A Statistical Model for Predicting Generalization in Few-Shot Classification

Yassir Bendou[1], Vincent Gripon[1], Bastien Pasdeloup[1], Giulia Lioi[1], Lukas Mauch[2],
Stefan Uhlich[2], Fabien Cardinaux[2], Ghouthi Boukli Hacene[23] and Javier Alonso Garcia[2]

[1]IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238 Brest, France
[2]Sony Europe, R&D Center, Stuttgart Laboratory 1, Germany
[3]Mila, Montréal, Canada
[1]name.surname@imt-atlantique.fr, [2]name.surname@sony.com

***Abstract*—The estimation of the generalization error of classifiers often relies on a validation set. Such a set is hardly available in few-shot learning scenarios, a highly disregarded shortcoming in the field. In these scenarios, it is common to rely on features extracted from pre-trained neural networks combined with distance-based classifiers such as nearest class mean. In this work, we introduce a Gaussian model of the feature distribution. By estimating the parameters of this model, we are able to predict the generalization error on new classification tasks with few samples. We observe that accurate distance estimates between class-conditional densities are the key to accurate estimates of the generalization performance. Therefore, we propose an unbiased estimator for these distances and integrate it in our numerical analysis. We empirically show that our approach outperforms alternatives such as the leave-one-out cross-validation strategy.**

***Index Terms*—few-shot learning, classification, deep learning, generalization**

I. Introduction

The problem of few-shot classification, where the number of training samples per class is small (typically less than ten [1]), has known a large number of contributions [2], [3]. Most of current state-of-the-art solutions consist in using a pre-trained deep feature extractor to embed samples in a feature space where classes are expected to be easier to discriminate, followed by a distance-based classifier such as the nearest-class mean (NCM) [4], [5]. However, measuring the performance of such classifier in the few-shot scenario is not straightforward. The classical approach is to perform a leave-one-out cross-validation, where one sample is arbitrarily removed from the training set to be used as a validation probe, this process being repeated a large number of times to obtain as average the expected accuracy [6].

Finding an alternative to cross-validation has been extensively studied in the literature for standard classification settings where a large number of training samples is available [7]. In this work, we are mainly interested in proposing an alternative to cross-validation and to existing generalization prediction methods in the context of few-shot classification. The proposed method uses a statistical model of class-conditional densities in the feature space to estimate the probability of error, as represented in Figure 1.

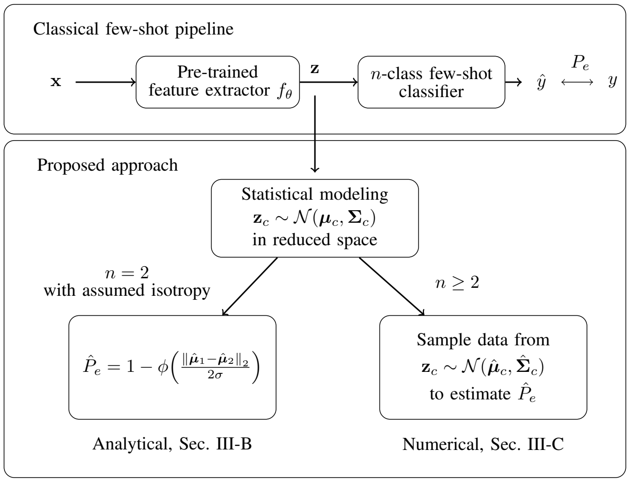

Classical few-shot pipeline
$\mathbf{x}$
Pre-trained feature extractor f_θ
$\mathbf{z}$
n-class few-shot classifier
$\hat{y}$
P_e
y
Proposed approach
Statistical modeling $\mathbf{z}_c \sim \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ in reduced space
$n = 2$ with assumed isotropy
$n \geq 2$
$\hat{P}_e = 1 - \phi\left(\frac{\|\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2\|_2}{2\sigma}\right)$
Sample data from $\mathbf{z}_c \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_c, \hat{\boldsymbol{\Sigma}}_c)$ to estimate $\hat{P}_e$
Analytical, Sec. III-B
Numerical, Sec. III-C

Fig. 1: The classical few-shot pipeline uses a pre-trained feature extractor to embed data $\mathbf{x}$ into features $\mathbf{z} = f_\theta(\mathbf{x})$ followed by a classifier. Our approach models the class-densities of the features as Gaussian distributions in a lower dimensional subspace and predicts the probability error $\hat{P}_e$ of the classifier, either analytically or by sampling.

