Meta-free few-shot learning via representation learning with weight averaging

Kuilin Chen
University of Toronto
Toronto, Canada
kuilin.chen@mail.utoronto.ca

Chi-Guhn Lee
University of Toronto
Toronto, Canada
cglee@mie.utoronto.ca

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.

Index Terms—Few-shot learning, stochastic weight averaging, low-rank representation

I. INTRODUCTION

Currently, the vast majority of few-shot learning methods are within the general paradigm of meta-learning (a.k.a. learning to learn) [1]–[3], which learns multiple tasks in an episodic manner to distill transferrable knowledge [4]–[6]. Although many episodic meta-learning methods report state-of-the-art (SOTA) performance, recent studies show that simple transfer learning methods with fixed embeddings [7], [8] can achieve similar or better performance in few-shot learning. 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 [9]. The quality of the presentation is not quantitatively defined, except for some empirical case studies [10]. Recent machine learning theories [11] indicate that low-rank representation leads to better sample efficiency in learning a new task. However, those theoretical studies 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 [7], [12], compared with state-of-the-art meta-learning methods [13], [14]. 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. 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 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 [16]–[19]. Few-shot learning models can be used in risk-averse applications such as medical diagnosis [20]. 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 [21]. 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 develop a simple transfer learning method as our own baseline to allow easy regularization towards more generalizable representation and calibration of prediction uncertainty. The regularization in the proposed transfer learning method works for regression and classification problems so that we can handle both problems within a common architecture. The calibration procedure is easily integrated into the developed transfer learning method to obtain few-shot learning models with good uncertainty quantification. Therefore, the resulting method, called Meta-Free Representation Learning (MFRL), overcomes the aforementioned limitations in existing transfer learning methods for few-shot learning. Our empirical studies demonstrate that the relatively overlooked transfer learning method can achieve high accuracy and well-calibrated
uncertainty in few-shot learning when it is combined with the proper regularization and calibration. Those two tools are also portable to meta-learning methods to improve accuracy and calibration, but the improvement is less significant compared with that of transfer learning.

We use stochastic weight averaging (SWA) [22], which is agnostic to loss function types, as implicit regularization to improve the generalization capability of the representation. We also shed light on that the effectiveness of SWA is due to its bias towards low-rank 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.

II. 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 [23], cosine similarity [4], and Euclidean distance to class-mean representation [6]. Instead of relying on predefined metrics, learnable similarity metrics are introduced to improve the few-shot classification performance [24], [25]. 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 [13], [26]–[28], graph neural networks [29], implicit class representation [30], and task-dependent conditioning [24], [31], [32]. 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 [5]. 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 [33]–[35]. 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 [36]. Meanwhile, R2-D2 [37] and MetaOptNet [38] 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 [7]. 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 [39], knowledge distillation [8], and auxiliary self-supervised learning [15]. 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 [11]. Furthermore, MFRL is the first transfer learning method that can handle both few-shot regression and classification problems and make predictions with well-calibrated uncertainty.

III. BACKGROUND

A. Episodic meta-learning

In episodic meta-learning, the meta-training data contains $T$ episodes or tasks, where the $t^{th}$ episode consists of data $D_t = \{(x_{t,j}, y_{t,j})\}_{j=1}^{N_{t}}$ with $N_{t}$ samples. Tasks and episodes are used interchangeably in the rest of the paper. Episodic meta-learning algorithms aim to find common model parameters $\theta$ which can be quickly adapted to task-specific parameters $\phi_{t}$ ($t = 1, \ldots, T$). For example, MAML-type algorithms assume $\phi_{t}$ is one or a few gradient steps away from $\theta$ [5], [16]–[18], while other meta-learning approaches assume that $\phi_{t}$ and $\theta$ share the parameters in the feature extractor and only differ in the top layer [6], [37], [38].
B. Stochastic weight averaging

The idea of stochastic weight averaging (SWA) along the trajectory of SGD goes back to Polyak–Ruppert averaging [40]. Theoretically, weight averaging results in faster convergence for linear models in supervised learning and reinforcement learning [41], [42]. In deep learning, we are more interested in tail stochastic weight averaging [43], which averages the weights after \( T \) training epochs. The averaged model parameters \( \theta_{\text{SWA}} \) 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 \( \theta_i \) 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 [22].

