Alleviating the Incompatibility between Cross Entropy Loss and Episode Training for Few-shot Skin Disease Classification

Wei Zhu[1], Haofu Liao[1], Wenbin Li[2], Weijian Li[1], and Jiebo Luo[1]

[1] University of Rochester
[2] Nanjing University

Abstract. Skin disease classification from images is crucial to dermatological diagnosis. However, identifying skin lesions involves a variety of aspects in terms of size, color, shape, and texture. To make matters worse, many categories only contain very few samples, posing great challenges to conventional machine learning algorithms and even human experts. Inspired by the recent success of Few-Shot Learning (FSL) in natural image classification, we propose to apply FSL to skin disease identification to address the extreme scarcity of training sample problem. However, directly applying FSL to this task does not work well in practice, and we find that the problem can be largely attributed to the incompatibility between Cross Entropy (CE) and episode training, which are both commonly used in FSL. Based on a detailed analysis, we propose the Query-Relative (QR) loss, which proves superior to CE under episode training and is closely related to recently proposed mutual information estimation. Moreover, we further strengthen the proposed QR loss with a novel adaptive hard margin strategy. Comprehensive experiments validate the effectiveness of the proposed FSL scheme and the possibility to diagnosis rare skin disease with a few labeled samples.

Keywords: Skin Disease Classification · Few-Shot Learning · Query-Relative Loss.

1 Introduction

As a key step in the dermatological diagnosis, skin disease classification is quite challenging due to the extremely scarce annotations for a large number of categories. Such complexity in skin disease taxonomy requires a great deal of expertise. In addition, the diagnosis is often subjective and inaccurate even by human experts, which necessitates the research for computer-aided diagnosis [10,14]. Motivated by the unprecedented success of deep neural networks (DNNs), many researchers resort to deep learning technologies to handle this task [2,7,8]. For example, Esteva *et al.* adopt GoogleNet Inception V3 [16] to train a large-scale skin disease classification network [2]. Liao *et al.* jointly train skin lesion and body location classifiers using a multi-task network [8]. However, since DNN-based methods usually require a significant number of training samples for each

category, categories with only a few number of samples are often discarded [7]. This reduces the applicability of DNN-based methods, especially for infrequent skin disease diagnosis.

Shi *et al.* propose to adopt active learning to reduce the annotation cost [12], but still need up to 50% of labeled samples to train their model. Alternatively, Few-Shot Learning (FSL) is usually leveraged to address such tasks with only a few training samples [13,15,5,6]. By assuming the availability of a large-scale auxiliary training set, one can learn generalized patterns and knowledge which facilitate the learning for unseen tasks. Formally, for each few-shot task, we are provided with a support set S, a query set Q, and an auxiliary set A, where the support set S contains C different categories and each category has K training samples, *i.e.*, C-way K-shot, and Q contains unlabeled query data. Instead of conventional minibatch training, FSL is always trained with the episode training mechanism [13]. Basically, at each training iteration, we generate an episode by drawing samples from C different categories of the auxiliary set A, with K samples in each category as support samples S_{train} and others as query samples Q_{train}. As a crucial step, we need to randomly shuffle the labels for all categories from episode to episode. Episode training mechanism benefits FSL in at least two aspects. First, it enables FSL to be trained under similar scenarios as testing tasks. Second, the labels are randomly shuffled during episode training, which enables the model to learn category-agnostic representation for a better generalization ability.

Generally, FSL employs the Cross Entropy (CE) loss as an objective for classification. Although CE is useful for conventional classification, we find that it is somewhat incompatible with the episode mechanism. Well-designed FSL methods trained with CE even perform significantly worse than the baseline methods [1]. As we will see, CE classifies the query samples individually and relies highly on well-trained category-wise representation, a.k.a. proxies in proxy-based metric learning methods which share the similar formulation as CE [9,11]. The proxy is an category-wise *aggregation* of labeled support samples, e.g., the center used in Prototypical Network (PN). However, accurate proxies could only be obtained by a large-scale unified labeled dataset under the conventional minibatch training mechanism. This is hardily fulfilled under the episode training mechanism since we are only provided with a few training samples with randomly shuffled labels in each iteration.

