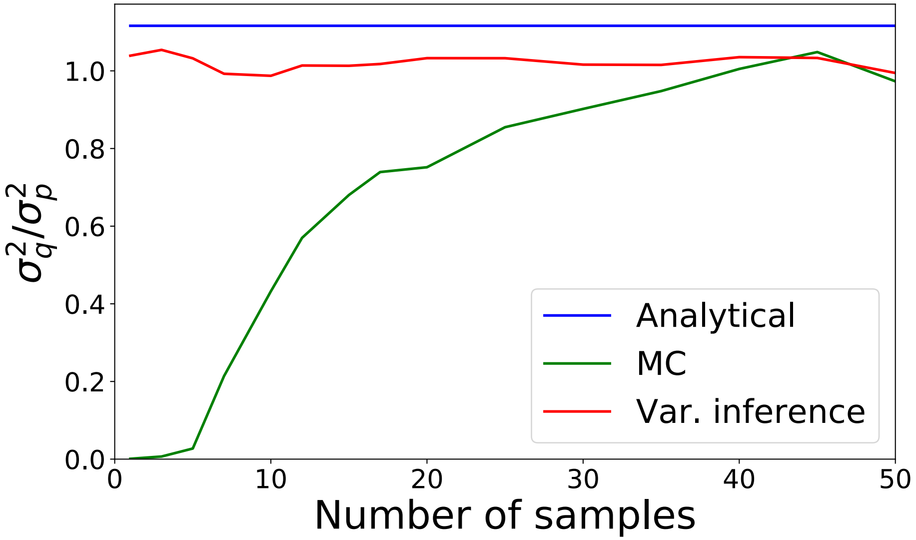
- Hierarchical generative process: $p(w^t) = \mathcal{N}(0, 1)$, and $p(y^t|w^t) = \mathcal{N}(w^t, \sigma_y^2)$.
- $T = 250$ sampled tasks, $K = 5$ support observations $D^t = \{y_k^t\}_{k=1}^K$ and $M = 15$ query observations $\widetilde{D}^t = \{\tilde{y}_m^t\}_{k=1}^M$.
- Posterior over latent variable $q_\phi(w^t|\, D^t) = \mathcal{N}(\mu_q, \sigma_q^2)$.
- Parameters of the posterior are obtained via inference network with parameters ϕ:

$$\begin{pmatrix} \mu_q \\ \log \sigma_q^2 \end{pmatrix} = \phi_1 \sum_{k=1}^{K} y_k^t + \phi_2$$

Results: synthetic task

- Exact marginal log-likelihood $\log p(\widetilde{D}^t|D^t)$.
- Monte Carlo estimation of $\log p(\widetilde{D}^t|D^t)$ with L samples from the prior.
- Variational inference for $\log p(\widetilde{D}^t|D^t)$ with L samples from the posteror.

estimated / true



(b) $\sigma_y = 0.5$

- Monte Carlo requires large sample sets compared to variational inference.

Experimental setup for real data

- 5-shot and 1-shot, 5-way classification tasks.
- Test data contains 15 query samples per class.
- Evaluation is performed on 5,000 randomly sampled tasks.
- We report the mean accuracy over these tasks, and 95% confidence intervals.

Comparison with VERSA on miniImageNet

- SAMOVAR-base and VERSA train the same meta-learning model.
- SAMOVAR with separate prior and posterior is inferior to other models.
- SAMOVAR is comparable with VERSA on 1-shot task, and outperforms it on 5-shot task.

	5-SHOT	1-SHOT
VERSA (OUR IMPLEM.)	67.97 ± 0.23	52.45 ± 0.30
SAMOVAR-BASE	69.86 ± 0.23	52.36 ± 0.29
SAMOVAR-BASE (SEPARATE)	66.60 ± 0.23	50.80 ± 0.29

VERSA – Gordon et al. ICLR'19

Comparison with VERSA on miniImageNet

- SAMOVAR-base and VERSA train the same meta-learning model.
- SAMOVAR with separate prior and posterior is inferior to other models.
- SAMOVAR is comparable with VERSA on 1-shot task, and outperforms it on 5-shot task.

