META-FREE FEW-SHOT LEARNING VIA REPRESENTATION LEARNING WITH WEIGHT AVERAGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent studies on few-shot classification using transfer learning pose challenges to the effectiveness and efficiency of episodic meta-learning algorithms. Transfer learning approaches are a natural alternative, but they are restricted to few-shot classification. Moreover, little attention has been on the development of probabilistic models with well-calibrated uncertainty from few-shot samples, except for some Bayesian episodic learning algorithms. To tackle the aforementioned issues, we propose a new transfer learning method to obtain accurate and reliable models for few-shot regression and classification. The resulting method does not require episodic meta-learning and is called meta-free representation learning (MFRL). MFRL first finds low-rank representation generalizing well on meta-test tasks. Given the learned representation, probabilistic linear models are fine-tuned with few-shot samples to obtain models with well-calibrated uncertainty. The proposed method not only achieves the highest accuracy on a wide range of few-shot learning benchmark datasets but also correctly quantifies the prediction uncertainty. In addition, weight averaging and temperature scaling are effective in improving the accuracy and reliability of few-shot learning in existing meta-learning algorithms with a wide range of learning paradigms and model architectures.

1 INTRODUCTION

Currently, the vast majority of few-shot learning methods are within the general paradigm of meta-learning (a.k.a. learning to learn) (Bengio et al., 1991; Schmidhuber, 1987; Thrun & Pratt, 1998), which learns multiple tasks in an episodic manner to distill transferrable knowledge (Vinyals et al., 2016; Finn et al., 2017; Snell et al., 2017). Despite the success of episodic meta-learning in few-shot learning tasks, they are slow to converge, prone to over-fitting, and tricky to implement (Antoniou et al., 2018). Furthermore, recent studies on transfer learning (Chen et al., 2019; Tian et al., 2020) cast doubt on whether it is the episodic meta-learning algorithm or the learned representation that is responsible for fast adaption to new tasks. It is found that the effectiveness of optimization-based meta-learning algorithms is due to reusing high-quality representation, instead of rapid learning of task-specific representation (Raghu et al., 2020). The quality of the presentation is not quantitatively defined, except for some empirical case studies (Goldblum et al., 2020). Recent machine learning theories (Saunshi et al., 2021) indicate that low-rank representation leads to better sample efficiency in learning a new task. However, those theoretical studies are within the paradigm of meta-learning and do not reveal how to obtain low-rank representation for few-shot learning outside the realm of meta-learning. This motivates us to investigate ways to improve the representation for adapting to new few-shot tasks in a meta-free manner by taking the advantage of simplicity and robustness in transfer learning.

In parallel, existing transfer learning methods also have limitations. That is, the existing transfer learning methods may not find representation generalizing well to unseen few-shot tasks (Chen et al., 2019; Dhillon et al., 2020) , compared with state-of-the-art meta-learning methods (Ye et al., 2020; Zhang et al., 2020). Although some transfer learning methods utilize knowledge distillation and self-supervised training to achieve strong performance in few-shot classification, they are restricted to few-shot classification problems (Mangla et al., 2020; Tian et al., 2020). To the best of our knowledge, no transfer learning method is developed to achieve similar performance to meta-learning in few-shot regression. As such, it is desirable to have a transfer learning method that finds high-quality representation generalizing well to unseen classification and regression problems.

The last limitation of the existing transfer learning methods in few-shot learning is the lack of uncertainty calibration. Uncertainty quantification is concerned with the quantification of how likely certain outcomes are. Despite a plethora of few-shot learning methods (in fact, machine learning in general) to improve the point estimation accuracy, few methods are developed to get probabilistic models with improved uncertainty calibration by integrating Bayesian learning into episodic meta-training (Grant et al., 2018; Finn et al., 2018; Yoon et al., 2018; Snell & Zemel, 2021). Few-shot learning models can be used in risk-averse applications such as medical diagnosis (Prabhu et al., 2019). The diagnosis decision is made on not only point estimation but also probabilities associated with the prediction. The risk of making wrong decisions is significant when using uncalibrated models (Begoli et al., 2019). Thus, the development of proper fine-tuning steps to achieve well-calibrated models is the key towards practical applications of transfer learning in few-shot learning.

In this paper, we propose a new method that does not rely on meta-training to overcome the limitations of commonly used episodic meta-learning approaches in few-shot learning. The resulting meta-free method, called Meta-Free Representation Learning (MFRL), is easy to implement, fast to converge, and robust to over-fitting on meta-training data. MFRL is a transfer learning method but overcomes the aforementioned limitations in existing transfer learning methods for few-shot learning. In the proposed MFRL, we develop a common architecture for both regression and classification with a variation only in the definition of loss functions. Moreover, we use stochastic weight averaging (SWA) (Izmailov et al., 2018), which is agnostic to loss function types, as implicit regularization towards low-rank representation to improve the generalization capability of the representation.

To address the issue of uncertainty quantification, we fine-tune appropriate linear layers during the meta-test phase to get models with well-calibrated uncertainty. In MFRL, hierarchical Bayesian linear models are used to properly capture the uncertainty from very limited training samples in few-shot regression, whereas the softmax output is scaled with a temperature parameter to make the few-shot classification model well-calibrated. Our method is the first one to achieve well-calibrated few-shot models by only fine-tuning probabilistic linear models in the meta-test phase, without any learning mechanisms related to the meta-training or representation learning phase.

Our contributions in this work are summarized as follows:

- We propose a transfer learning method that can handle both few-shot regression *and* classification problems with performance exceeding SOTA.
- For the first time, we empirically find the implicit regularization of SWA towards low-rank representation, which is a useful property in transferring to few-shot tasks.
- The proposed method results in well-calibrated uncertainty in few-shot learning models while preserving SOTA accuracy.
- The implicit regularization of SWA and temperature scaling factor can be applied to existing meta-learning methods to improve their accuracy and reliability in few-shot learning.

2 Related Work

Episodic meta-learning approaches can be categorized into metric-based and optimization-based methods. Metric-based methods project input data to feature vectors through nonlinear embeddings and compare their similarity to make the prediction. Examples of similarity metrics include the weighted $L1$ metric (Koch et al., 2015), cosine similarity (Qi et al., 2018; Vinyals et al., 2016), and Euclidean distance to class-mean representation (Snell et al., 2017). Instead of relying on predefined metrics, learnable similarity metrics are introduced to improve the few-shot classification performance (Oreshkin et al., 2018; Sung et al., 2018). Recent metric-based approaches focus on developing task-adaptive embeddings to improve few-shot classification accuracy. Those task-adaptive embeddings include attention mechanisms for feature transformation (Fei et al., 2021; Gidaris & Komodakis, 2018; Ye et al., 2020; Zhang et al., 2021), graph neural networks (Garcia & Estrach, 2018), implicit class representation (Ravichandran et al., 2019), and task-dependent conditioning (Oreshkin et al., 2018; Yoon et al., 2020; 2019). Although metric-based approaches achieve strong performance in few-shot classification, they cannot be directly applied to regression problems.

Optimization-based meta-learning approaches try to find transferrable knowledge and adapt to new tasks quickly. An elegant and powerful meta-learning approach, termed model-agnostic meta-

learning (MAML), solves a bi-level optimization problem to find good initialization of model parameters (Finn et al., 2017). However, MAML has a variety of issues, such as sensitivity to neural network architectures, instability during training, arduous hyperparameter tuning, and high computational cost. On this basis, some follow-up methods have been developed to simplify, stabilize and improve the training process of MAML (Antoniou et al., 2018; Flennerhag et al., 2020; Lee & Choi, 2018; Nichol et al., 2018; Park & Oliva, 2019). In practice, it is very challenging to learn high-dimensional model parameters in a low-data regime. Latent embedding optimization (LEO) attempts to learn low-dimensional representation to generate high-dimensional model parameters (Rusu et al., 2019). Meanwhile, R2-D2 (Bertinetto et al., 2019) and MetaOptNet (Lee et al., 2019) reduce the dimensionality of trainable model parameters by freezing feature extraction layers during inner loop optimization. Note that the proposed method is fundamentally different from R2-D2 and MetaOptNet because our method requires neither episodic meta-learning nor bi-level optimization.

