Cross-Domain Few-Shot Sparse-Quantization Aware Learning for Lymphoblast Detection in Blood Smear Images

Dina Aboutahoun[1](✉), Rami Zewail[1], Keiji Kimura[2], and Mostafa I. Soliman[1]

[1] Department of Computer Science and Engineering, Egypt-Japan University for Science and Technology, Alexandria, Egypt
{dina.aboutahoun,rami.zewail,mostafa.soliman}@ejust.edu.eg
[2] Department of Computer Science and Engineering, Waseda University, Tokyo, Japan
kimura@apal.cs.waseda.ac.jp

Abstract. Deep learning for medical image classification has enjoyed increased attention. However, a bottleneck that prevents it from widespread adoption is its dependency on very large, annotated datasets, a condition that cannot always be satisfied. Few-shot learning in the medical domain is still in its infancy but has the potential to overcome these challenges. Compression is a way for models to be deployed on resource-constrained machines. In an attempt to tackle some of the challenges imposed by limited data and high computational resources, we present a few-shot sparse-quantization aware meta-training framework (FS-SQAM). The proposed framework aims to exploit the role of sparsity and quantization for improved adaptability in a low-resource cross-domain setting for the classification of acute lymphocytic leukemia (ALL) in blood cell images. Combining these strategies enables us to approach two of the most common problems that encounter deep learning for medical images: the need for extremely large datasets and high computational resources. Extensive experiments have been conducted to evaluate the performance of the proposed framework on the ALL-IDB2 dataset in a cross-domain few-shot setting. Performance gains in terms of accuracy and compression have been demonstrated, thus serving to realize the suitability of meta-learning on resource-constrained devices. Future advancements in the domain of efficient deep learning computer-aided diagnosis systems will facilitate their adoption in clinical medicine.

Keywords: Few-Shot Learning · Medical Image Analysis · Compression

1 Introduction

The problem of restricted availability of labeled medical data remains a hindrance for conventional deep learning methods as they are dependent on a high volume of data. Conventionally, there are common challenges that plague deep learning for medical applications [1], one is the dataset size, with medical datasets being in the order of hundreds or thousands of samples [2]. Second, is the class imbalance problem of rarer

H. Lu et al. (Eds.): ACPR 2023, LNCS 14408, pp. 213–226, 2023.
https://doi.org/10.1007/978-3-031-47665-5_18

diseases, where some diseases manifest more rarely in a population which is reflected in classes with much fewer samples [3]. Overcoming the data problem in medical applications allows for better diagnostic accuracy and potentially improves pathologists' accuracy [4].

Few-shot learning (FSL) as a concept has been developed to overcome these hurdles by imitating how humans can learn from a few examples. Likewise, when applied to scarce medical datasets, few-shot learning seeks to learn from just a few samples. A more general term, meta-learning, is an approach conventionally described as learning to learn, that aims to generalize to new tasks that are different from the originally trained tasks. Learning to learn from a few examples is a powerful technique that is heavily investigated by the research community. Few-shot meta-learning methods are suited for applications where encountering novel classes with limited samples is a common occurrence as with the case in the field of cancer diagnostics.

Deep learning has been used extensively in various medical image applications such as the detection of abnormalities [5] and segmentation of areas of interest [6], encouraged by the advent of high-quality images and the increase in computational resources. As a result, numerous computer-aided diagnosis (CAD) systems are dependent on deep learning [7].

Encouraged by the efforts in deep learning and few-shot learning healthcare targeted applications, we investigate the role of sparsity and quantization in meta-training in a limited resource medical cross-domain setting. In this work, we exploit sparsity and quantization for the purpose of detecting the presence of blast cells in blood cell images through training on the task of detecting malignant tumors in breast histopathological images.

Our main contributions in this work are as follows:

- We present FS-SQAM, a joint sparse-quantization aware meta-training scheme for the purpose of lymphoblast detection in blood smear images in a cross-domain setting with limited data and computational resources.
- We conduct experiments to assess the influence of sparsity and quantization and the impact of regularization on the performance and efficiency of lymphoblast detection to assist in leukemia diagnosis in a low-resource setting. We observe that the results demonstrate that FS-SQAM allows for strong generalization on the ALL-IDB2 [8] dataset in addition to the gains on the memory footprint reduction front.
- We hope the presented work can encourage more research into approaches that are well-suited for use in challenging clinical environments where both the computational resources and data are limited.

