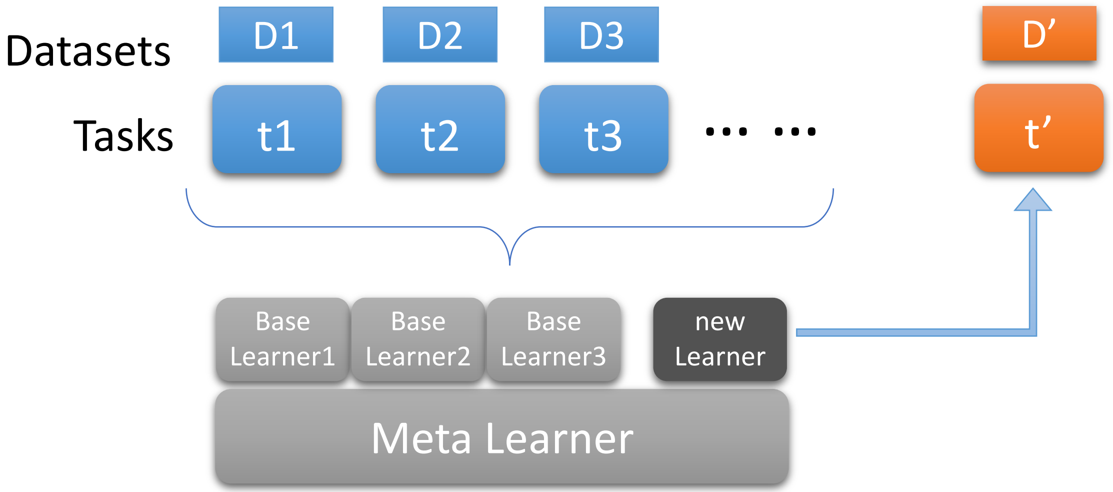
Learning to Learn Kernels with Variational Random Features

Presenter : Haoliang Sun

Xiantong Zhen*, Haoliang Sun*, Yingjun Du*, Jun Xu, Yilong Yin, Ling Shao, Cees Snoek

Meta-Learning (Leaning to Learn)



Meta-Learning.

- Extract prior (meta) knowledge from related tasks (meta learner)
- Fast adaptation to a new task (base learner)

Meta Knowledge：

- Good parameter initialization (Finn et al., 2017)
- Efficient optimization update rules (Ravi et al., 2017)
- General feature extractors (Vinyals et al., 2016)

…

Few-Shot Learning (FSL) with Meta-Learning (ML)

- The episodic training-testing strategy

  -- **meta-training**: a meta-learner is trained to enhance base-learners' performance on the meta-training set with a batch of few-shot learning tasks