IV. 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.

A. Representation Learning

Common representation can be learned via maximization of the likelihood of all training data with respect to \( \theta \) rather than following episodic meta-learning. To do so, we group the data \( \mathcal{D}_t = \{ (x_{\tau,j}, y_{\tau,j}) \}_{j=1}^{N_{\tau}} \) from all meta-training tasks into a single dataset \( \mathcal{D}_t \). Given aggregated training data \( \mathcal{D}_t = \{ X, Y \} \), representation can be learned by maximizing the likelihood \( p(\mathcal{D}_t | \theta) \) with respect to \( \theta \). Let \( \theta = \{ \theta_f, \mathbf{W} \} \), where \( \theta_f \) represents parameters in the feature extractor and \( \mathbf{W} \) denotes the parameters in the top linear layer. The feature extractor \( h(x) \in \mathbb{R}^p \) is a neural network parameterized by \( \theta_f \) 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:

\[
\mathcal{L}_{\text{RP}}(\theta) = -\log p(\mathcal{D}_t | \theta) = \begin{cases} 
\mathcal{L}_{\text{MSE}}(\theta), & \text{regression} \\
\mathcal{L}_{\text{CE}}(\theta), & \text{classification}
\end{cases}
\]

where

\[
\mathcal{L}_{\text{MSE}}(\theta) = \frac{1}{2N} \sum_{\tau=1}^{T} \sum_{j=1}^{N_{\tau}} (y_{\tau,j} - \mathbf{w}^\top h(x_{\tau,j}))^2, \tag{2}
\]

\[
\mathcal{L}_{\text{CE}}(\theta) = - \sum_{j=1}^{N} \sum_{c=1}^{C} y_{j,c} \log \left( \frac{\exp(\mathbf{w}_c^\top h(x_j))}{\sum_{c=1}^{C} \exp(\mathbf{w}_c^\top h(x_j))} \right) \tag{3}
\]

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

Post-processing via SWA Minimizing the loss functions in (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 [11]. We aim to find low-rank representation \( \mathbf{F} = h(x) \) without episodic meta-learning, which is equivalent to finding the conjugate kernel \( K^C = \mathbf{F} \mathbf{F}^\top \) 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 \( \theta_T \) and get the finite width neural tangent kernel (NTK) \( K_{\text{NTK}}(x, x) = \mathbf{J}(x) \mathbf{J}(x)^\top \), where \( \mathbf{J}(x) = \nabla_{\theta} f_0(x) \in \mathbb{R}^{N \times |\theta|} \) is the Jacobian matrix, and \( K_{\text{NTK}} \) is a composite kernel containing \( K^C \) [44]. The distributions of eigenvalues for \( K_{\text{NTK}} \) and \( K^C \) are empirically similar. Analyzing \( K_{\text{NTK}} \) could shed light on the properties of \( K^C \). In parallel, \( K_{\text{NTK}} \) shares the same eigenvalues of the Gauss-Newton matrix \( G = \frac{1}{N'} \mathbf{J}(x)^\top \mathbf{J}(x) \). For linearized networks with squared loss, the Gauss-Newton matrix \( G \) well approximates the Hessian matrix \( H \) when \( y \) is well-described by \( f_0(x) \) [45]. This is the case when SGD converges to \( \theta_T \) 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 [46]. It is known that averaging the weights after SGD convergence in a local minimum basin pushes \( \theta_T \) towards the flat side of the loss valley [47]. 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 experiment Section.

