
Meta-learning Amidst Heterogeneity and Ambiguity

Kyeongryeol Go
 Graduate School of AI
 KAIST
 Daejeon, South Korea
 kyeongryeol.go@kaist.ac.kr

Seyoung Yun
 Graduate School of AI
 KAIST
 Daejeon, South Korea
 yunseyoung@kaist.ac.kr

Abstract

Meta-learning aims to learn a model that can handle multiple tasks generated from an unknown but shared distribution. However, typical meta-learning algorithms have assumed the tasks to be similar such that a single meta-learner is sufficient to aggregate the variations in all aspects. In addition, there has been less consideration on uncertainty when limited information is given as context. In this paper, we devise a novel meta-learning framework, called Meta-learning Amidst Heterogeneity and Ambiguity (MAHA), that outperforms previous works in terms of prediction based on its ability on task identification. By extensively conducting several experiments in regression and classification, we demonstrate the validity of our model, which turns out to be robust to both task heterogeneity and ambiguity.

1 Introduction

Although deep learning models have shown remarkable performance in various domains, they have consistently been criticized because of their sensitivity to the amount of data [8, 39, 9, 56, 21]. Despite all available public data, the data scarcity issue is still not negligible. In many cases, the actual data that is worth analyzing is quite limited for many different reasons, for example, concerns about data privacy [36] and noisy data with anomalies [50]. Along with transfer learning, few-shot learning, and multi-task learning, meta-learning has recently been highlighted as a way to overcome this deficiency with its adaptive behavior using a few data points [60, 23].

Meta-learning aims to handle multiple tasks by efficiently organizing the acquired knowledge. However, typical algorithms have been assessed based on a solid assumption which lacks the representative potential in real-world scenarios. Among many tackles [58, 32], we mainly focus on the following two assumptions. First of all, the tasks are regarded to be similar such that a single meta-learner is sufficient to aggregate the variations in all aspects. It implies that there has been little effort to compactly abstract notions within heterogeneity, one of the essential factors characterizing human intelligence, which is advantageous in decision-making to query the associated information to solve the problem. In addition, there has been less consideration on uncertainty for identifying particular task with a few data points. It is therefore not easy to analyze or transfer the acquired knowledge of the model, which is critical in the growing AI industries, such as a medical diagnosis [1, 5, 62] and autonomous vehicles [26, 52, 6], because a certain level of interpretability is required for greater safety.

In this respect, we hypothesize that a disentanglement in task representation is advantageous, which frequently appears in studies to analyze the inherent factors of variation within the dataset. This is to *i) uncover the distinctive properties as a tool for interpretability and to ii) explicitly separate the dataset into several clusters, which would have been detrimental when trained altogether*. However, as a trade-off for interpretability, the overconfident nature of deep learning may strictly assign the tasks into certain clusters without considering ambiguity, which requires an additional treatment to cope with the anomalies.

To this end, we propose a new meta-learning framework, Meta-learning Amidst Heterogeneity and Ambiguity (MAHA), that performs robustly on the following two huddles. **Task heterogeneity**: there is no clear discrimination between the tasks that are sampled from the faraway modes of task distribution [64, 67, 68]. **Task ambiguity**: too few data points are given to infer the task identity [13, 49]. Specifically, we devise a pre-task built upon the neural processes [15, 16, 25] to obtain well-clustered and interpretable representation. Then, an agglomerative clustering is applied to the representation without any external knowledge such as the number of clusters and separately train a different model for each cluster. Please refer to Figure 7 for the overall training process of MAHA.

To summarize, the main contributions of this paper are the following 4-folds:

- We propose a simple yet powerful architecture design for the neural processes to better leverage the latent variables and be applicable in classification. (See Section 5.1)
- We resolve the information asymmetry in the neural processes and construct well-clustered and interpretable representations. (See Section 5.2)
- We validate MAHA through both regression and classification, by which the experimental results demonstrate its ability to cope with the heterogeneity and ambiguity. (See Section 6)
- We devise an additional regularization term for the low-shot regime that distills an obtainable knowledge from relatively various training samples and variations. (See Appendix B)

2 Related work

Gradient-based meta-learning, represented by MAML [12], aims to learn the prior parameters that can quickly adapt to certain tasks through several gradient steps. It consists of the inner-loop for the task adaptation and the outer-loop for the meta-update over tasks. Many variants have emerged to balance generalization and customization in a task-adaptive manner. To begin with a generalization perspective, [13, 27] suggested probabilistic extensions through the hierarchical Bayesian model and Stein variational gradient descent (SVGD) [37]. In addition, [49] conducted the inner-loop on the low-dimensional latent embedding space, and [70] proposed the meta-regularization that was built on information theory. From a customization perspective, [35] divided the parameters into two categories, one of which is shared across tasks, and the other can be modulated task-specifically. [74] was informed by the layer-wise adaptive units, and [64, 67, 68, 69] considered the auxiliary networks that modulate the initial parameter before the inner-loop.

The family of neural processes, also known as contextual meta-learning, is devised to imitate the flexibility of the Gaussian Process [44] while resolving the scalability issue. Rather than explicitly modeling the kernel to conduct the Bayesian inference like [65], it learns an implicit kernel directly from data which overcomes the design restrictions. Task-specific information is extracted from the subset of data through an encoder, which is then aggregated for utilization in the decoder to predict the corresponding outputs of the remaining data. Starting from the conditional neural process (CNP) [15], which was built solely on a deterministic path, the neural process (NP) [16] applies the addition of a stochastic path. The attentive neural process (ANP) [25] further applies an attention mechanism to resolve the underfitting issue in NP by enlarging the locally adaptive behavior. More complex modules, such as a graph structure [38] and recurrent neural network [30, 53], were further considered to capture the dependencies on latent variables and the complex temporal dynamics.

However, many problems remain unsolved. Firstly, the neural processes yet rely on a complex feature extractor to enable task-specific modulation, which requires various regularization techniques with additional hyperparameters [47]. Furthermore, whereas the neural processes are able to obtain an explicit task representation, the existing approaches have investigated little regarding interpretability. Finally, the performance analysis has been mainly focused on regression [31, 25, 53, 57, 19], and some are not even directly applicable for classification [34].