Transfer learning approaches first learn a feature extractor on all training data through standard supervised learning, and then fine-tune a linear predictor on top of the learned feature extractor in a new task (Chen et al., 2019). However, vanilla transfer learning methods for few-shot learning do not take extra steps to make the learned representation generalizing well to unseen meta-test tasks. Some approaches in this paradigm are developed to improve the quality of representation and boost the accuracy of few-shot classification, including cooperative ensembles (Dvornik et al., 2019), knowledge distillation (Tian et al., 2020), and auxiliary self-supervised learning (Mangla et al., 2020). Nevertheless, those transfer learning methods are restricted to few-shot classification. MFRL aims to find representation generalizing well from the perspective of low-rank representation learning, which is supported by recent theoretical studies (Saunshi et al., 2021). Furthermore, MFLR is the first transfer learning method that can handle both few-shot regression and classification problems and make predictions with well-calibrated uncertainty.

3 BACKGROUND

3.1 EPISODIC META-LEARNING

In episodic meta-learning, the meta-training data contains $\mathcal{T}$ episodes or tasks, where the τth episode consists of data $\mathcal{D}_\tau = \{(\mathbf{x}_{\tau,j}, \mathbf{y}_{\tau,j})\}_{j=1}^{N_\tau}$ with N_τ samples. Tasks and episodes are used interchangeably in the rest of the paper. Episodic meta-learning algorithms aim to find common model parameters θ which can be quickly adapted to task-specific parameters ϕ_τ ($\tau = 1, ..., \mathcal{T}$). For example, MAML-type algorithms assume ϕ_τ is one or a few gradient steps away from θ (Finn et al., 2017; 2018; Grant et al., 2018; Yoon et al., 2018), while other meta-learning approaches assume that ϕ_τ and θ share the parameters in the feature extractor and only differ in the top layer (Bertinetto et al., 2019; Lee et al., 2019; Snell et al., 2017).

3.2 STOCHASTIC WEIGHT AVERAGING

The idea of stochastic weight averaging (SWA) along the trajectory of SGD goes back to Polyak–Ruppert averaging (Polyak & Juditsky, 1992). Theoretically, weight averaging results in faster convergence for linear models in supervised learning and reinforcement learning(Bach & Moulines, 2013; Lakshminarayanan & Szepesvari, 2018). In deep learning, we are more interested in tail stochastic weight averaging (Jain et al., 2018), which averages the weights after T training epochs. The averaged model parameters θ_{SWA} can be computed by running s additional training epochs using SGD

$$\theta_{\text{SWA}} = \frac{1}{s} \sum_{i=T+1}^{T+s} \theta_i, \tag{1}$$

where θ_i denotes the model parameters at the end of the i-th epoch. SWA has been applied to supervised learning of deep neural neural networks to achieve higher test accuracy (Izmailov et al., 2018).

4 METHODOLOGY

The proposed method is a two-step learning algorithm: meta-free representation learning followed by fine-tuning. We employ SWA to make the learned representation low-rank and better generalize to meta-test data. Given a meta-test task, a new top layer is fine-tuned with few-shot samples to obtain probabilistic models with well-calibrated uncertainty. Note that MFRL can be used for both regression and classification depending on the loss function. The pseudocode of MFRL is presented in Appendix A.1.

4.1 REPRESENTATION LEARNING

Common representation can be learned via maximization of the likelihood of all training data with respect to θ rather than following episodic meta-learning. To do so, we group the data $\mathcal{D}_\tau = \{(\mathbf{x}_{\tau,j}, \mathbf{y}_{\tau,j})\}_{j=1}^{N_\tau}$ from all meta-training tasks into a single dataset $\mathcal{D}_{\text{tr}}$. Given aggregated training data $\mathcal{D}_{\text{tr}} = \{\mathbf{X}, \mathbf{Y}\}$, representation can be learned by maximizing the likelihood $p\left(\mathcal{D}_{\text{tr}} \mid \theta\right)$ with respect to θ. Let $\theta = [\theta_f, \mathbf{W}]$, where θ_f represents parameters in the feature extractor and $\mathbf{W}$ denotes the parameters in the top linear layer. The feature extractor $h(\mathbf{x}) \in \mathbb{R}^p$ is a neural network parameterized by θ_f and outputs a feature vector of dimension p. The specific form of the loss function depends on whether the task is regression or classification and can be given as follows:

$$\mathcal{L}_{RP}(\theta) = -\log p\left(\mathcal{D}_{\text{tr}} \mid \theta\right) = \begin{cases} \mathcal{L}_{MSE}(\theta), & \text{regression} \\ \mathcal{L}_{CE}(\theta), & \text{classification} \end{cases}$$

where

$$\mathcal{L}_{MSE}(\theta) = \frac{1}{2N'} \sum_{\tau=1}^{\mathcal{T}} \sum_{j=1}^{N_\tau} \left(y_{\tau,j} - \mathbf{w}_\tau^\top h\left(\mathbf{x}_{\tau,j}\right)\right)^2, \tag{2}$$

$$\mathcal{L}_{CE}(\theta) = -\sum_{j=1}^{N'} \sum_{c=1}^{\mathcal{C}} y_{j,c} \log \frac{\exp(\mathbf{w}_c^\top h(\mathbf{x}_j))}{\sum_{c'=1}^{\mathcal{C}} \exp(\mathbf{w}_{c'}^\top h(\mathbf{x}_j))} \tag{3}$$

For regression problems, the model learns $\mathcal{T}$ regression tasks ($\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_\mathcal{T}] \in \mathbb{R}^{(p+1)\times\mathcal{T}}$) simultaneously using the loss function $\mathcal{L}_{MSE}$ given in Eq. 2, whereas the model learns a $\mathcal{C}$-class classification model [1] ($\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_\mathcal{C}] \in \mathbb{R}^{(p+1)\times\mathcal{C}}$) for classification problems using the loss function $\mathcal{L}_{CE}$ in Eq. 3. The loss function - either Eq. 2 or 3 - can be minimized through standard stochastic gradient descent, where $N' = \sum_{\tau=1}^{\mathcal{T}} N_\tau$ is the total number of training samples.

Post-processing via SWA Minimizing the loss functions in Eq. 2 and 3 by SGD may not necessarily result in representation that generalizes well to few-shot learning tasks in the meta-test set. The last hidden layer of a modern deep neural network is high-dimensional and may contain spurious features that over-fit the meta-training data. Recent meta-learning theories indicate that better sample complexity in learning a new task can be achieved via low-rank representation, whose singular values decay faster (Saunshi et al., 2021). We aim to find low-rank representation $\mathbf{\Phi} = h(\mathbf{X})$ without episodic meta-learning, which is equivalent to finding the conjugate kernel $K^{\text{C}} = \mathbf{\Phi}\mathbf{\Phi}^\top$ with fast decaying eigenvalues. To link the representation with the parameter space, we can linearize the neural network by the first-order Taylor expansion at θ_T and get the finite width neural tangent kernel (NTK) $K^{\text{NTK}}(\mathbf{X}, \mathbf{X}) = \mathbf{J}(\mathbf{X})\mathbf{J}(\mathbf{X})^\top$, where $\mathbf{J}(\mathbf{X}) = \nabla_\theta f_\theta(\mathbf{X}) \in \mathbb{R}^{N' \times |\theta|}$ is the Jacobian matrix, and K^{NTK} is a composite kernel containing K^{C} (Fan & Wang, 2020). The distributions of eigenvalues for K^{NTK} and K^{C} are empirically similar. Analyzing K^{NTK} could shed light on the properties of K^{C}. In parallel, K^{NTK} shares the same eigenvalues of the Gauss-Newton matrix $G = \frac{1}{N'}\mathbf{J}(\mathbf{X})^\top\mathbf{J}(\mathbf{X})$. For linearized networks with squared loss, the Gauss-Newton matrix G well approximates the Hessian matrix H when y is well-described by $f_\theta(x)$ (Martens, 2020). This is the case when SGD converges to θ_T within a local minimum basin. A Hessian matrix with a lot of small eigenvalues corresponds to a flat minimum, where the loss function is less sensitive to the perturbation of model parameters (Keskar et al., 2017). It is known that averaging the weights after

[1] $\mathcal{C}$ is the total number of classes in $\mathcal{D}_{\text{tr}}$. Learning a $\mathcal{C}$-class classification model solves all possible tasks in the meta-training dataset because each task $\mathcal{D}_\tau$ only contains a subset of $\mathcal{C}$ classes.

SGD convergence in a local minimum basin pushes θ_T towards the flat side of the loss valley (He et al., 2019). As a result, SWA could result in a faster decay of eigenvalues in the kernel matrix, and thus low-rank representation. Our conjecture about SWA as implicit regularization towards low-rank representation is empirically verified in Section 5.

