ST-MAML Paper at *UAI 2022*

ST-MAML: A stochastic-task based method for task-heterogeneous meta-learning

Zhe Wag, Jake Grigsby, Yanjun Qi

May 26, 2023

Outline

Content

Few-shot meta-learning is important

Few-shot meta-learning requires model to be a quick learner.

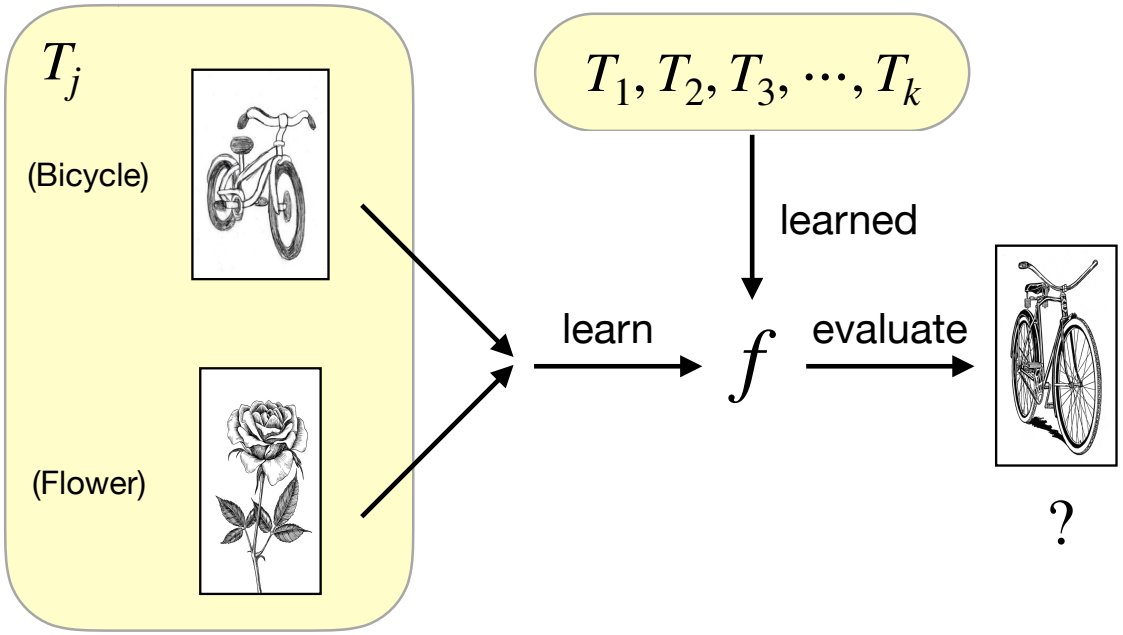

Figure 1: 2-Way 1-Shot meta-learning example.[1]

[1] pictures are downloaded from Pinterest

Task-heterogeneous meta-learning is more practical

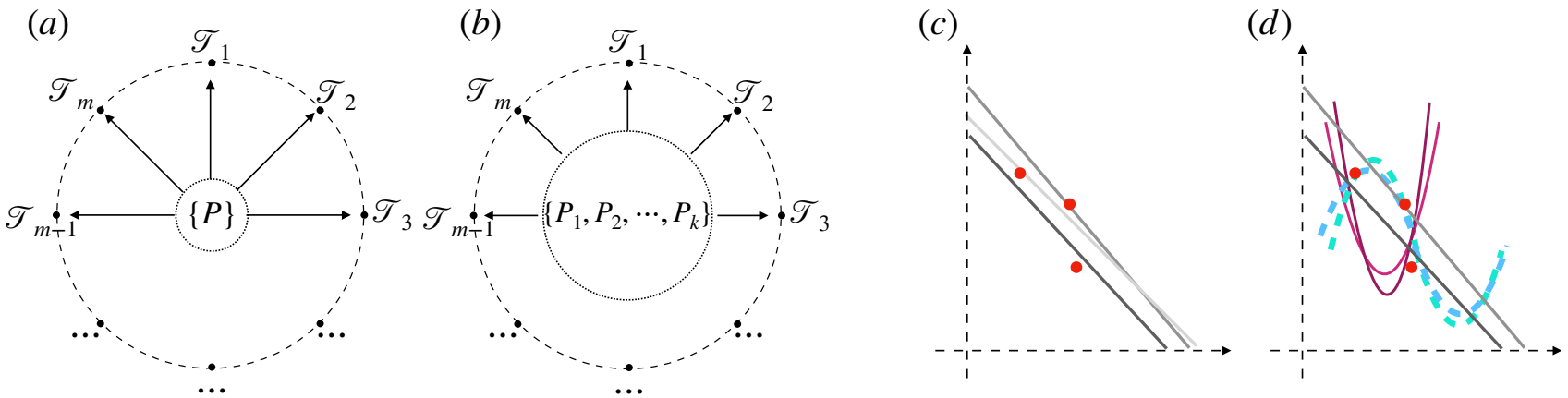

Figure 2: Two critical challenges in meta-learning. (a, b): The figures show the difference between task homogeneity and task heterogeneity in meta-learning. The solid line with arrow represents the uniformly random sampling from meta distributions (inner circle). (c, d): The figures demonstrate the task ambiguity in meta-learning. In heterogeneous setup, the task ambiguity is more critical due to the distributional uncertainty. The red dots represent the available training data, the dashed and solid curves are potential explanations of the data.