B. Fine-Tuning

After representation learning is complete, \( \mathbf{W} \) is discarded and \( \theta_f \) 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 \( y = \mathbf{w}^\top h(x) + \epsilon \) on a fixed feature extractor \( h(x) \in \mathbb{R}^p \) with few-shot training data \( \mathcal{D} = \{ (x_i, y_i) \}_{i=1}^{n} \), where \( \mathbf{w} \) denotes the model parameters and \( \epsilon \) is Gaussian noise with zero mean and variance \( \sigma^2 \). To avoid interpolation on few-shot training data \( (n \ll p) \), a Gaussian prior

\(^1\) \( C \) is the total number of classes in \( \mathcal{D}_t \). Learning a \( C \)-class classification model solves all possible tasks in the meta-training dataset because each task \( \mathcal{D}_t \) only contains a subset of \( C \) classes.
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 $\lambda$ and $\sigma^2$ are defined as $p(\lambda) = \text{Gamma}(\lambda | a, b)$ and $p(\sigma^{-2}) = \text{Gamma}(\sigma^{-2} | c, d)$, respectively. The hyperpriors 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 $p(\mathbf{w}, \lambda, \sigma^2 | \mathbf{X}, \mathbf{y})$, where $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ and $\mathbf{y} = \{y_i\}_{i=1}^n$. However, the posterior distribution $p(\mathbf{w}, \lambda, \sigma^2 | \mathbf{X}, \mathbf{y})$ is intractable. The iterative optimization based approximate inference [48] is chosen because it is highly efficient. The point estimation for $\lambda$ and $\sigma^2$ is obtained by maximizing the marginal likelihood function $p(\mathbf{y} | \mathbf{X}, \lambda, \sigma^2)$. The posterior of model parameters $p(\mathbf{w} | \mathbf{X}, \mathbf{y}, \lambda, \sigma^2)$ is calculated using the estimated $\lambda$ and $\sigma^2$. Previous two steps are repeated alternately until convergence.

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

$$p(\mathbf{y}_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}, \lambda, \sigma^2) = \frac{\int p(\mathbf{y}_* | \mathbf{x}_*, \mathbf{w}, \sigma^2) p(\mathbf{w} | \mathbf{X}, \mathbf{y}, \lambda, \sigma^2) d\mathbf{w}}{\int p(\mathbf{y}_* | \mathbf{x}_*, \mathbf{w}, \sigma^2) p(\mathbf{w} | \mathbf{X}, \mathbf{y}, \lambda, \sigma^2) d\mathbf{w}},$$

which can be computed analytically because both distributions on the right hand side of (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

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

where $\lambda$ is the regularization coefficient, which affects the prediction accuracy and uncertainty. It is difficult to select an appropriate value of $\lambda$ in each of the meta-test tasks due to the lack of validation data in $\mathcal{D}$. We instead treat $\lambda$ as a global hyper-parameter so that the value of $\lambda$ should be determined based on the accuracy on meta-validation data. Note that the selected $\lambda$ with high validation accuracy does not necessarily lead to well calibrated classification models. As such, we introduce the temperature scaling factor [49] as another global hyper-parameter to scale the softmax output.

Given a test sample $\mathbf{x}_*$, the predicted probability for class $c$ becomes

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

where $T$ is the temperature scaling factor. In practice, we select the $L2$ regularization coefficient $\lambda$ and the temperature scaling factor $T$ as follows. At first, we set $T$ to 1, and do grid search on the meta-validation data to find the $\lambda$ resulting in the highest meta-validation accuracy. However, fine-tuning $\lambda$ does not ensure good calibration. It is the temperature scaling factor that ensures the good uncertainty calibration. Similarly, we do grid search of $T$ on the meta-validation set, and choose the temperature scaling factor resulting in the lowest expected calibration error [49]. Note that different values of $T$ do not affect the classification accuracy because temperature scaling is accuracy preserving.

V. 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.

A. Few-shot regression results

Sine waves [5] and head pose estimation [50] datasets are used to evaluate the performance of MFRL in few-shot regression. We use the same backbones in literature [50] to make fair comparisons.

![Fig. 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.](image-url)
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 [52]. 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.

B. Few-shot classification results

We conduct few-shot classification experiments on four widely used few-shot image recognition benchmarks: miniImageNet [53], tieredImageNet [54], CIFAR-FS [37], and FC100 [24]. In addition, we test our approach on a cross-domain few-shot classification task from the miniImageNet to CUB. The proposed method is applied to two widely used network architectures: ResNet-12 [30], [38] and wide ResNet (WRN-28-10) [12], [36].