To alleviate the problem, we propose a Query-Relative (QR) loss, which works much better with the episode training mechanism than CE for FSL.

We highlight our main contributions as follows:

- Upon an insightful analysis of the CE loss and episode mechanism, we propose a Query-Relative (QR) loss to better utilize the cross sample information and avoid possible sub-optimal aggregation of negative support samples, which significantly boosts the FSL performance;
- We develop an adaptive hard margin method for the QR loss to further penalize the categories with more error similarity connections;

– We evaluate our methods against a benchmark FSL suite [1], and the experiments strongly validate our analysis and the proposed methods.
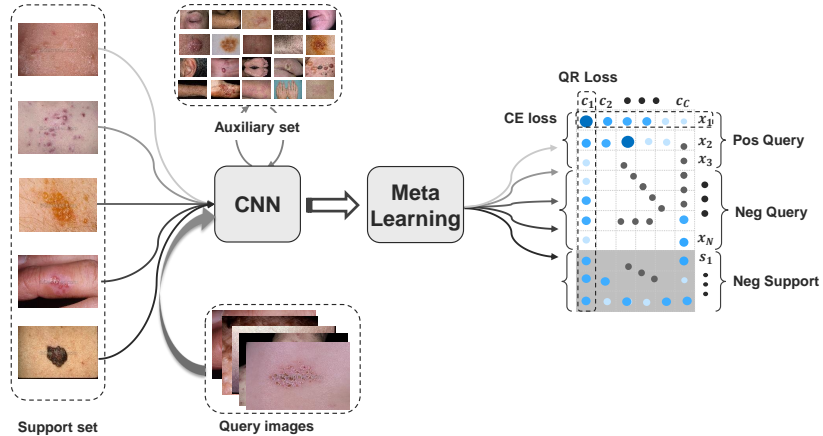

Fig. 1. Block diagram of few-shot learning-based skin disease classification and the difference between QR and CE loss. CE considers queries individually, while QR takes the relation across samples into consideration. Moreover, CE aggregates the support samples into proxies with possible information loss, while QR allows the model to fully exploit the information of negative support samples guided by the training objective.

2 Methodology

2.1 Discussions on FSL

Cross Entropy (CE) loss is often jointly used with episode mechanism to solve the FSL tasks. It can be generally formulated as

$$\mathcal{L}_{\text{CE}} = -\sum_i \log \frac{e^{s(c_{y_i}, x_i)}}{\sum_j e^{s(c_j, x_i)}}. \tag{1}$$

Here, $\{x_1, x_2, \ldots, x_N\} \in \mathbb{R}^{d \times N}$ are the query embeddings, and $\{c_1, c_2, \ldots, c_C\} \in \mathbb{R}^{d \times C}$ are the representations for the support categories, where N, C, and d denote the number of queries, support categories, and feature dimensions, respectively. $s(c_j, x_i)$ denotes the similarity between the support category proxy c_j and query sample x_i. Different FSL methods have different formulations of similarity measurement $s(\cdot, \cdot)$ and category proxy c aggregated by the support samples. For example, PN (Prototypical Network) uses the centers of the support samples from the j-th category as c_j and the Euclidean distance as $s(\cdot, \cdot)$,

Matching Net employs an FCE (Fully Context Embedding) layer to encode the support samples and chooses cosine similarity for $s(\cdot,\cdot)$, and MAML implements $s(\cdot,\cdot)$ as a Fully Connected (FC) layer where the j-th weight vector of the FC layer corresponds to c_j. To unify these methods, we normalize $x_i = \frac{x_i}{\|x_i\|_2}$ and $c_j = \frac{c_i}{\|c_i\|_2}$ which leads to better performance shown in the recent literature [20]. Eq. (1) can then be rewritten as $\min -\sum_i \log \frac{e^{c_{y_i}^T x_i}}{\sum_j e^{c_j^T x_i}}$.