Challenges in Few-shot Heterogeneous Meta-learning

While being more practical, task-heterogeneous meta-learning is more challenge:

- The i.i.d assumption on task level is broken.
- The task is more ambiguous due to more uncertainty factors.
  - The uncertainty comes from limited training data
  - The uncertainty also exists in meta-distribution.

Illustrative example for Challenges

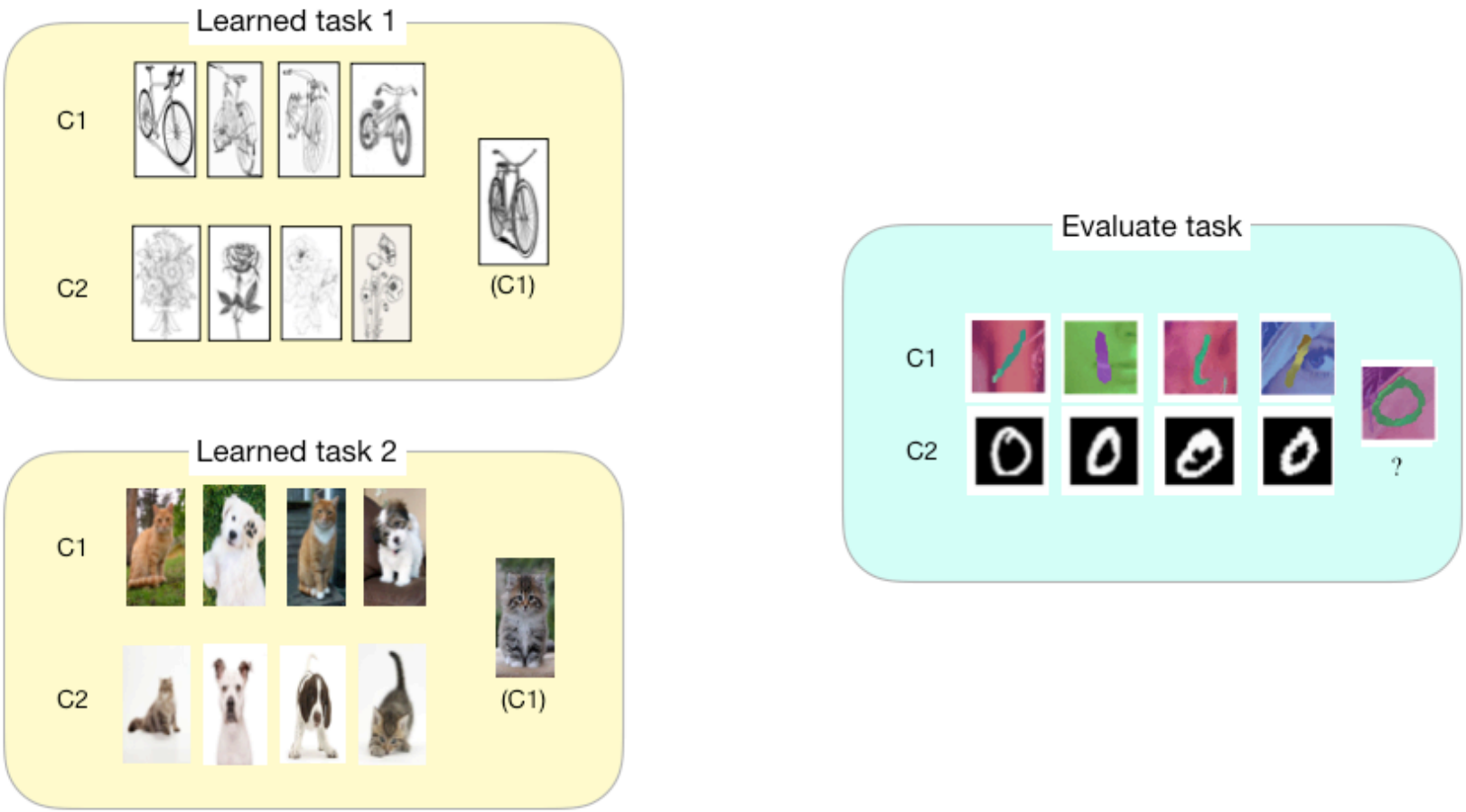

Figure 3: 2-Way 4-Shot task-heterogeneous meta-learning with strong ambiguity.

Notations and descriptions

Suppose we have a meta-distribution set from which we sample tasks, every task $\mathcal{T}$ consists of a limited training set $D_{\mathcal{T}}^{tr} = \{X_{\mathcal{T}}^{tr}, Y_{\mathcal{T}}^{tr}\}$ and a test set $D_{\mathcal{T}}^{te}$,

- During meta-training, $D_{\mathcal{T}}^{te} = \{X_{\mathcal{T}}^{te}, Y_{\mathcal{T}}^{te}\}$ are fully observed.
- During meta-testing, $D_{\mathcal{T}}^{te} = \{X_{\mathcal{T}}^{te}\}$, the target is not available.