The results of the proposed method and previous SOTA methods using similar backbones are reported in Table II and III. 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++ [7] and fine-tuning on logits [12]. 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 C-class classification model on all training data and then apply highly sophisticated meta-learning techniques to the pretrained model to achieve SOTA performance [36], [59]. 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 [8], 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 [8], [15].

MFRL is also applied to the cross-domain few-shot classification task as summarized in Table IV. MFRL outperforms other methods in this challenging task, indicating that the learned representation has strong generalization capability. We use the same hyperparameters (training epochs, learning rate, learning rate in SWA, SWA epoch, etc.) as in Table 2. The strong results indicate that MFRL is robust to hyperparameter choice. Surprisingly, meta-learning methods with adaptive embeddings do not outperform simple transfer learning methods like Baseline++ when the domain gap between base classes and novel classes is large. We notice that [8] also reports similar results that transfer learning methods show superior performance on a large-scale cross-domain few-shot classification dataset. We still believe that adaptive embeddings

### Table I

10-shot Regression on Sine Waves and Head Pose Estimation.

| Model          | MSE     |
|----------------|---------|
| MAML [3]       | 0.87 ± 0.06 |
| Bayesian MAML [18] | 0.84 ± 0.05 |
| ALPCA [57]     | 0.64 ± 0.09 |
| R2D2 [17]      | 0.48 ± 0.05 |
| DKT-RBF [50]   | 1.38 ± 0.03 |
| DKT-Spectral [50] | 0.68 ± 0.06 |
| MFRL (no SWA)  | 0.023 ± 0.016 |
| MFRL           | 0.016 ± 0.008 |

### Table II

Few-Shot Classification Results on miniImageNet and tieredImageNet.

| Method          | Baseline 5-way | Baseline 1-way |
|-----------------|---------------|---------------|
| ResNet-12       | 69.8 ± 0.67   | 61.2 ± 0.46   |
| WRN-28-10       | 72.0 ± 0.67   | 66.8 ± 0.56   |
| MAML            | 70.4 ± 0.67   | 69.0 ± 0.67   |
| MAML + RBF      | 71.4 ± 0.67   | 72.4 ± 0.67   |
| DKT             | 72.0 ± 0.67   | 66.8 ± 0.56   |
| DKT + RBF       | 73.5 ± 0.67   | 72.4 ± 0.67   |
| DKT + Spectral  | 72.5 ± 0.67   | 72.4 ± 0.67   |
| MFRL (no SWA)   | 74.0 ± 0.67   | 74.0 ± 0.67   |
| MFRL            | 75.0 ± 0.67   | 75.0 ± 0.67   |

### Table III

Few-Shot Classification Results on CIFAR-FS and FC100.

| Method          | CIFAR-10 5-way | FC100 5-way |
|-----------------|---------------|-------------|
| ResNet-12       | 71.2 ± 0.67   | 66.8 ± 0.56 |
| WRN-28-10       | 72.0 ± 0.67   | 66.8 ± 0.56 |
| MAML            | 70.4 ± 0.67   | 69.0 ± 0.67 |
| MAML + RBF      | 71.4 ± 0.67   | 72.4 ± 0.67 |
| DKT             | 72.0 ± 0.67   | 66.8 ± 0.56 |
| DKT + RBF       | 73.5 ± 0.67   | 72.4 ± 0.67 |
| DKT + Spectral  | 72.5 ± 0.67   | 72.4 ± 0.67 |
| MFRL (no SWA)   | 74.0 ± 0.67   | 74.0 ± 0.67 |
| MFRL            | 75.0 ± 0.67   | 75.0 ± 0.67 |

### Table IV

Cross-Domain Few-Shot Classification Results.