There are two key difficulties here: 1) We need to find a model that is expressive enough to capture the data distribution while depending on as few parameters as possible to accurately estimate them in the very low data regime. 2) Our derived expression for the probability of error depends on the distances between class centers. We observe that the naive estimate for the distances is biased, which leads to underestimating the probability of error especially when working in high-dimensional spaces with very few samples.

The main contributions[1] of our work are as follows:

- We introduce a statistical model of class-conditional densities in the feature space and propose an unbiased estimator for the distances between class centers;
- We demonstrate that our method outperforms other model-free generalization error predictors on standardized few-shot classification benchmarks.

[1] Our code is available at: https://github.com/ybendou/fs-generalization.

II. Related Work

A. Classification in few-shot learning

In this paper, we refer to P for probabilities and p for probability density functions. We formalize few-shot classification as follows:

Definition 1 (Few-shot classification). *Let $\mathcal{D}_B = \{(\mathbf{x}'_i, y'_i)\}_{i=1}^m$ be a large base dataset where $\forall i, (\mathbf{x}'_i, y'_i)$ are i.i.d samples drawn from the true joint probability distribution $p_{\mathcal{X}_B,\mathcal{Y}_B}$ and $\forall i, y'_i \in \mathcal{Y}_B$ is the class associated with $\mathbf{x}'_i$. We are also given a small few-shot training dataset $\mathcal{D} = \{(\mathbf{x}_j, y_j)\}_{j=1}^\ell$, where $\forall j, y_j \in \mathcal{Y} \neq \mathcal{Y}_B$ and $\forall j, (\mathbf{x}_j, y_j) \sim p_{\mathcal{X},\mathcal{Y}}$.*

The goal of few-shot classification is to train a classifier C using $\mathcal{D}$, with the potential help of $\mathcal{D}_B$.

Since there are few training samples in few-shot classification tasks, it is not feasible to train a deep neural network architecture with a large number of parameters. Most methods consist in pre-training a deep feature extractor f_θ with parameter set θ using $\mathcal{D}_B$. This feature extractor is then either adapted or used as it is on $\mathcal{D}$ to produce feature vectors $\mathbf{z} = f_\theta(\mathbf{x})$ in an Euclidean space. The few-shot classifier then works with the modified training dataset $f_\theta(\mathcal{D}) = \{(\mathbf{z}_j, y_j)\}_{j=1}^\ell$.

There are multiple strategies on how to train f_θ that can roughly be classified as optimization-based approaches and distance metric approaches.

Optimization-based methods aim to effectively adapt f_θ to new few-shot tasks. Meta-learning has been a popular method especially with the introduction of MAML [8] and its variants [9]. On the other hand, distance-based approaches aim to learn a good feature extractor [10].

In the distance-based category, the feature extractor can be trained in two different ways. The first one relies on episodic training where the idea is to reproduce the same conditions of the few-shot adaptation phase during the pre-training of the feature extractor [11]. The second way to train a feature extractor is to use a standard cross-entropy loss. This is usually referred to as transfer learning, which has been successful in recent years and largely adopted due to its competitive performance compared to episodic training, while being relatively simple to implement [4], [5].

Various few-shot classifiers have been proposed in the literature such as fine tuning a multi-layer perceptron with a cross-entropy loss [10], which has been criticized for being biased in few-shot regimes [12]. Other distance-based approaches such as using a nearest class mean (NCM) classifier [5] or an earth distance metric using optimal transport [13] have also been studied. We adhere to the NCM approach due to its well established performance and its simplicity.

