Meta-Learning without Memorization

Mingzhang Yin[1,2], **George Tucker**[2], **Sergey Levine**[2,3], **Chelsea Finn**[2,4]
mzyin@utexas.edu, gjt@google.com, svlevine@eecs.berkeley.edu
cbfinn@cs.stanford.edu
[1]UT Austin, [2]Google Research, Brain team, [3]UC Berkeley, [4]Stanford

Abstract

Meta-learning is a promising technique to learn new concepts with small amounts of data. However, most meta-learning algorithms implicitly require that the meta-training tasks be *mutually-exclusive*, such that no single model can solve all of the tasks at once. If this is not done, the meta-learner can ignore the task training data and learn a single model that performs all of the meta-training tasks zero-shot, but does not adapt effectively to new image classes. This requirement limits the domains that meta-learning can be effectively applied on. In this paper, we address this challenge by designing a meta-regularization objective using information theory that places precedence on data-driven adaptation. By doing so, our algorithm can successfully use data from *non-mutually-exclusive* tasks to efficiently adapt to novel tasks. We demonstrate its applicability to both contextual and gradient-based meta-learning algorithms, and apply it in practical settings where applying standard meta-learning has been difficult. Our approach substantially outperforms standard meta-learning algorithms in these settings.

1 Introduction

The ability to learn new concepts and skills with small amounts of data is a critical aspect of intelligence that many machine learning systems lack. Meta-learning (Schmidhuber, 1987) accomplishes this by explicitly optimizing for few-shot generalization across a set of meta-training tasks. While these methods have shown promising results, current methods require careful design of the meta-training tasks to prevent a subtle form of *task overfitting*, distinct from standard overfitting in supervised learning. If the task can be accurately inferred from the test input alone, the task training data can be ignored while still achieving low meta-training loss. In effect, the model will collapse to one that makes zero-shot decisions. This presents an opportunity for overfitting where the meta-learner generalizes on meta-training tasks, but fails to adapt when presented with training data from novel tasks. We call this form of overfitting the *memorization problem* in meta-learning because the meta-learner memorizes a function that solves all of the meta-training tasks, rather than adapting.

Existing meta-learning algorithms implicitly resolve this problem by carefully designing the meta-training tasks such that no single model can solve all tasks zero-shot; we call tasks constructed in this way *mutually-exclusive*. This ensures that the task-specific class-to-label assignment cannot be inferred from a test input alone. However, the mutually-exclusive tasks requirement places a substantial burden on the user to cleverly design the meta-training setup (e.g., by shuffling labels or omitting goal information). While shuffling labels provides a reasonable mechanism to force tasks to be mutually-exclusive with standard few-shot image classification datasets such as MiniImageNet (Ravi & Larochelle, 2016), this solution cannot be applied to all domains where we would like to utilize meta-learning, as examples shown in the next section.

Implementation and examples available here: `https://github.com/google-research/google-research/tree/master/meta_learning_without_memorization`.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

In this work, we identify and formalize the memorization problem in meta-learning, and propose a meta-regularizer (MR) using information theory as a general approach for mitigating this problem *without*. In a series of experiments on non-mutually-exclusive task distributions, we find that memorization poses a significant challenge for both gradient-based (Finn et al., 2017) and contextual (Garnelo et al., 2018) meta-learning methods, resulting in near random performance on test tasks in some cases. Our meta-regularization approach enables both of these methods to achieve efficient adaptation and generalization, leading to substantial performance gains across the board on non-mutually-exclusive tasks.