| Method          | Baseline 5-way | Baseline 1-way |
|-----------------|---------------|---------------|
| ResNet-12       | 69.8 ± 0.67   | 61.2 ± 0.46   |
| WRN-28-10       | 72.0 ± 0.67   | 66.8 ± 0.56   |
| MAML            | 70.4 ± 0.67   | 69.0 ± 0.67   |
| MAML + RBF      | 71.4 ± 0.67   | 72.4 ± 0.67   |
| DKT             | 72.0 ± 0.67   | 66.8 ± 0.56   |
| DKT + RBF       | 73.5 ± 0.67   | 72.4 ± 0.67   |
| DKT + Spectral  | 72.5 ± 0.67   | 72.4 ± 0.67   |
| MFRL (no SWA)   | 74.0 ± 0.67   | 74.0 ± 0.67   |
| MFRL            | 75.0 ± 0.67   | 75.0 ± 0.67   |

**LEO [8]**

| Method          | Baseline 5-way | Baseline 1-way |
|-----------------|---------------|---------------|
| ResNet-12       | 69.8 ± 0.67   | 61.2 ± 0.46   |
| WRN-28-10       | 72.0 ± 0.67   | 66.8 ± 0.56   |
| MAML            | 70.4 ± 0.67   | 69.0 ± 0.67   |
| MAML + RBF      | 71.4 ± 0.67   | 72.4 ± 0.67   |
| DKT             | 72.0 ± 0.67   | 66.8 ± 0.56   |
| DKT + RBF       | 73.5 ± 0.67   | 72.4 ± 0.67   |
| DKT + Spectral  | 72.5 ± 0.67   | 72.4 ± 0.67   |
| MFRL (no SWA)   | 74.0 ± 0.67   | 74.0 ± 0.67   |
| MFRL            | 75.0 ± 0.67   | 75.0 ± 0.67   |

**Fine-tune [12]**

| Method          | Baseline 5-way | Baseline 1-way |
|-----------------|---------------|---------------|
| ResNet-12       | 69.8 ± 0.67   | 61.2 ± 0.46   |
| WRN-28-10       | 72.0 ± 0.67   | 66.8 ± 0.56   |
| MAML            | 70.4 ± 0.67   | 69.0 ± 0.67   |
| MAML + RBF      | 71.4 ± 0.67   | 72.4 ± 0.67   |
| DKT             | 72.0 ± 0.67   | 66.8 ± 0.56   |
| DKT + RBF       | 73.5 ± 0.67   | 72.4 ± 0.67   |
| DKT + Spectral  | 72.5 ± 0.67   | 72.4 ± 0.67   |
| MFRL (no SWA)   | 74.0 ± 0.67   | 74.0 ± 0.67   |
| MFRL            | 75.0 ± 0.67   | 75.0 ± 0.67   |
should be helpful when the domain gap between base and novel classes is large. Nevertheless, how to properly train a model to obtain useful adaptive embeddings in novel tasks is an open question.

### TABLE IV

| Method       | Backbone | miniImageNet 5-way | CUB 5-way |
|--------------|----------|--------------------|-----------|
| MAML [5]     | WRN-28-10| 56.34 ± 0.30       | 78.91 ± 0.30 |
| LEO [36]     | WRN-28-10| 57.66 ± 0.48       | 79.64 ± 0.48 |
| MTL [39]     | WRN-28-10| 60.12 ± 0.30       | 80.21 ± 0.30 |
| Matching Net [4] | WRN-28-10| 68.60 ± 0.47       | 80.86 ± 0.47 |
| SIB [15]     | WRN-28-10| 68.49 ± 0.48       | 80.49 ± 0.48 |
| Baseline [7] | WRN-28-10| 68.49 ± 0.48       | 80.49 ± 0.48 |
| Baseline++ [7] | WRN-28-10| 68.49 ± 0.48       | 80.49 ± 0.48 |
| MFRL (w/o SWA) | WRN-28-10| 46.98 ± 0.51       | 66.92 ± 0.42 |
| MFRL         | WRN-28-10| 46.98 ± 0.51       | 66.92 ± 0.42 |

### 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.

### 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 [62]. Fig. 3 shows the classification reliability diagrams along with widely used metrics for uncertainty calibration, including expected calibration error (ECE) [49], maximum calibration error (MCE) [63], and Brier score (BRI) [64]. 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.

### Application in meta-learning

Meanwhile, we also apply SWA to episodic meta-learning methods, such as Proto Net, MAML, and Matching Net, to improve their classification accuracy. The results in Table V 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 V

| Method     | Proto Net | MAML | Matching Net |
|------------|-----------|------|--------------|
| w/o SWA    | 60.37     | 78.92| 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 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.

