December 13, 2023

CS 502 - Project report option 1

Zoé Majeux Moritz Rau

Abstract

Accurate tumor detection is vital in cancer diagnosis, but some rare cancers lack sufficient data for traditional deep learning. Few-Shot Learning offers a promising solution. This project expands a benchmark, using a breast tumor dataset, to improve few-shot histology image classification across varying microscope magnifications with three different meta-learning and two baseline methods.

1. Introduction

Effective tumor detection and classification are crucial for cancer diagnosis and treatment. Traditional histological analysis faces challenges like subjective interpretation and high costs. There exists a diverse spectrum of cancer types affecting different anatomical structures. Thus, leveraging machine learning, particularly Few-Shot Learning, offers a promising alternative. In this project, we extend the existing benchmark by adding a new public data set named BreakHis (Spanhol et al., 2016).

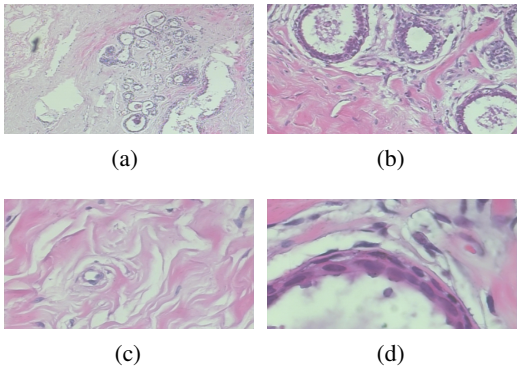
Various machine learning research has already been conducted on histology image classification. Most notably, Shakeri et al. presented a public benchmark, designed to quantitatively evaluate few-shot histology image classification methods. They consider three different scenarios as a function of the domain shifts between the source and target histology data (near-domain, middle-domain, and out-domain). The study utilizes a dataset of 100'000 images of colorectal cancer, the benchmark's largest, as the base training source. Three adaptation scenarios are defined: near-domain, testing on a different smaller colon dataset; middle-domain, evaluating LC25000 (lung and colon dataset); and out-domain, assessing BreakHis (breast dataset).

In this study the BreakHis dataset is integrated into a benchmark that has already demonstrated success in addressing various themes using Few-Shot learning. Our study differs by not relying on knowledge from a broader base dataset encompassing various tumors across different body parts. Instead, for training images of common breast tumor types are used to improve the accuracy of classifying less common tumor types. The research objective is to evaluate multiple algorithms on BreakHis, utilizing different pretrained networks as feature extractors. The experiment will initially assess algorithm performance with varying parameters and backbones, followed by a comparative analysis of the fine-tuned algorithms. Ultimately, the study aims to identify the optimal accuracy across different image magnifications.

1.1. Break His

The Breast Cancer Histopathological Image Classification (BreakHis) dataset comprises 9,109 microscopic images of breast tumor tissue obtained from 82 patients at various magnification levels (40X, 100X, 200X, and 400X). The dataset includes 2,480 benign and 5,429 malignant samples, each with dimensions of 700 by 460 pixels, in 3-channel RGB format with 8-bit depth per channel, and stored in PNG format. The data was collected in collaboration with the P& D Laboratory – Pathological Anatomy and Cytopathology in Parana, Brazil.

BreakHis is categorized into two groups: benign tumors and malignant tumors. Benign tumors exhibit no malignancy criteria and are generally slow-growing and localized. Malignant tumors, synonymous with cancer, can invade adjacent structures and metastasize. The dataset consists of four histologically distinct types of benign tumors (adenosis, fibroadenoma, phyllodes tumor, and tubular adenoma) and four malignant tumors (carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma).



(a) (b) (c) (d)

Figure 1: Tumors histology with magnifications (a) 40x, (b) 200X, (c) 300X, (d) 400X.

2. Method

Our approach is based on the use of a pre-trained convolutional neural network (CNN) as a fixed feature extractor, combined with trainable linear layers that serve as a classifier. This combination forms the backbone of our model.

The pretrained CNN, trained on a large-scale dataset (e.g. ImageNet), is utilized as a feature extractor. This network has already learned to identify a wide range of features from its pretraining phase, which allows us to leverage these learned features for our specific task. We freeze the weights of this network, meaning that they are not updated during the training of our model. This allows us to maintain the useful features that the CNN has already learned, while significantly reducing the number of parameters that need to be learned by our Few-Shot algorithms. For our experiments, we consider two different CNNs: convnext_base from Liu et al. and efficientnet_b0 from Tan & Le. They have been chosen due to their relatively high accuracy, (96.87 and 93.532) and their high difference in the number of parameters (88.6M and 5.3M).