4.2 FINE-TUNING

After representation learning is complete, $\mathbf{W}$ is discarded and θ_f is frozen in a new few-shot task. Given the learned representation, we train a new probabilistic top layer in a meta-test task using few-shot samples. The new top layer will be configured differently depending on whether the few-shot task is a regression or a classification problem.

In a few-shot regression task, we learn a new linear regression model $y = \mathbf{w}^\top h(\mathbf{x}) + \epsilon$ on a fixed feature extractor $h(\mathbf{x}) \in \mathbb{R}^p$ with few-shot training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{w}$ denotes the model parameters and ϵ is Gaussian noise with zero mean and variance σ^2. To avoid interpolation on few-shot training data ($n \ll p$), a Gaussian prior $p(\mathbf{w} \mid \lambda) = \prod_{i=0}^p \mathcal{N}(w_i \mid 0, \lambda)$ is placed over $\mathbf{w}$, where λ is the precision in the Gaussian prior. However, it is difficult to obtain an appropriate value for λ in a few-shot regression task because no validation data is available in $\mathcal{D}$.

Hierarchical Bayesian linear models can be used to obtain optimal regularization strength and grounded uncertainty estimation using few-shot training data only. To complete the specification of the hierarchical Bayesian model, the hyperpriors on λ and σ^2 are defined as $p(\lambda) = \mathrm{Gamma}(\lambda \mid a, b)$ and $p(\sigma^{-2}) = \mathrm{Gamma}(\sigma^{-2} \mid c, d)$, respectively. The hyper-priors become very flat and non-informative when a, b, c and d are set to very small values. The posterior over all latent variables given the data is $p\left(\mathbf{w}, \lambda, \sigma^2 \mid \mathbf{X}, \mathbf{y}\right)$, where $\mathbf{X} = \{\mathbf{x}\}_{i=1}^n$ and $\mathbf{y} = \{y_i\}_{i=1}^n$. However, the posterior distribution $p\left(\mathbf{w}, \lambda, \sigma^2 \mid \mathbf{X}, \mathbf{y}\right)$ is intractable. The iterative optimization based approximate inference (Tipping, 2001) is chosen because it is highly efficient. The point estimation for λ and σ^2 is obtained by maximizing the marginal likelihood function $p\left(\mathbf{y} \mid \mathbf{X}, \lambda, \sigma^2\right)$. The posterior of model parameters $p\left(\mathbf{w} \mid \mathbf{X}, \mathbf{y}, \lambda, \sigma^2\right)$ is calculated using the estimated λ and σ^2. Previous two steps are repeated alternately until convergence.

The predictive distribution for a new sample $\mathbf{x}_*$ is

$$p\left(y_* \mid \mathbf{x}_*, \mathbf{X}, \mathbf{y}, \lambda, \sigma^2\right) = \int p\left(y_* \mid \mathbf{x}_*, \mathbf{w}, \sigma^2\right) p\left(\mathbf{w} \mid \mathbf{X}, \mathbf{y}, \lambda, \sigma^2\right) d\mathbf{w}, \tag{4}$$

which can be computed analytically because both distributions on the right hand side of Eq. 4 are Gaussian. Consequently, hierarchical Bayesian linear models avoids over-fitting on few-shot training data and quantifies predictive uncertainty.

In a few-shot classification task, a new logistic regression model is learned with the post-processed representation. A typical K-way n-shot classification task $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{nK}$ consists of K classes (different from meta-training classes) and n training samples per class. Minimizing an un-regularized cross-entropy loss results in a significantly over-confident classification model because the norm of logistic regression model parameters $\mathbf{W} \in \mathbb{R}^{(p+1)\times K}$ becomes very large when few-shot training samples can be perfectly separated in the setting $nK \ll p$. A weighted $L2$ regularization term is added to the cross-entropy loss to mitigate the issue

$$\mathcal{L}(\mathbf{W}) = -\sum_{i=1}^{nK}\sum_{c=1}^{K} y_{i,c} \log \frac{\exp(\mathbf{w}_c^\top h(\mathbf{x}_i))}{\sum_{c'=1}^{K} \exp(\mathbf{w}_{c'}^\top h(\mathbf{x}_i))} + \lambda \sum_{c=1}^{K} \mathbf{w}_c^\top \mathbf{w}_c, \tag{5}$$

where λ is the regularization coefficient, which affects the prediction accuracy and uncertainty. It is difficult to select an appropriate value of λ in each of the meta-test tasks due to the lack of validation data in $\mathcal{D}$. We instead treat λ as a global hyper-parameter so that the value of λ should be determined based on the accuracy on meta-validation data. Note that the selected λ with high validation accuracy does not necessarily lead to well calibrated classification models. As such, we introduce the temperature scaling factor (Guo et al., 2017) as another global hyper-parameter to scale the softmax output. Given a test sample $\mathbf{x}_*$, the predicted probability for class c becomes

$$p_c = \frac{\exp(\mathbf{w}_c^\top h(\mathbf{x}_*)/T)}{\sum_{c'=1}^{K} \exp(\mathbf{w}_{c'}^\top h(\mathbf{x}_*)/T)}, \tag{6}$$

where T is the temperature scaling factor. The value of T is selected using meta-validation data, after the value of λ is determined. Different values of T do not affect the classification accuracy because temperature scaling is accuracy preserving.

5 EXPERIMENTS

We follow the standard setup in few-shot learning literature. The model is trained on a meta-training dataset and hyper-parameters are selected based on the performance on a meta-validation dataset. The final performance of the model is evaluated on a meta-test dataset. The proposed method is applied to few-shot regression and classification problems and compared against a wide range of alternative methods.

5.1 FEW-SHOT REGRESSION RESULTS

Sine waves (Finn et al., 2017) and head pose estimation (Patacchiola et al., 2020) datasets are used to evaluate the performance of MFRL in few-shot regression. We use the same backbones in literature (Patacchiola et al., 2020) to make fair comparisons. Details of the few-shot regression experiments can be found in Appendix A.2.

The results for few-shot regression are summarized in Table 1. In the sine wave few-shot regression, MFRL outperforms all meta-learning methods, demonstrating that high-quality representation can be learned in supervised learning, without episodic meta-learning. Although DKT with a spectral mixture (SM) kernel achieves high accuracy, the good performance should be attributed to the strong inductive bias to periodic functions in the SM kernel (Wilson & Adams, 2013). Additional results for MFRL with different activation functions are reported in Appendix A.4. In the head pose estimation experiment, MFRL also achieves the best accuracy. In both few-shot regression problems, SWA results in improved accuracy, suggesting that SWA can improve the quality of features and facilitate the learning of downstream tasks. In Fig. 1, uncertainty is correctly estimated by the hierarchical Bayesian linear model with learned features using just 10 training samples.

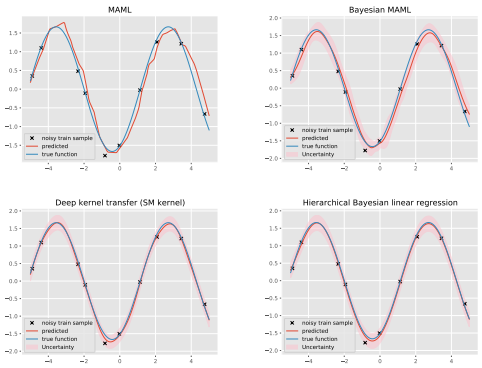


Figure 1: Sine wave regression and uncertainty quantification (10 training samples). The true and the estimated (by MFRL) standard deviation of data generation noise are 0.1, and 0.093.

Sine wave (2-layer MLP)	MSE
MAML (Finn et al., 2017)	0.67 ± 0.06
Bayesian MAML (Yoon et al., 2018)	0.54 ± 0.05
ALPaCA (Harrison et al., 2018)	0.14 ± 0.09
R2D2 (Bertinetto et al., 2019)	$0.46 \pm$ NA
DKT+RBF (Patacchiola et al., 2020)	1.38 ± 0.03
DKT+Spectral (Patacchiola et al., 2020)	0.08 ± 0.06
MFRL (w.o. SWA)	0.023 ± 0.016
MFRL	$\mathbf{0.016 \pm 0.008}$
Head pose (3-layer Conv Net)	MSE
MAML (Finn et al., 2017)	0.21 ± 0.01
Bayesian MAML (Yoon et al., 2018)	0.18 ± 0.01
DKT+Spectral (Patacchiola et al., 2020)	0.10 ± 0.01
MFRL (w.o. SWA)	0.033 ± 0.006
MFRL	$\mathbf{0.027 \pm 0.005}$

Table 1: 10-shot regression on sine waves and head pose estimation.