According to Eq. (1), CE individually classifies the query samples and completely relies on the category-wise representation c_j to train the model. For the conventional classification task trained with minibatch SGD, such a mechanism could prompt c_j to learn high-level representative features of each category by exposing them to a large and balanced dataset. Unfortunately, this is not the case for FSL due to the episode training mechanism. Although episode training is important for FSL since it empowers FSL with the ability to learn generalized class agnostic representation and provides similar training scenarios as testing scenarios, it is also a double-edged sword: it makes c_j inevitably biased and inaccurate. The reasons are two aspects: first, c_j is learned from a few samples in each episode, e.g., 1 and 5 for 1-shot and 5-shot respectively, and it is difficult to learn to aggregate support samples to obtain c_j without losing useful information with so few training samples; second, the labels are randomly shuffled for each episode which limits c_j to be consistently trained across episodes. Therefore, c_j cannot be fully relied on under the episode mechanism and training the model with CE loss will eventually degrade the performance for FSL. The suboptimal performance has been observed and experimentally validated by several recent benchmark papers for natural images, where well-designed baselines could achieve similar and even better performance than CE-trained FSL counterparts [1,17]. Similar results are also disappointingly observed in the skin disease tasks according to our experiments in Section 3.

2.2   Query-Relative Loss

We alleviate the above problem from two aspects. First, instead of classifying the query separately, we unify all samples into a joint objective to allow them to mutually share information cross samples. Second, we avoid using negative category proxies which are aggregated by the negative support samples with a manually designed strategy (e.g., the center of support samples in PN), and the information of the support samples can then be largely preserved and extracted with the guidance of the training objective. To this end, we propose the Query-Relative (QR) loss as follows

$$\mathcal{L}_{QR} = \sum_j \log(1 + \frac{1}{2|P_j|} \sum_{x_i^+ \in P_j} e^{-s(c_j, x_i^+)} + \frac{1}{2|N_j|} \sum_{x_i^- \in N_j} e^{s(c_j, x_i^-)}), \quad (2)$$

where P_j denotes the set of positive query samples that belong to the j-th category and N_j denotes the set of *negative query and support* samples that are not from the j-th category. $|\cdot|$ denotes the number of samples in the set.

We then present an analysis on how our objective improves CE from the two aforementioned aspects. First of all, Eq. (2) implicitly utilizes the cross sample information to re-weight each sample. Specifically, taking the derivation w.r.t. $s(c_j, x_p^+)$ and $s(c_j, x_n^-)$, we have

$$\left| \frac{\partial \mathcal{L}}{\partial s(c_j, x_p^+)} \right| = \frac{\frac{1}{2|P_j|} e^{-s(c_j, x_p^+)}}{1 + \frac{1}{2|P_j|} \sum_{x_i^+ \in P_j} e^{-s(c_j, x_i^+)} + \frac{1}{2|N_j|} \sum_{x_i^- \in N_j} e^{s(c_j, x_i^-)}} \tag{3}$$

$$\left| \frac{\partial \mathcal{L}}{\partial s(c_j, x_n^-)} \right| = \frac{\frac{1}{2|N_j|} e^{s(c_j, x_n^-)}}{1 + \frac{1}{2|P_j|} \sum_{x_i^+ \in P_j} e^{-s(c_j, x_i^+)} + \frac{1}{2|N_j|} \sum_{x_i^- \in N_j} e^{s(c_j, x_i^-)}} \tag{4}$$

Here, we only focus on the absolute value of the gradient. According to Eq. (3), $s(c_j, x_p^+)$ will induce a large gradient and will be punished if (i) $s(c_j, x_p^+)$ is small; (ii) $s(c_j, x_p^+)$ is smaller than $s(c_j, x_i^+)$ where $x_i^+ \in P_j$, $p \neq i$; or (iii) $s(c_j, x_i^-)$ is small so that we could focus on intra-class relation. Moreover, a large $s(c_j, x_i^-)$ will provide $s(c_j, x_i^+)$ with tolerance to some extent, which allows our model to focus on reducing the large similarity of $s(c_j, x_i^-)$. Similar analysis can be performed with $s(c_j, x_n^-)$ based on Eq. (4), and we omit the detail here. Therefore, in contrast to CE which deals with each sample separately, QR allows the query and support samples to share information across each other and category-wisely re-weights their importance.