Definition 2 (Nearest class mean classifier). *A nearest class mean classifier C_{NCM} is the optimal classifier when class-conditional densities follow a Gaussian distribution with equal isotropic covariance and uniform prior across classes [14]:*

$$p(\mathbf{z} \mid y = c) = \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}_c, \sigma^2\mathbf{I}) \; , \tag{1}$$

where $\boldsymbol{\mu}_c$ is the center of class $c \in \mathcal{Y}$, σ is the standard deviation and $\mathbf{I}$ the identity matrix. The classification of a new sample $\mathbf{z}$ is performed according to:

$$C_{NCM}(\mathbf{z}) = \arg\min_{c\in\mathcal{Y}} \|\mathbf{z} - \boldsymbol{\mu}_c\|_2 \; . \tag{2}$$

In practice we estimate the class centers from the training data $\mathcal{D}$ using the empirical average of each class.

Once the class centers are estimated, there are a maximum of $(n-1)$ dimensions of interest in the considered Euclidean space, which correspond to directions between class centers, where n is the cardinal of $\mathcal{Y}$. Remaining ones can be disregarded, as they produce contributions that are orthogonal to the axes between class centers. Projecting the data onto this lower dimensional subspace is preferable in few-shot classification as it allows to work with lower dimensions [15]. Such a projection can be performed using for example a QR decomposition [16] to reduce data dimension. Indeed, as shown in [15], a subspace of dimension $(n-1)$ does not impact the boundary decisions of a NCM classifier.

B. Predicting Generalization

Predicting generalization is one of the most important topics in machine learning. It was brought into focus by [17], who asked the question of how one can measure the generalization from training data, showing that neural networks can easily fit randomly labeled data with high accuracy but with low generalization capabilities.

Many works have been proposed on generalization of neural networks trained on large training datasets such as ImageNet [18]. The proposed methods in the literature can be summarized into few different families. The first one is PAC-Bayes methods where the generalization behavior of a model is described by probably approximately correct (PAC) bounds [19]; these methods often provide an upper-bound on the generalization error and the results are often restricted to a small set of models (*e.g.*, no depth variations). The second family is norm-based methods, which analyzes the neural network weights. These methods have shown to perform poorly [7]. The last family of methods aims to analyze the intermediate representation of the training data in the feature space such as using the Davies-Bouldin Index [20] which is a clustering measure of the training data. Note that the main focus of these methods is to predict the generalization of a model trained for a certain task where large training data is available. Predicting generalization when working with few labeled samples has mainly been addressed when using meta-learning methods [21], [22]. The closest work to ours is [23], where some of the strategies for predicting generalization mentioned before have been tested for few-shot classification using transfer learning.

Differently to previously mentioned works, in this paper we aim at deriving a statistical model of the class-conditional densities in the feature space and to use this model to estimate the generalization error. As we will demonstrate in the experiments, the proposed methodology can outperform the previously mentioned ones in few-shot settings.

III. Methodology

A. Statistical model

The first step of our proposed methodology consists in proposing a statistical model for class-conditional densities in the feature space. Let us assume that each class follows a Gaussian distribution with a uniform prior across the classes, i.e, $p(y_{\mathbf{z}} = c) = \frac{1}{n}, \forall c \in \mathcal{Y}$, where n is the cardinal of $\mathcal{Y}$, which is reasonable to assume in a few-shot setting [15]. The conditional densities are defined as:

$$p(\mathbf{z} \mid y = c) = \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \ , \tag{3}$$

where $\boldsymbol{\Sigma}_c$ is the covariance matrix of class c.

Our hypothesis stands from the fact that given a well pretrained feature extractor, each class in the feature space should follow a multivariate Gaussian distribution centered around a class center. This assumption has been largely adopted in the few-shot literature [24]–[27]. Furthermore, the performance we obtain through our experimental results in Section IV show that this model fits our data.

Predicting generalization can be defined as predicting the probability of error from the classifier C. Let $R_c = \{\mathbf{z} \mid C(\mathbf{z}) = c\}$ be the decision region for class c using the classifier C and $R = \cup_{c \in \mathcal{Y}} R_c$. The theoretical error of our problem is defined by the sum of integrals:

$$P_e = \sum_{c \in \mathcal{Y}} \int_{R \setminus R_c} p(\mathbf{z} \mid y = c) p(y = c) \, d\mathbf{z} \ . \tag{4}$$

B. Analytical insight

A closed form solution of Equation 4 when $n > 2$ is often intractable. In this section we focus on the case of binary classification and derive an analytical expression for P_e.

