Meta-Learning without Data via Wasserstein Distributionally-Robust Model Fusion (Supplementary material)

Zhenyi Wang[1] **Xiaoyang Wang**[2] **Li Shen**[3] **Qiuling Suo**[2] **Kaiqiang Song**[2] **Dong Yu**[2] **Yan Shen**[1] **Mingchen Gao**[1]

[1]Department of Computer Science and Engineering., State University of New York at Buffalo, NY, USA
[2]Tencent AI Lab, Seattle, WA, USA
[3]JD Explore Academy, Beijing, China

1 EXPERIMENTS

1.1 BASELINES

To show the effectiveness of the proposed methods, we construct various baseline methods and compare them in the following.

Finetuning We first randomly initialize the parameters for the network, then use few-shot labeled data to finetune this randomly initialized network. This method performs the worst because it does not incorporate the information from the pre-trained models.

Vanilla averaging (VA). We average all the models in a layerwise manner. We average the parameter values elementwise across all the pre-trained models as the fused parameters for each averaged layer. This method assumes that all the pre-trained models are solving the same task, and there is correspondence for the same position parameters across all the pre-trained models. However, this property does not hold in our data-free meta-learning setting since each pre-trained model is to solve a different task. Thus, there is no correspondence among different pre-trained models.

MAML [Finn et al., 2017], which meta trains all the tasks with available training and testing data together. This setting is entirely different from ours. We use these datasets to train a MAML as in standard meta-learning. This baseline gives us a sense of how MAML performs with available training and testing data compared to the data-free setting. MAML with available training data does not perform well in this setting because the number of tasks (100), the same number as the pre-trained models, is relatively much smaller than that of standard data-based meta-learning. Thus, it learns weak domain knowledge.

Optimal transport averaging (OTA) [Singh and Jaggi, 2020], Step 1: following [Singh and Jaggi, 2020], assume we are at layer l and that neurons in the previous layers have already been aligned.

Step 2: we use uniform distributions to initialize the histogram for this layer probability measures.

Step 3: we use layer l of one randomly sampled pre-trained model as the estimate of the fused model for layer l. We then calculate the aligned model with respect to this estimate for each pre-trained model.

Step 4: we calculate the average of all the aligned models as the fused model for layer l.

This method also assumes that the different pre-trained models solve the same task. Thus, different model parameters can be aligned. However, in the data-free meta-learning scenario, different models solve different tasks. Second, they did not consider and optimize the generalization to the unseen tasks.

Model fusion with Gaussian process (MFGP) [Lam et al., 2021]

There are three modules for MFGP.

[1] Base Module network. This module is to compute the mean vector and diagonal covariance matrix of the outer multivariate Gaussian that distributes $\boldsymbol{w}_{\oslash}$ is a 100-dimension vector generated from a 100-dimensional noise vector.

[2] Task-Specific Module Gaussian process parameterization. This module consists of 10 independent sparse Gaussian processes (GPs), which represent the 10 independent priors over 10 random functions mapping from the task embedding to a scalar.

[3] Crossing Module $P(\boldsymbol{\theta}|\boldsymbol{w}_{\oslash}, \boldsymbol{w})$ Parameterization.

This module is to compute the mean vector and diagonal covariance matrix of the outer multivariate Gaussian that models the distributions of $\boldsymbol{\theta}$. The above parameterization describes the generative process of $\boldsymbol{\theta}$ from $\boldsymbol{w}$ and $\boldsymbol{w}_{\oslash}$ for a single task $\mathcal{T}_i$. The fusion model is trained with a variational lower bound.

During meta testing, we adapt MFGP to fuse pre-trained

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

models in the following way with our proposed method:

$\boldsymbol{e}_{init} = \frac{1}{N}\sum_{i=1}^{i=N} \boldsymbol{e}_i$

$\boldsymbol{\theta}_{init} = f_{\boldsymbol{\phi}_{meta}}(\boldsymbol{e}_{init})$

Where $\boldsymbol{e}_{init}$ is the average embedding of all the pre-trained models, and $\boldsymbol{\phi}_{meta}$ is the optimal solution to the Eq 8 (main text).

This method uses the Gaussian process, which can only handle simple networks, such as MLP, to fuse standard pre-trained models, and can be hard to scale to more complex problems, e.g., our setting. Furthermore, they did not consider and optimize the generalization to unseen tasks.

1.2 MORE RESULTS

In this section, we give several ablation studies to verify the effective and stability of our proposed framework on the offline DFL2L task.

Ablation Study We evaluate the effectiveness of DRO for model fusion by ablating the component of DRO. The results are shown in Table 1. We can observe that with DRO, the performance can be improved by 1.2% and 1.5% for 10-shot and 20-shot on CIFAR-FS, respectively.