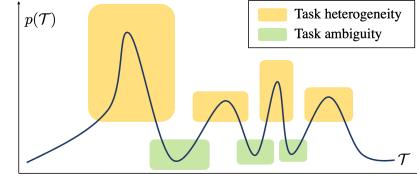


Figure 1: Heterogeneity and ambiguity occurred in task distribution. Those are not independent concepts, but the ambiguity naturally comes after the heterogeneity.

3 Problem setting

Let $C = \{C_x, C_y\}$ be the context set, and let $T = \{T_x, T_y\}$ be the target set, where both C and T are sampled from the same task $\mathcal{T} \sim p(\mathcal{T})$. A common goal in meta-learning is to devise an algorithm for the model $f(\cdot)$ that appropriately uses the model parameter θ to obtain the task-specific parameter ϕ according to the input-output pairs in C such that when T_x is given, T_y can be accurately estimated with high confidence. For example, in MAML [12], a task-specific parameter can be computed by using a gradient step $\phi = \theta - \alpha \cdot \nabla_\theta L(f(C_x; \theta), C_y)$. On the other hand, in CNP [15], θ and ϕ no longer share the same parameter space. Here, the model parameter is divided into an encoder and a decoder part $\theta = \{\theta_{\text{enc}}, \theta_{\text{dec}}\}$, and the task-specific parameter can be computed by the encoder output $\phi = f_{\text{enc}}(C; \theta_{\text{enc}})$. Hereafter, we omit θ for brevity.

For model training, ϕ is iteratively updated using *batchs*. Here, each *batch* is constructed through multiple tasks that are characterized by *way* and *shot*. If there are N classes, each of which contains K input-output pairs, we call it an N -way K -shot problem. The class labels are shuffled in classification whenever a task instance is created, which encourages a meta-learning algorithm to learn how to classify images even when the configuration of unseen classes occurs.

4 Preliminary : (Attentive) Neural Process

In Figure 2, we summarize how a basic family of neural processes has evolved in terms of the graphical model. The encoder comprises a deterministic path and stochastic path computing the task-specific parameter $\phi = \{r, z\}$ of the variational distributions which we denote by $q(r|X, Y) = \mathcal{N}(r, 0)$ and $q(z|X, Y) = \mathcal{N}(\mu_z, 0.1 + 0.9 \cdot \text{sigmoid}(\omega_z))$.¹ Here, $\{X, Y\}$ indicates a set of input-output pairs and a reparameterization trick is applied at the end of the stochastic path for differentiable non-centered parameterization [28].

For both paths, NP is constructed by:

$$\begin{aligned} r &= \text{MeanPool}_{\text{shot}}(\text{rFF}(\{X, Y\})) \\ [\mu_z, \omega_z] &= \text{MeanPool}_{\text{shot}}(\text{rFF}(\{X, Y\})) \end{aligned}$$

where $\text{MeanPool}(\cdot)$ is a mean-pooling operation along the subscripted dimension, $\text{rFF}(\cdot)$ can be any row-wise feedforward layer, such as Multi-Layer Perceptron (MLP), and $[\cdot]$ denotes the concatenation. On the other hand, ANP exploits the multi-head attention, connecting T_x to r in graphical model, and self-attention, both of which are proposed in [61]. As in NP, the value of z is same for every *shot* of T_x , however, based on the attention score with each element of X , r is now computed in *shot*-dependent manner.

Then, conditioned on the encoder outputs, r and z , with the target input T_x , the decoder computes the parameters of predictive distribution on the target output T_y :

$$[\mu_{T_y}, \omega_{T_y}] = \text{rFF}([T_x, r, z]) \quad (1)$$

where the predictive distribution is expressed as $p(T_y|T_x, r, z) = \mathcal{N}(\mu_{T_y}, 0.1 + 0.9 \cdot \text{softplus}(\omega_{T_y}))$. Eventually, relying on the variational inference, one can obtain the loss function which approximates the negative ELBO by replacing an intractable $p(z|C)$ with the variational distribution $q(z|C)$ following [16]:

$$\mathcal{L}_{(A)NP} = -\mathbb{E}_{q(r|C)q(z|T)} [\log p(T_y|T_x, r, z)] + \beta_1 KL(q(z|T)\|q(z|C)) \quad (2)$$

As a result, based on the Kolmogorov extension and de-Finetti theorems, the neural processes become a stochastic process that satisfies the exchangeability and consistency [16]. However, when trained using the deterministic path, the neural processes with latent variables is empirically shown to have difficulty capturing the variability of the stochastic process [31], of which causes are investigated and resolved in Section 5.2.

¹Note that r is deterministic with zero variance.

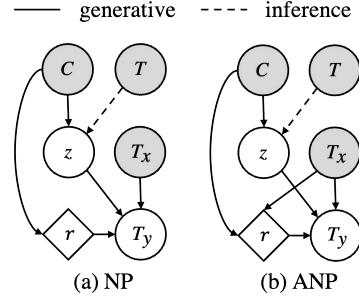


Figure 2: Graphical model of the related baselines. Circles denote random variables, whereas diamonds denote deterministic variables. Shaded variables are observed during the test phase, and every in-between edge is implemented as a neural network.

5 Meta-learning Amidst Heterogeneity and Ambiguity

This section describes our algorithm MAHA whose primary focus is to devise a pre-task to cope with task heterogeneity and ambiguity in meta-learning. We first introduce an encoder-decoder pipeline of MAHA, namely FELD, of which effects are examined by substituting the correspondent within NP in Section 6. Then, a dimension-wise pooling and an auto-encoding structure are proposed to obtain well-clustered and interpretable representation. Finally, the training process of MAHA is described, which applies to both regression and classification.

5.1 Encoder-decoder pipeline

Flexible Encoder Although the attention mechanism proposed in ANP was a key to resolve the underfitting in NP, there is less incentive for r to focus on task identity that is shared across $shots$. As a result, in Figure 3, ANP appears to strongly fit the given input-output pairs, which leads to a wiggly prediction. Particularly within task heterogeneity and ambiguity where the prediction space is prone to be highly variable, the wiggly prediction of ANP leads to a poor generalization performance (See Figure 8). Therefore, the graphical model of NP is rather considered in MAHA since its latent variables are $shot$ -independent. Then, based on analysis in [10], the problematic underfitting is dealt with by substituting the encoder with the flexible and permutation-invariant Set Transformer (ST) [33]. Note that the Set Transformer can incorporate the $rFF(\cdot)$ and $\text{MeanPool}_{shot}(\cdot)$ in the encoder of NP. See Appendix A for a more detailed explanation about the modules in Set Transformer.