2 Related Works

In this section we start by focusing on previous works that target lymphoblast detection, and then we mention examples in the literature that have utilized FSL for medical applications. We end this section by mentioning some of the influential research directions in efficient few-shot learning and relevant works that examine medical cross-domain FSL.

Acute lymphoblastic leukemia or ALL is a cancer of blood cells and the most common childhood cancer. Early diagnosis is crucial as the disease is characterized by rapid progression. One of the ways to diagnose ALL is through blood testing which requires manual examination by a pathologist. In order to expedite the diagnosis process, many works have proposed approaches for the detection of target cells in microscopic blood smear images that could be incorporated into CAD systems. The literature on ALL detection mainly employs deep learning with transfer learning [9–11] being a dominant approach as training from scratch prerequires the availability of a large dataset. Whilst most studies using deep learning achieve a performance of $>95\%$ diagnostic accuracy, efficient resource utilization that targets resource-constrained devices is not taken into account by the majority.

Few-shot learning aims to learn new classes from a few examples, making it a suitable approach for data limited settings, which is the case in most medical applications, making it a method with immense potential benefits for medical classification problems. Metric-learning is a class of FSL approaches where classification is dependent on a learning similarity metric that is capable of discerning similar instances. Among the approaches in this class are Prototypical networks [12], Matching networks [13], and Relation networks [14]. While still a nascent approach, few-shot learning has been gaining interest in the past few years with works applying it to various medical datasets such as histopathological [15], X-rays [16], and cell [17] image datasets among others. A specific subset in few-shot learning problems is cross-domain few-shot learning [18] where the classes used in training and testing are drawn from different domains.

In cross-domain few-shot learning, extreme shifts in the source and target domains are detrimental to performance and provide a challenging problem to overcome. This becomes even more important in medical images, where the data on some diseases are rarer than others, which enables learning for these understudied categories. It has been suggested in [19] that an updatable feature extractor leads to better generalization ability on hard few-shot tasks, as the model can modify its parameters to better suit the task through fine-tuning on the support set. This improves upon using a frozen feature extractor that relies solely on distance calculations to classify.

Compression for neural networks has been studied extensively with the goal of achieving models suitable for devices with limited resources, pruning, and quantization being popular methods. Their combination has been experimented with in general computer vision tasks [20]. For example in [21], the authors devise a quantized sparse training regime that leverages a combined pruning-quantization function that determines the optimal pruning and quantization parameters for each layer.

Further investigation is required to understand how compression techniques can be effectively integrated with few-shot meta-learning, especially in the context of medical applications where models need to generalize quickly and operate within computational constraints. Meta-learning can be memory intensive as reported in [22] due to the requirement of loading the support images to memory in order to obtain a task adapted model. To counteract this effect, the authors devised a training scheme by restricting backpropagation to only a random subset from the support set, making this approach friendly to computationally limited devices. Previous works [23, 24] utilized quantization-aware training (QAT) as a way to quantize the weights of the feature extractor and incorporated both quantization loss and classification loss calculated on the query set. Similarly, incorporating iterative pruning into meta-learning has been attempted in [25] for the purpose of limiting meta-overfitting through pruning the least significant weights and then fine-tuning via a meta algorithm. Another consequence of sparsifying the model is the reduced number of parameters. Thus, if the model allows updates to the feature extractor, this will lead to a task-adapted model with compressed weights.

Since the literature on applying FSL to ALL detection is scarce, we mention a few works that apply cross-domain FSL approaches on histopathological datasets, a related cancer detection task. Examining FSL on histopathological datasets reveals that accuracy ranges from 36.4 ± 0.51% to 70.0 ± 0.49% for out-domain CRC-TP (colorectal tissues) [26] → BreakHis [27] (breast tissues) classification with an average accuracy of 60% as demonstrated by [28]. The authors stipulate that this is a good starting performance for out-domain settings. Another approach proposed in [29] investigates contrastive learning in various cross-domain scenarios. The approach achieves an accuracy of 67.56% in the out-domain setting (NCT-CRC → PAIP) where the source dataset is NCT-CRC-HE-100K (colorectal tissues) [30] and the target dataset is PAIP (liver tissues) [31].

It is well established that deep learning, especially transfer learning, performs well on medical datasets including ALL datasets, however, little work has been done on the topic resource efficient few-shot learning for ALL images and how sparsity and quantization impact meta-learning in general.

