Meta-Learning without Memorization

Mingzhang Yin, George Tucker, Mingyuan Zhou, Sergey Levine, Chelsea Finn

Supervised Meta-Learning

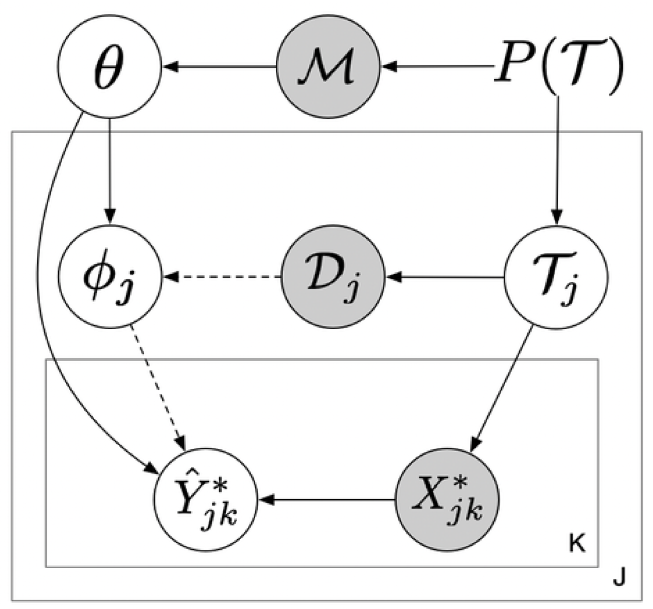
- Assume task $\mathcal{T}_i \sim p(\mathcal{T})$; Each task consists of task training data $\mathcal{D}_i = (\boldsymbol{x}_i, \boldsymbol{y}_i)$ and validation data $\mathcal{D}_i^* = (\boldsymbol{x}_i^*, \boldsymbol{y}_i^*)$.
- $\boldsymbol{x}_i = (x_{i1}, \dots, x_{iK}), \boldsymbol{y}_i = (y_{i1}, \dots, y_{iK}) \sim p(x, y|\mathcal{T}_i)$ and similarly for $\mathcal{D}_i^*$.
- Entire meta-training set is $\mathcal{M} = \{\mathcal{D}_i, \mathcal{D}_i^*\}_{i=1}^N$
- The objective is

$$-\frac{1}{N}\sum_i \mathbb{E}_{q(\theta|\mathcal{M})q(\phi|\mathcal{D}_i,\theta)}\left[\frac{1}{K}\sum_{(x^*,y^*)\in\mathcal{D}_i^*} \log q(\hat{y}^* = y^*|x^*, \phi, \theta)\right].$$

where $q(\theta|\mathcal{M})$ summarizes meta-training data, $q(\phi|\mathcal{D}, \theta)$ summarizes the per-task training set and $q(\hat{y}^*|x^*, \phi, \theta)$ is the predictive distribution.

The Memorization Problem

Definition 1 (Complete Meta-Learning Memorization)**.** *Complete memorization in meta-learning is when the learned model ignores the task training data such that* $I(\hat{y}^*; \mathcal{D}|x^*, \theta) = 0$ *(i.e.,* $q(\hat{y}^*|x^*, \theta, \mathcal{D}) = q(\hat{y}^*|x^*, \theta) = \mathbb{E}_{\mathcal{D}'|x^*}[q(\hat{y}^*|x^*, \theta, \mathcal{D}')]$*).*



Without either one of the dashed arrows, $\hat{Y}^*$ is conditionally independent of $\mathcal{D}$ given θ and X^*, which we refer to as complete memorization.