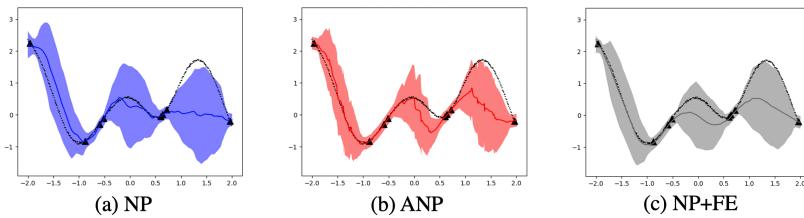


Figure 3: Qualitative comparison between NP, ANP, and NP with the flexible encoder (NP+FE) on functions generated from Gaussian Process. The shaded areas correspond to the ± 2 standard deviations. Prediction of ANP turns out to be wiggly, while NP and NP+FE are relatively smooth following Occam’s razor. Note that quantitative comparison can be looked up in Table 1.

Linear Decoder We avoid using a complex decoder such as [41] and apply feature-wise linear modulation to the target input T_x . Inspired by [72], we composite the latent variables using a skip connection. Among the many normalization techniques, a layer normalization [3] is applied since the statistic is computed independently for each *batch* instance such that only z can capture the heterogeneity in accordance with the pooling proposed in Section 5.2.

$$[\mu_{T_y}, \omega_{T_y}] \quad \text{or} \quad \text{logit} = g(T_x) \cdot W \quad \text{where} \quad W = w(r, z) = \text{LN}(r + rFF(z))^T \quad (3)$$

Here, $g(\cdot)$ implies any feature extractor, $\text{LN}(\cdot)$ indicates a layer normalization, and the transpose operation T permutes the last two dimensions of the tensor. It is aligned with the previous approaches [4, 51, 66, 20] which weaken the decoder to allow the latent variables to be appropriately leveraged. Also, it relates to studies on few-shot classification [18, 47] where each column of W is computed by $shots$ within the same *way*. However, when accompanied with the pooling in Section 5.2, the columns are no more independent by one another and share information across *way*.

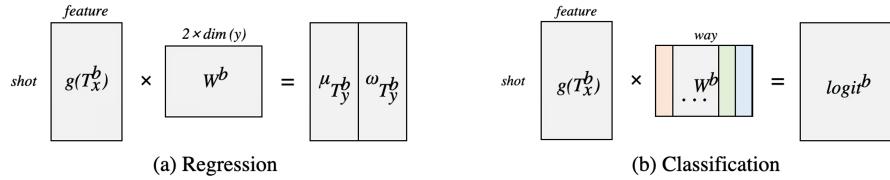


Figure 4: Prediction on output distribution. Superscript b indicates the b -th batch instance.

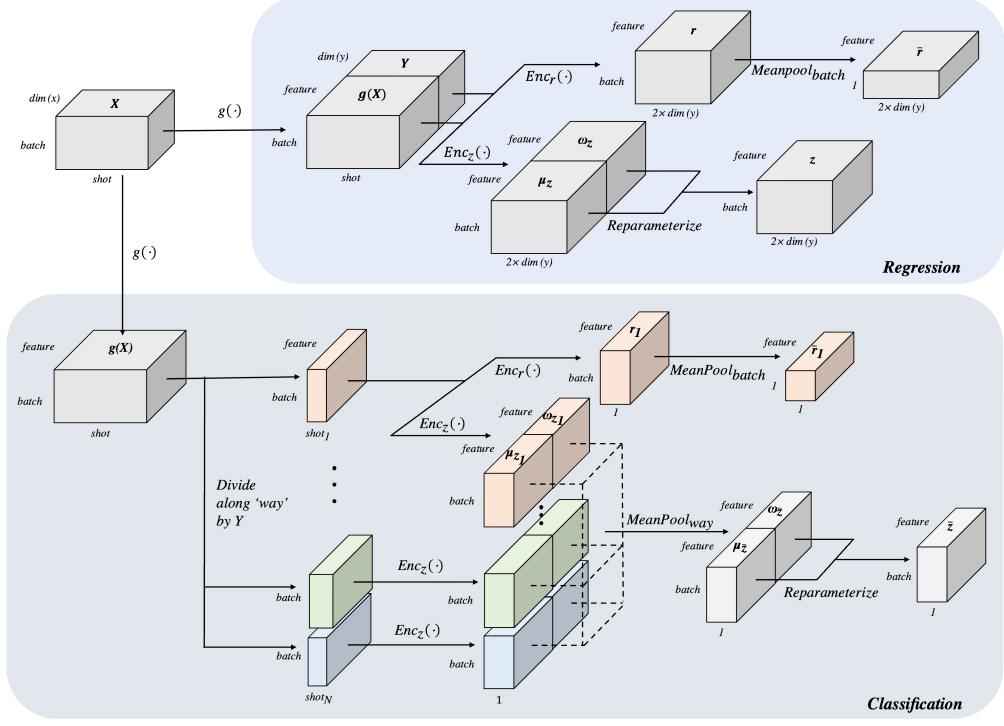


Figure 6: Computational diagram for \bar{r} and \bar{z} . For visual comfort, every block of the encoder outputs in regression is reshaped from $[batch, 1, 2 \times dim(y) \times feature]$ into $[batch, 2 \times dim(y), feature]$. In classification, $shot$ dimension is divided along ‘way’ with subscript $\{1, \dots, N\}$ and $\bar{r} = [\bar{r}_1, \dots, \bar{r}_N]$.

5.2 Inducing disentanglement on z

For NP and ANP trained on functions generated from GP, we illustrate the weight norm of the decoding layer right behind the latent variables in Figure 5. The sparsely-coded decoder implies the redundancy of the stochastic path due to the component collapsing behavior referred to in [40, 24]. This phenomenon can be explained by the information preference problem [7, 73] where the information flow is concentrated on the deterministic path with the tendency to ignore the stochastic path.