5.2 FEW-SHOT CLASSIFICATION RESULTS

We conduct few-shot classification experiments on four widely used few-shot image recognition benchmarks: miniImageNet (Ravi & Larochelle, 2017), tieredImageNet (Ren et al., 2018), CIFAR-FS (Bertinetto et al., 2019), and FC100 (Oreshkin et al., 2018). The experiment details about the few-shot classification datasets can be found in Appendix A.2. The proposed method is applied to three widely used network architectures: ResNet-12 (Lee et al., 2019; Ravichandran et al., 2019), wide ResNet (WRN-28-10) (Dhillon et al., 2020; Rusu et al., 2019), and a 4-layer convolutional neural network with 64 channels (Chen et al., 2019; Patacchiola et al., 2020) (in Appendix A.4).

Table 2: Few-shot classification results on miniImageNet and tieredImageNet.

Method	Backbone	miniImageNet 5-way		tieredImageNet 5-way	
		1-shot	5-shot	1-shot	5-shot
Matching Net (Vinyals et al., 2016)	ResNet-12	63.08 ± 0.80	75.99 ± 0.60	68.50 ± 0.92	80.60 ± 0.71
Proto Net (Snell et al., 2017)	ResNet-12	60.37 ± 0.83	78.02 ± 0.57	65.65 ± 0.92	83.40 ± 0.65
Proto Net + SWA	ResNet-12	63.51 ± 0.82	81.98 ± 0.58	67.95 ± 0.85	84.76 ± 0.66
MAML (Finn et al., 2017)	ResNet-12	56.58 ± 1.84	70.85 ± 0.91	-	-
MAML + SWA	ResNet-12	58.21 ± 1.86	72.47 ± 0.87	-	-
AdaResNet (Munkhdalai et al., 2018)	ResNet-12	56.88 ± 0.62	71.94 ± 0.57	-	-
TADAM (Oreshkin et al., 2018)	ResNet-12	58.50 ± 0.30	76.70 ± 0.30	-	-
Baseline++ (Chen et al., 2019)	ResNet-12	60.83 ± 0.81	77.81 ± 0.76	68.64 ± 0.86	80.47 ± 0.67
TapNet (Yoon et al., 2019)	ResNet-12	61.65 ± 0.15	76.36 ± 0.10	63.08 ± 0.15	80.26 ± 0.12
MetaOptNet (Lee et al., 2019)	ResNet-12	62.64 ± 0.61	78.63 ± 0.46	65.99 ± 0.72	81.56 ± 0.53
Ensemble (Dvornik et al., 2019)	ResNet-18	59.48 ± 0.65	75.62 ± 0.42	-	-
DSN (Simon et al., 2020)	ResNet-12	62.64 ± 0.66	78.83 ± 0.45	66.22 ± 0.75	82.79 ± 0.48
DKT (Patacchiola et al., 2020)	ResNet-12	61.29 ± 0.57	76.25 ± 0.51	67.21 ± 0.52	79.69 ± 0.53
FEAT (Ye et al., 2020)	ResNet-12	66.78 ± 0.20	82.05 ± 0.14	70.80 ± 0.23	84.79 ± 0.16
DeepEMD (Zhang et al., 2020)	ResNet-12	65.91 ± 0.82	82.41 ± 0.56	71.16 ± 0.87	86.03 ± 0.58
Distill (Tian et al., 2020)	ResNet-12	64.82 ± 0.60	82.14 ± 0.43	71.52 ± 0.69	86.03 ± 0.49
MFRL (w.o. SWA)	ResNet-12	62.27 ± 0.86	80.23 ± 0.57	70.03 ± 0.77	84.42 ± 0.64
MFRL	ResNet-12	$\mathbf{67.18 \pm 0.79}$	$\mathbf{83.81 \pm 0.53}$	$\mathbf{71.58 \pm 0.79}$	$\mathbf{86.87 \pm 0.62}$
LEO (Rusu et al., 2019)	WRN-28-10	61.76 ± 0.08	77.59 ± 0.12	66.33 ± 0.03	81.44 ± 0.09
Fine-tune (Dhillon et al., 2020)	WRN-28-10	57.73 ± 0.62	78.17 ± 0.49	66.58 ± 0.70	85.55 ± 0.48
Inductive SIB (Hu et al., 2020)	WRN-28-10	60.12 ± 0.56	78.17 ± 0.35	69.20 ± 0.58	84.96 ± 0.36
MetaFun (Xu et al., 2020)	WRN-28-10	64.13 ± 0.13	80.82 ± 0.17	67.27 ± 0.14	83.28 ± 0.12
MFRL (w.o. SWA)	WRN-28-10	61.83 ± 0.82	80.12 ± 0.88	69.89 ± 0.79	84.42 ± 0.66
MFRL	WRN-28-10	$\mathbf{66.42 \pm 0.80}$	$\mathbf{82.26 \pm 0.61}$	$\mathbf{71.47 \pm 0.84}$	$\mathbf{86.34 \pm 0.65}$

Table 3: Few-shot classification results on CIFAR-FS and FC100.

Method	Backbone	CIFAR-FS 5-way		FC100 5-way	
		1-shot	5-shot	1-shot	5-shot
Proto Net (Snell et al., 2017)	ResNet-12	72.2 ± 0.7	83.5 ± 0.5	41.5 ± 0.7	57.0 ± 0.7
TADAM (Oreshkin et al., 2018)	ResNet-12	-	-	40.1 ± 0.4	56.1 ± 0.4
Baseline++ (Chen et al., 2019)	ResNet-12	72.2 ± 0.9	84.2 ± 0.6	43.1 ± 0.7	55.7 ± 0.7
MetaOptNet (Lee et al., 2019)	ResNet-12	72.8 ± 0.7	85.0 ± 0.5	41.1 ± 0.6	55.5 ± 0.6
MTL (Sun et al., 2019)	ResNet-12	-	-	45.1 ± 1.8	57.6 ± 0.9
Shot-free (Ravichandran et al., 2019)	ResNet-12	$69.1 \pm$ NA	$84.7 \pm$ NA	-	-
TEAM (Qiao et al., 2019)	ResNet-12	$70.4 \pm$ NA	$81.3 \pm$ NA	-	-
SIB (Hu et al., 2020)	ResNet-12	70.0 ± 0.5	83.5 ± 0.4	-	-
DSN (Simon et al., 2020)	ResNet-12	72.3 ± 0.8	85.1 ± 0.6	-	-
MABAS (Kim et al., 2020)	ResNet-12	73.5 ± 0.9	85.6 ± 0.6	42.3 ± 0.7	58.1 ± 0.7
Distill (Tian et al., 2020)	ResNet-12	73.9 ± 0.8	86.9 ± 0.5	44.6 ± 0.7	60.9 ± 0.6
MFRL (w.o. SWA)	ResNet-12	71.4 ± 0.8	86.1 ± 0.5	42.5 ± 0.7	59.1 ± 0.7
MFRL	ResNet-12	$\mathbf{74.0 \pm 0.8}$	$\mathbf{87.4 \pm 0.6}$	$\mathbf{45.3 \pm 0.8}$	$\mathbf{61.1 \pm 0.7}$
Fine-tune (Dhillon et al., 2020)	WRN-28-10	68.7 ± 0.7	86.1 ± 0.6	38.2 ± 0.5	57.2 ± 0.6
MFRL (w.o. SWA)	WRN-28-10	71.7 ± 0.9	86.2 ± 0.9	41.5 ± 0.7	57.3 ± 0.7
MFRL	WRN-28-10	$\mathbf{76.7 \pm 0.9}$	$\mathbf{88.6 \pm 0.5}$	$\mathbf{45.1 \pm 0.8}$	$\mathbf{61.0 \pm 0.8}$

The results of the proposed method and previous SOTA methods using similar backbones are reported in Table 2 and 3. The proposed method achieves the best performance in most of the experiments when compared with previous SOTA methods. Our method is closely related to Baseline++ (Chen et al., 2019) and fine-tuning on logits (Dhillon et al., 2020). Baseline++ normalizes both classification weights and features, while the proposed method only normalizes features. It allows our method to find a more accurate model in a more flexible hypothesis space, given high-quality representation. Compared with fine-tuning on logits, our method obtains better results by learning a new logistic regression model on features, which store richer information about the data. Some approaches pretrain a $\mathcal{C}$-class classification model on all training data and then apply highly sophisticated meta-learning techniques to the pretrained model to achieve SOTA performance (Rusu et al., 2019; Sun et al., 2019). Our approach with SWA outperforms those pretrained-then-meta-learned models, which demonstrates that SWA obtains high-quality representation that generalizes well to unseen tasks. Compared with improving representation quality for few-shot classification via self-distillation (Tian et al., 2020), the computational cost of SWA is significantly smaller because it does not require training models from scratch multiple times. Moreover, SWA can be applied to find good representation for both few-shot regression and classification, while previous transfer learning approaches can only handle few-shot classification problems (Mangla et al., 2020; Tian et al., 2020).