2.1. Few-Shot Algorithms

While an in depth description of the different Few-Shot algorithms is ommited a brief overview is provided. In literature often it is distinguished between meta-learning methods (MatchingNet, ProtoNet, MAML) and Baseline methods. In general, meta-learning methods are designed to learn how to learn, i.e., they aim to learn a strategy for quickly adapting to new tasks. On the other hand, methods like Baseline and Baseline++ focus more on learning good feature representations that can be used for classification.

2.2. Train-test split

Our dataset contains eight distinct tumour types as described above. Notably, the prevalence of these tumor types varies, with certain classes being significantly less common than others (Makki, 2015). This variability in occurrence positions the less common tumor types as ideal candidates for Few-Shot Learning. In our approach, we designate the more prevalent tumor types as our training classes, making use of the abundance of data. We assign the rarer tumor types as validation and test classes, aiming to extrapolate knowledge to these less common instances. Also, we assigned benign and malignant classes in equal proportion to all splits.

Training:	Adenosis, Fibroadenoma, Ductal Carcinoma, Lobular Carcinoma
Validation:	Phyllodes Tumor, Papillary Carcinoma
Test:	Tubular Adenoma, Mucinous Carcinoma

Table 1: Dataset split

2.3. Hyperparameter Tuning

To find optimal model performance, we have to tune hyperparameters, notably these are the learning rate and backbone dimensions. Employing a grid search, we systematically run various combinations to fine-tune these parameters. The number of epochs is not fine-tuned, but left to the default values, as the train loss converged to near zero values. Future methods should imperatively consider the validation loss and batch size. The hyperparameter space is kept small as we assess five different algorithms. The tested linear backbone layers were $1024x64$ (two layers) and 1024 (single layer) combined with learning rates of 0.01 and 0.001. The best parameters are chosen based on the accuracy on the validation set. The choice of the pretrained CNN is not considered a hyperparameter but as a different experiment.

2.4. N-way, n-shot

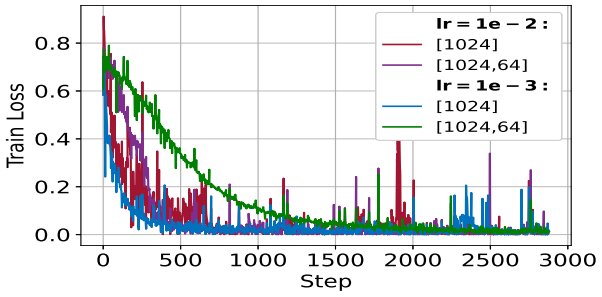
In literature a common setting for Few-Shot experiments is 5-way 1-shot/5-shot/10-shot (Shakeri et al., 2022)(Chen et al., 2020). In configuring our Few-Shot Learning parameters for our dataset, we were constrained by the limited number of classes (8). N-way is set to 2, which is a small number, however, should make the problem more manageable. Furthermore, we set n-shot to 10, determining the number of samples per class for learning, and n-query to 15, indicating the number of examples per class for evaluation. The minimum number of samples per class is thus n-shot + n-query $= 15$, which is fulfilled by all classes.

3. Experiments

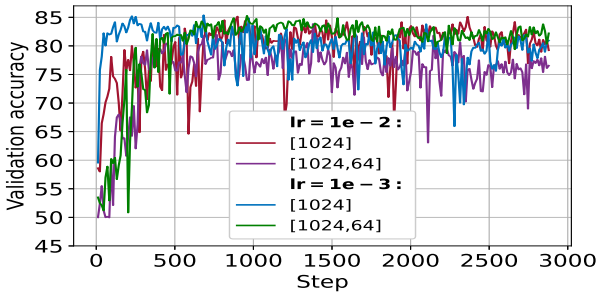
For each pretrained CNN and algorithm the hyper-parameter tuning is done as described in 2.3. The loss and accuracy curves for MAML and Baseline++ are illustrated in Fig.2 to highlight the distinctions in behavior between different parameters and algorithms. For each algorithm the optimal model is chosen based on the accuracy on the validation set. However, for some configurations the accuracy on the validation set fluctuated enormously with each gradient step. Consequently these configurations were not considered, even though their training loss converged. Then the performance of the different algorithms is compared, as shown in Fig. 3. The accuracies on the test sets are reported in Tab. 2.