Task-homogeneous meta-learning: $\mathcal{T}_i \sim \{P\}$.
Task-heterogeneous meta-learning: $\mathcal{T}_i \sim \{P_1, \cdots, P_k\}$.

Content

Background: MAML

MAML[2] trains an initialization which is close to all tasks:

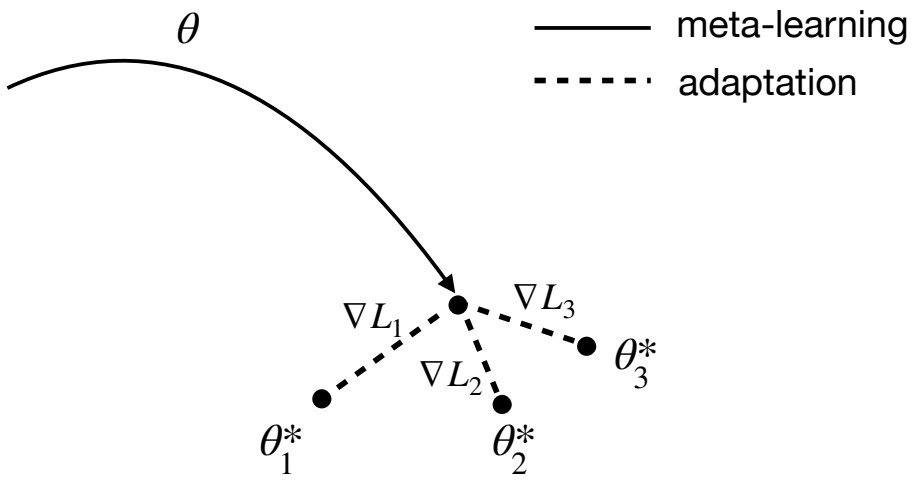

Figure 4: Diagram of MAML

[2]Finn, et.al. Model-agnostic meta-learning for fast adaptation of deep networks, ICML, 2017

Background: MAML

Training objective for MAML:

$$\min_{\theta} \mathop{\mathbf{E}}_{\mathcal{T}\sim P(\mathcal{T})}[\mathcal{L}oss(\theta^1_{\mathcal{T}}, D^{te}_{\mathcal{T}})],$$
$$\text{where} \quad \theta^1_{\mathcal{T}} = \theta^0_{\mathcal{T}} - \alpha\nabla_{\theta}[\mathcal{L}oss(\theta^0_{\mathcal{T}}, D^{tr}_{\mathcal{T}})], \quad \theta^0_{\mathcal{T}} = \theta. \tag{1}$$

Equivalently:

$$\max_{\theta} \mathop{\mathbf{E}}_{\mathcal{T}\sim P(\mathcal{T})}[\mathcal{L}(\mathcal{T})] = \prod_{\mathcal{T}\sim P(\mathcal{T})} p(Y^{te}_{\mathcal{T}}|X^{te}_{\mathcal{T}}, D^{tr}_{\mathcal{T}}, \theta)$$
$$= \prod_{\mathcal{T}\sim P(\mathcal{T})} \sum_{\theta^1_{\mathcal{T}}} p(Y^{te}_{\mathcal{T}}|X^{te}_{\mathcal{T}}, \theta^1_{\mathcal{T}}) p(\theta^1_{\mathcal{T}}|D^{tr}_{\mathcal{T}}, \theta), \tag{2}$$

where $p(\theta^1_{\mathcal{T}}|D^{tr}_{\mathcal{T}}, \theta)$ is a Dirac distribution derived by minimizing the negative log-likelihood(NLL) on $D^{tr}_{\mathcal{T}}$ with gradient descent.

Limitations of MAML

Limitations of MAML:

- Designed for task-homogeneous meta-learning.
- Task ambiguity?
- Meta-knowledge set contains parameters only.