5.3 EFFECTIVE RANK OF THE REPRESENTATION

The rank of representation defines the number of independent bases. For deep learning, noise in gradients and numerical imprecision can cause the resulting matrix to be full-rank. Therefore, simply counting the number of non-zero singular values may not be an effective way to measure the rank of the representation. To compare the effective ranks, we plot the normalized singular values of the representation of meta-test data in Fig. 2, where the representation with SWA has a faster decay in singular values, thus indicating the lower effective rank of the presentation with SWA. The results empirically verify our conjecture that SWA is an implicit regularizer towards low-rank representation.

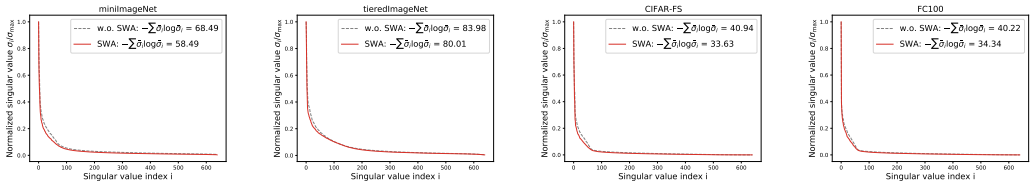


Figure 2: Normalized singular values for representation with and without SWA. The metric $-\sum\bar{\sigma}_i\log\bar{\sigma}_i$ is used to measure the effective rank of the representation, where $\bar{\sigma}_i = \sigma_i/\sigma_{\max}$. Faster decay in singular values indicates that fewer dimensions capture the most variation in all dimensions, thus lower effective rank.

5.4 FEW-SHOT CLASSIFICATION RELIABILITY

The proposed method not only achieves high accuracy in few-shot classification but also makes the classification uncertainty well-calibrated. A reliability diagram can be used to check model calibration visually, which plots an identity function between prediction accuracy and confidence when the model is perfectly calibrated (DeGroot & Fienberg, 1983). Fig. 3 shows the classification reliability diagrams along with widely used metrics for uncertainty calibration, including expected calibration error (ECE) (Guo et al., 2017), maximum calibration error (MCE) (Naeini et al., 2015), and Brier score (BRI) (Brier, 1950). ECE measures the average binned difference between confidence and accuracy, while MCE measures the maximum difference. BRI is the squared error between the predicted probabilities and one-hot labels. MAML is over-confident because tuning a deep neural network on few-shot data is prone to over-fitting. Meanwhile, Proto Net and Matching Net are better calibrated than MAML because they do not fine-tune the entire network during testing. Nevertheless, they are still slightly over-confident. The results indicate that MFRL with a global temperature scaling factor can learn well-calibrated models from very limited training samples.

5.5 APPLICATION IN META-LEARNING

Meanwhile, we also apply SWA to episodic meta-learning methods, such Prot Net, MAML and Matching Net, to improve their classification accuracy. The results in Table 4 indicate that SWA can improve the few-shot classification accuracy in both transfer learning and episodic meta-learning. SWA is orthogonal to the learning paradigm and model architecture. Thus, SWA can be applied to a wide range of few-shot learning methods to improve accuracy.

Table 4: Application of SWA on meta-learning methods for the miniImageNet dataset

Method	Proto Net		MAML		Matching Net	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
w.o. SWA	60.37	78.02	56.58	70.85	63.08	75.99
SWA	63.51	81.98	58.21	72.47	63.76	76.78

Furthermore, the temperature scaling factor can be applied to calibrate meta-learning methods, including MAML, Proto Net, and Matching Net. The reliability diagrams in Fig. 3 indicate that

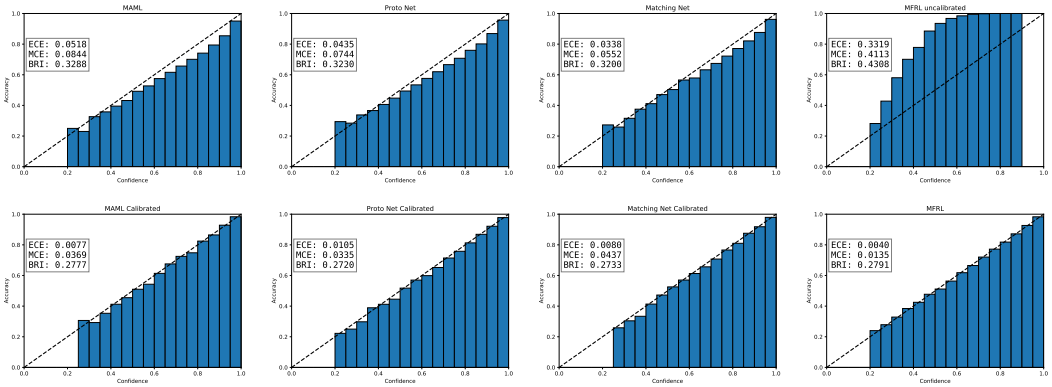


Figure 3: Study on the temperature scaling factor for 5-way 5-shot classification using ResNet-12 for the proposed MFRL and existing episodic meta learning methods. Uncalibrated models are in the first row, and calibrated models with the temperature scaling factor are in the second row.

the temperature scaling factor not only calibrates classification uncertainty of transfer learning approaches, such as the proposed MFRL, but also makes the classification uncertainty well-calibrated in episodic meta-learning methods. Therefore, the temperature scaling factor can be applied to a wide range of few-shot classification methods to get well-calibrated uncertainty, while preserving the classification accuracy.

6 Discussion

SWA has been applied to supervised learning of deep neural networks (Izmailov et al., 2018; Athiwaratkun et al., 2019) and its effectiveness was attributed to convergence to a solution on the flat side of an asymmetric loss valley (He et al., 2019). However, it does not explain the effectiveness of SWA in few-shot learning because the meta-training and meta-testing losses are not comparable after the top layer is retrained by the few-shot support data in a meta-test task. The effectiveness of SWA in few-shot learning must be related to the property of the representation. Although our results empirically demonstrate that SWA results in low-rank representation, further research about their connection is needed.

Explicit regularizers can also be used to obtain simple input-output functions in deep neural networks and low-rank representation, including L1 regularization, nuclear norm, spectral norm, and Frobenius norm (Bartlett et al., 2017; Neyshabur et al., 2018; Sanyal et al., 2020). However, some of those explicit regularizers are not compatible with standard SGD training or are computationally expensive. In addition, it is difficult to choose the appropriate strength of explicit regularization. Too strong explicit regularization can bias towards simple solutions that do not fit the data. In comparison, SWA is an implicit regularizer that is completely compatible with the standard SGD training without much extra computational cost. Thus, it can be easily combined with transfer learning and meta-learning to obtain more accurate few-shot learning models. In parallel, SWA is also robust to the choice of the hyperparameters - the learning rate and training epochs in the SWA stage (see details in Appendix A.5).

7 Conclusions

In this article, we propose MFRL to obtain accurate and reliable few-shot learning models. SWA is an implicit regularizer towards low-rank representation, which generalizes well to unseen meta-test tasks. The proposed method can be applied to both classification and regression tasks. Extensive experiments show that our method not only outperforms other SOTA methods on various datasets but also correctly quantifies the uncertainty in prediction.

REFERENCES

Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your maml. In *International Conference on Learning Representations*, 2018.

Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. There are many consistent explanations of unlabeled data: Why you should average. In *International Conference on Learning Representations*, 2019.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 1*, pp. 773–781, 2013.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 30:6240–6249, 2017.

Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, 2019.

Y Bengio, S Bengio, and J Cloutier. Learning a synaptic learning rule. In *IJCNN-91-Seattle International Joint Conference on Neural Networks*, volume 2, pp. 969–vol. IEEE, 1991.