In order to handle the information asymmetry, several solutions were proposed in studies on the generative model, such as the KL annealing scheduler [4, 14] and expressive posterior approximation [48, 29], but these are generally not robust to changes in model architecture. Instead, we propose a simple method to avoid redundancy of the stochastic path by encouraging it to acquire multi-modality within heterogeneity and ambiguity.

Dimension-wise pooling We explicitly capture the distinct variations within the information flow by pooling each path across different dimensions, $batch$ for r and way for z :

$$\bar{r} = \text{MeanPool}_{batch}(r) \quad \text{and} \quad [\mu_{\bar{z}}, \omega_{\bar{z}}] = \text{MeanPool}_{way}([\mu_z, \omega_z]) \quad (4)$$

Then, the deterministic representation \bar{r} becomes identical not only across $shot$, but also across $batch$. Then, whenever it is insufficient to handle all variations across tasks within the same $batch$ i.e., facing task heterogeneity, the model should resort to the stochastic representation \bar{z} since the deterministic representation only captures the average properties. On the other hand, the stochastic representation \bar{z} allows the different way to share information and becomes class-invariant. We illustrate how the latent variables \bar{r} and \bar{z} are computed in Figure 6. Note that the value of way is set to 1 in regression such that pooling on z is negligible.

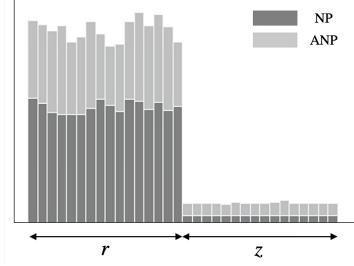


Figure 5: Stacked bar plot for the weight norm of the decoding layer

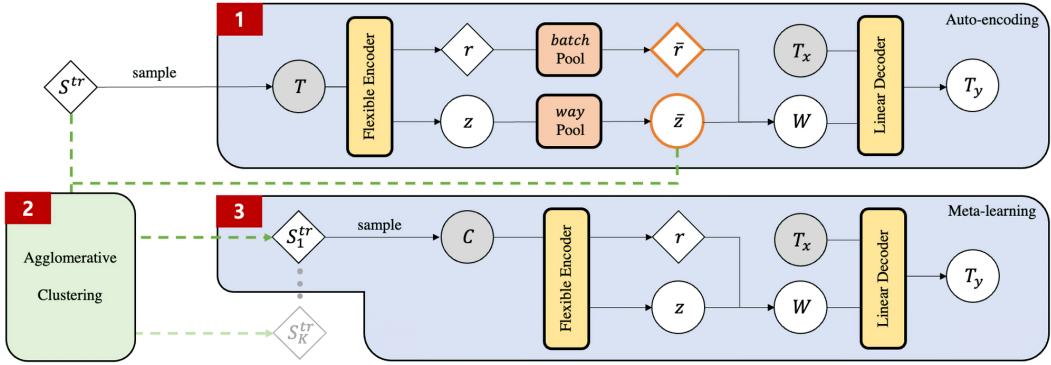


Figure 7: MAHA. K is the number of estimated clusters such that the meta-train set $S^{tr} = \bigcup_{k=1}^K S_k^{tr}$.

Auto-encoding structure Empirically, we observe that the KL collapse [4, 2, 55, 73] does not occur whenever the pooling operations is used (see Appendix D). This implies that the posterior $q(\bar{z}|T)$ does not simply converge to the approximate prior $q(\bar{z}|C)$ so that the decoder gets dependent on the stochastic path. However, there is still an incentive for \bar{r} to be underutilized during the decoding because it is inferred by small C not large T [22] and neural networks exploiting set representation is known to poorly perform in low-shot regime [11, 71] i.e., facing task ambiguity.

Thereby, we resort to the conditional auto-encoding structure [54] on top of the dimension-wise pooling to cope with the lack of training samples. As a result, the following loss function is derived which differs from Equation 2 on *i) whether the pooling operations are used or not* and *ii) which set is used to compute the deterministic representation*, each of which is the result of the dimension-wise pooling and the auto-encoding structure:

$$\mathcal{L}_{pre} = -\mathbb{E}_{q(\bar{r}|T)q(\bar{z}|T)} [\log p(T_y|T_x, \bar{r}, \bar{z})] + \beta_2 KL(q(\bar{z}|T)\|q(\bar{z}|C)) \quad (5)$$

5.3 Training process

See Figure 7. Initially, the dimension-wise pooling and the auto-encoding structure proposed in Section 5.2 are used along with FELD to minimize the loss function in Equation 5. Next, an agglomerative clustering is applied to the disentangled representation from the stochastic path to estimate the number of clusters with the highest purity value.² Finally, for each cluster, separate FELD is trained from the beginning by Equation 2 where the tasks are no longer uniformly sampled but statistically skewed based on the ratio of heterogeneous tasks within the cluster. According to the Euclidean distance to the cluster centers, FELD in correspondence to the closest cluster is exploited for evaluation.

6 Experiment

We first experiment on frequently appearing benchmark datasets in meta-learning and investigate the role of the encoder-decoder pipeline (FELD) by gradually adjusting NP. Those datasets are generally regarded to be homogeneous such that the MAHA is equivalent to FELD when assuming a single cluster as noted in Section 5.3. After that, MAHA is evaluated on heterogeneous datasets following the experimental setting of [67] with the dimension-wise pooling and the auto-encoding structure in Section 5.2, of which roles are examined in both quantitative and qualitative manner. Please refer to Appendix C for details about the data-split, architecture design, and the hyperparameter search.

Overall, we are to answer the following three questions:

- Does MAHA outperform the previous baselines in terms of prediction? (See Table 1 to 5)
- What are the benefits of using the flexible encoder and the linear decoder? (See Section 6.1)
- How does the dimension-wise pooling and the auto-encoding structure contribute to obtaining well-clustered representation within heterogeneity? (See section 6.2)

²For a homogeneous dataset, a single cluster is available such that the previous steps can be omitted.