VI. DISCUSSION

SWA has been applied to supervised learning of deep neural networks [22], [65] and its effectiveness was attributed to convergence to a solution on the flat side of an asymmetric loss valley [47]. 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 [66]–[68]. However, some of these 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.

VII. 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

[1] Y. Bengio, S. Bengio, and J. Cloutier, “Learning a synaptic learning rule,” in IJCNN-91-Seattle International Joint Conference on Neural Networks, vol. 2. IEEE, 1991, pp. 969–vol.
[2] J. Schmidhuber, “Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook,” Ph.D. dissertation, Technische Universität München, 1987.
[3] S. Thrun and L. Pratt, Learning to learn. Springer Science & Business Media, 1998.
[4] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra et al., “Matching networks for one shot learning,” in Advances in neural information processing systems, 2016, pp. 3630–3638.
[5] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR.org, 2017, pp. 1126–1135.
[6] J. Snell, K. Swersky, and R. Zemel, “Prototypical networks for few-shot learning,” in Advances in Neural Information Processing Systems, 2017, pp. 4077–4087.
[7] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, “A closer look at few-shot classification,” in International Conference on Learning Representations, 2019.
[8] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola, “Rethinking few-shot image classification: a good embedding is all you need?” in European conference on computer vision. Springer, 2020.
[9] A. Raghu, M. Raghu, S. Bengio, and O. Vinyals, “Rapid learning or feature reuse? towards understanding the effectiveness of maml,” in International Conference on Learning Representations, 2020.
[10] M. Goldblum, S. Reich, L. Fowl, R. Ni, V. Cherepanova, and T. Goldstein, “Unraveling meta-learning: Understanding feature representations for few-shot tasks,” in Proceedings of the 37th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. Virtual: PMLR, 13–18 Jul 2020, pp. 3607–3616.
[11] N. Sausnhai, A. Gupta, and W. Hu, “A representation learning perspective on the importance of train-validation splitting in meta-learning,” in International Conference on Machine Learning. PMLR, 2021, pp. 9333–9343.
[12] G. S. Dhillon, P. Chaudhari, A. Ravichandran, and S. Soatto, “A baseline for few-shot image classification,” in International Conference on Learning Representations, 2020.
[13] H.-J. Ye, H. Hu, D.-C. Zhan, and F. 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, 2020, pp. 8808–8817.
[14] C. Zhang, Y. Cai, G. Lin, and C. Shen, “Deeppend: 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, 2020, pp. 12203–12213.
[15] P. Mangla, M. Singh, A. Sinha, N. Kumari, V. N. Balasubramanian, and B. Krishnamurthy, “Charting the right manifold: Manifold mixup for few-shot learning,” in 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2020, pp. 2207–2216.
[16] E. Grant, C. Finn, S. Levine, T. Darrell, and T. Griffiths, “Recasting gradient-based meta-learning as hierarchical bayes,” in International Conference on Learning Representations, 2018.
[17] C. Finn, K. Xu, and S. Levine, “Probabilistic model-agnostic meta-learning,” in Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018, pp. 9537–9548.
[18] J. Yoon, T. Kim, O. Dia, S. Kim, Y. Bengio, and S. Ahn, “Bayesian model-agnostic meta-learning,” in Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018, pp. 7343–7353.
[19] J. Snell and R. Zemel, “Bayesian few-shot classification with ones-each polya-gamma augmented gaussian processes,” in International Conference on Learning Representations, 2021.
[20] V. Prabhu, A. Kannan, M. Ravuri, M. Chaplain, D. Sontag, and X. Amatriain, “Few-shot learning for dermatological disease diagnosis,” in Machine Learning for Healthcare Conference. PMLR, 2019, pp. 532–552.
[21] E. Begoli, T. Bhattacharya, and D. Kusnezov, “The need for uncertainty quantification in machine-assisted medical decision making,” Nature Machine Intelligence, vol. 1, no. 1, pp. 20–23, 2019.
[22] P. Izmiaiov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, “Averaging weights leads to wider optimas and better generalization,” in 34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018. Association For Uncertainty in Artificial Intelligence (AUAI), 2018, pp. 876–885.
[23] G. Koch, R. Zemel, and R. Salakhutdinov, “Siamese neural networks for one-shot image recognition,” in ICML deep learning workshop, vol. 2. Lille, 2015.
[24] B. Oreshkin, P. R. López, and A. Lacoste, “Tadam: Task dependent adaptive metric for improved few-shot learning,” in Advances in Neural Information Processing Systems, 2018, pp. 721–731.
[25] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, “Learning to compare: Relation network for few-shot learning,” in Advances in Neural Information Processing Systems, 2017, pp. 4077–4087.
Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1199–1208.