Table 1: Ablation study on offline DFL2L CIFAR-FS 5-way classification

	10-shot	20-shot
Ours (w/o DRO)	49.23 ± 1.7	53.35 ± 1.4
Ours (w/ DRO)	50.42 ± 1.5	54.86 ± 1.2

Hyperparameter Sensitivity We evaluate the model performance sensitivity with different values γ in Table 2. For the considered γ value, the proposed model performance is not very sensitive to γ value variations, although there are some variations among different γ values.

Table 2: Hyperparameter sensitivity on offline DFL2L Mini-imageNet 5-way classification

γ	10-shot	20-shot
$\gamma = 10.0$	37.09 ± 1.8	43.37 ± 1.5
$\gamma = 2.0$	37.36 ± 1.7	43.67 ± 1.6
$\gamma = 0.5$	37.57 ± 1.5	43.31 ± 1.4

1.3 HYPERPARAMETER SELECTION

As mentioned in the main text, we convert the Wasserstein ball constraint into the objective functions; after using Lagrangian duality, the optimization becomes:

$\max_{\boldsymbol{\phi}} \inf_{\nu \in \mathcal{P}} \mathbb{E}_{\nu}[\mathcal{F}(\boldsymbol{\phi}) + \gamma(W(\mu, \nu) - \delta)]$.

Since δ is not an optimization variable (constant) and does not affect optimization, the constraint is implicitly regularized by the Lagrange multiplier γ. That is to say, with or without δ does not affect the optimization. Therefore, the above optimization can be equivalently formulated as follows:

$\max_{\boldsymbol{\phi}} \inf_{\nu \in \mathcal{P}} \mathbb{E}_{\nu}[\mathcal{F}(\boldsymbol{\phi}) + \gamma W(\mu, \nu)]$.

In this case, the γ controls the regularization. The problem of choosing δ becomes choosing γ. For selecting γ, as mentioned in the main text, we have a validation set of pre-trained models that can be used for determining γ.

First, we calculate the meta initialization for the validation set of pre-trained models as follows:

$\boldsymbol{e}_{init} = \frac{1}{N}\sum_{i=1}^{i=N} \boldsymbol{e}_i$

$\boldsymbol{\theta}_{init} = f_{\boldsymbol{\phi}_{meta}}(\boldsymbol{e}_{init})$

Where $\boldsymbol{e}_{init}$ is the average embedding of all the pre-trained models, and $\boldsymbol{\phi}_{meta}$ is the optimal solution to Eq (8) (main text).

Then, we calculate the likelihood of the validation-set pre-trained models based on the following equations. The likelihood function of the validation pre-trained model $\boldsymbol{\theta}_i$ follows the following Gaussian likelihood function:

$P(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{init}) = exp(-\frac{||\boldsymbol{\theta}_{init} - \boldsymbol{\theta}_i||^2}{\sigma^2})$

Then, we can use grid search to select γ with the highest likelihood on the validation-set of pre-trained models as the best γ. Suppose we want to work with δ directly instead of the γ regularization. We can use projected gradient descent to project the gradient update into the Wasserstein ball constraint; the best δ can be selected similarly to the above procedures for selecting γ.

1.4 MORE DISCUSSION

Model Fusion vs Transfer Learning The number of pre-trained models determines which method should be adopted, classical transfer learning or meta-learning. If the number of pre-trained models is small, then meta-learning is unnecessary. If we only have one pre-trained model, transfer learning would be enough and well-studied in existing works. If we only have very few pre-trained models, how to use them depends on the downstream tasks, practical deployment requirements, etc. For example, if we have both GPT and BERT, then using which one depends on downstream tasks. If the downstream task is text generation, we can choose GPT. If the task is language understanding, we can use BERT. However, our focus is on the meta-learning scenario, i.e., there are many available pre-trained models, but we have to design a general method for learning how to use them. Thus, the research focus is entirely different.

For how to use big models, such as BERT and GPT, fusing them would be more challenging. However, most existing works still focus on much smaller and simpler networks, such as MLP and CNN. One solution for fusing such big models is that, we can first divide large layers into smaller blocks, then apply our method to fuse models in a blockwise manner. This would simplify the fusion process.

References

Durmus Alp Emre Acar, Ruizhao Zhu, and Venkatesh Saligrama. Memory efficient online meta learning. *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Sudarshan Babu, Pedro Henrique Pamplona Savarese, and Michael Maire. Online meta-learning via learning with layer-distributed memory. *Advances in Neural Information Processing Systems*, 2021.

Samy Bengio, Yoshua Bengio, Jocelyn Cloutier, and Jan Gecsei. On the optimization of a synaptic learning rule, 1997.

Luca Bertinetto, João F. Henriques, Philip H. S. Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *International Conference on Learning Representations*, 2019.

Jose Blanchet and Karthyek R. A. Murthy. Quantifying distributional model risk via optimal transport. *https://arxiv.org/abs/1604.01446*, 2017.