Second, note that the negative set N_j of each category contains not only the negative query samples but also the support samples from other categories. This avoids the information loss caused by the possibly sub-optimal support sample aggregation and allows the model to learn to utilize the negative support samples directly by the objective.

It turns out that the QR loss is closely related to Deep Mutual Information (MI) maximization recently proposed by [4]. Without loss of generality, following [4], the JSD-based (Jensen-Shannon Divergence) MI estimator between c_j and x can be formulated as

$$\begin{aligned}
\mathcal{L}_{\text{JSD MI}} &= \max \frac{1}{|P_j|} \sum_{x_i^+ \in P_j} -\log(1 + e^{-s(c_j, x_i^+)}) - \frac{1}{|N_j|} \sum_{x_i^- \in N_j} \log(1 + e^{s(c_j, x_i^-)}) \\
&\geq \max -\log(1 + \frac{1}{2|P_j|} \sum_{x_i^+ \in P_j} e^{-s(c_j, x_i^+)} + \frac{1}{2|N_j|} \sum_{x_i^- \in N_j} e^{s(c_j, x_i^-)}) \\
&= \mathcal{L}_{\text{QR}}
\end{aligned} \tag{5}$$

Here we use the fact that $-\log(1 + x)$ is convex and the Jensen's inequality. We can thus derive that the QR loss is actually a lower bound of the JSD MI. The reason why we do not directly optimize $\mathcal{L}_{\text{JSD MI}}$ is that the re-weighting mechanism of $\mathcal{L}_{\text{JSD MI}}$ does not take both P_j and N_j into consideration for each $s(c_j, x_i)$. We experimentally verify the superiority of our formulation in Sec. 3.

2.3   Adaptive Hard Margin

The adaptive hard margin is built upon the fact that the cosine similarity between uniformly distributed normalized samples approaches $\mathcal{N}(0, \frac{1}{2d})$ [19] and is thus likely to be zero. Therefore, $s(c_j, x_i^+)$ should be at least larger than E_j^- and $s(c_j, x_i^-)$ should be at least smaller than E_j^+, where E_j^+ and E_j^- denote the average of $s(c_j, x_i^+)$ with $s(c_j, x_i^+) < 0$ and average of $s(c_j, x_i^-)$ with $s(c_j, x_i^+) > 0$, respectively. Based on this observation, we propose a QR loss with online Adaptive Hard Margin which can be written as

$$\mathcal{L}_{\text{QR+margin}} = \sum_j \log(1 + \frac{1}{2|P_j|} \sum_{x_i^+ \in P_j} e^{-s(c_j, x_i^+) + E_j^-} + \frac{1}{2|N_j|} \sum_{x_i^- \in N_j} e^{s(c_j, x_i^-) - E_j^+}).$$

(6)

Basically, Eq. (6) imposes extra punishment on categories with more positive samples whose similarities are smaller than random or negative samples, and negative samples whose similarities are larger than random or positive samples.

3   Experiments

3.1   Datasets

We collect the dermatology images from the Dermnet atlas website [3]. To perform few-shot learning, we discard categories with less than 10 samples, which are required for the 5-way 5-shot setting. Finally, we obtain $20,230$ images in total belonging to 334 different categories. The largest category "seborrheic keratoses ruff" contains 516 images and the smallest categories contain 10 samples. Detailed statistics of the data can be found in the supplemental material. The data is manually split into 186 categories for training, 74 for validation, and 74 for testing, respectively. Moreover, to better simulate the scenario of few-shot learning, we deliberately choose categories with more than 120 samples (38 categories in total) as the training data.