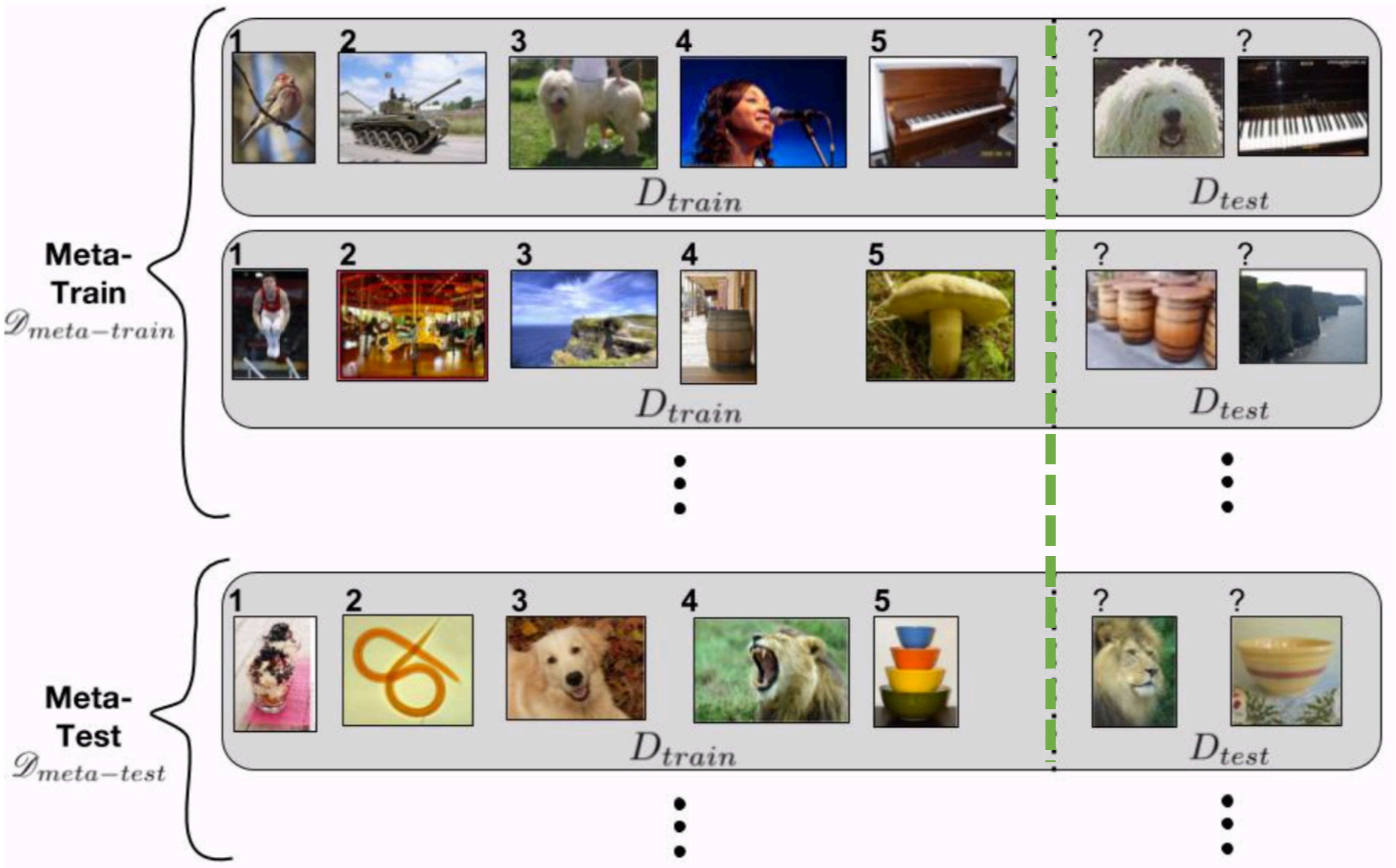
  -- **meta-testing**: base-learners are evaluated on the meta-test set with novel categories of data

- An episode (task)

  -- sample C-way k-shot classification tasks from the meta-training (testing) set

  -- k is the number of labelled examples for each of the C classes

Few-Shot Learning (FSL) with Meta-Learning (ML)



Episode 1

Episode 2

⋮

Example of few-shot learning setup (Ravi et al., 2017)

An Effective Meta-Learning Scenario

- Base-learner:
  - -- be powerful to solve individual tasks
  - -- be able to absorb common information
- Meta-learner:
  - -- extract valid prior knowledge

Key idea：

- integrate kernel learning with random features and variational inference (VI) into the ML framework for FSL
- formulate the optimization as a VI problem by deriving new ELBO
- a context inference puts the inference of random bases of the current task into the context of all previous, related tasks

Problem Statement

Meta-learning with kernels

$$\sum_{t}^{T} \sum_{(\tilde{\mathbf{x}},\tilde{\mathbf{y}})\in\mathcal{Q}^t} L\left(f_{\alpha^t}\left(\Phi^t(\tilde{\mathbf{x}})\right), \tilde{\mathbf{y}}\right), \text{s.t.}\ \alpha^t = \Lambda\left(\Phi^t(X), Y\right)$$

For task t, support set $\mathcal{S}^t = \{X, Y\}$, query set $\mathcal{Q}^t$, predictor f_{α^t},

base-learner Λ, loss L, mapping function Φ, $\mathrm{k}^t(\mathbf{x}, \mathbf{x}') = \langle \Phi^t(\mathbf{x}), \Phi^t(\mathbf{x}') \rangle$.

A practical base-learner (Kernel ridge regression)

$$\Lambda = \arg\min_{\alpha} \mathrm{Tr}[(Y - \alpha K)(Y - \alpha K)^\top] + \lambda\, \mathrm{Tr}[\alpha K \alpha^\top]$$

The closed-form solution $\alpha = Y(\lambda \mathrm{I} + K)^{-1}$. The predictor $\hat{Y} = f_\alpha(\tilde{X}) = \alpha \tilde{K}$.