2 The Memorization Problem in Meta-Learning

We consider the standard supervised meta-learning setup as in Finn et al. (2017). Briefly, we assume tasks $\mathcal{T}_i$ are sampled from an (unknown) distribution $p(\mathcal{T})$. For each task, we are given task training data $D_i = (X_i, Y_i)$ and validation data $D_i^* = (X_i^*, Y_i^*)$ with $X_i = (x_{i1}, \ldots, x_{iK}), Y_i = (y_{i1}, \ldots, y_{iK})$ and similarly for D_i^*. Denote M as all the data that have been used in the meta-training $M = \{D_i, D_i^*\}_{i=1}^N$. Following Grant et al. (2018); Gordon et al. (2018), we consider meta-learning algorithms that produce a distribution $q(\tau|D)$, which summarizes the task training data and a prediction distribution $q(y^*|x^*, \tau)$. This includes popular meta-learning algorithms such as MAML (Finn et al., 2017) and conditional neural processes (CNP) (Garnelo et al., 2018). For a test task with training data D, our goal is to maximize the log-likelihood of y^* given input x^* and D under the model (i.e., $\log \mathbb{E}_{q(\tau|D)}\left[q(y^*|x^*, \tau)\right]$).

Ideally, the meta-learning algorithm will learn to generalize to novel tasks. However, we find that unless tasks are carefully designed, current meta-learning algorithms can overfit to the tasks and end up ignoring the task training data (i.e., either $q(\tau|D)$ does not depend on D or $q(y^*|x^*, \tau)$ does not depend on τ) which can lead to poor generalization. This memorization phenomenon is best understood with an example.

Consider a 3D object pose prediction problem (illustrated in Figure 1 and described in detail below). Each task corresponds to a 3D object and a random canonical pose for that object. The (x, y) pairs for the task are 2D grey-scale images of the rotated object (x) and the rotation angle relative to the canonical pose for that object (y). When generating the dataset, we randomly sample the canonical pose rotation for each 3D object and maintain the same random canonical pose every time that object is used in a task. Hence, for an unseen 3D object, the task is impossible without using D because the canonical pose for the unseen object is unknown.

The number of objects in the dataset is small, so it is straightforward for a single network to infer the object from the input image and to memorize the canonical pose for each training object, thus achieving a low training error without using D. However, by construction, this solution will necessarily have poor generalization to new tasks. In practice, we find that MAML and CNP frequently converge to this solution (Table 2). For training tasks, the network generalizes to unseen (x, y) pairs, which distinguishes this from typical overfitting in supervised learning. We formally define (complete) memorization as:

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

3 Information-Theoretic Meta-Regularization

At a high level, the sources of information in the predictive distribution $q(\hat{y}^*|x^*, \theta, D)$ come from the input, the meta-parameters, and the data. The memorization problem occurs when the model encodes task information in the predictive network that is readily available from the task training set (i.e., it memorizes the task information for each meta-training task). We could resolve this problem by encouraging the model to minimize the training error and to rely on the task training dataset as much as possible for the prediction of y^* (i.e., to maximize $I(\hat{y}^*; D|x^*, \theta)$). Explicitly maximizing $I(\hat{y}^*; D|x^*, \theta)$ requires an intractable marginalization over task training sets to compute $q(\hat{y}^*|x^*, \theta)$. Instead, we can implicitly encourage it by restricting the information flow from other sources (x^* and θ) to $\hat{y}^*$. To achieve both low error and low mutual information between $\hat{y}^*$ and (x^*, θ), the model must use task training data D to make predictions, hence increasing the mutual information $I(\hat{y}^*; D|x^*, \theta)$, leading to reduced memorization.

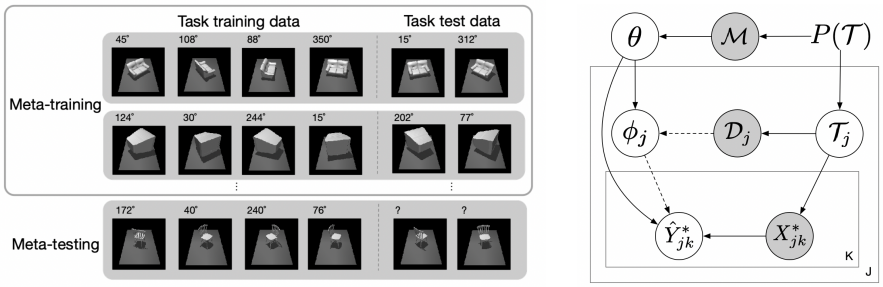

Figure 1: Left: An example of non-mutually-exclusive pose prediction tasks, which may lead to the memorization problem. The training tasks are non-mutually-exclusive because the test data label (right) can be inferred accurately without using task training data (left) in the training tasks, by memorizing the canonical orientation of the meta-training objects. For a new object and canonical orientation (bottom), the task cannot be solved without using task training data (bottom left) to infer the canonical orientation. Right: Graphical model for meta-learning. Observed variables are shaded. 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 (Definition 1).