3.2   Benchmark Methods and Experimental Settings

We benchmark the dataset on an FSL suite proposed by [1]. The suite contains 2 strong baseline methods (denoted as baseline and baseline++ following [1]) and 4 FSL methods including Relation Net[15], Model-Agnostic Meta-Learning (MAML) [3], Matching Net (MN)[18], and Prototypical Net (PN)[13]. The baseline methods are carefully designed and outperform FSL methods in some cases. We refer readers to [1] for details. The four FSL methods are regarded as the state-of-the-art FSL baselines in recent benchmark literature [17,1], and we train them with CE as our baselines except for the Relation Net, which is trained with Mean Square Error (MSE) Loss following the original paper. We apply the proposed QR loss to MN and PN since these two methods have proven to have

---

[3] www.dermnet.com

superior and stable performance in natural image classification [1]. The model trained with JSD-based MI maximization Eq. (5) is denoted as JSD MI, and models trained with the proposed QR loss Eq. (2) and QR loss with adaptive hard margin Eq. (6) are denoted as QR and QR+M, respectively.

For the network structure, we follow the commonly adopted FSL settings [15,1]. The feature embedding network used in this paper is a convolutional neural network which has four convolutional blocks with each block containing a sequence of a convolutional layer with 64 filters of size 3×3, a batch normalization layer, a 2×2 max-pooling layer and a Leaky ReLU layer. For the experimental settings, the episodic training mechanism is applied to all FSL models, and $60,000$ episodes are constructed in total during training for all methods. For validation and testing, 600 episodes are randomly constructed from the validation and test set, respectively. We conduct 5-way 1-shot and 5-way 5-shot classification tasks on the collected Dermnet dataset, and 5 query samples are provided for each category within each episode for either training and testing. For optimization, we adopt the Adam algorithm with a learning rate of 0.001. Experiments are run five times and we report the performance on test set corresponding to the best validation results. The average Accuracy, Precision, and F1 score with 95% confidence interval are reported.