Learning adaptive kernels $\mathrm{k}(\cdot)$ with data-driven random Fourier features

Problem Statement

Random Fourier Features (RFFs)

- learn adaptive kernels in a data-driven way
- leverage the shared knowledge by exploring dependencies among related tasks to generate rich features
- construct approximate translation-invariant kernels using explicit feature maps via random bases (Bochner's theorem)

Data-driven adaptive kernels is to find the posterior $p(\omega|\mathbf{y}, \mathbf{x}, \mathcal{S})$ for random bases ω

Formulated as a variational inference problem

Meta Variational Random Features (MetaVRF)

The objective function

- The posterior is intractable. Approximate it by using a meta variational distribution

$$D_{\mathrm{KL}}[\underbrace{q_\phi(\omega|\mathcal{S})}_{\text{Variational distribution}}||p(\omega|\mathbf{y},\mathbf{x},\mathcal{S})]$$

- The Evidence Lower Bound (ELBO)

$$\log p(\mathbf{y}|\mathbf{x},\mathcal{S}) \geq \underbrace{\mathbb{E}_{q_\phi(\omega|\mathcal{S})} \log p(\mathbf{y}|\mathbf{x},\mathcal{S},\omega) - D_{\mathrm{KL}}[q_\phi(\omega|\mathcal{S})||p(\omega|\mathbf{x},\mathcal{S})]}_{\text{ELBO}}$$

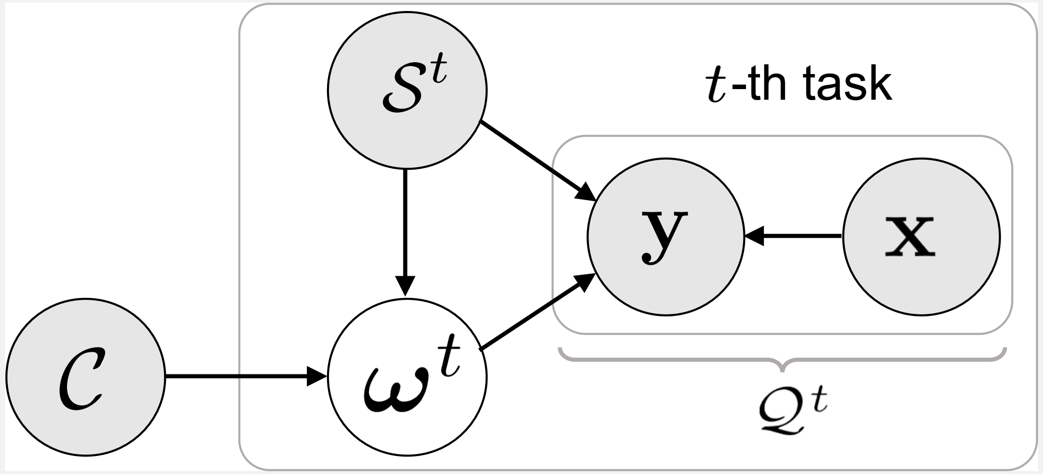
- The objective (maximizing ELBO *w.r.t.* T tasks)

$$\frac{1}{T}\sum_{t=1}^{T}\Big(\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{Q}^t}\mathbb{E}_{q_\phi(\omega^t|\mathcal{S}^t)} \log p(\mathbf{y}|\mathbf{x},\mathcal{S}^t,\omega^t) - D_{\mathrm{KL}}[q_\phi(\omega^t|\mathcal{S}^t)||p(\omega^t|\mathbf{x},\mathcal{S}^t)]\Big)$$

Context Inference

- generate rich random bases to build strong kernels
- put the inference of bases ω of the current task into the context of all previous, related tasks
- The context $\mathcal{C}$ of related tasks

$$q_\phi(\omega^t|\mathcal{S}^t) \longrightarrow q_\phi(\omega^t|\mathcal{S}^t, \mathcal{C})$$



The directed graphical model.