	5-SHOT	1-SHOT
VERSA (OUR IMPLEM.)	67.97 ± 0.23	52.45 ± 0.30
SAMOVAR-BASE	69.86 ± 0.23	52.36 ± 0.29
SAMOVAR-BASE (SEPARATE)	66.60 ± 0.23	50.80 ± 0.29

VERSA – Gordon et al. ICLR'19

Comparison with VERSA on miniImageNet

- SAMOVAR-base and VERSA train the same meta-learning model.
- SAMOVAR with separate prior and posterior is inferior to other models.
- SAMOVAR is comparable with VERSA on 1-shot task, and outperforms it on 5-shot task.

	5-SHOT	1-SHOT
VERSA (OUR IMPLEM.)	67.97 ± 0.23	52.45 ± 0.30
SAMOVAR-BASE	69.86 ± 0.23	52.36 ± 0.29
SAMOVAR-BASE (SEPARATE)	66.60 ± 0.23	50.80 ± 0.29

VERSA – Gordon et al. ICLR'19

Comparison with TADAM on miniImageNet

- Both models are trained with auxiliary co-training.
- SAMOVAR consistently improves TADAM across all ablations.

α: cosine scaling, AT: auxiliary co-training, TEN: task embedding network

			5-shot		1-shot	
α	AT	TEN	TADAM	SAMOVAR	TADAM	SAMOVAR
			74.5 ± 0.2	75.2 ± 0.2	56.5 ± 0.4	59.2 ± 0.3
✓	✓		75.6 ± 0.4	77.4 ± 0.2	58.0 ± 0.3	60.4 ± 0.3
✓	✓	✓	76.7 ± 0.3	77.9 ± 0.2	58.5 ± 0.3	60.8 ± 0.3

Additional ablations can be found in the paper.

TADAM – Oreshkin et al. NeurIPS'18

Comparison with TADAM on miniImageNet

- Both models are trained with auxiliary co-training.
- SAMOVAR consistently improves TADAM across all ablations.

α: cosine scaling, AT: auxiliary co-training, TEN: task embedding network

			5-SHOT		1-SHOT	
α	AT	TEN	TADAM	SAMOVAR	TADAM	SAMOVAR
			74.5 ± 0.2	75.2 ± 0.2	56.5 ± 0.4	59.2 ± 0.3
✓	✓		75.6 ± 0.4	77.4 ± 0.2	58.0 ± 0.3	60.4 ± 0.3
✓	✓	✓	76.7 ± 0.3	77.9 ± 0.2	58.5 ± 0.3	60.8 ± 0.3

Additional ablations can be found in the paper.

TADAM – Oreshkin et al. NeurIPS'18

Comparison with state of the art on miniImageNet

- SAMOVAR demonstrates competitive results with and without data augmentation.
- SAMOVAR is complementary to approaches like CTM [Li et al. CVPR'19] or [Gidaris et al. ICCV'19]

†: Transductive methods.

Method	Features	5-shot	1-shot
without data augmentation			
MTL (Sun et al., 2019)	ResNet-12	75.50 ± 0.80	61.20 ± 1.80
TADAM (Oreshkin et al., 2018)	ResNet-12	76.70 ± 0.30	58.50 ± 0.30
SAMOVAR-SC-AT-TEN (ours)	**ResNet-12**	$\mathbf{77.89 \pm 0.23}$	$\mathbf{60.76 \pm 0.29}$
with data augmentation			
MetaOptNet-SVM (Lee et al., 2019)	ResNet-12	78.63 ± 0.46	62.64 ± 0.61
SIB (Hu et al., 2020)	WRN-28-10†	79.20 ± 0.40	70.00 ± 0.60
SAMOVAR-SC-AT-TEN (ours)	**ResNet-12**	$\mathbf{79.85 \pm 0.20}$	$\mathbf{62.33 \pm 0.28}$
(Gidaris et al., 2019)	WRN-28-10	79.87 ± 0.33	62.93 ± 0.45
CTM (Li et al., 2019)	ResNet-18†	80.51 ± 0.13	64.12 ± 0.82
(Dvornik et al., 2019)	WRN-28-10	80.63 ± 0.42	63.06 ± 0.61