3 Methodology

This section describes the methodology adopted in the proposed FS-SQAM framework. The goals of this approach are two-fold: 1) Produce a lightweight model. 2) Equipping the classifier with better generalization ability in a medical cross-domain setting with reasonable performance on a novel dataset with disjoint classes belonging to the medical domain. We test FS-SQAM on the following cross-domain setting BreakHis → ALL-IDB2, which belongs to the domain of histopathological images (breast tissue) and blood cell (blood smear) images respectively. Even though the domain gap between the training dataset and the test dataset adds to the problem's difficulty, it presents a realistic scenario where uncommon diseases that manifest less frequently in a population may be encountered by medical practitioners. The described setting can be considered an extreme cross-domain scenario (out-domain) given the great domain shift between these two domains.

The stages underlying the framework are as follows: 1) the pre-trained model undergoes sparse-aware meta-training and then quantization-aware meta-training on the BreakHis dataset. 2) Afterwards, the model is evaluated on test tasks sampled from the ALL-IDB2 dataset to assess the compressed classifier's performance as shown in Fig. 1. To achieve the goal of strong generalization, that is generalization on classes foreign to the training dataset, we used a network capable of fine-tuning on the sampled support sets in order to allow for adaptation for out-domain tasks.

In brief, our framework utilizes an updatable Prototypical based model [12] that undergoes sparse-aware meta-training, and quantization-aware meta-training for blast cell detection we extend the proposal in [19] to achieve an efficient flexible classifier for this task. Although this approach incurs additional training overhead compared to the simpler methods such as one-shot pruning and post-training quantization, it offers

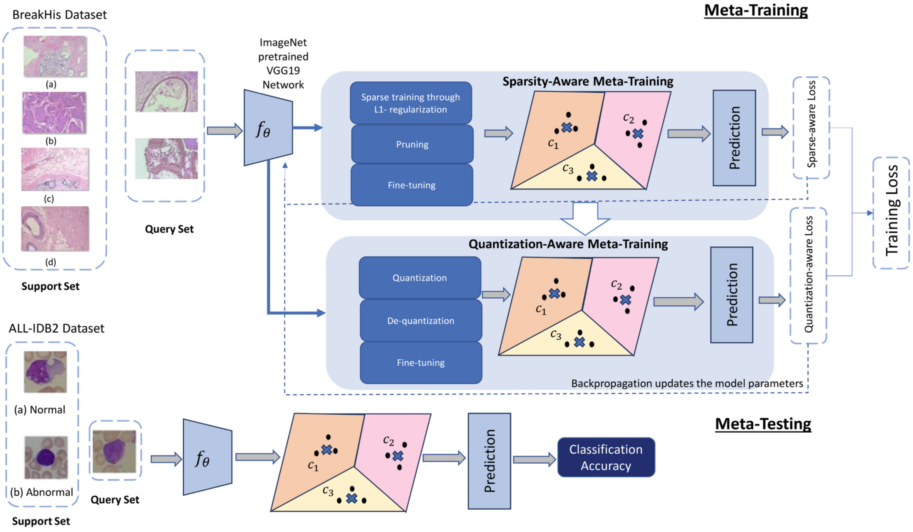

Fig. 1. Training and testing scheme of FS-SQAM, support, and query sets are sampled from BreakHis in the meta-training phase and ALL-IDB2 in the meta-testing phase. In the meta-training stage, the model undergoes 1) sparse training (through L_1 regularization) 2) pruning (part of sparse-aware meta-training 3) quantization-aware meta-training. The support and query samples are projected to the embedded space in order for the query samples be classified.

the significant benefit of preserving accuracy. This is particularly important in medical applications where diagnostic accuracy is crucial.

3.1 Prototypical Networks

Given a situation where there is a training dataset that has a set of classes C_{train} and a set of test classes C_{test}, where C_{train} and C_{test} are disjoint. The goal is to produce a model f_θ capable of adapting to and classifying examples belonging to C_{test} from just a few labeled examples [32]. Episodic learning has been proposed as a way to simulate conditions at test time by sampling examples from the larger training dataset. The episodes are constructed by sampling K-shots from each of the N classes. The K examples are form a support set $S = \{(x_1, y_1), ..., (y_K, y_K)\}_{i=1}^{N}$, alongside a query set Q of different examples from those sampled in the support set, drawn from the same N classes. This setup is referred to as K-shot N-way episode. The model adapts to the support set and the performance is evaluated on the query set with the goal of minimizing loss. Repeating this process on multiple episodes augments the model with the ability to generalize to the examples in the query set. In the literature, the episodic learning setup is termed meta-learning.