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

Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2019.

Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. In *International Conference on Learning Representations*, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Nikita Dvornik, Cordelia Schmid, and Julien Mairal. Diversity with cooperation: Ensemble methods for few-shot classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3723–3731, 2019.

Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

Nanyi Fei, Zhiwu Lu, Tao Xiang, and Songfang Huang. {MELR}: Meta-learning via modeling episode-level relationships for few-shot learning. In *International Conference on Learning Representations*, 2021.

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.

Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 9537–9548, 2018.

Sebastian Flennerhag, Andrei A Rusu, Razvan Pascanu, Francesco Visin, Hujun Yin, and Raia Hadsell. Meta-learning with warped gradient descent. In *International Conference on Learning Representations*, 2020.

Victor Garcia and Joan Bruna Estrach. Few-shot learning with graph neural networks. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4367–4375, 2018.

Micah Goldblum, Steven Reich, Liam Fowl, Renkun Ni, Valeriia Cherepanova, and Tom Goldstein. Unraveling meta-learning: Understanding feature representations for few-shot tasks. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3607–3616, Virtual, 13–18 Jul 2020. PMLR.

Shaogang Gong, Stephen McKenna, and John J Collins. An investigation into face pose distributions. In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, pp. 265–270. IEEE, 1996.

J Gordon, J Bronskill, M Bauer, S Nowozin, and RE Turner. Meta-learning probabilistic inference for prediction. In *International Conference on Learning Representations (ICLR 2019)*. OpenReview. net, 2019.

Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. In *International Conference on Learning Representations*, 2018.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

James Harrison, Apoorva Sharma, and Marco Pavone. Meta-learning priors for efficient online bayesian regression. In *International Workshop on the Algorithmic Foundations of Robotics*, pp. 318–337. Springer, 2018.

Haowei He, Gao Huang, and Yang Yuan. Asymmetric valleys: Beyond sharp and flat local minima. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 2549–2560, 2019.

Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

Shell Xu Hu, Pablo Garcia Moreno, Yang Xiao, Xi Shen, Guillaume Obozinski, Neil Lawrence, and Andreas Damianou. Empirical bayes transductive meta-learning with synthetic gradients. In *International Conference on Learning Representations*, 2020.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pp. 876–885. Association For Uncertainty in Artificial Intelligence (AUAI), 2018.

Prateek Jain, Sham Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 18, 2018.

Ghassen Jerfel, Erin Grant, Thomas L Griffiths, and Katherine Heller. Reconciling meta-learning and continual learning with online mixtures of tasks. *Advances in Neural Information Processing Systems*, 32, 2019.

Nitish Shirish Keskar, Jorge Nocedal, Ping Tak Peter Tang, Dheevatsa Mudigere, and Mikhail Smelyanskiy. On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

Jaekyeom Kim, Hyoungseok Kim, and Gunhee Kim. Model-agnostic boundary-adversarial sampling for test-time generalization in few-shot learning. In *European conference on computer vision*. Springer, 2020.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015.

Chandrashekar Lakshminarayanan and Csaba Szepesvari. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, pp. 1347–1355. PMLR, 2018.

Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10657–10665, 2019.

Yoonho Lee and Seungjin Choi. Gradient-based meta-learning with learned layerwise metric and subspace. In *International Conference on Machine Learning*, pp. 2927–2936, 2018.

Puneet Mangla, Mayank Singh, Abhishek Sinha, Nupur Kumari, Vineeth N Balasubramanian, and Balaji Krishnamurthy. Charting the right manifold: Manifold mixup for few-shot learning. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2207–2216. IEEE, 2020.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21:1–76, 2020.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. In *International Conference on Learning Representations*, 2018.

Tsendsuren Munkhdalai, Xingdi Yuan, Soroush Mehri, and Adam Trischler. Rapid adaptation with conditionally shifted neurons. In *International Conference on Machine Learning*, pp. 3664–3673. PMLR, 2018.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 721–731, 2018.

Eunbyung Park and Junier B Oliva. Meta-curvature. *Advances in Neural Information Processing Systems*, 32:3314–3324, 2019.

Massimiliano Patacchiola, Jack Turner, Elliot J Crowley, Michael O'Boyle, and Amos J Storkey. Bayesian meta-learning for the few-shot setting via deep kernels. *Advances in Neural Information Processing Systems*, 33, 2020.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

Viraj Prabhu, Anitha Kannan, Murali Ravuri, Manish Chaplain, David Sontag, and Xavier Amatriain. Few-shot learning for dermatological disease diagnosis. In *Machine Learning for Healthcare Conference*, pp. 532–552. PMLR, 2019.

Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5822–5830, 2018.

Limeng Qiao, Yemin Shi, Jia Li, Yaowei Wang, Tiejun Huang, and Yonghong Tian. Transductive episodic-wise adaptive metric for few-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3603–3612, 2019.

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*, pp. 7229–7238, 2018.

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.

Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=rJY0-Kcll`.

Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Few-shot learning with embedded class models and shot-free meta training. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 331–339, 2019.

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. In *International Conference on Learning Representations*, 2018.

Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*, 2019.

Amartya Sanyal, Philip H Torr, and Puneet K Dokania. Stable rank normalization for improved generalization in neural networks and gans. In *International Conference on Learning Representations*, 2020.

Nikunj Saunshi, Arushi Gupta, and Wei Hu. A representation learning perspective on the importance of train-validation splitting in meta-learning. In *International Conference on Machine Learning*, pp. 9333–9343. PMLR, 2021.

Jürgen Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. PhD thesis, Technische Universität München, 1987.

Christian Simon, Piotr Koniusz, Richard Nock, and Mehrtash Harandi. Adaptive subspaces for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4136–4145, 2020.

Jake Snell and Richard Zemel. Bayesian few-shot classification with one-vs-each pólya-gamma augmented gaussian processes. In *International Conference on Learning Representations*, 2021.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 4077–4087, 2017.

Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 403–412, 2019.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1199–1208, 2018.

Sebastian Thrun and Lorien Pratt. *Learning to learn*. Springer Science & Business Media, 1998.

Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *European conference on computer vision*. Springer, 2020.

Michael E Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pp. 3630–3638, 2016.

Florian Wenzel, Kevin Roth, Bastiaan Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the bayes posterior in deep neural networks really? In *International Conference on Machine Learning*, pp. 10248–10259. PMLR, 2020.

Andrew Wilson and Ryan Adams. Gaussian process kernels for pattern discovery and extrapolation. In *International conference on machine learning*, pp. 1067–1075, 2013.

Jin Xu, Jean-Francois Ton, Hyunjik Kim, Adam Kosiorek, and Yee Whye Teh. Metafun: Meta-learning with iterative functional updates. In *International Conference on Machine Learning*, pp. 10617–10627. PMLR, 2020.

Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8808–8817, 2020.

Jaesik Yoon, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 7343–7353, 2018.

Sung Whan Yoon, Jun Seo, and Jaekyun Moon. Tapnet: Neural network augmented with task-adaptive projection for few-shot learning. In *ICML 2019 (International Conference on Machine Learning)*. ICML, 2019.

Sung Whan Yoon, Do-Yeon Kim, Jun Seo, and Jaekyun Moon. Xtarnet: Learning to extract task-adaptive representation for incremental few-shot learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10852–10860. PMLR, 2020.

Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12203–12213, 2020.

Manli Zhang, Jianhong Zhang, Zhiwu Lu, Tao Xiang, Mingyu Ding, and Songfang Huang. {IEPT}: Instance-level and episode-level pretext tasks for few-shot learning. In *International Conference on Learning Representations*, 2021.

A APPENDIX

A.1 PSEUDO CODE FOR MFRL

A.2 EXPERIMENT DETAILS

Sine waves are generated by $y = A\sin(x-\varphi)+\epsilon$, where amplitude $A \in [0.1, 5.0]$, phase $\varphi \in [0, \pi]$ and ϵ is white noise with standard deviation of 0.1 (Finn et al., 2017). Each sine wave contains 200 samples by sampling x uniformly from $[-5.0, 5.0]$. We generate 500 waves for training, validation and testing, respectively. All sine waves are different from each other. We use the same backbone network described in MAML (Finn et al., 2017): a two-layer MLP with 40 hidden units in each layer. We use the SGD optimizer with a learning rate of 10^{-3} over 8×10^4 training iterations and run SWA over 2×10^4 training iterations with a learning rate of 0.05.