Properties

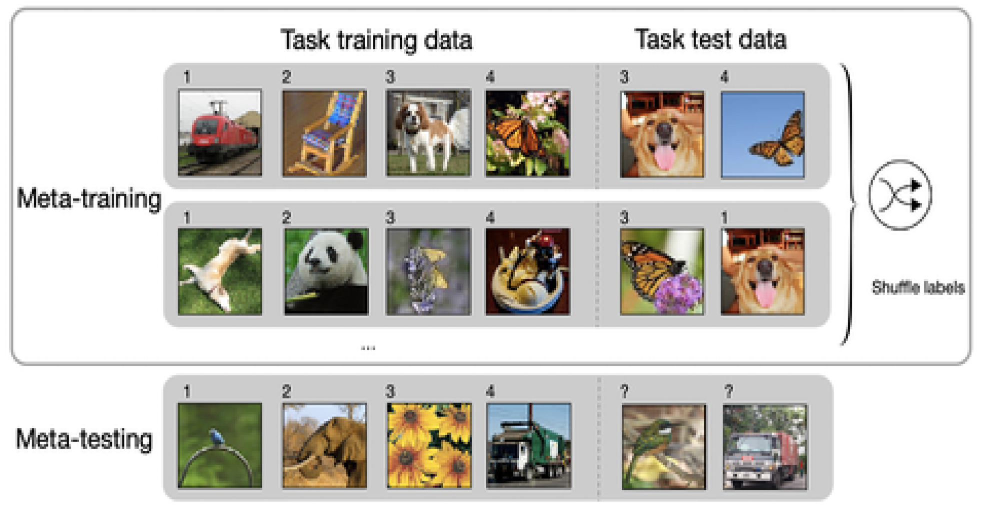
- Memorization means one model can solve all training tasks.
- Memorized model generalizes to unseen points in training tasks, but cannot generalize to unseen tasks (task-level overfitting).
- Memorization occurs in many meta-learning algorithms:
  MAML: Loss $\mathcal{L}(x, y, \theta) \approx 0$ for $(x, y) \in \mathcal{D}$ and $\mathcal{D}^*$ can result in minimal task adaptation i.e. $\phi \approx \theta$;
  Conditional Neural Process (CNP): $q(\hat{y}^*|x^*, \phi, \theta)$ can achieve low training error without using the task training summary statistics ϕ.

Examples

- Pose Regression. Predict pose of object from 2D image; can overfit to training objects.
- Automated precision medicine system. Each patient represents a separate task. $\mathcal{D}$ is the patient's medical history, input x is the symptom and patient's identity infomation; output $\hat{y}$ is the recommended medication.

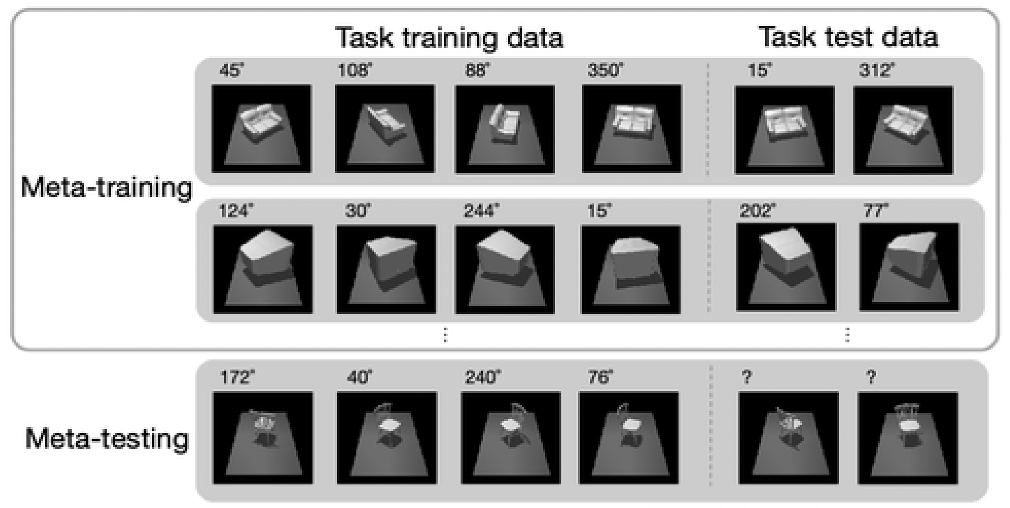
Why does it happen in Meta-Learning?

Mutually-exclusive Task Distribution



Random label permutation for few-shot classification.