6.1 Homogeneous dataset

Gaussian Process Following the basic neural processes [15, 16, 25], we consider functions generated from GP with squared exponential kernel $k(x, x') = \sigma^2 \exp(-0.5(x - x')^2/l^2)$. The experimental result in Table 1 states that although ANP performs better than NP in terms of flexibility, the dominance no longer holds when NP is equipped with the flexible encoder. However, a degradation in performance is shown when using the linear decoder in NP. This is empirical evidence that NP strongly relies on the complexity of the decoder in regression, by which the model is prone to ignore the latent variables [7, 73]. By exploiting the flexible encoder to obtain more informative latent variables by themselves such that the (shallow) linear decoder is just enough for prediction, FELD performs better than any other models with the (deep) conventional decoder. We find the Set Transformer is the perfect choice whose improvement can not be caught up by simply stacking MLPs. Moreover, it is noticeable that FELD outperforms NP+FE despite a decreased model capacity.

Mini-ImageNet, Tiered-ImageNet Similar tendency can be observed in classification. We consider mini-ImageNet [63] and tiered-ImageNet [46], which are frequently used large-scale datasets for few-shot image classification. For mini-ImageNet, we follow the split of [45], which assigns 64 classes for the meta-train set, 16 classes for the meta-valid set, and 20 classes for the meta-test set. For tiered-ImageNet, 608 classes are first grouped into 34 higher-level nodes, divided into 20, 6, and 8 nodes to construct the meta-train set, meta-valid set, and meta-test set. We use the feature provided by [49], which is obtained by pre-training a deep residual network in a supervised manner as in [17, 42, 43]. However, unlike [43, 49], the meta-valid set is used for early stopping and hyperparameter search but not utilized to update the parameters.

In Table 2, 3, accuracy on mini-ImageNet and tiered-ImageNet is reported. We collect the score of various baselines that use either convolutional networks or deep residual networks and do not exploit any data augmentation for a fair comparison. While NP performs no better than a random guess when following [15], NP+LD results in a comparable score to the recent models in gradient-based meta-learning, verifying the validity of the linear decoder in classification. FELD achieves even better performance than the state-of-the-art, which is remarkable in the sense that the attention modules in Set Transformers can not be fully utilized in low-shot regime.

6.2 Heterogeneous dataset

Sine & Polynomial To verify the performance on the family of functions, we experiment on the toy 1D regression as in [64, 67, 68]. In particular, we follow the exact setting of [67] where each task is randomly chosen to be one of the following one-dimensional functions where the coefficients are uniformly sampled from the prefixed intervals summarized in Appendix C.1: (sine) $y = A_s \sin(B_s x) + C_s$, (line) $y = A_l x + B_l$, (quad) $y = A_q x^2 + B_q x + C_q$, (cubic) $y = A_c x^3 + B_c x^2 + C_c x + D_c$. A small number of data points are given as context, requiring the model to appropriately interpolate and extrapolate in a highly variable prediction space.

Table 1: MSE on Gaussian Process

MODEL	FE	LD	MSE
NP			0.166 ± 0.002
ANP			0.142 ± 0.002
NP+FE	✓		0.138 ± 0.002
NP+LD		✓	0.312 ± 0.002
FELD	✓	✓	0.130 ± 0.002

Table 2: Accuracy on mini-ImageNet

MODEL	5-WAY 1-SHOT	5-WAY 5-SHOT
MATCHING NET	$43.40 \pm 0.78\%$	$51.09 \pm 0.71\%$
META-LSTM	$43.44 \pm 0.77\%$	$60.60 \pm 0.71\%$
MAML	$48.70 \pm 1.84\%$	$63.11 \pm 0.92\%$
PROTONET	$49.42 \pm 0.78\%$	$68.20 \pm 0.66\%$
REPTILE	$49.97 \pm 0.32\%$	$65.99 \pm 0.58\%$
RELATION NET	$50.44 \pm 0.82\%$	$65.32 \pm 0.70\%$
CAVIA	$51.82 \pm 0.65\%$	$65.85 \pm 0.55\%$
VERSA	$53.40 \pm 1.82\%$	$67.37 \pm 0.86\%$
TPN	$55.51 \pm 0.86\%$	$69.86 \pm 0.65\%$
META-SGD	$54.24 \pm 0.03\%$	$70.86 \pm 0.04\%$
SNAIL	$55.71 \pm 0.99\%$	$68.88 \pm 0.92\%$
NP+LD	$57.30 \pm 0.06\%$	$75.10 \pm 0.04\%$
TADAM	$58.50 \pm 0.30\%$	$76.70 \pm 0.30\%$
LEO	$61.76 \pm 0.08\%$	$77.59 \pm 0.12\%$
FELD	$62.77 \pm 0.05\%$	$81.15 \pm 0.03\%$

Table 3: Accuracy on tiered-ImageNet

MODEL	5-WAY 1-SHOT	5-WAY 5-SHOT
MAML	$51.67 \pm 1.81\%$	$70.30 \pm 0.08\%$
PROTONET	$53.31 \pm 0.89\%$	$72.69 \pm 0.74\%$
RELATION NET	$54.48 \pm 0.93\%$	$71.32 \pm 0.78\%$
WARP-MAML	$57.20 \pm 0.90\%$	$74.10 \pm 0.70\%$
TPN	$57.41 \pm 0.94\%$	$71.55 \pm 0.74\%$
META-SGD	$62.95 \pm 0.03\%$	$79.34 \pm 0.06\%$
NP+LD	$63.36 \pm 0.06\%$	$80.50 \pm 0.04\%$
LEO	$66.33 \pm 0.05\%$	$81.44 \pm 0.09\%$
FELD	$66.87 \pm 0.06\%$	$83.54 \pm 0.04\%$