For the case of a binary classifier of isotropic Gaussian data with equal standard deviation σ and class centers $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_b$, P_e has a closed form which only depends on the distance between the class centers $r = \|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|_2$ and σ:

$$P_e = 1 - \phi\Big(\frac{r}{2\sigma}\Big) \ , \tag{5}$$

where ϕ is the cumulative distribution function of $\mathcal{N}(0, 1)$.

To estimate P_e, we typically estimate r and σ using i.i.d samples such that $\hat{r} = \|\hat{\boldsymbol{\mu}}_a - \hat{\boldsymbol{\mu}}_b\|_2$, where $\hat{\boldsymbol{\mu}}$ is the empirical mean estimate. Under this analytical form, we can derive a statistical bound for $\hat{P}_e = 1 - \phi\big(\frac{\hat{r}}{2\hat{\sigma}}\big)$ and prove that this bound is of $\mathcal{O}(\frac{1}{\sqrt{k}})$. More details on this statistical bound can be found in the long version of this paper [28].

Estimating the probability error depends on estimating the distance between class centers. The naive approach for estimating distances is usually performed by estimating the means of each distribution using the empirical mean estimate and computing $\hat{r} = \|\hat{\boldsymbol{\mu}}_a - \hat{\boldsymbol{\mu}}_b\|_2$ to which we refer to as the naive estimator. However, this estimation of the distance between the class centers is biased.

Lemma 1. *Let $(\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_k)$ and $(\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_k)$ be two sequences of i.i.d random variables drawn from their respective multivariate probability distributions p_a and p_b assumed independent with finite expected values $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_b$ and finite second order moment with covariance matrices $\boldsymbol{\Sigma}_a$ and $\boldsymbol{\Sigma}_b$. Let $\hat{r}$ be the naive estimator for the distances using $\hat{\boldsymbol{\mu}}_a = \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{a}_i$ and $\hat{\boldsymbol{\mu}}_b = \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{b}_i$ the mean estimator of each of the two sequences, then:*

$$\mathbb{E}_{\substack{\boldsymbol{a} \sim p_a \\ \boldsymbol{b} \sim p_b}}\big(\hat{r}^2\big) - r^2 = \frac{\mathrm{Tr}(\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_b)}{k} \ . \tag{6}$$

The bias is a function of the noise and the number of samples k. Our numerical approach includes a bias reduction step. The correction is performed in the original high dimensional space. The proof of Lemma 1 as well as experiments to show the extent of this bias are included in the long version of this paper [28]. We also provide experiments with and without the bias correction to demonstrate its importance.

C. Numerical insight

Computing P_e analytically in equation 4 for $n > 2$ is hard. In practice we can approximate P_e using a Monte Carlo method. For each class, we draw a large number of data points from Gaussian distributions fitted to the few-shot training dataset to artificially enrich it and compute the classifier's decisions on a *virtual validation set*. We sample in the reduced subspace with $(n - 1)$ dimensions.

We take into account the positively biased distances between class centers which lead to underestimated P_e (as demonstrated in long version of this paper [28]). Correcting this bias is key to accurately estimating P_e. In fact, for a distance-based classifier, the absolute positioning of the class centers do not affect its decisions. In order to perform the sampling, we generate a set of points which respect the new estimated distances using the Nonmetric Multidimensional Scaling algorithm (MDS) [29].

Estimating P_e by sampling means that there are no restrictions for choosing the covariance of the data. In section IV we compare the performance for different covariance matrices and run an experiment to validate our choice, i.e.: 1) using the identity matrix, 2) using a shared isotropic covariance matrix across classes, 3) using isotropic covariance matrix per class, 4) using the full covariance matrix per class.

IV. Experiments

A. Datasets and implementation details

We use two standardized few-shot vision classification benchmarks: Tiered-ImageNet [30] and Meta-dataset [31] (inc. ImageNet and VGG-Flower). We sample 10^3 problems from each dataset which contains at least 500 samples per class. We pre-train a 512-dimensional ResNet-18 architecture using the standard procedure from [4]. Details can be found in the long version of this paper [28].