To achieve both low error and low mutual information between $\hat{y}^*$ and (x^*, θ), the model must use task training data D to make predictions, hence increasing the mutual information $I(\hat{y}^*; D|x^*, \theta)$, leading to reduced memorization. In this section, we describe two tractable ways to achieve this.

Given θ, the statistical dependency between x^* and $\hat{y}^*$ is controlled by the direct path from x^* to $\hat{y}^*$ and the indirect path through $\mathcal{D}$ (see Figure 1), where the latter is desirable. We can control the information flow between x^* and $\hat{y}^*$ by introducing an intermediate stochastic bottleneck variable z^* such that $q(\hat{y}^*|x^*, \phi, \theta) = \int q(\hat{y}^*|z^*, \phi, \theta) q(z^*|x^*, \theta)\, dz^*$ (Alemi et al., 2016) as shown in Figure 2. Now, we would like to maximize $I(\hat{y}^*; \mathcal{D}|z^*, \theta)$ to prevent memorization. We can lower bound this mutual information by

$$I(\hat{y}^*; \mathcal{D}|z^*, \theta) \geq I(x^*; \hat{y}^*|D, \theta) - \mathbb{E}\left[D_{\mathrm{KL}}(q(z^*|x^*, \theta)||r(z^*))\right] \tag{1}$$

where $r(z^*) = \mathcal{N}(z^*; 0, I)$ is a variational approximation to the marginal (see Figure 2 and Appendix A.1 for the proof). In practice, replacing the maximization of $I(x^*; \hat{y}^*|D, \theta)$ with minimization of the training loss, we have

$$\mathbb{E}_{M,D,D^*}\big[\log q(\hat{y}^* = y^*|x^*, \phi, \theta) - \beta D_{\mathrm{KL}}(q(z^*|x^*, \theta)||r(z^*))\big], \tag{2}$$

which is in the form of an information bottleneck (Tishby et al., 2000). Instead of the activation z^*, we can also view the predictor network weights θ as random variables that depend on the stochasticity of training dynamics (Hinton & Van Camp, 1993). Following the decomposition of the cross-entropy loss as in (Achille & Soatto, 2018), we can add $I(y^*_{1:N}, D_{1:N}; \theta|x^*_{1:N})$ as a regularizer to the loss function which measures the amount of information memorized in the weights about the labels that is unrelated to the data distribution. The regularizer can be upper bounded by

$$I(y^*_{1:N}, D_{1:N}; \theta|x^*_{1:N}) = \mathbb{E}_M[\log \tfrac{q(\theta|M)}{q(\theta|x^*_{1:N})}] \leq \mathbb{E}_M D_{\mathrm{KL}}(q(\theta|M)||p(\theta)),$$

with $p(\theta) = \mathcal{N}(\theta; 0, I)$ and $q(\theta|M) = \mathcal{N}(\theta_\mu, \theta_\sigma)$, the meta-regularized objective is

$$\mathbb{E}_{M,D,D^*}\mathbb{E}_{\theta\sim q(\theta|M),\phi\sim q(\phi|D,\theta)}\big[\log q(\hat{y}^* = y^*|x^*, \phi, \theta) - \beta D_{\mathrm{KL}}(q(\theta|M)||p(\theta))\big]. \tag{3}$$

4 Experiments

We evaluate MAML and CNP and their meta-regularized versions MR-MAML(A), MR-CNP(A), MR-MAML(W), and MR-CNP(W). (A) denotes meta-regularization of the activations (Eq. 2) and (W) denotes meta-regularization of the weights (Eq. 3).