Methods	5-way 1-shot			5-way 5-shot		
	ACC%	Precision%	F1%	ACC%	Precision%	F1%
Baseline	$39.89_{\pm 0.89}$	$40.57_{\pm 1.12}$	$37.16_{\pm 0.91}$	$59.87_{\pm 0.94}$	$62.37_{\pm 1.10}$	$58.19_{\pm 1.01}$
Baseline++	$42.47_{\pm 0.94}$	$43.70_{\pm 1.11}$	$40.34_{\pm 0.93}$	$63.37_{\pm 0.95}$	$65.80_{\pm 1.07}$	$61.75_{\pm 1.01}$
MAML	$45.95_{\pm 1.06}$	$44.82_{\pm 1.29}$	$42.18_{\pm 1.08}$	$66.93_{\pm 0.96}$	$69.24_{\pm 1.11}$	$64.92_{\pm 1.05}$
Relation Net	$45.50_{\pm 1.07}$	$46.36_{\pm 1.18}$	$44.00_{\pm 1.07}$	$62.53_{\pm 1.02}$	$64.90_{\pm 1.11}$	$62.26_{\pm 1.05}$
MN	$44.59_{\pm 0.97}$	$44.96_{\pm 1.19}$	$41.52_{\pm 1.00}$	$61.21_{\pm 0.90}$	$63.15_{\pm 1.13}$	$58.29_{\pm 0.99}$
MN+JSD MI	$43.28_{\pm 1.04}$	$43.26_{\pm 1.25}$	$40.00_{\pm 1.05}$	$58.99_{\pm 0.94}$	$60.23_{\pm 1.20}$	$55.78_{\pm 1.02}$
MN+QR	$48.01_{\pm 1.09}$	$48.87_{\pm 1.13}$	$44.30_{\pm 1.13}$	$67.09_{\pm 0.97}$	$69.18_{\pm 1.16}$	$64.53_{\pm 1.08}$
MN+QR+M	$49.29_{\pm 1.31}$	$49.95_{\pm 1.05}$	$45.64_{\pm 1.09}$	$66.83_{\pm 0.95}$	$69.10_{\pm 1.16}$	$64.25_{\pm 1.05}$
MN+QR*	$48.66_{\pm 1.07}$	$48.86_{\pm 1.30}$	$44.98_{\pm 1.11}$	-	-	-
MN+QR+M*	$49.76_{\pm 1.07}$	$49.52_{\pm 1.32}$	$46.01_{\pm 1.13}$	-	-	-
PN	$46.77_{\pm 1.04}$	$46.82_{\pm 1.06}$	$43.58_{\pm 1.07}$	$62.06_{\pm 1.02}$	$63.39_{\pm 1.22}$	$59.50_{\pm 1.10}$
PN+JSD MI	$47.55_{\pm 1.00}$	$47.90_{\pm 1.25}$	$44.33_{\pm 1.05}$	$61.15_{\pm 0.94}$	$61.74_{\pm 1.16}$	$58.34_{\pm 1.02}$
PN+QR	$49.85_{\pm 1.11}$	$49.53_{\pm 1.32}$	$46.34_{\pm 1.14}$	$70.38_{\pm 0.96}$	$72.13_{\pm 1.08}$	$68.50_{\pm 1.05}$
PN+QR+M	$\mathbf{52.41}_{\pm 1.09}$	$\mathbf{53.21}_{\pm 1.27}$	$\mathbf{49.52}_{\pm 1.12}$	$\mathbf{71.99}_{\pm 0.87}$	$\mathbf{74.23}_{\pm 0.98}$	$\mathbf{70.30}_{\pm 0.94}$
PN+QR*	$50.62_{\pm 1.10}$	$50.83_{\pm 1.32}$	$47.16_{\pm 1.13}$	-	-	-
PN+QR+M*	$\underline{53.30}_{\pm 1.11}$	$\underline{53.69}_{\pm 1.35}$	$\underline{50.45}_{\pm 1.17}$	-	-	-

Table 1. Experimental results on the Derment skin disease classification dataset. * denotes that the model is trained with 9 query samples per episode. - denotes that the setting is not applicable. M denotes our methods with an adaptive hard margin.

3.3   Result Analysis

The experimental results are reported in Table 1, and we draw several interesting points from the results as follows. First of all, the baseline methods with minibatch training and CE loss perform reasonably well in practice. The FSL methods trained with CE loss have comparable or slightly better performance. In contrast, FSL methods trained with the proposed QR loss significantly outperform the baseline methods and the FSL methods with CE. For Matching Net, our QR loss achieves 3.42 % and 5.88 % improvements compared with the CE loss in terms of accuracy for 5-way 1-shot and 5-way 5-shot tasks. Significant improvements are also observed for PN, and our QR loss outperforms CE 3.08 % and 8.32 % for 5-way 1-shot and 5-way 5-shot, respectively. The improvements are obtained by fully utilizing the cross-sample information and avoiding the information loss caused by manually designed support sample aggregation during training. Second, we compare the QR loss with JSD MI. Although the formulations are similar, QR is significantly better than JSD MI. The reason should be attributed to the fact that JSD MI does not mutually utilize the information in P_j and N_j. Third, the adaptive hard margin consistently boosts the performance of the models trained by QR. For example, the adaptive hard margin improves PN trained with QR 2.56 % and 1.61 % for 5-way 1-shot and 5-way 5-shot, respectively. Finally, our method could be further boosted by increasing the number of queries for both training and testing. Overall, our proposed FSL methods classify skin disease with only a few available training samples and makes it possible to diagnose rare diseases using modern neural networks.

3.4   Influence of the number of shots and ways

For simplicity, we only conduct experiments on PN with various ways and shots and report the accuracy. As shown in Tables 2 and 3, QR has clear advantages over CE when more samples are available per episode, suggesting that QR can better utilize the cross sample information.