4. Discussion

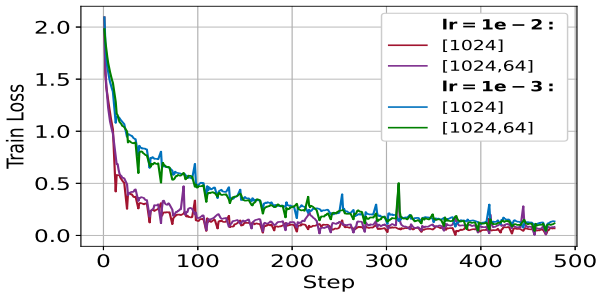
In our experiments, we observed two notable trends across all algorithms, with the exception of MAML. The first trend is the rapid reduction of training loss to near-zero values within a few epochs. However, this minimized training loss exhibits only partial stability, as at some rare epochs the loss significantly spikes up (see 3).



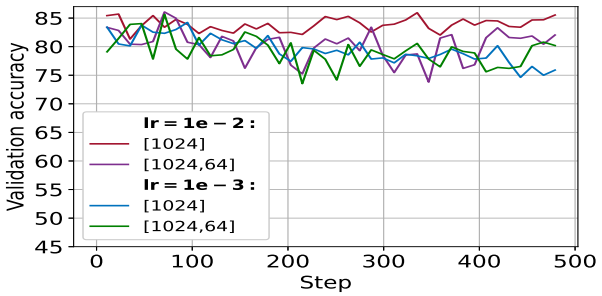
(a)



(b)



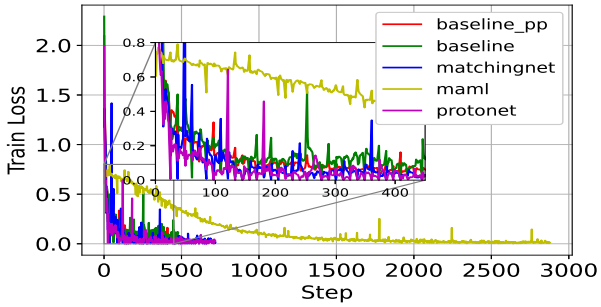
(c)



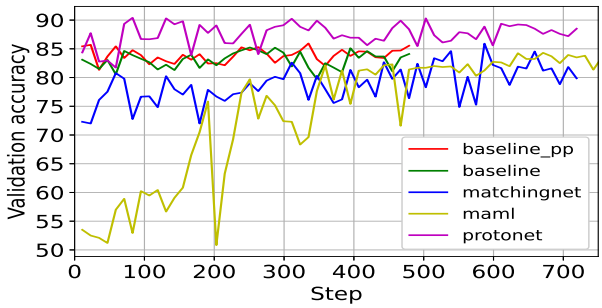
(d)

Figure 2: Loss train (a), (c) for the best validation accuracy (b), (d) for respectively MAML and Baseline++ aglorithms.

The second trend is the attainment of high validation accuracies post the initial epoch (see 3(b)). Despite these high accuracies, they display significant instability. It's highly remarkable that these trends are not observed in the MAML algorithm and will thus be discussed separately.



(a)



(b)

Figure 3: Loss train (a) for the best validation accuracy (b) of different algorithm and ConvNext base as pretrained CNN.

		convnext base	efficient net b0
100X	baseline	87 ± 7	82 ± 8
	baseline++	83 ± 8	79 ± 8
	matchingnet	81 ± 9	74 ± 8
	maml	70 ± 9	65 ± 9
	protonet	76 ± 8	82 ± 7
400X	baseline	81 ± 8	78 ± 8
	baseline++	81 ± 8	69 ± 10
	matchingnet	76 ± 10	70 ± 10
	maml	70 ± 10	68 ± 10
	protonet	80 ± 8	75 ± 8

Table 2: Best accuracy for different algorithms, pretrained CNNs, and magnifications

Following these observations, we encountered challenges in hyperparameter tuning based on validation accuracy. The instability of the validation accuracy, despite its high values, introduced significant difficulty in determining optimal hyperparameters. The fluctuations made it hard to discern whether changes in hyperparameters were leading to gen-

uine improvements or if the observed changes were merely due to the inherent instability. This instability thus complicated the process of refining our models for optimal performance.