4.1 Sinusoid Regression

First, we consider a toy sinusoid regression problem that is mutually-inclusive. The data for each task is created in the following way: the amplitude A of the sinusoid is uniformly sampled from a finite set $\mathcal{A} = \{0.1, 0.3, \ldots, 4\}$ of 20 equally-spaced points, u is sampled uniformly from $[-5, 5]$, and y is sampled from $\mathcal{N}(A\sin(u), 0.1^2)$. For the mutually-inclusive sinusoid regression problem,

we provide both u and the amplitude A (as a one-hot vector) as input. It is straightforward to learn a single network that uses the input A to achieve low training error without using the task training data. At the test time, we expand the range of tasks by sampling the data-generating amplitude A uniformly randomly from a $[0.1, 4]$ and use a random one-hot vector as input to the network. The meta-training tasks are a proper subset of the meta-testing tasks.

Without the additional amplitude input, both MAML and CNP can easily solve the task. However, once we add the additional amplitude input, both MAML and CNP converge to the memorization solution and fail to generalize well to test data (Table 1 and Appendix Figures 4 and 5). Our proposed meta-regularizers encourage the algorithm to use the task training data during meta-training. At the test-time, MR-MAML and MR-CNP greatly outperform the unregularized methods.

Table 1: Test MSE for the non-mutually-exclusive sinusoid regression problem. We compare MAML and CNP against meta-regularized MAML (MR-MAML) and meta-regularized CNP (MR-CNP) where regularization is either on the activations (A) or the weights (W). We report the mean and the standard deviation over 5 trials.

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)**

4.2 Pose Prediction

Next, we created a multi-task regression dataset based on the Pascal 3D data (Xiang et al., 2014). The dataset consists of 10 classes of 3D objects such as "aeroplane", "sofa", "TV monitor", etc. Each class has multiple different objects and there are 65 objects in total. We randomly select 50 objects for meta-training and the other 15 objects for meta-testing. For each object, we randomly select a canonical pose for the object, and use MuJoCo to render images with random orientations of the object on a table (see Figure 1 for an illustration of the problem). The meta-learning algorithm takes the image as input and predicts the orientation relative to the canonical pose.

Table 2: Meta-test MSE for the pose prediction problem. We compare MR-MAML (ours) with conventional MAML and fine-tuning (FT). We report the average over 5 trials and standard deviation in parentheses.

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)

The results on the meta-test set are shown in Table 2. We additionally include fine-tuning as baseline, which trains a single network on all the instances jointly, and then fine-tunes on the task training data. We also compare with weight decay regularization on all the weights. We choose the learning rate from $\{10^{-4}, 5 \times 10^{-4}, 10^{-3}\}$ and β from $\{10^{-6}, 10^{-5}, \cdots, 1\}$ for meta-regularization and report the best result for each method. Meta-learning with meta-regularization (on weights) outperforms all the competing methods by a large margin and has better stability for different training dynamics. We plot the test MSE as a function of β in Appendix Figure 6. Modulating β can shift the optimal solution from the memorization solution to the adaptation solution.

When the meta-regularization is on the activations, the solution that the algorithms converge to depends on the learning rate. Our hypothesis is that the information content of the prediction $\hat{y}^*$ is not large, high likelihood can be achieved with small $I(x^*; \hat{y}^*|\theta)$ which is smaller than the variational bottleneck bound. We find that meta-regularization on the weights does not suffer from this pathology and is robust to different learning rates.

5 Conclusion

We identify the memorization problem and an information-theoretic meta-learning objective that places precedence on data-driven adaptation. This causes the meta-learner to decide what should be learned from data and what must be inferred from the input. By doing so, the algorithm can successfully use experience across mutually-inclusive tasks to quickly adapt to new tasks. We combine our approach with both contextual and gradient-based meta-learning algorithms and apply it in practical settings where meta-learning has not previously been applied, substantially outperforming traditional meta-learning algorithms and transfer learning methods.

References

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1126–1135. JMLR. org, 2017.

Marta Garnelo, Dan Rosenbaum, Chris J Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo J Rezende, and SM Eslami. Conditional neural processes. *arXiv preprint arXiv:1807.01613*, 2018.

Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard E Turner. Meta-learning probabilistic inference for prediction. *arXiv preprint arXiv:1805.09921*, 2018.

Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. *arXiv preprint arXiv:1801.08930*, 2018.

Geoffrey Hinton and Drew Van Camp. Keeping neural networks simple by minimizing the description length of the weights. In *in Proc. of the 6th Ann. ACM Conf. on Computational Learning Theory*. Citeseer, 1993.

Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR 2016*, 2016.

Jurgen Schmidhuber. Evolutionary principles in self-referential learning. *On learning how to learn: The meta-meta-... hook.) Diploma thesis, Institut f. Informatik, Tech. Univ. Munich*, 1:2, 1987.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014.

A Appendix

A.1 Derivation of Meta Regularization on Activation

First we show that $I(x^*; \hat{y}^*|\mathcal{D}, z^*, \theta) \leq I(\hat{y}^*; \mathcal{D}|z^*, \theta)$. By Figure 2, we have that $I(\hat{y}^*; x^*|\theta, \mathcal{D}, z^*) = 0$. By the chain rule of mutual information we have

$$\begin{aligned} I(\hat{y}^*; \mathcal{D}|z^*, \theta) =& I(\hat{y}^*; \mathcal{D}|z^*, \theta) + I(\hat{y}^*; x^*|\mathcal{D}, \theta, z^*) \\ =& I(\hat{y}^*; x^*, \mathcal{D}|\theta, z^*) \\ =& I(x^*; \hat{y}^*|D, \theta, z^*) + I(\hat{y}^*; D|\theta, z^*) \\ \geq& I(x^*; \hat{y}^*|D, \theta, z^*) \end{aligned} \tag{4}$$

Then with the dependency structure as shown in Figure 2, the derivation follows

$$\begin{aligned} I(\hat{y}^*; \mathcal{D}|z^*, \theta) &\geq I(x^*; \hat{y}^*|D, \theta, z^*) = I(x^*; \hat{y}^*|D, \theta) - I(x^*; z^*|D, \theta) + I(x^*; z^*|\hat{y}^*, D, \theta) \\ &\geq I(x^*; \hat{y}^*|D, \theta) - I(x^*; z^*|D, \theta) = I(x^*; \hat{y}^*|D, \theta) - \mathbb{E}_{p(x^*)q(z^*|x^*, \theta, D)} \left[\log \frac{q(z^*|x^*, \theta, D)}{q(z^*|\theta, D)} \right] \\ &= I(x^*; \hat{y}^*|D, \theta) - \mathbb{E}_{p(x^*)q(z^*|x^*, \theta)} \left[\log \frac{q(z^*|x^*, \theta)}{q(z^*|\theta, D)} \right] \\ &\geq I(x^*; \hat{y}^*|D, \theta) - \mathbb{E} \left[\log \frac{q(z^*|x^*, \theta)}{r(z^*)} \right] = I(x^*; \hat{y}^*|D, \theta) - \mathbb{E} \left[D_{D_{\mathrm{KL}}}(q(z^*|x^*, \theta)||r(z^*)) \right] \end{aligned} \tag{5}$$

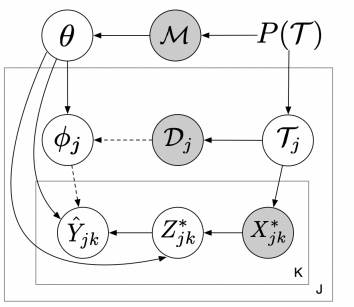

Figure 2: Graphical model of the regularization on activations. Observed variables are shaded and Z is bottleneck variable. The complete memorization corresponds to the graph without the dashed arrows.