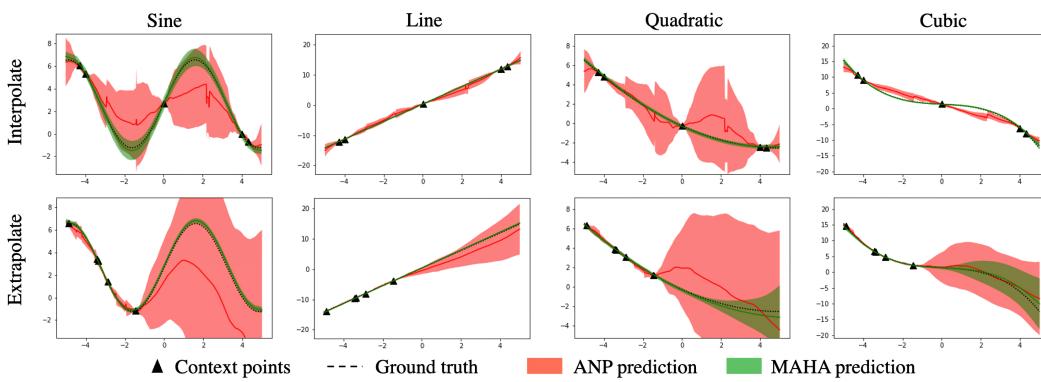


Figure 8: Qualitative comparison of ANP and MAHA on various function types. The context points are selected from 40% of the entire domain for extrapolation.

In Table 4, MSE over 4000 tasks are presented with 95% confidence interval. Generally, all the gradient-based meta-learning algorithms are outperformed by the neural processes, and a noticeable gain is again observed by solely exploiting the encoder-decoder pipeline, FELD. By adjusting FELD to MAHA by task clustering and MAHA to MAHA* by knowledge distillation, a monotonic improvement is observed.³

In Figure 8, we illustrate the interpolation and extrapolation of MAHA in comparison to ANP. As noted in Section 5.1, the main interest of ANP is shown to fitting the context points, which poorly perform in predicting the target outputs whose corresponding inputs are located farther away from that of the context points. This tendency can be observed during interpolation and extrapolation, leading to a wiggly prediction with significant variance. By contrast, MAHA can correctly infer the functional shape, which can be confirmed through a consistently low variance.

Multi-dataset Four distinct fine-grained image classification datasets are combined to construct the multi-dataset proposed in [67]: (Bird) CUB-200-2011, (Texture) Describable Textures Dataset, (Aircraft) FGVC of Aircraft, and (Fungi) FGVCx-Fungi. Compared to a homogeneous setting, this is more challenging since overfitting to a particular dataset can critically harm the performance. For the feature extractor, we followed [67] where 2-Conv blocks are used for task clustering, and 4-Conv blocks are used for prediction.

In Figure 9, for 1-shot setting, mean value of the variational distribution $q(\bar{z}|C)$ is visualized through t-SNE [59]. Without external knowledge, such as the number of true clusters, the embeddings get interpretable when using both the dimension-wise pooling and the auto-encoding structure. The distinct datasets are no more clearly discriminated without either of them, which is quantitatively demonstrated by the estimated purity values in the bottom table. Note that the validity of the methodologies stands out particularly in low-shot regime which implies the difficulty of task identification within ambiguity.

Table 4: MSE on Sine & Polynomial

MODEL	5-SHOT	10-SHOT
BMAML	2.435 ± 0.130	0.967 ± 0.056
MAML	2.205 ± 0.121	0.761 ± 0.068
META-SGD	2.053 ± 0.117	0.836 ± 0.065
MT-NET	2.016 ± 0.019	0.698 ± 0.054
MUMOMAML	1.096 ± 0.085	0.256 ± 0.028
HSMIL	0.856 ± 0.073	0.161 ± 0.021
NP	0.514 ± 0.051	0.089 ± 0.015
ANP	0.415 ± 0.046	0.058 ± 0.016
FELD	0.118 ± 0.015	0.008 ± 0.002
MAHA	0.077 ± 0.006	0.003 ± 0.001
MAHA*	0.056 ± 0.003	0.002 ± 0.001

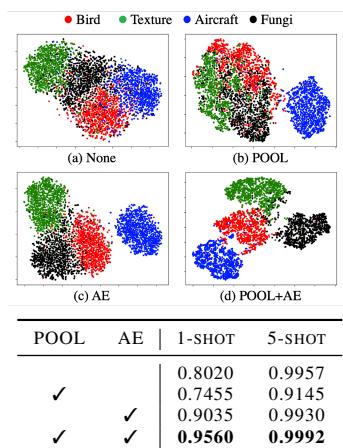


Figure 9: t-SNE of $\mu_{\bar{z}}$ from $q(\bar{z}|C)$ and the estimated purity values

³We handle the overconfident nature of deep learning to better cope with the ambiguity by distilling an obtainable knowledge from T to C . Please refer to Appendix B for a more detailed explanation.

Table 5: Accuracy on multi-dataset

	MODEL	BIRD	TEXTURE	AIRCRAFT	FUNGI	AVERAGE
5-WAY 1-SHOT	MAML	53.94 \pm 1.45%	31.66 \pm 1.31%	51.37 \pm 1.38%	42.12 \pm 1.36%	44.77%
	META-SGD	55.58 \pm 1.43%	32.38 \pm 1.32%	52.99 \pm 1.36%	41.74 \pm 1.34%	45.67%
	MT-NET	58.72 \pm 1.43%	32.80 \pm 1.35%	47.72 \pm 1.46%	43.11 \pm 1.42%	45.59%
	BMAML	54.89 \pm 1.48%	32.53 \pm 1.33%	53.63 \pm 1.37%	42.50 \pm 1.33%	45.89%
	MUMOMAML	56.82 \pm 1.49%	33.81 \pm 1.36%	53.14 \pm 1.39%	42.22 \pm 1.40%	46.50%
	HSML	60.98 \pm 1.50%	35.01 \pm 1.36%	57.38 \pm 1.40%	44.02 \pm 1.39%	49.35%
	ARML	62.33 \pm 1.47%	35.65 \pm 1.40%	58.56 \pm 1.41%	44.82 \pm 1.38%	50.34%
	FELD	56.17 \pm 0.64%	35.86 \pm 0.41%	53.03 \pm 0.58%	45.41 \pm 0.58%	47.61%
	MAHA	63.89 \pm 0.34%	37.22 \pm 0.23%	58.90 \pm 0.44%	47.95 \pm 0.34%	51.99%
	MAHA*	64.45 \pm 0.36%	37.83 \pm 0.23%	59.18 \pm 0.43%	48.33 \pm 0.33%	52.41%
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%	57.78%
	META-SGD	67.87 \pm 0.74%	45.49 \pm 0.68%	66.84 \pm 0.70%	52.51 \pm 0.81%	58.18%
	MT-NET	69.22 \pm 0.75%	46.57 \pm 0.70%	63.03 \pm 0.69%	53.49 \pm 0.83%	58.08%
	BMAML	69.01 \pm 0.74%	46.06 \pm 0.69%	65.74 \pm 0.67%	52.43 \pm 0.84%	58.31%
	MUMOMAML	70.49 \pm 0.76%	45.89 \pm 0.69%	67.31 \pm 0.68%	53.96 \pm 0.82%	59.41%
	HSML	71.68 \pm 0.73%	48.08 \pm 0.69%	73.49 \pm 0.68%	56.32 \pm 0.80%	62.39%
	ARML	73.34 \pm 0.70%	49.67 \pm 0.67%	74.88 \pm 0.64%	57.55 \pm 0.82%	63.86%
	FELD	77.63 \pm 0.46%	55.80 \pm 0.38%	75.88 \pm 0.41%	63.68 \pm 0.50%	68.24%
	MAHA	75.04 \pm 0.26%	54.39 \pm 0.21%	79.98 \pm 0.20%	65.09 \pm 0.25%	68.62%
	MAHA*	75.82 \pm 0.26%	54.28 \pm 0.22%	79.91 \pm 0.19%	65.18 \pm 0.25%	68.79%