Proposed model

Stochastic-task based method for task-heterogeneous meta-learning

Framework of ST-MAML

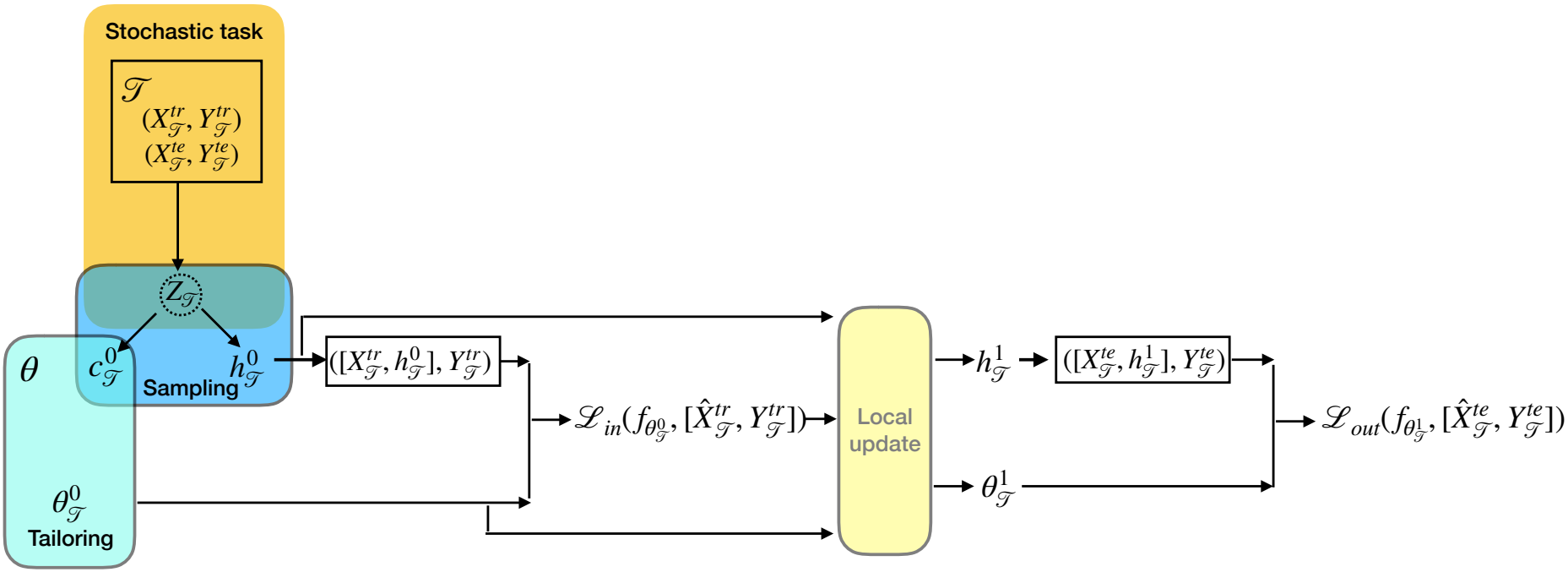

Figure 5: Framework of ST-MAML.

In practice, we apply multiple iterations for inner update.

Stochastic task module

We propose to model task $\mathcal{T}$ via a stochastic variable $Z_{\mathcal{T}}$. With Bayes rules we can rewrite the per task likelihood as:

$$\mathcal{L}(\mathcal{T}) = \sum_{Z_{\mathcal{T}}} p(Y_{\mathcal{T}}^{te} | X_{\mathcal{T}}^{te}, D_{\mathcal{T}}^{tr}, Z_{\mathcal{T}}, \theta) p(Z_{\mathcal{T}} | D_{\mathcal{T}}^{tr}). \tag{3}$$

We model the conditional prior $p(Z_{\mathcal{T}} | D_{\mathcal{T}}^{tr})$ as a Gaussian distribution via deep set operator:

$$p(Z_{\mathcal{T}} | D_{\mathcal{T}}^{tr}) = \mathcal{N}(\mu(r_{\mathcal{T}}), \sigma(r_{\mathcal{T}})). \tag{4}$$

where,

$$r_{\mathcal{T}} = \frac{1}{|D_{\mathcal{T}}^{tr}|} \sum_{j=1}^{|D_{\mathcal{T}}^{tr}|} r_{\mathcal{T},j}, \quad r_{\mathcal{T},j} = g_{\phi}^{Enc}(x_{\mathcal{T},j}^{tr}, y_{\mathcal{T},j}^{tr}), \quad j = 1, \cdots, |D_{\mathcal{T}}^{tr}|$$

Stochastic task module

Due to the unknown task distribution, the posterior distribution of $Z_{\mathcal{T}}$ is intractable.

$$p(Z_{\mathcal{T}}|\mathcal{T}) = \frac{p(Z_{\mathcal{T}}|D_{\mathcal{T}}^{tr})p(Y_{\mathcal{T}}^{te}|Z_{\mathcal{T}}, X_{\mathcal{T}}^{te}, D_{\mathcal{T}}^{tr})}{p(\mathcal{T})} \tag{5}$$

To apply variational inference, we need to sample from $q(Z_{\mathcal{T}}|\mathcal{T})$. We approximate it with:

$$q(Z_{\mathcal{T}}|\mathcal{T}) = \mathcal{N}(\mu(r'_{\mathcal{T}}), \sigma(r'_{\mathcal{T}})). \tag{6}$$

where,

$$r'_{\mathcal{T}} = \frac{1}{|\mathcal{T}|}\sum_{j=1}^{|\mathcal{T}|} r_{\mathcal{T},j}, \quad r_{\mathcal{T},j} = g_{\phi}^{Enc}(x_{\mathcal{T},j}, y_{\mathcal{T},j}), \quad j = 1, \cdots, |D_{\mathcal{T}}^{tr}| + |D_{\mathcal{T}}^{te}|$$

Sampling, tailoring, and local update modules

$$z_{\mathcal{T}} \sim P(Z_{\mathcal{T}}),$$

$$c^0_{\mathcal{T}} = w_1 z_{\mathcal{T}} + w_0,$$
$$h^0_{\mathcal{T}} = \beta_1 z_{\mathcal{T}} + \beta_0,$$

Sampling

$$C^0_{\mathcal{T}} \quad \theta \xrightarrow{\sigma(C^0_{\mathcal{T}}) \cdot \theta} \theta^0_{\mathcal{T}}$$

Tailoring

$$\theta^0_{\mathcal{T}} \quad \mathcal{L}_{in} \quad h^0_{\mathcal{T}} \longrightarrow \theta^1_{\mathcal{T}} = \theta^0_{\mathcal{T}} - \frac{\partial \mathcal{L}_{in}}{\partial \theta^0_{\mathcal{T}}}, \quad h^1_{\mathcal{T}} = h^0_{\mathcal{T}} - \frac{\partial \mathcal{L}_{in}}{\partial h^0_{\mathcal{T}}}$$

Local Update

Figure 6: Sampling module, tailoring module, and local update module.