An LSTM-Based Context Inference Network

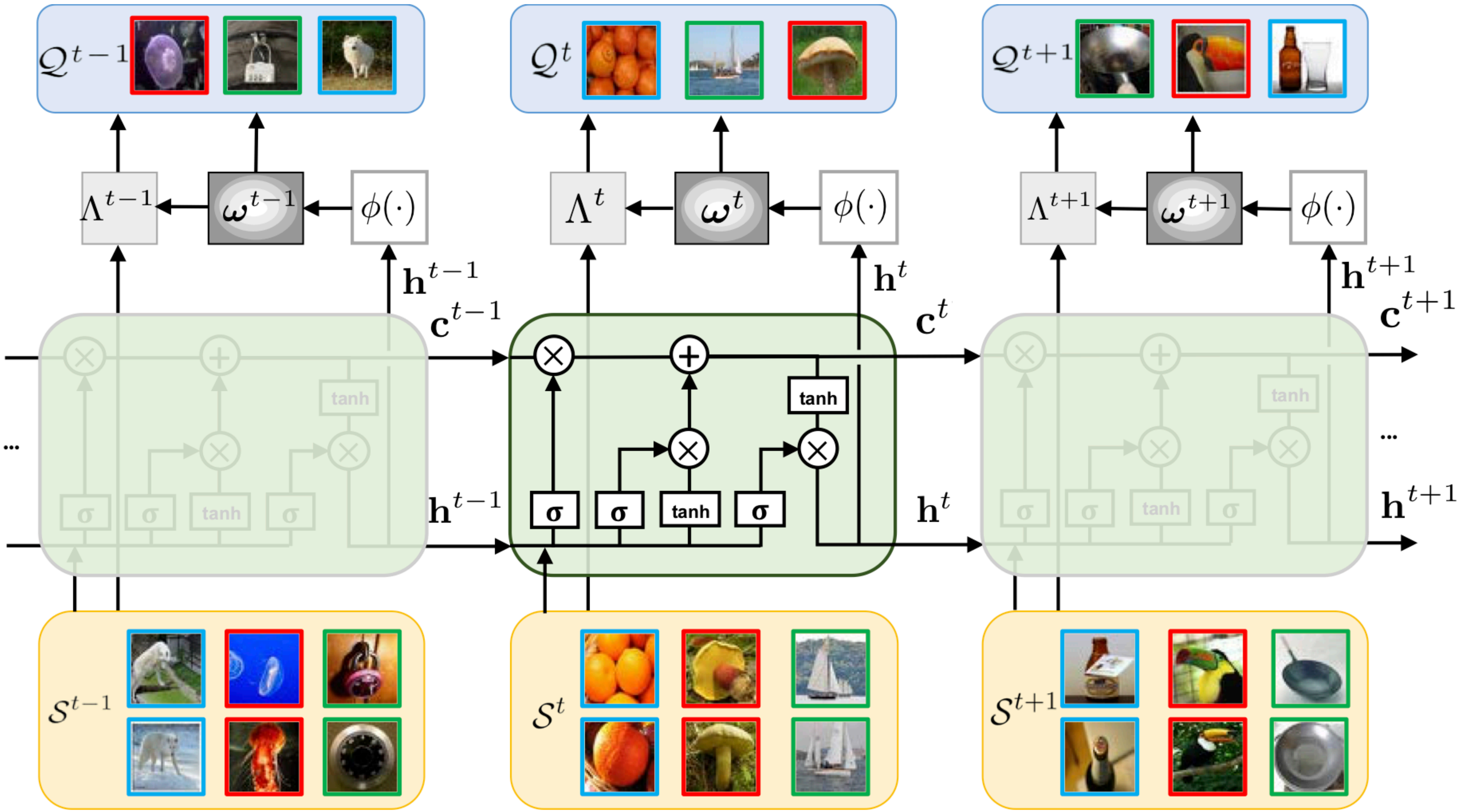
- LSTM transformation with input of the support set and previous cell states

$$[\mathbf{h}^t, \mathbf{c}^t] = g_{\text{LSTM}}(\bar{\mathcal{S}}^t, \mathbf{h}^{t-1}, \mathbf{c}^{t-1})$$

- shared MLPs for inference $\phi(\mathbf{h}^t)$ outputs the parameter of the variational distribution

- The optimization objective with the context inference

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T}\Big(\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{Q}^t}\mathbb{E}_{q_\phi(\omega^t|\mathbf{h}^t)}\log\, p(\mathbf{y}|\mathbf{x},\mathcal{S}^t,\omega^t) - D_{\text{KL}}[q_\phi(\omega^t|\mathbf{h}^t)||p(\omega^t|\mathbf{x},\mathcal{S}^t)]\Big)$$



$\mathcal{Q}^{t-1}$
$\mathcal{Q}^{t}$
$\mathcal{Q}^{t+1}$
Λ^{t-1}
ω^{t-1}
$\phi(\cdot)$
Λ^{t}
ω^{t}
Λ^{t+1}
ω^{t+1}
$\mathbf{h}^{t-1}$
$\mathbf{c}^{t-1}$
$\mathbf{h}^{t}$
$\mathbf{c}^{t}$
$\mathbf{h}^{t+1}$
$\mathbf{c}^{t+1}$
tanh
σ
$\mathcal{S}^{t-1}$
$\mathcal{S}^{t}$
$\mathcal{S}^{t+1}$