# shots	1	2	3	4	5
CE	46.77	54.04	57.15	59.65	62.06
QR	49.85	62.37	66.87	68.95	70.38

Table 2. 5-way different-shot. (ACC%)

# ways	2	3	5	10	20
CE	69.80	59.42	46.77	35.78	24.30
QR	72.02	61.57	49.85	40.37	31.31

Table 3. Different-way 1-shot. (ACC%)

4   Conclusions

We propose to apply Few-Shot Learning to address the classification for rare skin diseases. We find that existing FSL methods do not perform significantly better than the baseline methods. Through careful analysis, we believe the problem

should be largely attributed to the incompatibility between the episode training mechanism and cross entropy loss. Therefore, we propose a novel QR loss for FSL to make fully use of the information across samples and also allow the model to learn to extract information of the support samples guided by the training objective. With the proposed QR loss, the state-of-the-art FSL methods perform consistently better than methods training with the conventional CE loss. Our work demonstrates the promise of diagnosing rare skin diseases with one or a few labeled samples. In the future, we will investigate extensions to other medical classification problems or even natural image classification.

References

1. Chen, W.Y., Liu, Y.C., Kira, Z., Wang, Y.C., Huang, J.B.: A closer look at few-shot classification. In: International Conference on Learning Representations (2019)
2. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115 (2017)
3. Finn, C., Abbeel, P., Levine, S.: 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)
4. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670 (2018)
5. Li, W., Wang, L., Xu, J., Huo, J., Gao, Y., Luo, J.: Revisiting local descriptor based image-to-class measure for few-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7260–7268 (2019)
6. Li, W., Xu, J., Huo, J., Wang, L., Gao, Y., Luo, J.: Distribution consistency based covariance metric networks for few-shot learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8642–8649 (2019)
7. Liao, H., Li, Y., Luo, J.: Skin disease classification versus skin lesion characterization: Achieving robust diagnosis using multi-label deep neural networks. In: 2016 23rd International Conference on Pattern Recognition (ICPR). pp. 355–360. IEEE (2016)
8. Liao, H., Luo, J.: A deep multi-task learning approach to skin lesion classification. arXiv preprint arXiv:1812.03527 (2018)
9. Movshovitz-Attias, Y., Toshev, A., Leung, T.K., Ioffe, S., Singh, S.: No fuss distance metric learning using proxies. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 360–368 (2017)
10. Okuboyejo, D.A., Olugbara, O.O., Odunaike, S.A.: Automating skin disease diagnosis using image classification. In: proceedings of the world congress on engineering and computer science. vol. 2, pp. 850–854 (2013)
11. Qian, Q., Shang, L., Sun, B., Hu, J., Li, H., Jin, R.: Softtriple loss: Deep metric learning without triplet sampling. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6450–6458 (2019)
12. Shi, X., Dou, Q., Xue, C., Qin, J., Chen, H., Heng, P.: An active learning approach for reducing annotation cost in skin lesion analysis
13. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems. pp. 4077–4087 (2017)

14. Sumithra, R., Suhil, M., Guru, D.: Segmentation and classification of skin lesions for disease diagnosis. Procedia Computer Science **45**, 76–85 (2015)
15. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
16. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
17. Triantafillou, E., Zhu, T., Dumoulin, V., Lamblin, P., Evci, U., Xu, K., Goroshin, R., Gelada, C., Swersky, K., Manzagol, P.A., et al.: Meta-dataset: A dataset of datasets for learning to learn from few examples. arXiv preprint arXiv:1903.03096 (2019)
18. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: Advances in neural information processing systems. pp. 3630–3638 (2016)
19. Wu, C.Y., Manmatha, R., Smola, A.J., Krahenbuhl, P.: Sampling matters in deep embedding learning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2840–2848 (2017)
20. Ye, H.J., Chen, H.Y., Zhan, D.C., Chao, W.L.: Identifying and compensating for feature deviation in imbalanced deep learning. arXiv preprint arXiv:2001.01385 (2020)