Algorithm 1 Meta-free representation learning for few-shot learning

Merge all training tasks $\mathcal{D}_{\mathrm{tr}} = \{\mathcal{D}_\tau\}_{\tau=1}^{\mathcal{T}}$
Initialize model parameters $\theta = [\theta_f, \mathbf{W}]$
Maximize the likelihood on all training data $p\left(\mathcal{D}_{\mathrm{tr}} \mid \theta\right)$ using SGD
  Minimize the squared loss for regression problems
  Minimize the cross-entropy loss for classification problems
Run SWA to obtain θ_{SWA}
Discard $\mathbf{W}$ and freeze θ_f
Learn a new top layer using support data $\mathcal{D}$ in a test task:
  Learn a hierarchical Bayesian linear model for a regression task
  Learn a logistic regression model with the temperature scaling factor for a classification task

Head pose regression data is derived from the Queen Mary University of London multi-view face dataset (Gong et al., 1996). It contains images from 37 people and 133 facial images per person. Facial images cover a view sphere of 90° in yaw and 120° in tilt. The dataset is divided into 3192 training samples (24 people), 1064 validation samples (8 people), and 665 test samples (5 people). We use the same feature extractor described in literature (Patacchiola et al., 2020): a three-layer convolutional neural network, each with 36 output channels, stride 2, and dilation 2. We train the model on the training people set for 300 epochs using the SGD optimizer with a learning rate of 0.01 and run 25 epochs of SWA with a learning rate of 0.01.

miniImageNet is a 100-class subset of the original ImageNet dataset (Deng et al., 2009) for few-shot learning (Vinyals et al., 2016). Each class contains 600 images in RGB format of the size 84 $\times$ 84. miniImageNet is split into 64 training classes, 16 validation classes, and 20 testing classes, following the widely used data splitting protocol (Ravi & Larochelle, 2017).

tieredImageNet is another subset of the ImageNet dataset for few-shot learning (Ren et al., 2018). It contains 608 classes grouped into 34 categories, which are split into 20 training categories (351 classes), 6 validation categories (97 classes), and 8 testing categories (160 classes). Compared with miniImageNet, training classes in tieredImageNet are sufficiently distinct from test classes, making few-shot classification more difficult.

CIFAR-FS is a derivative of the original CIFAR-100 dataset by randomly splitting 100 classes into 64, 16, and 20 classes for training, validation, and testing, respectively (Bertinetto et al., 2019).

FC100 is another derivative of CIFAR-100 with minimized overlapped information between train classes and test classes by grouping the 100 classes into 20 superclasses (Oreshkin et al., 2018). They are further split into 60 training classes (12 superclasses), 20 validation classes (4 superclasses), and 20 test classes (4 superclasses).

The backbone model is trained on all training classes using $\mathcal{C}$-class cross-entropy loss by the SGD optimizer (momentum of 0.9 and weight decay of 1e-4) with a mini-batch size of 64. The learning rate is initialized as 0.05 and is decayed by 0.1 after 60, 80, and 90 epochs (100 epochs in total). After the SGD training converges, we run 100 epochs of SWA with a learning rate of 0.02. Note that MFRL is not sensitive to training epochs and learning rates in SWA (see Appendix A.5). The training images are augmented with random crop, random horizontal flip, and color jitter.

During testing, we conduct 5 independent runs of 600 randomly sampled few-shot classification tasks from test classes and calculate the average accuracy. Each task contains 5 classes, 1×5 or 5×5 support samples, and 75 query samples. A logistic regression model is learned using only the support samples. The classification accuracy is evaluated on the query samples.

A.3 Hierarchical Bayesian linear classification model

Similar to the hierarchical Bayesian linear regression model, the prior distribution over $\mathbf{w}$ is $p(\mathbf{w} \mid \lambda) = \prod_{i=0}^{p} \mathcal{N}\left(w_i \mid 0, \lambda\right)$, where λ is the precision in the Gaussian prior. The hyperprior on λ is defined as $p(\lambda) = \mathrm{Gamma}\left(\lambda \mid a, b\right)$. The posterior over all latent variables given the data is

$$p\left(\mathbf{w}, \lambda \mid \mathbf{X}, \mathbf{y}\right) = \frac{p\left(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \lambda\right) p\left(\mathbf{w} \mid \lambda\right) p(\lambda)}{p(\mathbf{y} \mid \mathbf{X})} \tag{7}$$

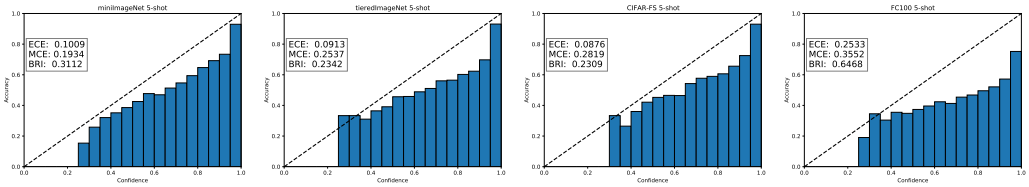


Figure 4: Reliability diagrams of hierarchical Bayesian linear classification models with flat and non-informative hyperpriors. The backbone is ResNet-12.

MCMC sampling (Hoffman & Gelman, 2014) is used to avoid potential deterioration in predictive performance due to approximated inference. A flat and non-informative hyperprior ($a = b = 10^{-6}$) is used because no prior knowledge is available. In Table 5, the hierarchical Bayesian linear classification model achieves slightly worse performance than the logistic regression model. However, the classification model is not well calibrated, as shown in Fig. 4.

Table 5: Comparison between logistic regression and hierarchical Bayesian linear models on 5-way 5-shot classification benchmarks using ResNet-12 backbone.

Top layer	miniImageNet	tieredImageNet	CIFAR-FS	FC-100
Logistic regression	83.81	86.87	87.4	61.1
Hierarchical Bayesian linear classification	81.94	85.32	85.9	59.2

After fine-tuning a and b using the meta-validation data, it is possible to get better calibrated classification models on test tasks. Besides, the classification accuracy is still slightly worse than logistic regression after hyperparameter tuning. Our observations align with a recent study, which shows that the Bayesian classification model cannot achieve similar performance to the non-Bayesian counterpart without tempering the posterior (Wenzel et al., 2020). We do not further experiment tempered posterior in the hierarchical Bayesian linear classification model because it introduces an extra temperature hyperparameter that requires tuning. The original purpose of introducing the hierarchical Bayesian model is to get an accurate and well calibrated classification model without hyperparameter tuning. Consequently, the hierarchical Bayesian model is not used in few-shot classification in that hierarchical Bayesian linear classification models cannot achieve high accuracy and good uncertainty calibration from a non-informative hyperprior. If hyperparameter tuning is inevitable, it is much easier to tune a logistic regression model with a temperature scaling factor, compared with tuning a hierarchical Bayesian model. Furthermore, the computational cost of learning a hierarchical Bayesian linear classification model via MCMC sampling is much larger than that of learning a logistic regression model.

A.4 Additional results on few-shot regression and classification

The additional results on few-shot regression using different activation functions are reported in Table 6. MFRL achieves high accuracy with different activation functions.

Table 6: 10-shot regression on sine waves with different activation functions.