[26] Fei, Z. Lu, T. Xiang, and S. Huang, “[MELR]: Meta-learning via modeling episode-level relationships for few-shot learning,” in International Conference on Learning Representations, 2021.

[27] S. Gidaris and N. Komodakis, “Dynamic few-shot visual learning without forgetting,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4367–4375.

[28] M. Zhang, J. Zhang, Z. Lu, T. Xiang, M. Ding, and S. Huang, “[IEPT]: Instance-level and episode-level pretext tasks for few-shot learning,” in International Conference on Learning Representations, 2021.

[29] V. Garcia and J. B. Estrach, “Few-shot learning with graph neural networks,” in 6th International Conference on Learning Representations, ICLR 2018, 2018.

[30] A. Ravichandran, R. Bhotika, and S. Soatto, “Few-shot learning with embedded class models and shot-free meta training,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 331–339.

[31] S. W. Yoon, D. Kim, J. Seo, and J. Moon, “Xtranet: 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, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 10852–10860.

[32] S. W. Yoon, J. Seo, and J. Moon, “Tapnet: Neural network augmented with task-adaptive projection for few-shot learning,” in ICML 2019 (International Conference on Machine Learning). ICML, 2019.

[33] A. Antoniou, H. Edwards, and A. Storkey, “How to train your maml,” in International Conference on Learning Representations, 2018.

[34] Y. Lee and S. Choi, “Gradient-based meta-learning with learned layerwise metric and subspace,” in International Conference on Machine Learning, 2018, pp. 2927–2936.

[35] A. Nichol, J. Achiam, and J. Schulman, “On first-order meta-learning algorithms,” arXiv preprint arXiv:1803.02999, 2018.

[36] A. A. Rusu, D. Bai, J. R. Sigmokowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell, “Meta-learning with latent embedding optimization,” in International Conference on Learning Representations, 2019.

[37] L. Bertinetto, J. F. Henriques, P. Torr, and A. Vedaldi, “Meta-learning with differentiable closed-form solvers,” in International Conference on Learning Representations, 2019.

[38] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, “Meta-learning with differentiable convex optimization,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10657–10665.

[39] N. Dvornik, C. Schmid, and J. Maial, “Diversity with cooperation: Ensemble methods for few-shot classification,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3723–3731.

[40] B. T. Polyak and A. B. Juditsky, “Acceleration of stochastic approximation by averaging,” SIAM journal on control and optimization, vol. 30, no. 4, pp. 838–855, 1992.

[41] F. Bach and E. 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, 2013, pp. 773–781.

[42] C. Lakshminarayanan and C. Szepesvari, “Linear stochastic approximation: How far does constant step-size and iterate averaging go? in International Conference on Artificial Intelligence and Statistics. PMLR, 2018, pp. 1347–1355.

[43] P. Jain, S. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford, “Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification,” Journal of Machine Learning Research, vol. 18, 2018.

[44] Z. Fan and Z. Wang, “Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks,” Advances in Neural Information Processing Systems, vol. 33, 2020.

[45] J. Martens, “New insights and perspectives on the natural gradient method,” Journal of Machine Learning Research, vol. 21, pp. 1–76, 2020.

[46] N. S. Keskar, J. Nocedal, P. T. P. Tang, D. Mudigere, and M. Smelyanskiy, “On large-batch training for deep learning: Generalization gap and sharp minima,” in 5th International Conference on Learning Representations, ICLR 2017, 2017.