The tendency can be observed by the performance measure presented in Table 5. Compared to *1-shot* setting where a noticeable gain is occurred by task clustering, in *5-shot* setting, there is almost no difference between FELD and MAHA. This is because the models can clearly identify the tasks regardless of whether the pooling or the auto-encoding structure is used or not, demonstrated by the high purity values. Accordingly, the knowledge distillation, which is fundamentally devised to regularize the model within ambiguity appropriately, has shown a worthwhile improvement from MAHA to MAHA* particularly in *1-shot* setting. Eventually, MAHA (and MAHA*) beats all the previous works with a fairly large margin and achieves state-of-the-art performance.

7 Conclusion

This paper proposes a new meta-learning framework, MAHA, that performs robustly amidst heterogeneity and ambiguity. We aim to disentangle the stochastic representation by the dimension-wise pooling and the auto-encoding structure based on the newly devised encoder-decoder pipeline to better leverage the latent variables. With the multi-step training process, comprehensive experiments are conducted on regression and classification. In the end, we argue that the proposed model captures the task identity with lower variance, leading to a noticeable improvement in performance. The potential limitation of MAHA would be the additional computational cost from the flexible encoder composed of multiple attention modules. However, by orthogonally applying to the existing work, the compatibility and the necessity are empirically verified. An interesting future work would be to apply our model to reinforcement learning. In particular, training a policy directly from well-clustered representations for sample-efficient exploration seems promising in an environment with sparse rewards.

Broader Impact

When training meta-learning models, there comes a customization process based on the problem at hand. If not using the benchmark datasets that frequently appear in academia, it becomes unclear to which extent the distinct datasets should be combined, expecting the model to be versatile on every possible task generation. MAHA, in this respect, can guide for a human to analyze and cluster the available data into separate clusters. Moreover, MAHA mainly benefits future AI industries where the limited communication between the decentralized servers is available as it can infer the global context even with a small amount of information. As a result, we do not expect any negative societal impacts, but we believe that MAHA possesses many implications in more realistic scenarios.

References

- [1] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pages 559–560, 2018.
- [2] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.
- [3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [4] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.
- [5] Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*, 28(3):231–237, 2019.
- [6] Jianyu Chen, Shengbo Eben Li, and Masayoshi Tomizuka. Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning. *arXiv preprint arXiv:2001.08726*, 2020.
- [7] Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.
- [8] Xue-Wen Chen and Xiaotong Lin. Big data deep learning: challenges and perspectives. *IEEE access*, 2: 514–525, 2014.
- [9] Junghwan Cho, Kyewook Lee, Ellie Shin, Garry Choy, and Synho Do. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? *arXiv preprint arXiv:1511.06348*, 2015.
- [10] Chris Cremer, Xuechen Li, and David Duvenaud. Inference suboptimality in variational autoencoders. *arXiv preprint arXiv:1801.03558*, 2018.
- [11] Harrison Edwards and Amos Storkey. Towards a neural statistician. *arXiv preprint arXiv:1606.02185*, 2016.
- [12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017.
- [13] Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. In *Advances in Neural Information Processing Systems*, pages 9516–9527, 2018.
- [14] Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145*, 2019.
- [15] 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.
- [16] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018.
- [17] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4367–4375, 2018.
- [18] Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard E Turner. Meta-learning probabilistic inference for prediction. *arXiv preprint arXiv:1805.09921*, 2018.
- [19] Jonathan Gordon, Wessel Bruinsma, Andrew YK Foong, James Requeima, Yann Dubois, and Richard E Turner. Convolutional conditional neural processes. 2020.
- [20] Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging inference networks and posterior collapse in variational autoencoders. *arXiv preprint arXiv:1901.05534*, 2019.
- [21] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