Sine wave	MSE
MFRL (ReLU activation)	0.16 ± 0.51
MFRL (tanh activation)	0.018 ± 0.011
MFRL (erf activation)	0.016 ± 0.008

The few-shot classification results using a 4-layer convolutional neural network (or similar architectures) are reported in Table 7 and 8. Similar to the results using ResNet-12 and WRN-28-10, the proposed method outperforms a wide range of meta-learning approaches. Our method is only second to few-shot embedding adaptation with transformer (FEAT) (Ye et al., 2020) on miniImageNet dataset. In FEAT, a meta-learned transformer is built on top of the convolutional neural network

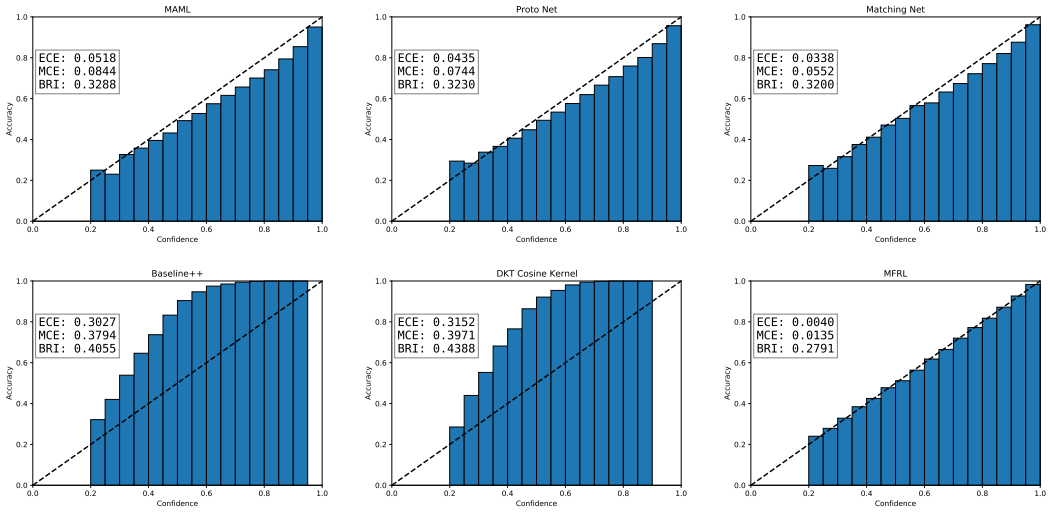


Figure 5: Reliability diagrams for 5-way 5-shot classification on miniImageNet

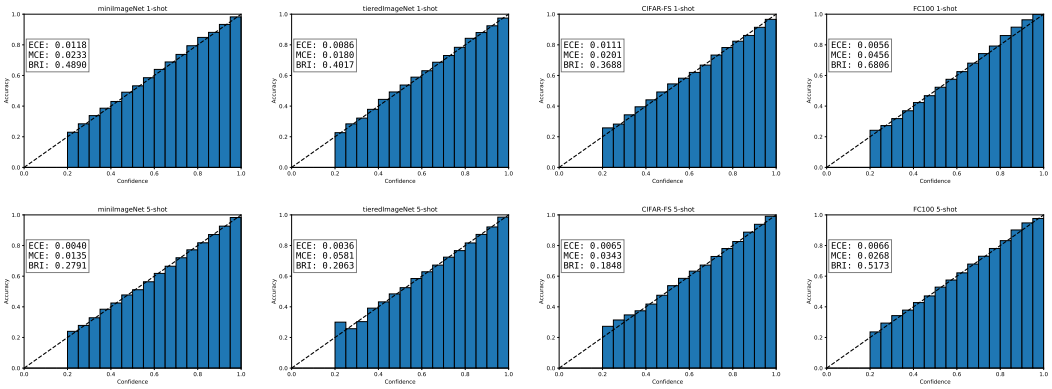


Figure 6: Reliability diagrams for 5-way few-shot classification using ResNet-12 backbone

Table 7: Few-shot classification results on miniImageNet and tieredImageNet.

Method	Backbone	miniImageNet 5-way		tieredImageNet 5-way	
		1-shot	5-shot	1-shot	5-shot
Matching Net (Vinyals et al., 2016)	Conv-4	43.56 ± 0.84	55.31 ± 0.73	54.48 ± 0.93	71.32 ± 0.78
Proto Net (Snell et al., 2017)	Conv-4	49.42 ± 0.78	68.20 ± 0.66	53.31 ± 0.89	72.69 ± 0.74
MAML (Finn et al., 2017)	Conv-4	48.70 ± 1.75	63.11 ± 0.92	-	-
SNAIL (Mishra et al., 2018)	Conv-4	$45.10 \pm$ NA	$55.20 \pm$ NA	-	-
VERSA (Gordon et al., 2019)	Conv-5	53.40 ± 1.82	67.37 ± 0.86	-	-
Meta Mixture (Jerfel et al., 2019)	Conv-4	49.60 ± 1.50	64.60 ± 0.92	-	-
RelationNet (Sung et al., 2018)	Conv-4	50.44 ± 0.82	65.32 ± 0.70	54.48 ± 0.93	71.32 ± 0.78
FPA (Qiao et al., 2018)	Conv-4	54.53 ± 0.40	67.87 ± 0.20	-	-
Shot-free (Ravichandran et al., 2019)	Conv-4	49.07 ± 0.43	65.73 ± 0.36	48.19 ± 0.43	65.50 ± 0.39
Baseline++ (Chen et al., 2019)	Conv-4	47.15 ± 0.49	66.18 ± 0.18	54.67 ± 0.61	72.37 ± 0.67
FEAT (Ye et al., 2020)	Conv-4	$\mathbf{55.15 \pm 0.20}$	$\mathbf{71.61 \pm 0.16}$	-	-
DKT (Patacchiola et al., 2020)	Conv-4	49.73 ± 0.07	64.00 ± 0.09	-	-
MFRL	Conv-4	53.62 ± 0.71	71.52 ± 0.60	$\mathbf{56.24 \pm 0.84}$	$\mathbf{72.88 \pm 0.76}$

Table 8: Few-shot classification results on CIFAR-FS and FC100.

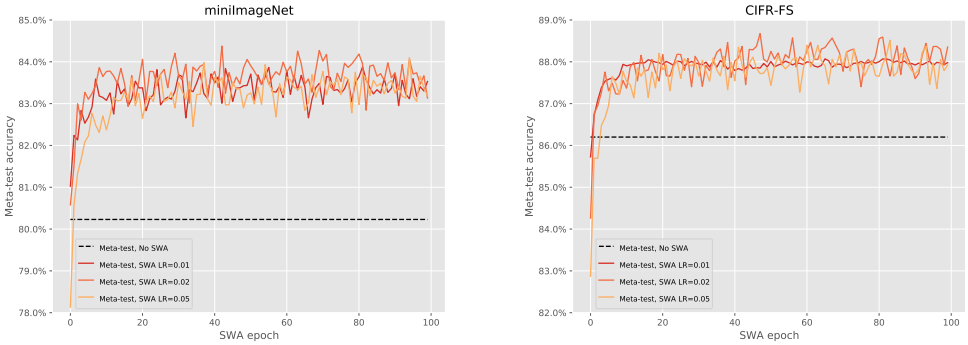
Method	Backbone	CIFAR-FS 5-way		FC100 5-way	
		1-shot	5-shot	1-shot	5-shot
Proto Net (Snell et al., 2017)	Conv-4	55.5 ± 0.7	72.0 ± 0.6	35.3 ± 0.6	48.6 ± 0.6
MAML (Finn et al., 2017)	Conv-4	58.9 ± 1.9	71.5 ± 1.0	38.1 ± 1.7	50.4 ± 1.0
RelationNet (Sung et al., 2018)	Conv-4	55.0 ± 1.0	69.3 ± 0.8	-	-
Shot-free (Ravichandran et al., 2019)	Conv-4	55.1 ± 0.5	71.7 ± 0.4	-	-
Baseline++ (Chen et al., 2019)	Conv-4	55.1 ± 0.9	72.3 ± 0.8	35.2 ± 0.7	49.8 ± 0.7
MFRL	Conv-4	$\mathbf{64.3 \pm 0.9}$	$\mathbf{79.4 \pm 0.5}$	$\mathbf{40.1 \pm 0.8}$	$\mathbf{54.4 \pm 0.7}$

to get task specific representation. Direct comparison to FEAT or other methods with transformer (Fei et al., 2021; Zhang et al., 2021) may not be fair because recent studies show that transformer itself can achieve better results than convolutional neural networks in image classification (Dosovitskiy et al., 2021). It is difficult to determine whether the performance improvement is due to the meta-learning algorithm or the transformer architecture.

The uncertainty calibration results of MFRL with the temperature scaling factor are presented in Fig. 6. The prediction confidence aligns well with the prediction accuracy. It demonstrates that MFRL with the temperature scaling factor results in well calibrated models.

A.5 Sensitivity of MFRL

The performance of MFRL is not sensitive to learning rates in SWA. As shown in Fig. 7, the representation learned by SWA generalizes better than the one from standard SGD, as long as the learning rate in SWA is in a reasonable range. In addition, the prediction accuracy on meta-test tasks keeps stable even after running SWA for many epochs on the training data. Therefore, MFRL is not sensitive to training epochs. This desirable property makes the proposed method easy to use when solving few-shot learning problems in practice.



(a) (b)

Figure 7: Evaluation on different learning rates and training epochs in SWA. (a) 5-way 5-shot accuracy on miniImageNet with ResNet12 backbone; (b) 5-way 5-shot accuracy on CIFAR-FS with WRN-28-10 backbone.