Experiments

- Few-Shot Regression

  -- Fitting a target sine function

- Few-Shot Classification

  -- Three benchmarks

- Further analysis

  -- Deep embedding

  -- Efficiency

  -- Versatility

Evaluation: Few-Shot Regression

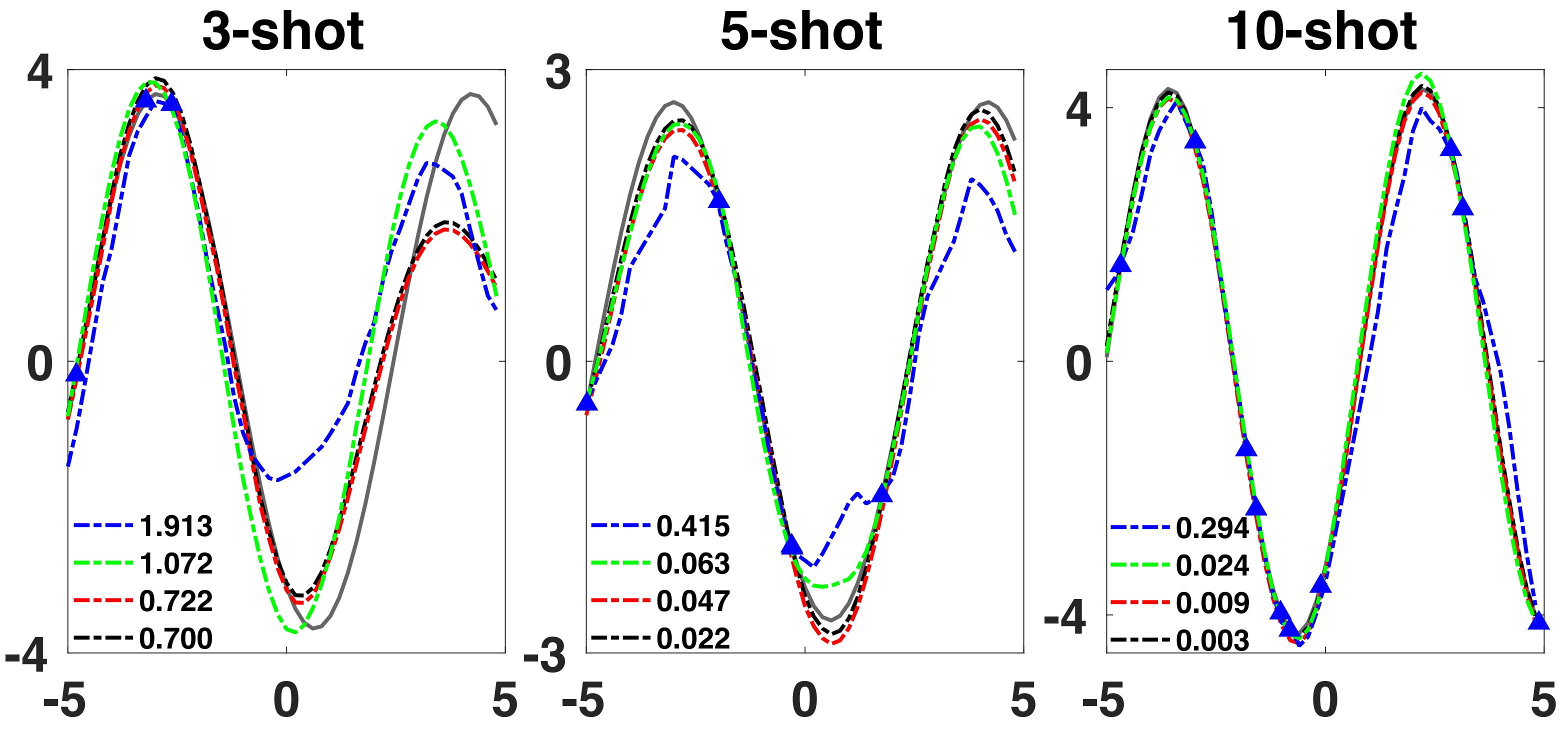


Figure 1: Performance (MSE) comparison for few-shot regression. Our MetaVRF fits the target function well, even with only three shots, and consistently outperforms regular RFFs and the counterpart MAML. (- - - MetaVRF with bi-LSTM; - - - MetaVRF with LSTM; - - - MetaVRF w/o LSTM; - - - MAML; —— Ground Truth; ▲ Support Samples.)