[47] H. He, G. Huang, and Y. Yuan, “Asymmetric valleys: Beyond sharp and flat local minima,” in Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 2540–2560.

[48] M. E. Tipping, “Sparse bayesian learning and the relevance vector machine,” Journal of machine learning research, vol. 1, no. Jun, pp. 211–244, 2001.

[49] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in International Conference on Machine Learning. PMLR, 2017, pp. 1321–1330.

[50] P. T. P. Tang, J. Turner, E. J. Crowley, M. O’Boyle, and A. J. Storkey, “Baysian meta-learning for the few-shot setting via deep kernels,” Advances in Neural Information Processing Systems, vol. 33, 2020.

[51] J. Harrison, A. Sharma, and M. Pavone, “Meta-learning priors for efficient online bayesian regression,” in International Workshop on the Algorithmic Foundations of Robotics. Springer, 2018, pp. 318–337.

[52] A. Wilson and R. Adams, “Gaussian process kernels for pattern discovery and extrapolation,” in International conference on machine learning, 2013, pp. 1067–1075.

[53] S. Ravi and H. 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. [Online]. Available: https://openreview.net/forum?id=BY0-Kell

[54] M. Sun, Y. Shi, J. Shu, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel, “Meta-learning for semi-supervised few-shot classification,” in International Conference on Learning Representations, 2019.

[55] T. Mukhaidalai, X. Yuan, S. Mehr, and A. Trischler, “Rapid adaptation with conditionally shifted neurons,” in International Conference on Machine Learning. PMLR, 2018, pp. 3664–3673.

[56] C. Simon, P. Koniusz, R. Nock, and M. Harandi, “Adaptive subspaces for few-shot learning,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4136–4145.

[57] S. X. Hu, P. G. Moreno, Y. Xiao, X. Shen, G. Obozinski, N. Lawrence, and A. Damianou, “Empirical bayes transductive meta-learning with synthetic gradients,” in International Conference on Learning Representations, 2020.

[58] J. Xu, J.-F. Ton, H. Kim, A. Kosiorek, and Y. W. Teh, “Metafun: Meta-learning with iterative functional updates,” in International Conference on Machine Learning, PMLR, 2020, pp. 10617–10627.

[59] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele, “Meta-transfer learning for few-shot learning,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2019, pp. 403–412.

[60] L. Qiao, Y. Shi, J. Li, Y. Wang, T. Huang, and Y. Tian, “Transductive episodic-wise adaptive metric for few-shot learning,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3603–3612.

[61] J. Kim, H. Kim, and G. Kim, “Model-agnostic boundary-adversarial sampling for test-time generalization in few-shot learning,” in European conference on computer vision. Springer, 2020.

[62] M. H. DeGroot and S. E. Fienberg, “The comparison and evaluation of forecasters,” Journal of the Royal Statistical Society: Series D (The Statistician), vol. 32, no. 1–2, pp. 12–22, 1983.

[63] M. P. Naen, G. Cooper, and M. Hauskrecht, “Obtaining well calibrated probabilities using bayesian binning,” in Proceedings of the AAAI Conference on Artificial Intelligence, 29, 2015.

[64] G. W. Brier, “Verification of forecasts expressed in terms of probability,” Monthly weather review, vol. 78, no. 1, pp. 1–3, 1950.

[65] B. Athiwaratkun, M. Finzi, P. Izmaiov, and A. G. Wilson, “There are many consistent explanations of unlabeled data: Why you should average,” in International Conference on Learning Representations, 2019.

[66] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky, “Spectrally-normalized margin bounds for neural networks,” Advances in Neural Information Processing Systems, vol. 30, pp. 6240–6249, 2017.

[67] B. Neyshabur, S. Bhojanapalli, and N. Srebro, “A pac-bayesian approach to spectrally-normalized margin bounds for neural networks,” in International Conference on Learning Representations, 2018.

[68] A. Sanyal, P. H. Torr, and P. K. Dokania, “Stable rank normalization for semi-supervised few-shot learning,” in International Conference on Learning Representations, 2020.