A.2 Additional Experimental Results

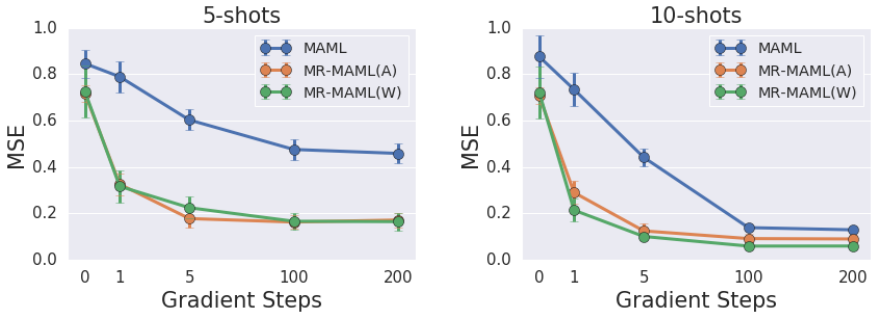

Figure 3: The trace plot of MAML and MR-MAML at the test-time in non-mutually-exclusive sinusoid problem. For each trial, we calculate mean MSE over 100 randomly generated meta-testing tasks. The trace plot show the mean and standard deviation of the results for 5 random trials.

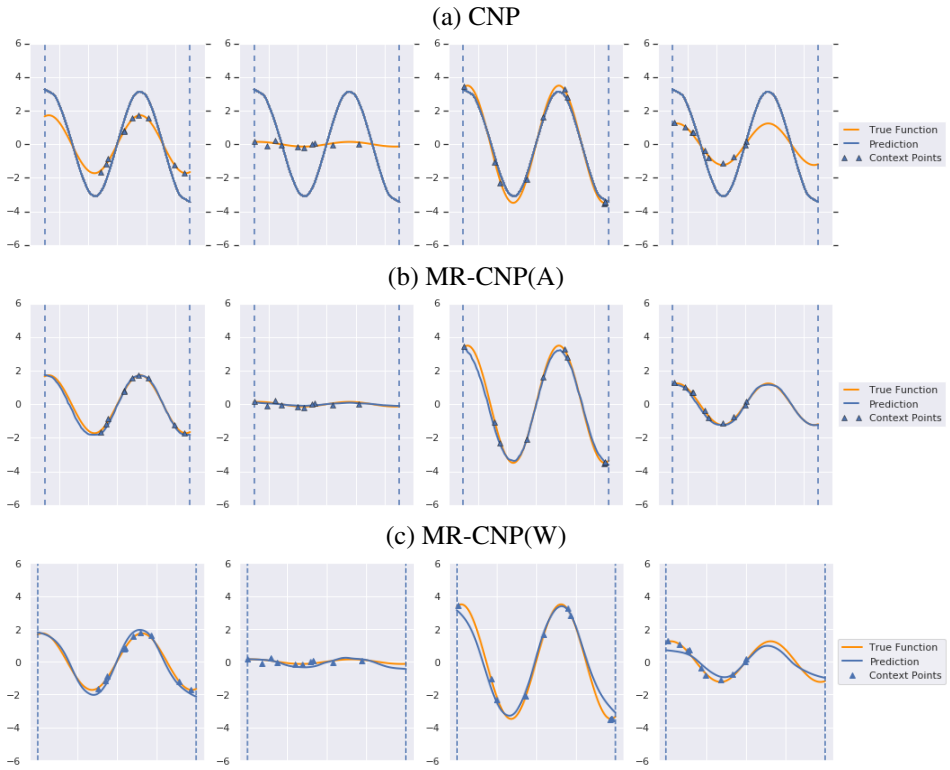

Figure 4: The illustrative results of non-mutually-exclusive sinusoid regression with neural processes at test-time. For each row, the amplitude of true function are set by four random samples from $[0.1, 4]$. The one-hot vector part of input at the test-time is fixed as $\boldsymbol{e}_{10}$ which is 20-way one-hot vector with the 10-th position as 1. (a): The prediction of vanilla CNP is largely determined by one-hot vector part of the input and cannot adapt to new support points at test-time which shows large generalization error for unseen tasks. (b) (c): Adding meta-regularization on both activation and weights can force the CNP to use the support data at meta-training and generalize well at the test-time for the unseen tasks.