Ondrej Bohdal, Yongxin Yang, and Timothy Hospedales. Evograd: Efficient gradient-based meta-learning and hyperparameter optimization. *Advances in Neural Information Processing Systems*, 2021.

John F Bronskill, Daniela Massiceti, Massimiliano Patacchiola, Katja Hofmann, Sebastian Nowozin, and Richard E Turner. Memory efficient meta-learning with large images. *Advances in Neural Information Processing Systems*, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

Massimo Caccia, P. Rodríguez, O. Ostapenko, Fabrice Normandin, Min Lin, L. Caccia, Issam H. Laradji, I. Rish, Alexande Lacoste, D. Vázquez, and Laurent Charlin. Online fast adaptation and knowledge accumulation: a new approach to continual learning. *advances in neural information processing systems*, 2020.

Qi CHEN, Changjian Shui, and Mario Marchand. Generalization bounds for meta-learning: An information-theoretic analysis. *Advances in Neural Information Processing Systems*, 2021.

Giulia Denevi, Dimitris Stamos, Carlo Ciliberto, and Massimiliano Pontil. Online-within-online meta-learning. In *Advances in Neural Information Processing Systems*, 2019.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Distributionally robust federated averaging. *Advances in Neural Information Processing Systems*, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *International Conference on Machine Learning*, 2017.

Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. *International Conference on Machine Learning*, 2019a.

Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. *Advances in Neural Information Processing Systems*, 2019b.

James Harrison, Apoorva Sharma, Chelsea Finn, and Marco Pavone. Continuous meta-learning without tasks. In *Advances in Neural Information Processing Systems*, 2020.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. *https://arxiv.org/abs/1908.08729*, 2019.

Thanh Chi Lam, Nghia Hoang, Bryan Kian Hsiang Low, and Patrick Jaillet. Model fusion for personalized learning. *Proceedings of the 38th International Conference on Machine Learning*, 2021.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. *International Conference on Artificial Intelligence and Statistics*, 2017.

D.K. Naik and R.J. Mammone. Meta-neural networks that learn by learning. In *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks*, 1992.

Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. In *International Conference on Learning Representations*, 2020.

Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *https://arxiv.org/abs/1908.05659*, 2019.

Aravind Rajeswaran, Chelsea Finn, Sham Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Advances in Neural Information Processing Systems*, 2019.

Sachin Ravi and Alex Beatson. Amortized bayesian meta-learning. *International Conference on Learning Representations*, 2019.

Jonas Rothfuss, Dominique Heyn, jinfan Chen, and Andreas Krause. Meta-learning reliable priors in the function space. *Advances in Neural Information Processing Systems*, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *International Conference on Learning Representations*, 2020.

Jurgen Schmidhuber. Evolutionary principles in self-referential learning. on learning now to learn: The meta-meta-meta...-hook. *Technische Universitat Munchen, Germany*, 1987.

Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. Personalized federated learning using hypernetworks. *International Conference on Machine Learning*, 2021.

Sidak Pal Singh and Martin Jaggi. Model fusion via optimal transport. *34th Conference on Neural Information Processing System*, 2020.

Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *International Conference on Learning Representations*, 2018.

Elena Smirnova, Elvis Dohmatob, and Jérémie Mary. Distributionally robust reinforcement learning. *https://arxiv.org/abs/1902.08708*, 2019.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 2017.

Yue Sun, Adhyyan Narang, Halil Ibrahim Gulluk, Samet Oymak, and Maryam Fazel. Towards sample-efficient overparameterized meta-learning. *Advances in Neural Information Processing Systems*, 2021.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *https://arxiv.org/pdf/1606.04080.pdf*, 2016.

Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. *International Conference on Learning Representations*, 2020a.

Zhenyi Wang, Yang Zhao, Ping Yu, Ruiyi Zhang, and Changyou Chen. Bayesian meta sampling for fast uncertainty adaptation. *International Conference on Learning Representations*, 2020b.

Zhenyi Wang, Tiehang Duan, Le Fang, Qiuling Suo, and Mingchen Gao. Meta learning on a sequence of imbalanced domains with difficulty awareness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8947–8957, October 2021.

Zhenyi Wang, Li Shen, Tiehang Duan, Donglin Zhan, Le Fang, and Mingchen Gao. Learning to learn and remember super long multi-domain task sequence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Huaxiu Yao, Yingbo Zhou, Mehrdad Mahdavi, Zhenhui Li, Richard Socher, and Caiming Xiong. Online structured meta-learning. *Advances in Neural Information Processing Systems*, 2020.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, and Trong Nghia Hoang. Statistical model aggregation via parameter matching. *Advances in Neural Information Processing Systems*, 2019a.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Trong Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. *International Conference on Machine Learning*, 2019b.

Yufan Zhou, Zhenyi Wang, Jiayi Xian, Changyou Chen, and Jinhui Xu. Meta-learning with neural tangent kernels. *International Conference on Learning Representations*, 2021.