Non-mutually-exclusive Task Distribution



Pose regression example: the training tasks are non-mutually-exclusive because the test data label (right) can be inferred accurately without using task training data (left) in the training tasks, by memorizing the canonical orientation of the meta-training objects.

Meta Regularization Using Information Theory

- Sources of information in the predictive distribution $q(\hat{y}^*|x^*, \theta, \mathcal{D})$ come from input, meta-parameters, and data.
- Encourage using task training data $\mathcal{D}$ by restricting the information flow from other sources (x^* and θ) to $\hat{y}^*$.

Meta Regularization on Activations

- Introduce an intermediate stochastic bottleneck variable z^* such that $q(\hat{y}^*|x^*, \phi, \theta) = \int q(\hat{y}^*|z^*, \phi, \theta)q(z^*|x^*, \theta)\, dz^*$.
- Optimize with the regularized training objective

$$\frac{1}{N}\sum_i \mathbb{E}_{q(\theta|\mathcal{M})q(\phi|\mathcal{D}_i,\theta)}\left[-\frac{1}{K}\sum_{(x^*,y^*)\in\mathcal{D}_i^*} \log q(\hat{y}^* = y^*|x^*, \phi, \theta) + \beta D_{\text{KL}}(q(z^*|x^*, \theta)||r(z^*))\right],$$

- In some cases, it can be sensitive to the initialization and learning rate.

Meta Regularization on Weights

- Limit the information about the training tasks stored in the meta-parameters θ by penalizing $I(y_{1:N}^*, \mathcal{D}_{1:N}; \theta|x_{1:N}^*)$.
- Related to a novel PAC Bayes bound for meta-learning (see paper for details).
- The objective is

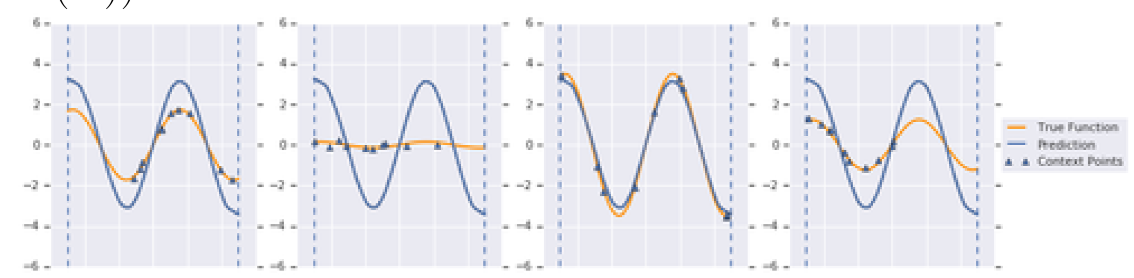
$$\frac{1}{N}\sum_i \mathbb{E}_{q(\theta)q(\phi|\mathcal{D}_i,\tilde{\theta})}\left[-\frac{1}{K}\sum_{(x^*,y^*)\in\mathcal{D}_i^*} \log q(\hat{y}^* = y^*|x^*, \phi, \theta, \tilde{\theta}) + \beta D_{\text{KL}}(q(\theta; \theta_\mu, \theta_\sigma)||r(\theta))\right]$$

Experiments

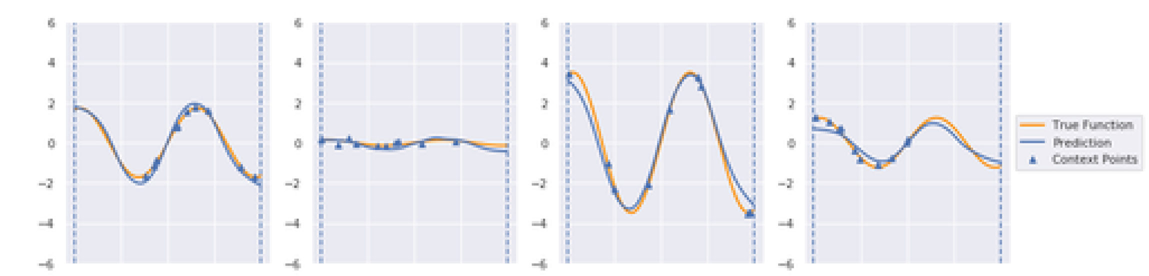
Sinusoid Regression

For each task, $u \sim U(-5, 5)$, $y \sim \mathcal{N}(A\sin(u), 0.1^2)$, $A \sim \{0.1, 0.3, \dots, 4\}$. Input $x = (u, \text{one-hot}(A))$.