Prototypical networks [12] depend on creating a prototype for each class and classifying an instance based on nearness to a class's prototype. The loss function of the training episode of a prototypical network is the cross-entropy loss function but applied to the query set examples, where T is the number of examples in the query set, y is the true class label and x is the query sample to be classified to the closest prototype P_c of class c. The prototype of a class P_c is the mean of the support set embeddings calculated as $\frac{1}{N}\sum_{i=1}^{N} f_\theta(x_i)$ where f_θ is the model that extracts the embeddings of S and Q. A query point is classified by comparing its distance to every class's prototype in the embedding space. The loss can be written as:

$$L_{CE} = -\frac{1}{T}\sum\nolimits_{i} logp(y_i|x_i, \{p_c\}) \tag{1}$$

Cross-Domain Learning. Violates the assumption that the meta-training and meta-testing tasks T are drawn from the same distribution where it becomes $P_{train}(T) \neq P_{test}(T)$. Hence, in order to account for the novel domain information through transfer learning. The model f_θ is fine-tuned on the classes of interest, specifically, the last k layers are fine-tuned on the target classes. This is in accordance with the idea proposed in [19] that an updatable metric feature extractor is better suited for adapting to unseen classes at test time.

3.2 Sparse-Aware Meta-training

The proposed framework FS-SQAM employs two compression methods, pruning, and quantization where their combination has been utilized in various computer vision tasks to produce an ultra-resource efficient model. In this section, we detail the specific pruning and quantization methods that are incorporated into the framework.

The first compression technique introduced is iterative pruning to enforce sparsity in the network in the sparse-aware meta-training phase through unstructured magnitude-based pruning. Pruning is preceded by sparse training through introducing regularization. Pruning depends on the hypothesis that within overparameterized networks there exists a sparser subnetwork, referred to as a winning ticket [33], that can match the accuracy of the original with the added benefits of fewer parameters and reduced size. We incorporate the L_1 regularization into the training to drive the weights w to zero through the penalization of the absolute sum of weights as indicated by the following loss function of a model f_θ with parameters θ:

$$L(y,\hat{y}) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f_\theta(x_i)) = L_{CE} + \lambda \sum\nolimits_{i=1}^{N} |w_i| \quad (2)$$

where y refers to the ground-truth and $\hat{y}$ is the predicted label. The regularization penalty term consists of the λ hyperparameter and the L$_1$-norm. We conduct experiments by applying FS-SQAM on a pre-trained VGG19 network with the following values for λ, where $\lambda = \{0.001, 0.005, 0.01\}$. The network is pruned iteratively with the goal of progressively revealing a subnetwork that can achieve comparable accuracy as the unpruned version. Iterative pruning is the preferred approach when accuracy is prioritized as training the network and then subsequently applying one-shot pruning can be harmful to accuracy.

3.3 Quantization-Aware Meta-training

In FS-SQAM, during the quantization-aware meta-training phase, we opt for QAT [34] to reduce the precision from 32-bit floating point to 8-bit integers. QAT relies on compensating for quantization error in the training. Given the input x and the corresponding labels y of a model f_θ and a quantization function q, the quantized value of x in the forward pass becomes $x_q = \min\left(q_{max}, \max\left(q_{min}, \frac{x}{s} + z\right)\right)$ which maps each input weight to an integer. q_{min} And q_{max} stand for the minimum and maximum values for the 8-bit range, s and z represent the scaling factor for the described quantization range and the zero-point respectively. Then dequantization happens through $\hat{x} = \left(x_q - z\right)s$ which maps the quantized input x_q back to floating-point. However, this operation $x_q \rightarrow \hat{x}$, yields $\hat{x}$ which is not exactly equal to the original input x, thus inducing quantization noise which is taken into account in the training. Re-training is required to recover accuracy degradation from quantization. The previously mentioned phases are detailed in Fig. 1.

The applied quantization loss, where q is the quantization operation, for the model can be described as:

$$L\left[y, f_\theta(x, q(\theta))\right] \quad (3)$$

Incorporating quantization loss into the training yields the following objective function to be minimized:

$$\min_\theta L\left(y,\hat{y}\right) + L\left[y, f_\theta(x, q(\theta))\right] \quad (4)$$

The workflow of FS-SQAM is listed in Algorithm 1, starting with sparsification through L$_1$ regularization then is proceeded by pruning and applying quantization through QAT.