B. Estimating the first and second order moments

We conduct a first experiment to determine the best type of covariance matrix for our Gaussian model. Namely, we can choose a simple identity covariance matrix, a shared isotropic model with a scaling parameter σ, an isotropic model with each class having its own scale, or a completely free model. We generate 10^3 5-class few-shot problems and measure the average Kullback-Leibler (KL) divergence between the obtained covariance matrices and the ground-truth one obtained with a large number of labelled samples from the same classes. Figure 2 shows a trade-off between model complexity and overfitting. For more samples, the free model with more parameters performs better, while for fewer samples it is better to use a shared isotropic model. That is why in our next experiments, we use a shared isotropic covariance matrix for $k \leq (n-1)^2$ (intersection between model 2 and 4 in Figure 2), and a completely free model otherwise.

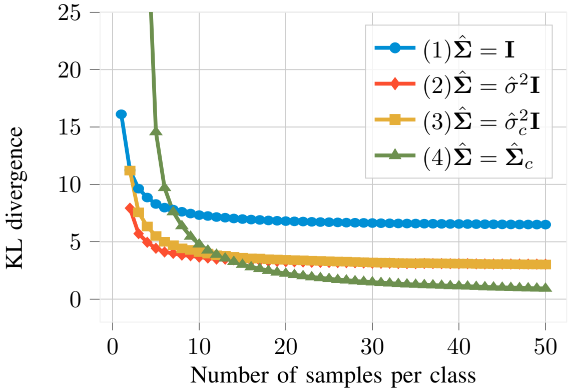
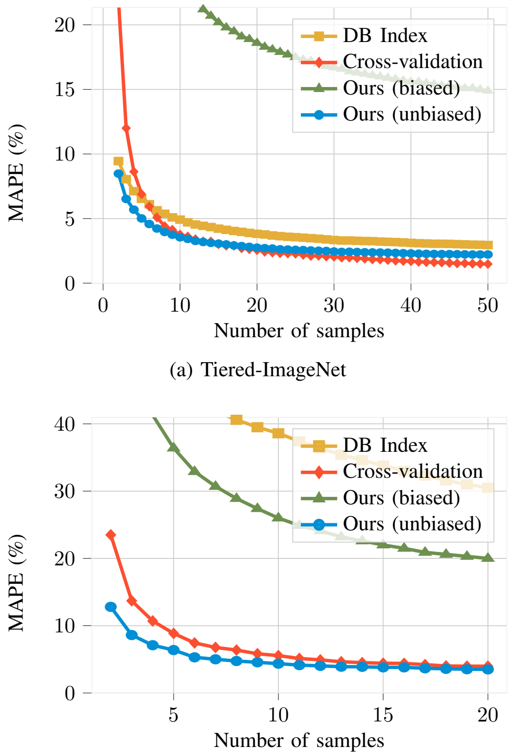


Fig. 2: Average KL divergence between a Gaussian distribution fitted with a limited number of samples and the closest Gaussian approximation using a larger number of samples.

C. Performance in predicting generalization

Next we compare our method to leave-one-out cross-validation and to the Davies-Bouldin (DB) Index, a clustering measure of inner-class and outer-class variances. For a fair comparison, we use the validation split of the original datasets to train a linear regression for the DB Index method and apply it to the few-shot tasks. For the cross-domain datasets, we use the validation split of ImageNet. For each few-shot problem, we predict the accuracy $(1 - \hat{P}_e)$ and compare it to the actual one. We use the Mean Absolute Percentage Error (MAPE) averaged over 10^3 problems as a metric. Our method (Ours unbiased) in Figure 3 outperforms cross-validation when few labeled samples are available, and outperforms the DB Index method on all datasets. Our method is also more efficient in predicting generalization and its predictions are more aligned with ground truth accuracies as shown in the scatter plot in Figure 4 for few-shot problems from ImageNet. Moreover, our method without the bias correction (Ours biased) does not yield good results, demonstrating the importance of the bias correction step.



(a) Tiered-ImageNet

(b) VGG-Flower

Fig. 3: Mean-Absolute-Percentage Error of different generalization predictors against the number of samples using 10^3 5-class few-shot classification problems. Figure (a) is an in-domain setting. Figure (b) is cross-domain. We compare our method (unbiased) to cross-validation or DB Index.