Comparison with state of the art on miniImageNet

- SAMOVAR demonstrates competitive results with and without data augmentation.
- SAMOVAR is complementary to approaches like CTM [Li et al. CVPR'19] or [Gidaris et al. ICCV'19]

†: Transductive methods.

Method	Features	5-shot	1-shot
Without data augmentation			
MTL (Sun et al., 2019)	ResNet-12	75.50 ± 0.80	61.20 ± 1.80
TADAM (Oreshkin et al., 2018)	ResNet-12	76.70 ± 0.30	58.50 ± 0.30
SAMOVAR-SC-AT-TEN (ours)	**ResNet-12**	$\mathbf{77.89} \pm \mathbf{0.23}$	$\mathbf{60.76} \pm \mathbf{0.29}$
With data augmentation			
MetaOptNet-SVM (Lee et al., 2019)	ResNet-12	78.63 ± 0.46	62.64 ± 0.61
SIB (Hu et al., 2020)	WRN-28-10†	79.20 ± 0.40	70.00 ± 0.60
SAMOVAR-SC-AT-TEN (ours)	**ResNet-12**	$\mathbf{79.85} \pm \mathbf{0.20}$	$\mathbf{62.33} \pm \mathbf{0.28}$
(Gidaris et al., 2019)	WRN-28-10	79.87 ± 0.33	62.93 ± 0.45
CTM (Li et al., 2019)	ResNet-18†	80.51 ± 0.13	64.12 ± 0.82
(Dvornik et al., 2019)	WRN-28-10	80.63 ± 0.42	63.06 ± 0.61

Comparison with state of the art on miniImageNet

- SAMOVAR demonstrates competitive results with and without data augmentation.
- SAMOVAR is complementary to approaches like CTM [Li et al. CVPR'19] or [Gidaris et al. ICCV'19]

†: Transductive methods.

Method	Features	5-shot	1-shot
Without data augmentation			
MTL (Sun et al., 2019)	ResNet-12	75.50 ± 0.80	61.20 ± 1.80
TADAM (Oreshkin et al., 2018)	ResNet-12	76.70 ± 0.30	58.50 ± 0.30
SAMOVAR-SC-AT-TEN (ours)	**ResNet-12**	$\mathbf{77.89 \pm 0.23}$	$\mathbf{60.76 \pm 0.29}$
With data augmentation			
MetaOptNet-SVM (Lee et al., 2019)	ResNet-12	78.63 ± 0.46	62.64 ± 0.61
SIB (Hu et al., 2020)	WRN-28-10†	79.20 ± 0.40	70.00 ± 0.60
SAMOVAR-SC-AT-TEN (ours)	**ResNet-12**	$\mathbf{79.85 \pm 0.20}$	$\mathbf{62.33 \pm 0.28}$
(Gidaris et al., 2019)	WRN-28-10	79.87 ± 0.33	62.93 ± 0.45
CTM (Li et al., 2019)	ResNet-18†	80.51 ± 0.13	64.12 ± 0.82
(Dvornik et al., 2019)	WRN-28-10	80.63 ± 0.42	63.06 ± 0.61

Summary

- Monte Carlo approximation underestimatea the variance in model parameters.
- We propose SAMOVAR, a meta-learning model based on shared amortized variational inference.
- Task on synthetic data shows that VI approach preserves stochasticity.
- SAMOVAR combined with TADAM shows competitive results on miniImageNet, FC100.

Thank you!

Université Grenoble Alpes

FACEBOOK AI

ICML | 2020
Thirty-seventh International Conference on Machine Learning