Evaluation: Few-Shot Classification

Table 1. Performance (%) on *mini*ImageNet and CIFAR-FS.

Method	*mini*ImageNet, 5-way		CIFAR-FS, 5-way	
	1-shot	5-shot	1-shot	5-shot
MATCHING NET (Vinyals et al., 2016)	44.2	57	—	—
MAML (Finn et al., 2017)	48.7±1.8	63.1±0.9	58.9±1.9	71.5±1.0
MAML (64C)	46.7±1.7	61.1±0.1	58.9±1.8	71.5±1.1
META-LSTM (Ravi & Larochelle, 2017)	43.4±0.8	60.6±0.7	—	—
PROTO NET (Snell et al., 2017)	47.4±0.6	65.4±0.5	55.5±0.7	72.0±0.6
RELATION NET (Sung et al., 2018)	50.4±0.8	65.3±0.7	55.0±1.0	69.3±0.8
SNAIL (32C) by (Bertinetto et al., 2019)	45.1	55.2	—	—
GNN (Garcia & Bruna, 2018)	50.3	66.4	61.9	75.3
PLATIPUS (Finn et al., 2018)	50.1±1.9	—	—	—
VERSA (Gordon et al., 2019)	53.3±1.8	67.3±0.9	62.5±1.7	75.1±0.9
R2-D2 (64C) (Bertinetto et al., 2019)	49.5±0.2	65.4±0.2	62.3±0.2	**77.4**±0.2
R2-D2 (Devos et al., 2019)	51.7±1.8	63.3±0.9	60.2±1.8	70.9±0.9
CAVIA (Zintgraf et al., 2019)	51.8±0.7	65.6±0.6	—	—
IMAML (Aravind Rajeswaran, 2019)	49.3±1.9	—	—	—
RFFS (2048d)	52.8±0.9	65.4±0.9	61.1±0.8	74.7±0.9
METAVRF (w/o LSTM, 780d)	51.3±0.8	66.1±0.7	61.1±0.7	74.3 ±0.9
METAVRF (vanilla LSTM, 780d)	53.1±0.9	66.8±0.7	62.1±0.8	76.0±0.8
METAVRF (bi-LSTM, 780d)	**54.2**±0.8	**67.8**±0.7	**63.1**±0.7	76.5±0.9

Evaluation: Few-Shot Classification

Table 2. Performance (%) on Omniglot.