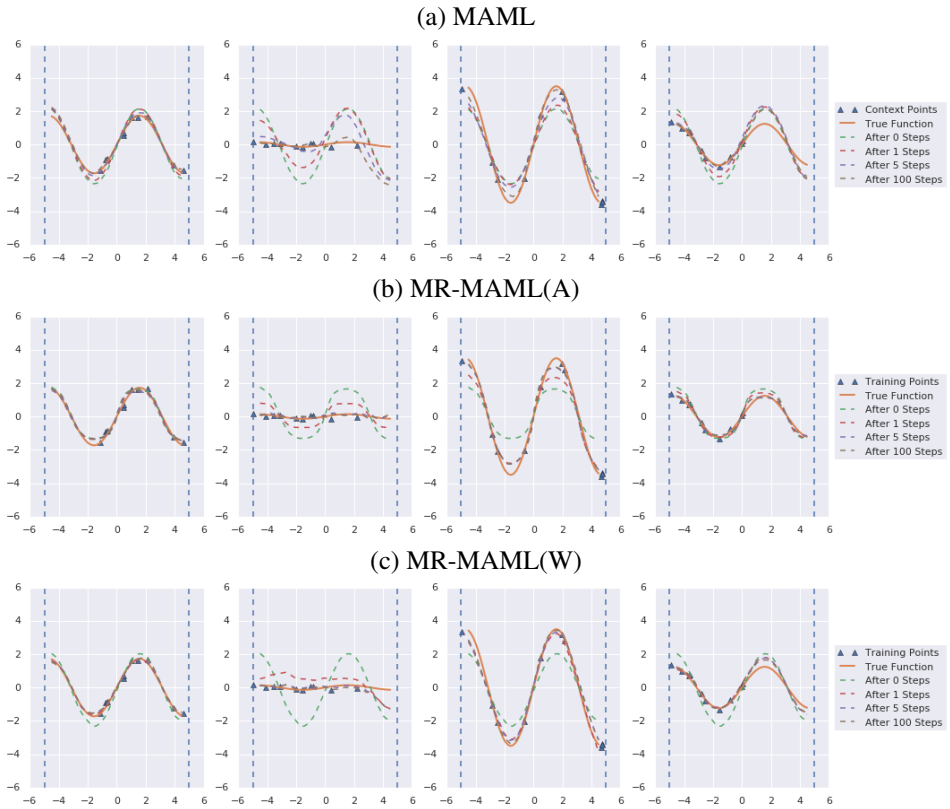

Figure 5: The illustrative results of non-mutually-exclusive sinusoid regression with MAML at test-time. For each row, the amplitude of true function are set by four random samples from $[0.1, 4]$. The one-hot vector part of input at the test-time is fixed as $\boldsymbol{e}_{10}$ which is 20-way one-hot vector with the 10-th position as 1. (a): Due to memorization, MAML adapts slowly at test-time and overfits when the number of data is small. (b) (c): Adding meta-regularization on both activation and weights can recover the capability of fast adaptation.

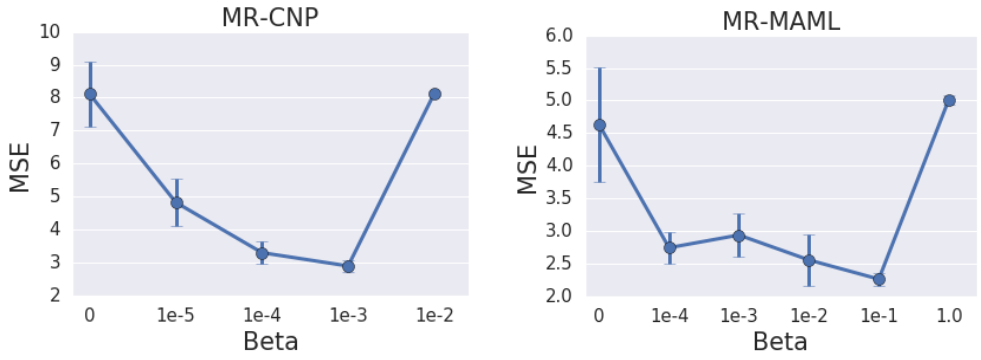

Figure 6: The change of test error with different β. The magnitude of β controls the amount of information in the weights. Small β leads to memorization problem which ignores the task training data while large β forces the weights to contain no information which ignores the meta-training data. A properly chosen β leads to the best generalization. The plot show the mean and standard deviation of the results for 5 random trials.