Algorithm 1. Sparse-QAT Aware Meta-training Algorithm

```
Input: A pretrained model $f_\theta$ with parameters $\theta$, total number of layers $L$, pruning mask $m$ and pruning percentage $p$, sampled episodes consisting of support set and query set $S = \{(x_1, y_1), \ldots, (x_{K \times N}, y_{K \times N})\}$ and $Q = \left\{(x_1, y_1), \ldots, \left(x_{K_q \times N}, y_{K_q \times N}\right)\right\}$
Output: Lightweight lymphoblast classifier
  Phase 1: Sparse-Aware Meta-Training
  /*Introduce sparsity to the network*/
Set the hyperparameter $\lambda$
Apply the regularization penalty to trainable layers in $f_\theta$
For each epoch do
  For each episode do
      Freeze the first $L - k$ layers of $f_\theta$
      Minimize the loss function $\min_\theta L_{CE}$
  end
end
  /*Iterative pruning*/
For each epoch do
  For each episode do
      Prune the smallest $p$ weights (in accordance with the pruning schedule) at every layer by applying a pruning mask $m$
      Fine-tune the masked model, $m \odot \theta$ through minimizing loss in $\min_\theta L_{CE} + \lambda \sum_{i=1}^{N} |w_i|$
      Repeat the previous two steps until the desired sparsity is reached
  end
end
  Phase 2: Quantization-Aware Meta-Training
  /*Quantization-aware training*/
For each epoch do
  For each episode do
      Freeze the first $L - k$ layers of $f_\theta$
      Train in simulated quantization
      Minimize the objective function $\min_\theta L(y, \hat{y}) + L\left[y, f\left(x, q(\theta)\right)\right]$
  end
end
Apply quantization to the model
```

4 Experimental Setup

Our initial model is an ImageNet pre-trained VGG19 network. The model undergoes compression as previously detailed. The first 3 convolutional blocks of the network are frozen during the sparse training and quantization-aware meta-training phases.

The meta-train dataset is the 40× magnification segment from the BreakHis [27] dataset which is where the training tasks are sampled from. The few-shot training tasks are 8-way 10-shots for each episode and the model is trained for 150 epochs where they are divided equally between the sparse training, sparse-aware, and quantization-aware meta-training phases. The model is evaluated on test episodes sampled from the ALL [8] dataset, specifically the ALL-IDB2 segment of the dataset.

4.1 Datasets

The first dataset used is BreakHis [27], which is a dataset of histological images of breast tumors. The dataset has eight classes, where the images of interest are the 40× magnification images which total 1,995 samples. The following table shows the number of images in each of the eight classes (Table 1).

Table 1. The Eight Classes of the BreakHis Dataset.

Classes	Magnification Factor				Total
	40×	100×	200×	400×	
Adenosis (A)	**114**	113	111	106	444
Fibroadenoma (F)	**253**	260	264	237	1014
Tubular Adenona (TA)	**109**	121	108	115	453
Phyllodes Tumor (PT)	**149**	150	140	130	569
Ductal Carcinoma (DC)	**864**	903	896	788	3451
Lobular Carcinoma (LC)	**156**	170	163	137	626
Mucinous Carcinoma (MC)	**205**	222	196	169	792
Papillary Carcinoma (PC)	**145**	142	135	138	500
Total	**1995**	2081	2013	1820	7909

The second dataset included is the ALL-IDB2 [8] dataset, which is a binary dataset consisting of images of blood smears. The image could either be normal or abnormal, reflecting the presence of normal or abnormal blast cells. The dataset is balanced with an equal number of normal and abnormal images, which sums up to a total of 260 images.

5 Results

We present the results of meta-training a prototypical network using FS-SQAM on the BreakHis dataset that is tested on the ALL-IDB2 dataset using a VGG19 backbone. The results of the model's performance are the mean accuracy of three test runs, where each result is the average performance of the classifier across the test tasks. The number of test tasks is set to 600 with the query set size set to four samples for both the training and testing. Experiments were repeated for sparsities of 50%, 70%, and 90% which are shown in Table 2. Results are reported for the use of pruning alone or followed by 8-bit quantization-aware training. The resulting sizes are also reported to provide a view of the accuracy-size tradeoffs for better assessment in Table 3. All of the reported sizes are of.*tflite* [35] files in MBs.