Iterative optimization process

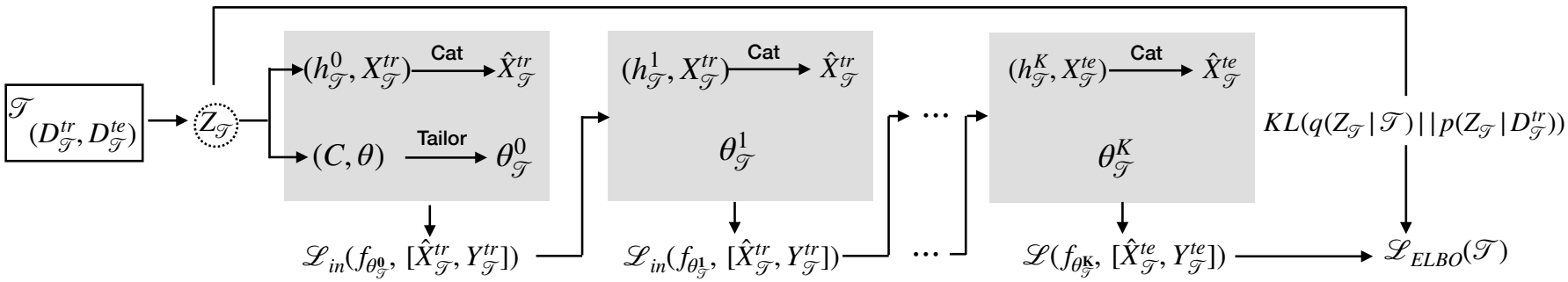

Figure 7: Iterative optimization process. In the inner loop, Starting from task-specific parameter initialization $\theta^0_{\mathcal{T}}$ and augmented features $h^0_{\mathcal{T}}$, their fine-tuned values $\theta^K_{\mathcal{T}}, h^K_{\mathcal{T}}$ are inferred by performing gradient descent on the training set $D^{tr}_{\mathcal{T}}$ for K iterations.

Objective to be maximized:

$$\mathcal{L}_{ELBO}(\mathcal{T}) = \mathbf{E}_{\Theta^1_{\mathcal{T}} \sim q(\Theta^1_{\mathcal{T}}|\mathcal{T})} \log p(Y^{te}_{\mathcal{T}}|X^{te}_{\mathcal{T}}, \Theta^1_{\mathcal{T}}) - KL(q(Z_{\mathcal{T}}|\mathcal{T})||p(Z_{\mathcal{T}}|D^{tr}_{\mathcal{T}}).$$

Main idea

ST-MAML extends MAML, it has the following abilities:

- Solve task-heterogeneous challenge without any prior about the meta-distribution set.
- Increase the diversity of meta-knowledge, which contains both model parameters and feature augmentation.
- The probabilistic framework alleviate the ambiguity challenge.

Information bottleneck explanation

Given task inputs $\mathbf{X}_{\mathcal{T}} = [X^{tr}_{\mathcal{T}}, X^{te}_{\mathcal{T}}]$, we are seeking a task-specific knowledge set $\Theta^1_{\mathcal{T}}$ that is maximally informative of test target $Y^{te}_{\mathcal{T}}$, while being mostly compressive of training target $Y^{tr}_{\mathcal{T}}$. The information bottleneck objective is:

$$\mathcal{L}_{IB}(\mathcal{T}) = I(Y^{te}_{\mathcal{T}}; \Theta^1_{\mathcal{T}} | \mathbf{X}_{\mathcal{T}}) - \beta I(Y^{tr}_{\mathcal{T}}; \Theta^1_{\mathcal{T}} | \mathbf{X}_{\mathcal{T}}). \tag{7}$$

We show the following lemma:

Lemma

Given a task $\mathcal{T}$, maximizing the information bottleneck loss $\mathcal{L}_{IB}$ defined in (7) is equivalent to maximizing the weighted ELBO :

$$\mathcal{L}_{wELBO}(\mathcal{T}) = \mathbf{E}_{\Theta^1_{\mathcal{T}} \sim q(\Theta^1_{\mathcal{T}} | \mathcal{T})} \log p(Y^{te}_{\mathcal{T}} | \Theta^1_{\mathcal{T}}, X^{te}_{\mathcal{T}}) - \beta KL(q(Z_{\mathcal{T}} | \mathcal{T}) || p(Z_{\mathcal{T}} | D^{tr}_{\mathcal{T}})). \tag{8}$$

Content

Experimental Design

Table 1: A summary of datasets, tasks and their properties.

Problems	Tasks	Heterogeneity	Ambiguity
Regression	2D regression	+	$10 \rightarrow 40$
	Weather prediction	++	$10 \rightarrow 100$
	Image completion	+	$40 \rightarrow 784$
Classification	PlainMulti classification	+	5way 5shot
	CelebA binary classification		2way 5shot

Regression experiments: 2D regression

In 2D curve regression task, meta-distribution contains 6 function families including: sinusoids, line, quadratic, cubic, quadratic surface and ripple.