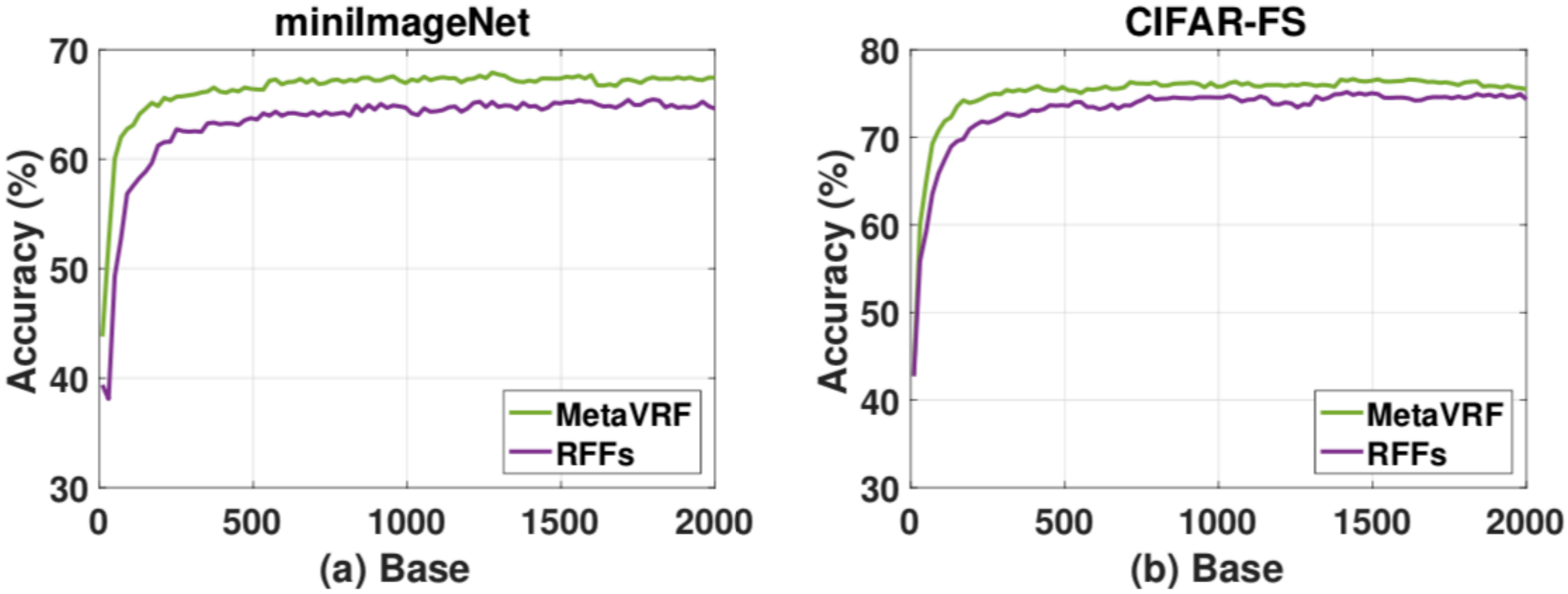
Method	Omniglot, 5-way 1-shot	Omniglot, 5-way 5-shot	Omniglot, 20-way 1-shot	Omniglot, 20-way 5-shot
SIAMESE NET (Koch, 2015)	96.7	98.4	88	96.5
MATCHING NET (Vinyals et al., 2016)	98.1	98.9	93.8	98.5
MAML (Finn et al., 2017)	98.7±0.4	**99.9**±0.1	95.8±0.3	98.9±0.2
PROTO NET (Snell et al., 2017)	98.5±0.2	99.5±0.1	95.3±0.2	98.7±0.1
SNAIL (Mishra et al., 2018)	99.1±0.2	99.8 ±0.1	97.6 ±0.3	**99.4** ±0.2
GNN (Garcia & Bruna, 2018)	99.2	99.7	97.4	99.0
VERSA (Gordon et al., 2019)	99.7±0.2	99.8±0.1	97.7±0.3	98.8±0.2
R2-D2 (Bertinetto et al., 2019)	98.6	99.7	94.7	98.9
IMP (Allen et al., 2019)	98.4±0.3	99.5±0.1	95.0±0.1	98.6±0.1
RFFS (2048d)	99.5±0.2	99.5±0.2	97.2±0.3	98.3±0.2
METAVRF (w/o LSTM, 780d)	99.6±0.2	99.6±0.2	97.0±0.3	98.4±0.2
METAVRF (vanilla LSTM, 780d)	99.7±0.2	99.8±0.1	97.5±0.3	99.0±0.2
METAVRF (bi-LSTM, 780d)	**99.8**±0.1	**99.9**±0.1	**97.8**±0.3	99.2±0.2

Further Analysis

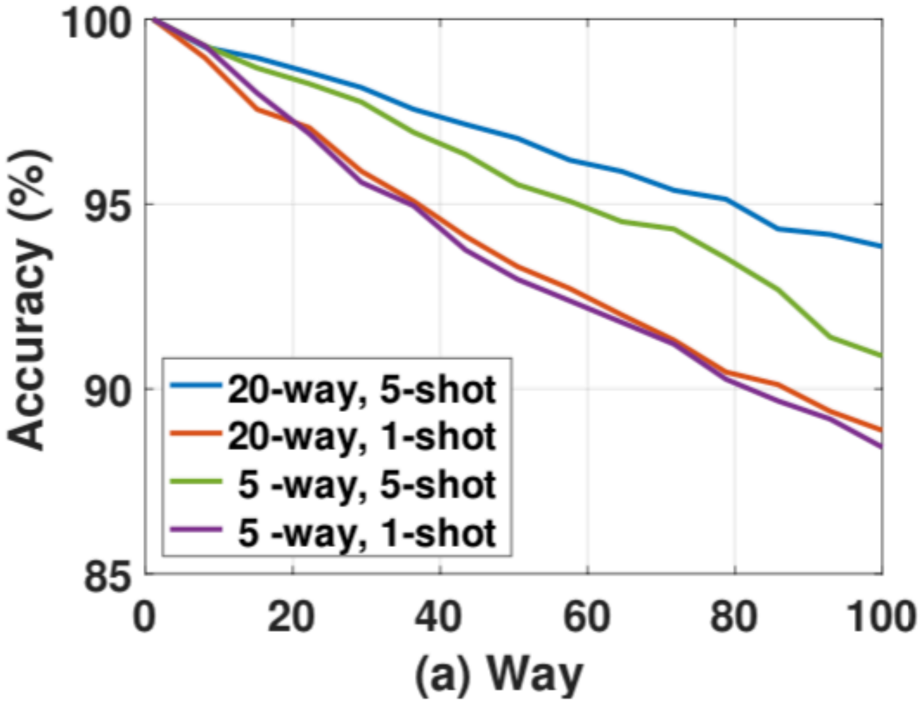
Table 3. Performance (%) on *mini*ImageNet (5-way)