The baseline result without any use of compression is 82.22% accuracy. It is observable that QAT slightly improves the accuracy of the 50% and 70% pruned networks. This leads to an accuracy of 77.95% at 70% sparsity and 8-bit quantized weights. With regards to the obtained compression, the network size is 19.60 MB which is a compression of 3.90× compared to the baseline.

Table 2. Meta-test accuracy for FS-SQAM on ALL-IDB2 dataset using VGG19 network for 2-way 10-shot classification (without regularization).

Sparsity				
	0%	50%	70%	90%
Pruning (Baseline)	0.8222	0.7813	0.7769	0.7485
With QAT	0.8127	0.7888	0.7795	0.7291

Table 3. Model sizes of VGG19 after FS-SQAM in MBs.

Sparsity				
	0%	50%	70%	90%
Sparse-Aware Meta-Training	76.41	55.84	33.60	11.34
FS-SQAM	29.28	32.39	19.60	6.79

Table 4. Meta-test accuracy for FS-SQAM with L_1 regularization applied with varying values of λ on ALL-IDB2 dataset using VGG19 network for 2-way 10-shot classification.

Sparsity			
	50%	70%	90%
$\lambda = 0.001$			
Sparse-Aware Meta-Training	0.7875	0.7743	0.7552
FS-SQAM	0.7864	0.7851	0.7251
$\lambda = 0.005$			
Sparse-Aware Meta-Training	0.7874	0.7867	0.7457
FS-SQAM	0.7898	0.7817	0.7335
$\lambda = 0.01$			
Sparse-Aware Meta-Training	0.7930	0.7826	0.7474
FS-SQAM	**0.7908**	0.7822	0.7338

Table 4 shows the results of using sparse-aware meta-training on its own and FS-SQAM with L_1 regularization applied using $\lambda = \{0.001, 0.005, 0.01\}$ for the sparsity ratios 50%, 70%, and 90%. It is worth noting that applying regularization improves the accuracy at some sparsities compared to not applying regularization. Notably at 50% sparsity, accuracy when $\lambda = 0.005$ and $\lambda = 0.01$ is 78.98% and 79.08% respectively, compared to 78.88% with no regularization applied with a compression ratio of 2.35×. Additionally in the FS-SQAM configuration, we find that $\lambda = 0.01$ yields the highest accuracy in this category for 50% and 70% sparsity ratios, which are 79.08% and 78.22%.

While the role of pruning and quantization remains largely unexplored in the context of few-shot learning, pruning has been used to prevent overfitting in meta-learning [36, 37]. The generalization ability of the classifier can be attributed to overcoming the over-parameterization in the network. Over-parameterization has the effect of hurting the generalization ability of the network to unseen examples in normal supervised learning [38]. Thus, expanding this concept to few-shot learning, it is possible to boost generalization to unseen classes through sparsification which also provides the added benefit of compression.

To enhance the compression of the model, QAT has been deemed a method that incurs minimal accuracy loss. Thus, the FS-SQAM framework provides an adaptable classifier that can achieve a fair accuracy-to-size trade-off in a cross-domain setting. Resource-wise, 90% sparsity in FS-SQAM provides an accuracy of 73.38% alongside resulting in the highest compressed model with a compression rate of 11.25×.

To provide context for the obtained results, we present some of the results from related works that address few-shot learning for medical datasets, specifically those that involve a significant domain shift in the following table to demonstrate the efficacy of FS-SQAM for lymphoblast detection in an out-domain setting (Table 5).

Table 5. Performance comparison with related works in out-domain setting.

Reference	Accuracy (%)	Out-Domain Setting
[28]	68.1 ± 0.51	CRC-TP → BreakHis
[29]	67.56	NCT → PAIP
[17]	55.12 ± 0.13	mini-ImageNet → HEp-2
50% FS-SQAM	79.08	BreakHis → ALL-IDB2

6 Conclusion

Rationing resources is important for real-life clinical scenarios where only restricted computational resources and scarce data are available. We empirically test our framework FS-SQAM which combines sparse-aware training and quantization-aware training in the meta-training process for efficient lymphoblast detection. The obtained results point to the possibility of applying resource-efficient few-shot learning in cross-domain medical settings. The utilization of compression in the meta-training process achieves the goal of minimizing the memory footprint and enabling adaptation to unseen classes which matches the reality of encountering uncommon classes of diseases with very few samples in clinical practice. We empirically investigate the effectiveness of FS-SQAM in BreakHis → ALL-IDB2 cross-domain setting, an extreme case of distribution shift, with a focus on resulting accuracy-size trade-offs. The application of sparse-aware training and quantization-aware training enhances performance and reduces the memory footprint. Additionally, the application of L_1 regularization as a preprocess before pruning yields notable sparse model performance increases at most λ values, especially at 50% sparsity. To further examine the generalizability of the framework, more evaluations need to be undertaken on other medical datasets and domains, we leave this as future work.