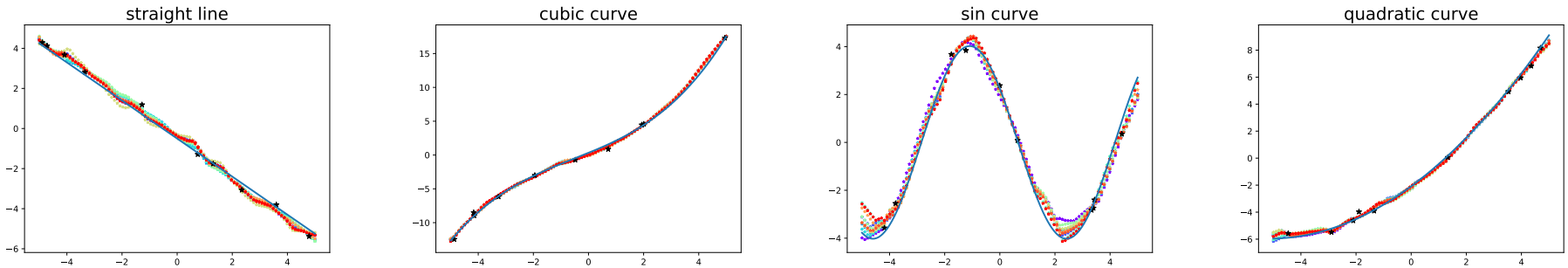

Figure 8: Qualitative Visualization of fitting curves. Black stars represent training set $D_{\mathcal{T}}^{tr}$, 10 different samples of fitting curves are shown as colored dotted lines. The blue solid line is the true mapping.

Regression experiments: 2D regression

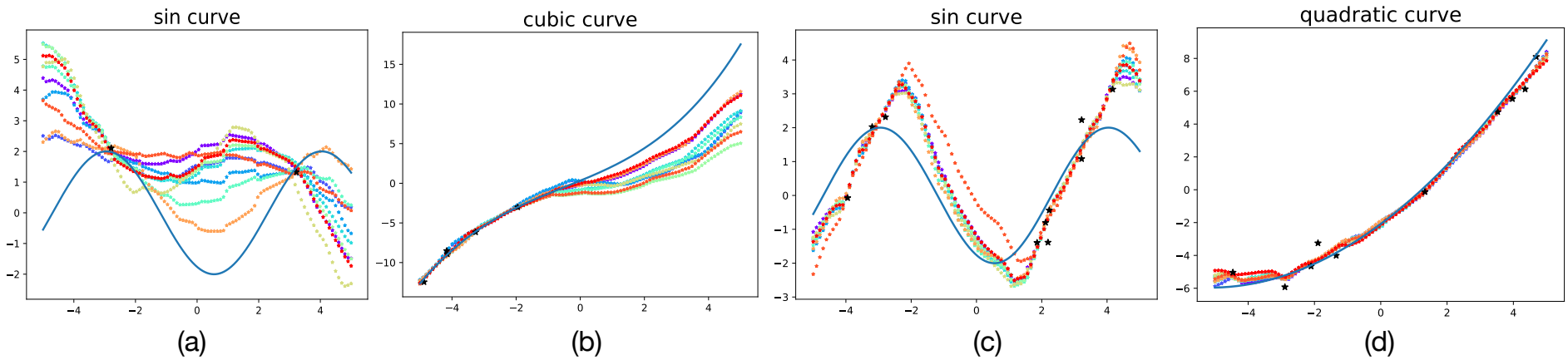

Figure 9: Few-shot 2D regression with various number of training data and noise level. (a) $|D_{\mathcal{T}}^{tr}| = 2, \sigma = 0.3$ (b) $|D_{\mathcal{T}}^{tr}| = 5, \sigma = 0.3$, (c) $|D_{\mathcal{T}}^{tr}| = 10, \sigma = 0.8$, (d) $|D_{\mathcal{T}}^{tr}| = 10, \sigma = 0.1$. Black star represents training data, dashed lines characterize different sampled models, the blue curve is the true mapping.

Table 2: Regression accuracy on 2D regression tasks.

Model	MAML	MetaSGD	BMAML	MMAML	ARML	ST-MAML
MSE	2.29 ± 0.16	2.91 ± 0.23	1.65 ± 0.10	0.52 ± 0.04	0.44 ± 0.03	$\mathbf{0.37 \pm 0.04}$