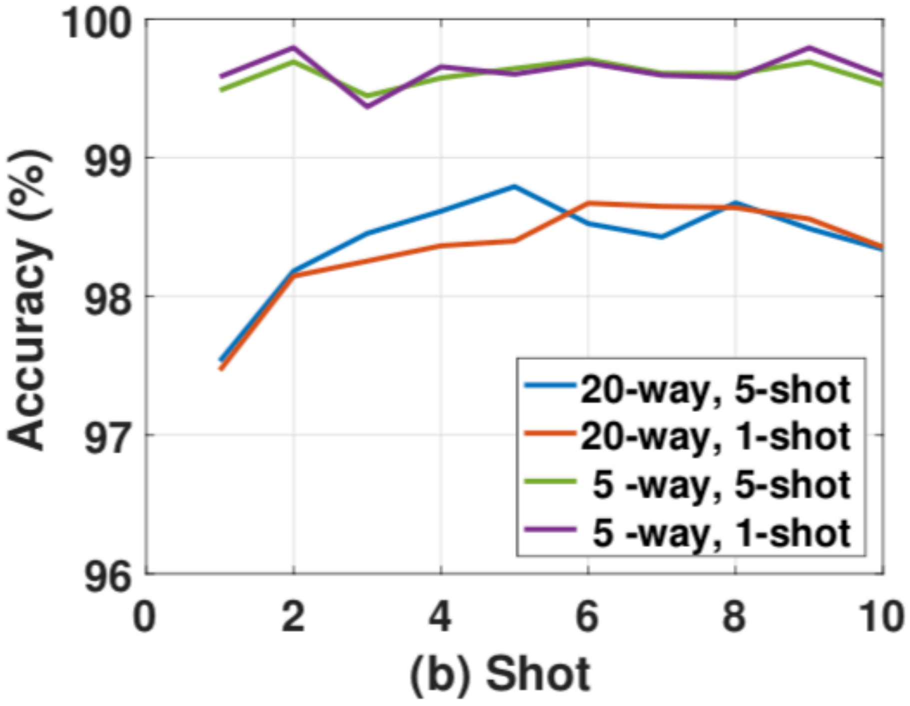
Method	1-shot	5-shot
META-SGD (Li et al., 2017)	54.24±0.03	70.86±0.04
(Gidaris & Komodakis, 2018)	56.20±0.86	73.00±0.64
(Bauer et al., 2017)	56.30±0.40	73.90±0.30
(Munkhdalai et al., 2017)	57.10±0.70	70.04±0.63
(Qiao et al., 2018)	59.60±0.41	73.54±0.19
LEO (Rusu et al., 2019)	61.76±0.08	77.59±0.12
SNAIL (Mishra et al., 2018)	55.71±0.99	68.88±0.92
TADAM (Oreshkin et al., 2018)	58.50±0.30	76.70±0.30
METAVRF (w/o LSTM, 780d)	**62.12**±0.07	77.05±0.28
METAVRF (vanilla LSTM, 780d)	**63.21**±0.06	**77.83**±0.28
METAVRF (bi-LSTM, 780d)	**63.80**±0.05	**77.97**±0.28

Further Analysis



Further Analysis





Conclusion

- A novel meta-learning framework, MetaVRF, introducing RFFs into the meta-learning framework and leveraging VI to infer the spectral distribution in a data-driven way.
- The LSTM-based context inference explores the shared knowledge and generates rich random features.
- Achieve the state-of-the-art performance.
- Learned kernels exhibit high representational power with a low spectral sampling rate.
- Robustness and flexibility to a great variety of testing conditions.

Thank you for your attention !