References

1. Ellis, R.J., Sander, R.M., Limon, A.: Twelve key challenges in medical machine learning and solutions. Intell. Based Med. **6**, 100068 (2022). https://doi.org/10.1016/j.ibmed.2022.100068
2. Willemink, M.J., et al.: Preparing medical imaging data for machine learning. Radiology **295**, 4–15 (2020). https://doi.org/10.1148/radiol.2020192224
3. Weng, W.-H., Deaton, J., Natarajan, V., Elsayed, G.F., Liu, Y.: Addressing the real-world class imbalance problem in dermatology. In: Proceedings of the Machine Learning for Health NeurIPS Workshop, pp. 415–429. PMLR (2020)
4. Litjens, G., et al.: Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci. Rep. **6**, 26286 (2016). https://doi.org/10.1038/srep26286
5. Rajpurkar, P., et al.: CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. NPJ Digit. Med. **3**, 1–8 (2020). https://doi.org/10.1038/s41746-020-00322-2
6. Altini, N., et al.: Liver, kidney and spleen segmentation from CT scans and MRI with deep learning: a survey. Neurocomputing **490**, 30–53 (2022). https://doi.org/10.1016/j.neucom.2021.08.157
7. Chan, H.-P., Hadjiiski, L.M., Samala, R.K.: Computer-aided diagnosis in the era of deep learning. Med. Phys. **47**, e218–e227 (2020). https://doi.org/10.1002/mp.13764
8. Labati, R.D., Piuri, V., Scotti, F.: All-IDB: The acute lymphoblastic leukemia image database for image processing. In: 2011 18th IEEE International Conference on Image Processing, pp. 2045–2048 (2011). https://doi.org/10.1109/ICIP.2011.6115881
9. Genovese, A.: ALLNet: acute lymphoblastic leukemia detection using lightweight convolutional networks. In: 2022 IEEE 9th International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), pp. 1–6 (2022). https://doi.org/10.1109/CIVEMSA53371.2022.9853691
10. Genovese, A., Hosseini, M.S., Piuri, V., Plataniotis, K.N., Scotti, F.: Histopathological transfer learning for acute lymphoblastic leukemia detection. In: 2021 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), pp. 1–6 (2021). https://doi.org/10.1109/CIVEMSA52099.2021.9493677
11. Maaliw, R.R., et al.: A multistage transfer learning approach for acute lymphoblastic leukemia classification. In: 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), pp. 0488–0495 (2022). https://doi.org/10.1109/UEMCON54665.2022.9965679
12. Snell, J., Swersky, K., Zemel, R.S.: Prototypical Networks for Few-shot Learning. arXiv preprint http://arxiv.org/abs/1703.05175 (2017). https://doi.org/10.48550/arXiv.1703.05175

13. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D.: Matching networks for one shot learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, pp. 3637–3645. Curran Associates Inc., Red Hook (2016)
14. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H.S., Hospedales, T.M.: Learning to Compare: Relation Network for Few-Shot Learning. arXiv preprint http://arxiv.org/abs/1711.06025 (2018). https://doi.org/10.48550/arXiv.1711.06025
15. Chao, S., Belanger, D.: Generalizing few-shot classification of whole-genome doubling across cancer types. Pac. Symp. Biocomput. **27**, 144–155 (2022)
16. Paul, A., Shen, T.C., Peng, Y., Lu, Z., Summers, R.M.: Learning few-shot chest X-ray diagnosis using images from the published scientific literature. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 344–348 (2021). https://doi.org/10.1109/ISBI48211.2021.9434059
17. Walsh, R., Abdelpakey, M.H., Shehata, M.S., Mohamed, M.M.: Automated human cell classification in sparse datasets using few-shot learning. Sci. Rep. **12**, 2924 (2022). https://doi.org/10.1038/s41598-022-06718-2
18. Guo, Y., et al.: A Broader Study of Cross-Domain Few-Shot Learning. arXiv preprint http://arxiv.org/abs/1912.07200 (2020)
19. Triantafillou, E., et al.: Meta-Dataset: A Dataset of Datasets for Learning to Learn from Few Examples. arXiv preprint http://arxiv.org/abs/1903.03096 (2020). https://doi.org/10.48550/arXiv.1903.03096
20. Zhang, X., Colbert, I., Kreutz-Delgado, K., Das, S.: Training Deep Neural Networks with Joint Quantization and Pruning of Weights and Activations. arXiv preprint http://arxiv.org/abs/2110.08271 (2021). https://doi.org/10.48550/arXiv.2110.08271
21. Park, J.-H., Kim, K.-M., Lee, S.: Quantized sparse training: a unified trainable framework for joint pruning and quantization in DNNs. ACM Trans. Embed. Comput. Syst. **21**, 60:1–60:22 (2022). https://doi.org/10.1145/3524066
22. Bronskill, J., Massiceti, D., Patacchiola, M., Hofmann, K., Nowozin, S., Turner, R.E.: Memory Efficient Meta-Learning with Large Images. arXiv preprint http://arxiv.org/abs/2107.01105 (2021). https://doi.org/10.48550/arXiv.2107.01105
23. Youn, J., Song, J., Kim, H.-S., Bahk, S.: Bitwidth-Adaptive Quantization-Aware Neural Network Training: A Meta-Learning Approach. arXiv preprint http://arxiv.org/abs/2207.10188 (2022). https://doi.org/10.48550/arXiv.2207.10188
24. Chauhan, J., Kwon, Y.D., Mascolo, C.: Exploring On-Device Learning Using Few Shots for Audio Classification **5**
25. Tian, H., Liu, B., Yuan, X.-T., Liu, Q.: Meta-learning with network pruning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12364, pp. 675–700. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58529-7_40
26. Javed, S., Mahmood, A., Werghi, N., Benes, K., Rajpoot, N.: Multiplex cellular communities in multi-gigapixel colorectal cancer histology images for tissue phenotyping. IEEE Trans. Image Process. **29**, 9204–9219 (2020). https://doi.org/10.1109/TIP.2020.3023795
27. Breast Cancer Histopathological Database (BreakHis). Laboratório Visão Robótica e Imagem. https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/. Accessed 20 Oct 2022
28. Shakeri, F., et al.: FHIST: A Benchmark for Few-shot Classification of Histological Images. http://arxiv.org/abs/2206.00092 (2022). https://doi.org/10.48550/arXiv.2206.00092
29. [2202.09059] Towards better understanding and better generalization of few-shot classification in histology images with contrastive learning. arXiv preprint https://arxiv.org/abs/2202.09059. Accessed 01 June 2023
30. 100,000 histological images of human colorectal cancer and healthy tissue. Zenodo. https://zenodo.org/record/1214456. Accessed 02 June 2023

31. Kim, Y.J., et al.: PAIP 2019: liver cancer segmentation challenge. Med. Image Anal. **67**, 101854 (2021). https://doi.org/10.1016/j.media.2020.101854
32. Ren, M., et al.: Meta-Learning for Semi-Supervised Few-Shot Classification. arXiv preprint http://arxiv.org/abs/1803.00676 (2018)
33. Frankle, J., Carbin, M.: The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. arXiv preprint http://arxiv.org/abs/1803.03635 (2019). https://doi.org/10.48550/arXiv.1803.03635
34. Jacob, B., et al.: Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. arXiv preprint http://arxiv.org/abs/1712.05877 (2017). https://doi.org/10.48550/arXiv.1712.05877
35. TensorFlow Lite. https://www.tensorflow.org/lite/guide. Accessed 26 Jan 2023
36. Chijiwa, D., Yamaguchi, S., Kumagai, A., Ida, Y.: Meta-ticket: finding optimal subnetworks for few-shot learning within randomly initialized neural networks. Presented at the Advances in Neural Information Processing Systems, 31 October (2022)
37. Liu, Z., et al.: MetaPruning: Meta Learning for Automatic Neural Network Channel Pruning. arXiv preprint http://arxiv.org/abs/1903.10258 (2019). https://doi.org/10.48550/arXiv.1903.10258
38. Hoefler, T., Alistarh, D., Ben-Nun, T., Dryden, N., Peste, A.: Sparsity in Deep Learning: Pruning and Growth for Efficient Inference and Training in Neural Networks. arXiv preprint http://arxiv.org/abs/2102.00554 (2021). https://doi.org/10.48550/arXiv.2102.00554