Regression experiments: Image completion

In image completion, meta-distribution:

$$\mathcal{T} = \{MNIST, FMNIST, KMNIST\}$$

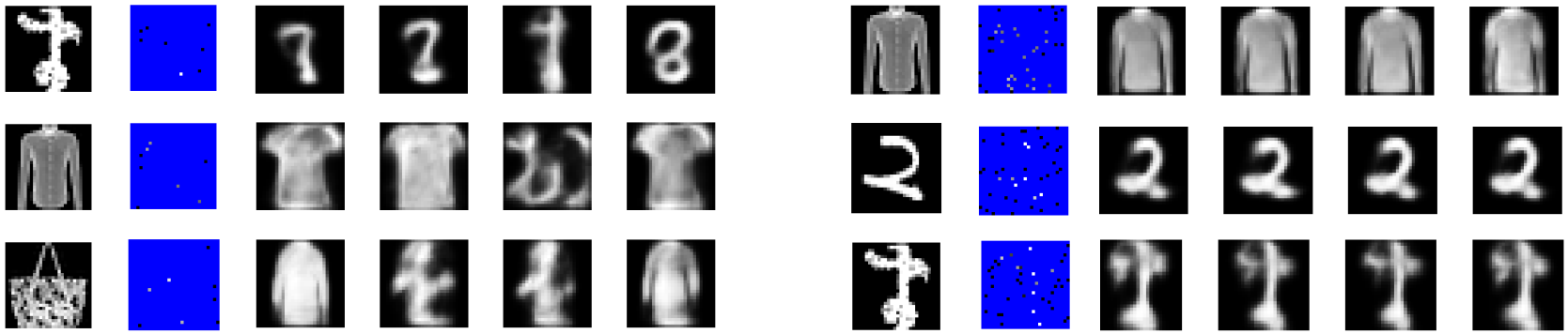

Figure 10: Visualization of completed images. First column contains original images, second column shows the observations which contains only 8 annotated pixels(left) and 40 annotated pixels(right).

Table 3: Image completion accuracy.

Model	NP	CNP	ST-MAML (deter)	ST-MAML
BCE	0.302	0.358	0.272	**0.268**

Regression experiments: Temperature prediction

The dataset contains temperature data collected from more than 9000 stations ranging from 1969 to 2019.

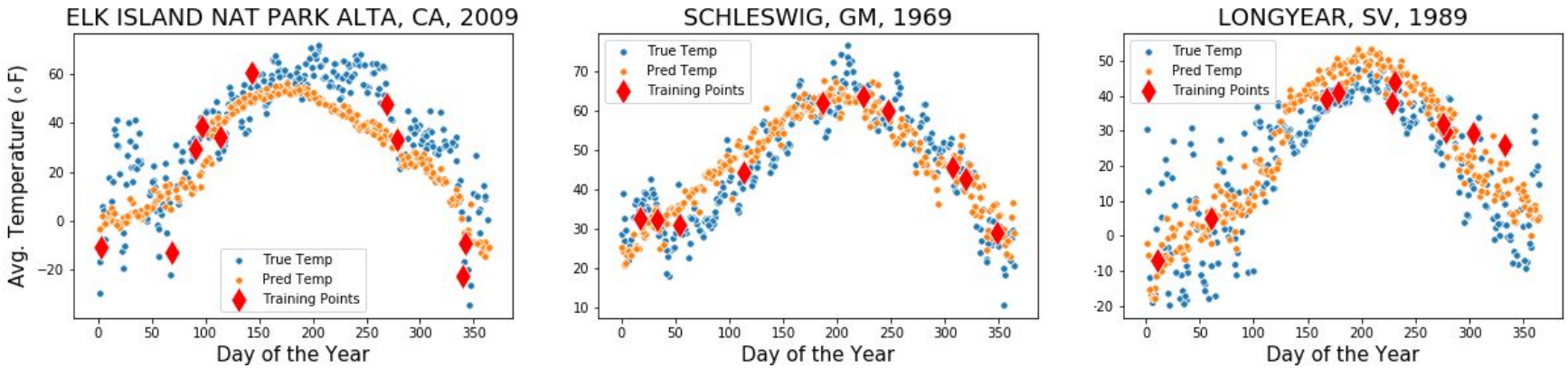

Figure 11: A visualization of trained ST-MAML on the NOAA-GSOD temperature prediction task. The model is given 10 training points (red) and predicts the remaining days of the year (orange). The true temperatures are shown in blue.

Table 4: 10-Shot temperature prediction.

Model	MAML	MetaSGD	ST-MAML	ST-MAML (w/o aug)	ST-MAML (w/o tailor)
MSE	$141.43 \pm 9.33\%$	$291.42 \pm 14.89\%$	$\mathbf{86.56 \pm 4.89\%}$	$100.27 \pm 5.87\%$	$106.37 \pm 5.77\%$

Classification experiments: Plain-Multi

Plain-Multi dataset[3] consists of four image classification datasets: Bird dataset, Textures Dataset, Aircraft Dataset, and FGVCx-Fungi dataset.

Table 5: 5-way 5-shot classification accuracy with 95% confidence interval on Plain-Multi dataset.