- [22] Luke B Hewitt, Maxwell I Nye, Andreea Gane, Tommi Jaakkola, and Joshua B Tenenbaum. The variational homoencoder: Learning to learn high capacity generative models from few examples. *arXiv preprint arXiv:1807.08919*, 2018.
- [23] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*, 2020.
- [24] Weonyoung Joo, Wonsung Lee, Sungrae Park, and Il-Chul Moon. Dirichlet variational autoencoder. *Pattern Recognition*, 107:107514, 2020.
- [25] Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. *arXiv preprint arXiv:1901.05761*, 2019.
- [26] Jinkyu Kim and John Canny. Interpretable learning for self-driving cars by visualizing causal attention. In *Proceedings of the IEEE international conference on computer vision*, pages 2942–2950, 2017.
- [27] Taesup Kim, Jaesik Yoon, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. *arXiv preprint arXiv:1806.03836*, 2018.
- [28] Diederik Kingma and Max Welling. Efficient gradient-based inference through transformations between bayes nets and neural nets. In *International Conference on Machine Learning*, pages 1782–1790. PMLR, 2014.
- [29] Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improving variational inference with inverse autoregressive flow. *arXiv preprint arXiv:1606.04934*, 2016.
- [30] Ananya Kumar, SM Eslami, Danilo J Rezende, Marta Garnelo, Fabio Viola, Edward Lockhart, and Murray Shanahan. Consistent generative query networks. *arXiv preprint arXiv:1807.02033*, 2018.
- [31] Tuan Anh Le, Hyunjik Kim, Marta Garnelo, Dan Rosenbaum, Jonathan Schwarz, and Yee Whye Teh. Empirical evaluation of neural process objectives. In *NeurIPS workshop on Bayesian Deep Learning*, 2018.
- [32] Hae Beom Lee, Hayeon Lee, Donghyun Na, Saehoon Kim, Minseop Park, Eunho Yang, and Sung Ju Hwang. Learning to balance: Bayesian meta-learning for imbalanced and out-of-distribution tasks. *arXiv preprint arXiv:1905.12917*, 2019.
- [33] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pages 3744–3753. PMLR, 2019.
- [34] Juho Lee, Yoonho Lee, Jungtaek Kim, Eunho Yang, Sung Ju Hwang, and Yee Whye Teh. Bootstrapping neural processes. *arXiv preprint arXiv:2008.02956*, 2020.
- [35] Yoonho Lee and Seungjin Choi. Gradient-based meta-learning with learned layerwise metric and subspace. *arXiv preprint arXiv:1801.05558*, 2018.
- [36] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. When machine learning meets privacy: A survey and outlook. *arXiv preprint arXiv:2011.11819*, 2020.
- [37] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29:2378–2386, 2016.
- [38] Christos Louizos, Xiahao Shi, Klamer Schutte, and Max Welling. The functional neural process. In *Advances in Neural Information Processing Systems*, pages 8746–8757, 2019.
- [39] Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald, and Edin Muharemagic. Deep learning applications and challenges in big data analytics. *Journal of big data*, 2(1):1–21, 2015.
- [40] Eric Nalisnick and Padhraic Smyth. Stick-breaking variational autoencoders. *arXiv preprint arXiv:1605.06197*, 2016.
- [41] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.
- [42] Boris N Oreshkin, Pau Rodriguez, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. *arXiv preprint arXiv:1805.10123*, 2018.

- [43] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. Few-shot image recognition by predicting parameters from activations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7229–7238, 2018.
- [44] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.
- [45] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016.
- [46] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.
- [47] James Requeima, Jonathan Gordon, John Bronskill, Sebastian Nowozin, and Richard E Turner. Fast and flexible multi-task classification using conditional neural adaptive processes. In *Advances in Neural Information Processing Systems*, pages 7959–7970, 2019.
- [48] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538. PMLR, 2015.
- [49] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*, 2018.
- [50] Hillary Sanders and Joshua Saxe. Garbage in, garbage out: How purport-edly great ml models can be screwed up by bad data. *Proceedings of Blackhat 2017*, 2017.
- [51] Stanislau Semeniuta, Aliaksei Severyn, and Erhardt Barth. A hybrid convolutional variational autoencoder for text generation. *arXiv preprint arXiv:1702.02390*, 2017.
- [52] Sina Shafaei, Stefan Kugele, Mohd Hafeez Osman, and Alois Knoll. Uncertainty in machine learning: A safety perspective on autonomous driving. In *International Conference on Computer Safety, Reliability, and Security*, pages 458–464. Springer, 2018.
- [53] Gautam Singh, Jaesik Yoon, Youngsung Son, and Sungjin Ahn. Sequential neural processes. In *Advances in Neural Information Processing Systems*, pages 10254–10264, 2019.
- [54] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28:3483–3491, 2015.
- [55] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *Advances in neural information processing systems*, pages 3738–3746, 2016.
- [56] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.
- [57] Anirudh Suresh and Srivatsan Srinivasan. Improved attentive neural processes.
- [58] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. Meta-dataset: A dataset of datasets for learning to learn from few examples. *arXiv preprint arXiv:1903.03096*, 2019.
- [59] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
- [60] Joaquin Vanschoren. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548*, 2018.
- [61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [62] Alfredo Vellido. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications*, pages 1–15, 2019.
- [63] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *arXiv preprint arXiv:1606.04080*, 2016.
- [64] Risto Vuorio, Shao-Hua Sun, Hexiang Hu, and Joseph J Lim. Toward multimodal model-agnostic meta-learning. *arXiv preprint arXiv:1812.07172*, 2018.

- [65] Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial intelligence and statistics*, pages 370–378, 2016.
- [66] Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. Improved variational autoencoders for text modeling using dilated convolutions. In *International conference on machine learning*, pages 3881–3890. PMLR, 2017.
- [67] Huaxiu Yao, Ying Wei, Junzhou Huang, and Zhenhui Li. Hierarchically structured meta-learning. *arXiv preprint arXiv:1905.05301*, 2019.
- [68] Huaxiu Yao, Xian Wu, Zhiqiang Tao, Yaliang Li, Bolin Ding, Ruirui Li, and Zhenhui Li. Automated relational meta-learning. *arXiv preprint arXiv:2001.00745*, 2020.
- [69] Huaxiu Yao, Yingbo Zhou, Mehrdad Mahdavi, Zhenhui Li, Richard Socher, and Caiming Xiong. Online structured meta-learning. *arXiv preprint arXiv:2010.11545*, 2020.
- [70] Mingzhang Yin, George Tucker, Mingyuan Zhou, Sergey Levine, and Chelsea Finn. Meta-learning without memorization. *arXiv preprint arXiv:1912.03820*, 2019.
- [71] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. *arXiv preprint arXiv:1703.06114*, 2017.
- [72] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Learning hierarchical features from generative models. *arXiv preprint arXiv:1702.08396*, 2017.
- [73] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Towards deeper understanding of variational autoencoding models. *arXiv preprint arXiv:1702.08658*, 2017.
- [74] Luisa Zintgraf, Kyriacos Shiarli, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. In *International Conference on Machine Learning*, pages 7693–7702. PMLR, 2019.