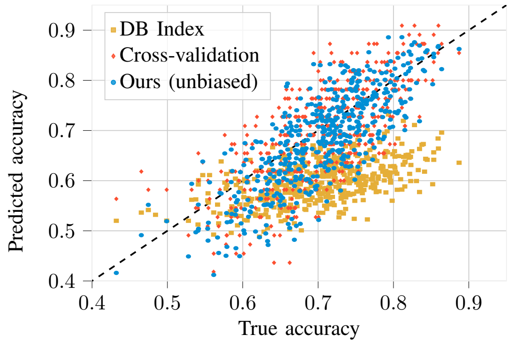


Fig. 4: Scatter plot of 5-class few-shot problems with 10 samples per class from ImageNet. Each point represents a different problem with a true ground-truth accuracy plotted against the predicted accuracy from the different methods.

V. Discussion and limitations

A first observation is that the DB Index performs poorly and tends to collapse on cross-domain datasets. We believe this is due to the large domain gap, causing a mismatch in the learned parameters of the linear regression. We can observe this behaviour in Figure 4 where the DB Index predictions are misaligned with the ground-truth accuracies. A linear regression was found to have the best results among various learned functions while our method and the cross-validation do not suffer from this behaviour.

Furthermore, when predicting generalization with limited samples, existing techniques, except for cross-validation, rely on clustering measures which depend on computing the distances between class centers and could benefit from the bias correction step when working with few samples. Other methods based on the analysis of the function space defined by the network [32] or its gradients [33] require training the network on the task which needs a large number of samples. For binary classification, our method is similar to the DB Index with a Gaussian kernel, but for multi-class classification, our method has the advantage of estimating class-conditional densities for computing the overlap between the classes.

Predicting error depends on the accuracy of the few-shot problem. Hard few-shot problems have a low SNR making the estimation of accuracy more difficult. On the other hand, good separation between the classes in the latent space leads to a better estimation of generalization. This explains the performance gap between in-domain and cross-domain datasets.

VI. Conclusion

This article proposes a model-based approach to estimate the generalization capability of few-shot classifiers. Our method outperforms the leave-one-out cross-validation and the Davis-Bouldin score-based estimator for different few-shot tasks with small number of labeled samples. This is especially important in the few-shot context. Our method strongly relies on unbiased estimates of the inter-class distances, which is a key contribution of this paper. Note that our method can be generalized to transfer-based few-shot learners with any distance-based classifier. Although we improve upon existing methods, we think that it opens up interesting new directions for further research.

References

[1] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, 2016.

[2] M. N. Rizve, S. Khan, F. S. Khan, and M. Shah, "Exploring complementary strengths of invariant and equivariant representations for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 836–10 846.

[3] Y. Hu, S. Pateux, and V. Gripon, "Squeezing backbone feature distributions to the max for efficient few-shot learning," *Algorithms*, vol. 15, no. 5, p. 147, 2022.

[4] Y. Bendou, Y. Hu, R. Lafargue, G. Lioi, B. Pasdeloup, S. Pateux, and V. Gripon, "Easy: Ensemble augmented-shot y-shaped learning: State-of-the-art few-shot classification with simple ingredients," 2022. [Online]. Available: http://arxiv.org/abs/2201.09699

[5] Y. Wang, W.-L. Chao, K. Q. Weinberger, and L. van der Maaten, "Simpleshot: Revisiting nearest-neighbor classification for few-shot learning," *arXiv preprint arXiv:1911.04623*, 2019.

[6] Q. F. Gronau and E.-J. Wagenmakers, "Limitations of bayesian leave-one-out cross-validation for model selection," *Computational brain & behavior*, vol. 2, no. 1, pp. 1–11, 2019.

[7] Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, and S. Bengio, "Fantastic generalization measures and where to find them," *arXiv preprint arXiv:1912.02178*, 2019.

[8] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *International Conference on Machine Learning*, pp. 1126–1135, 2017.

[9] J. Liu, F. Chao, and C.-M. Lin, "Task augmentation by rotating for meta-learning," *arXiv preprint arXiv:2003.00804*, 2020.

[10] G. S. Dhillon, P. Chaudhari, A. Ravichandran, and S. Soatto, "A baseline for few-shot image classification," *arXiv preprint arXiv:1909.02729*, 2019.