CNP

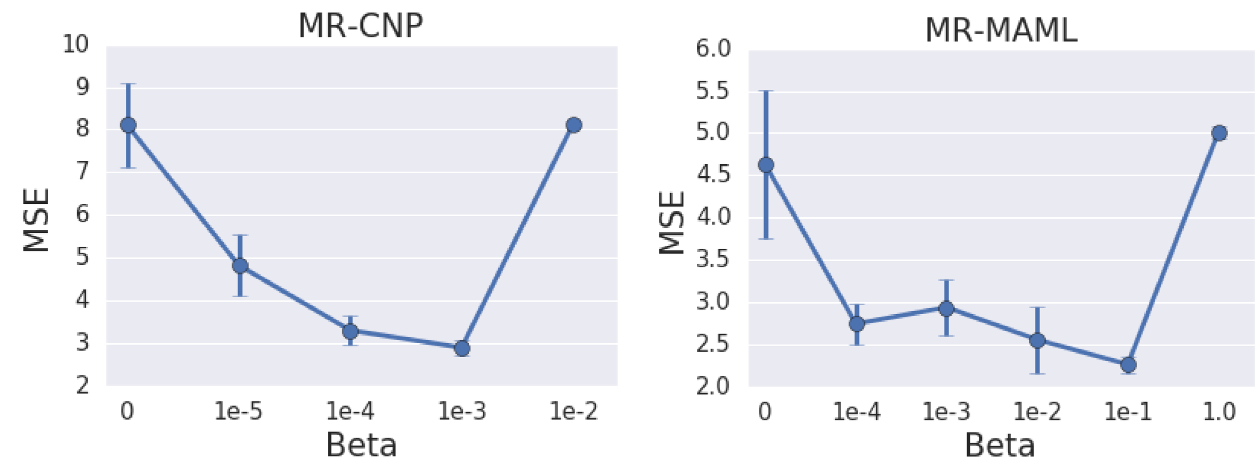


MR-CNP



Methods	MAML	MR-MAML (A) (ours)	MR-MAML (W) (ours)	CNP	MR-CNP (A) (ours)	MR-CNP (W) (ours)
5 shot	0.46 (0.04)	**0.17 (0.03)**	**0.16 (0.04)**	0.91 (0.10)	**0.10 (0.01)**	**0.11 (0.02)**
10 shot	0.13 (0.01)	**0.07 (0.02)**	**0.06 (0.01)**	0.92 (0.05)	**0.09 (0.01)**	**0.09 (0.01)**

Pose Regression



The performance of MAML and CNP with meta-regularization on the weights, as a function of the regularization strength β.

Method	MAML	MR-MAML (W) (ours)	CNP	MR-CNP (W) (ours)	FT	FT + Weight Decay
MSE	5.39 (1.31)	**2.26 (0.09)**	8.48 (0.12)	2.89 (0.18)	7.33 (0.35)	6.16 (0.12)

Non-mutually-exclusive Classification

Meta-test accuracy on non-mutually-exclusive (NME) classification.

NME Omniglot	20-way 1-shot	20-way 5-shot
MAML	7.8 (0.2)%	50.7 (22.9)%
TAML	9.6 (2.3)%	67.9 (2.3)%
MR-MAML (W)	**83.3 (0.8)**%	**94.1 (0.1)**%

NME MiniImagenet	5-way 1-shot	5-way 5-shot
Fine-tuning	28.9 (0.5))%	49.8 (0.8))%
Nearest-neighbor	41.1 (0.7)%	51.0 (0.7) %
MAML	26.3 (0.7)%	41.6 (2.6)%
TAML	26.1 (0.6)%	44.2 (1.7)%
MR-MAML (W)	**43.6 (0.6)**%	**53.8 (0.9)**%

SCAN ME