Settings	Algorithms	Data: Bird	Data: Texture	Data: Aircraft	Data: Fungi
5-way 5-shot	MAML	$68.52 \pm 0.79\%$	$44.56 \pm 0.68\%$	$66.18 \pm 0.71\%$	$51.85 \pm 0.85\%$
	MetaSGD	$67.87 \pm 0.74\%$	$45.49 \pm 0.68\%$	$66.84 \pm 0.70\%$	$52.51 \pm 0.81\%$
	BMAML	$69.01 \pm 0.74\%$	$46.06 \pm 0.69\%$	$65.74 \pm 0.67\%$	$52.43 \pm 0.84\%$
	MMAML	$70.49 \pm 0.76\%$	$45.89 \pm 0.69\%$	$67.31 \pm 0.68\%$	$53.96 \pm 0.82\%$
	HSML	$\mathbf{71.68 \pm 0.73\%}$	$\mathbf{48.08 \pm 0.69\%}$	$\mathbf{73.49 \pm 0.68\%}$	$\mathbf{56.32 \pm 0.80\%}$
	ST-MAML	$\mathbf{72.49 \pm 0.53\%}$	$46.51 \pm 0.42\%$	$\mathbf{72.64 \pm 0.44\%}$	$\mathbf{55.29 \pm 0.57\%}$
	ST-MAML (w/o aug)	$71.49 \pm 0.55\%$	$\mathbf{47.17 \pm 0.44\%}$	$71.62 \pm 0.43\%$	$54.91 \pm 0.56\%$
	ST-MAML (w/o tailor)	$71.48 \pm 0.55\%$	$46.07 \pm 0.40\%$	$70.46 \pm 0.44\%$	$54.59 \pm 0.56\%$

[3] Yao et.al, "Automated relational meta-learning", ICLR, 2020

Classification experiments: CelebA binary classification

Images can share multiple attributes. During meta-testing, there are three combinations of two attributes for classification.

Table 6: 5-Shot Ambiguous Binary Classification.

Model	Accuracy	Coverage number	NLL
MAML	77.924	1.00	0.454
ST-MAML	79.698	1.13	0.439

Figure 12: Sampled classifiers for ambiguous binary classification task.

Visualization

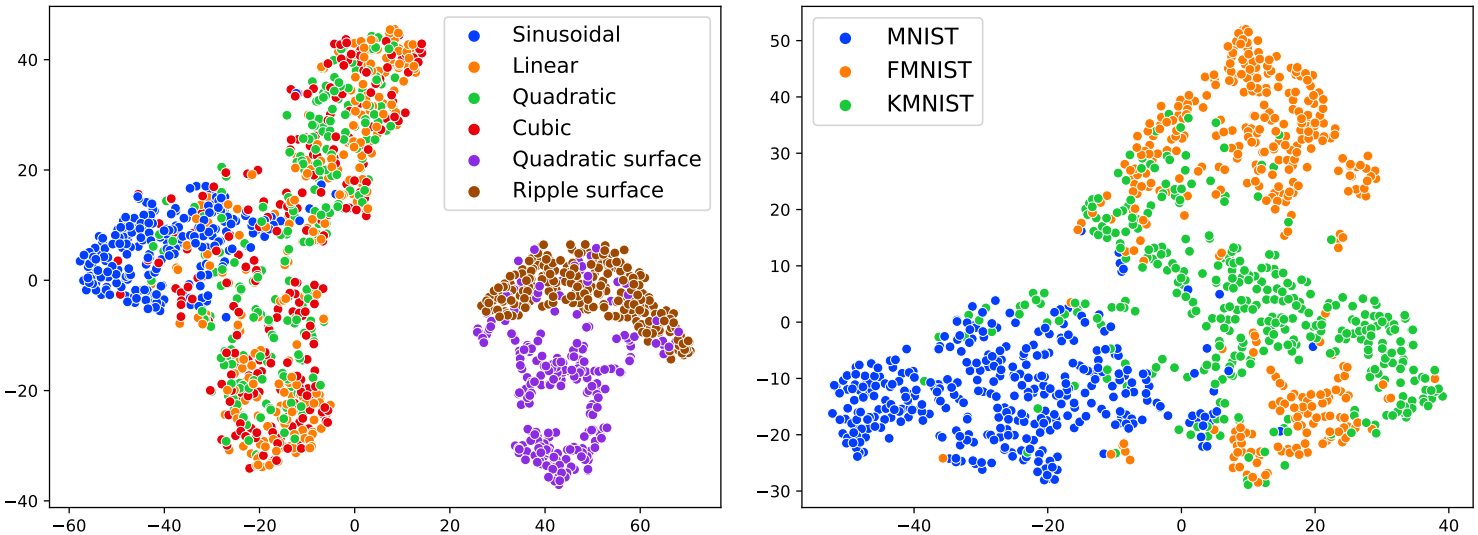

Figure 13: t-SNE plots of gate vectors for tasks randomly sampled from the meta-distributions of synthetic regression (left) and image completion (right)

.

Content

Related works

Table 7: Model comparison table. HoMAMLs are MAMLs designed for task homogeneity, and HeMAMLs are for heterogeneity. NPs describe methods in Neural Processes family. PMAMLs mean probabilistic extensions of MAML.

Category	Tasks	Knowledge Set	Tailoring	Sampling	Inference on
HoMAMLs	MAML	Initialization			
	MetaSGD	Initialization+lr			
HeMAMLs	MMAML	Initialization	✓		
	HSML	Initialization	✓		
NPs	NP	Aug feature		✓	Representation
	CNP	Aug feature			
PMAMLs	BMAML	Initialization		✓	Parameters
	PLATIPUS	Initialization		✓	Parameters
	ST-MAML	Initialization+Aug feature	✓	✓	Representation

Conclusion

Conclusions:

- Both task-heterogeneity and task-ambiguity are critical challenges in meta-learning.
- Customizing meta-knowledge is the key to task-heterogeneity challenge.
- Bayesian inference can better incorporate uncertainty into the ambiguous tasks.

Thank you for your lunch time!