[11] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[12] S. Ghaffari, E. Saleh, D. Forsyth, and Y.-X. Wang, "On the importance of firth bias reduction in few-shot classification," in *International Conference on Learning Representations*, 2022.

[13] C. Zhang, Y. Cai, G. Lin, and C. Shen, "Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12 203–12 213, 2020.

[14] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[15] Y. Hu, S. Pateux, and V. Gripon, "Adaptive dimension reduction and variational inference for transductive few-shot classification," *arXiv preprint arXiv:2209.08527*, 2022.

[16] W. Gander, "Algorithms for the qr decomposition," *Res. Rep*, vol. 80, no. 02, pp. 1251–1268, 1980.

[17] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization." *arXiv preprint arXiv:1611.03530*, 2016.

[18] Y. Jiang, P. Natekar, M. Sharma, S. K. Aithal, D. Kashyap, N. Subramanyam, C. Lassance, D. M. Roy, G. K. Dziugaite, S. Gunasekar *et al.*, "Methods and analysis of the first competition in predicting generalization of deep learning," in *NeurIPS 2020 Competition and Demonstration Track*. PMLR, 2021, pp. 170–190.

[19] N. Ding, X. Chen, T. Levinboim, S. Goodman, and R. Soricut, "Bridging the gap between practice and pac-bayes theory in few-shot meta-learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 506–29 516, 2021.

[20] P. Natekar and M. Sharma, "Representation based complexity measures for predicting generalization in deep learning," *arXiv preprint arXiv:2012.02775*, 2020.

[21] A. Farid and A. Majumdar, "Generalization bounds for meta-learning via pac-bayes and uniform stability," *Advances in Neural Information Processing Systems*, vol. 34, pp. 2173–2186, 2021.

[22] Q. Chen, C. Shui, and M. Marchand, "Generalization bounds for meta-learning: An information-theoretic analysis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 25 878–25 890, 2021.

[23] M. Bontonou, L. Béthune, and V. Gripon, "Predicting the accuracy of a few-shot classifier," *arXiv preprint arXiv:2007.04238*, 2020.

[24] T. Chobola, D. Vašata, and P. Kordík, "Transfer learning based few-shot classification using optimal transport mapping from preprocessed latent space of backbone neural network," in *AAAI Workshop on Meta-Learning and MetaDL Challenge*. PMLR, 2021, pp. 29–37.

[25] Y. Hu, V. Gripon, and S. Pateux, "Leveraging the feature distribution in transfer-based few-shot learning," in *International Conference on Artificial Neural Networks*. Springer, 2021, pp. 487–499.

[26] J. Xu and H. Le, "Generating representative samples for few-shot classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9003–9013.

[27] T. Cao, M. Law, and S. Fidler, "A theoretical analysis of the number of shots in few-shot learning," *arXiv preprint arXiv:1909.11722*, 2019.

[28] Y. Bendou, V. Gripon, B. Pasdeloup, L. Mauch, S. Uhlich, F. Cardinaux, G. B. Hacene, and J. A. Garcia, "A statistical model for predicting generalization in few-shot classification," 2022. [Online]. Available: https://arxiv.org/abs/2212.06461

[29] J. B. Kruskal, "Nonmetric multidimensional scaling: a numerical method," *Psychometrika*, vol. 29, no. 2, pp. 115–129, 1964.

[30] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel, "Meta-learning for semi-supervised few-shot classification," *arXiv preprint arXiv:1803.00676*, 2018.

[31] E. Triantafillou, T. Zhu, V. Dumoulin, P. Lamblin, U. Evci, K. Xu, R. Goroshin, C. Gelada, K. Swersky, P.-A. Manzagol *et al.*, "Meta-dataset: A dataset of datasets for learning to learn from few examples," *arXiv preprint arXiv:1903.03096*, 2019.

[32] G. Ortiz-Jiménez, S.-M. Moosavi-Dezfooli, and P. Frossard, "What can linearized neural networks actually say about generalization?" *Advances in Neural Information Processing Systems*, vol. 34, pp. 8998–9010, 2021.

[33] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," *arXiv preprint arXiv:1609.04836*, 2016.