4.1. High instable accuracies

The high accuracies observed early in training can be attributed to several factors. One key factor is the presence of multiple images from the same patient in the dataset. This introduces correlations within the data that the algorithms can quickly learn and exploit, leading to high accuracies. For instance, in the case of MatchingNet and ProtoNet, these algorithms are designed to learn a metric space where samples from the same class are closer to each other. If multiple images from the same patient are present, these images are likely to be very similar and thus close in the learned metric space. This can lead to high accuracy as the model can easily match or classify these images correctly. Despite these advantages, it was the Baseline method that ultimately outperformed the others. This suggests that, in this specific context, a simpler approach could be more effective at leveraging the correlations in the data and achieving high accuracy.

Using a pretrained CNN with frozen weights as a feature extractor simplifies the task for algorithms. The CNN already knows how to extract meaningful image features, giving algorithms a head start and potentially leading to high early accuracy. Training is further simplified by only training a linear layer head, which classifies based on the CNN-extracted features. But this approach has limitations. The frozen CNN isn't tailored to our task, so extracted features may not be optimal. Also, a linear layer might not capture complex feature-class relationships, limiting model performance.

As for the instability of the validation accuracy, it could be due to the inconsistent sampling of images from the same patient. In some epochs, if multiple images from the same patient are sampled, the model might perform exceptionally well due to the inherent correlations in these images. However, in epochs where the images from the same patient are not sampled together, the model might struggle to generalize, leading to a drop in accuracy. This inconsistent sampling could introduce significant variability in the model's performance, resulting in the observed instability in validation accuracy. The observed accuracy fluctuations could be due to the model falling into and escaping from local minima in the loss function. This can happen with a large learning rate or new data, causing temporary accuracy improvements. However, if the model re-enters the local minimum, accuracy can fluctuate.

The choice of pretrained CNN also impacts accuracy. For instance, EfficientNet B0 consistently yields lower accuracy than ConvNext base, likely due to ConvNext base having nearly 17 times more parameters. Differences in the CNNs' own accuracies on the tasks they were initially trained on (96.87 vs 93.532) might also play a role, suggesting future research should explore various pretrained CNNs.

Interestingly, images with 400x magnification generally have lower accuracy than those at 100x. This could be because lower magnification images cover a larger area, providing more information for analysis. However, optimizing the model for 400x magnification images could be beneficial, as higher zoom might reveal details unseen at lower magnifications, potentially improving tumor risk detection.

4.2. Rapid loss decay and MAML

The rapid decay of the loss function observed during training suggests that the model is quickly learning to fit the data. This typically occurs when the model effectively uses the data's features for accurate predictions. However, without access to the training loss, it becomes challenging to monitor for overfitting. A rapidly decaying loss without a corresponding decrease in validation loss could indicate overfitting, where the model memorizes the training data instead of learning generalizable patterns. In such cases, the lack of training loss data makes it difficult to assess and address potential overfitting.

MAML may not exhibit the same rapid decay in loss as other algorithms due to its unique approach to learning. Unlike traditional methods that optimize for a specific task, MAML is designed to find a model initialization that is suitable for fine-tuning on a variety of tasks. This means that instead of rapidly converging to a solution for a single task (which can lead to a quickly decaying loss), MAML seeks a more general solution that works well across multiple tasks. This can result in a slower, more stable decrease in loss during training.

5. Conclusion

In conclusion, our approach using a pre-trained CNN and trainable linear layers showed promising results. However, rapid training loss decay and unstable validation accuracies posed challenges in hyperparameter tuning. High early accuracies were due to multiple images from the same patient and the use of a pretrained CNN. The instability of validation accuracy, potentially due to inconsistent image sampling, complicated model refinement. The choice of pretrained CNN and image magnification also impacted accuracy, suggesting areas for future research. MAML, not exhibiting rapid loss decay, might be more suitable for this task, warranting further investigation. Future work should consider validation loss and batch size during hyperparameter tuning, and explore different pretrained CNNs and image magnifications.

6. Others

Paper (Langley, 2000).

References

Chen, W.-Y., Liu, Y.-C., Kira, Z., Wang, Y.-C. F., and Huang, J.-B. A closer look at few-shot classification, 2020.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s, 2022.

Makki, J. Diversity of breast carcinoma: Histological subtypes and clinical relevance. *Clinical Medicine Insights: Pathology*, 8:23–31, Dec 21 2015. doi: 10.4137/CPath.S31563.

Shakeri, F., Boudiaf, M., Mohammadi, S., Sheth, I., Havaei, M., Ayed, I. B., and Kahou, S. E. Fhist: A benchmark for few-shot classification of histological images, 2022.

Spanhol, F., Oliveira, L. S., Petitjean, C., and Heutte, L. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering (TBME)*, 63(7):1455–1462, 2016.

Tan, M. and Le, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.