High Accuracy, Low Data: Few-Shot Learning for Medical Image Diagnostics

Momchil Gavrilov
UC Davis
mggavrilov@ucdavis.edu

Tyler Culp
UC Davis
taculp@ucdavis.edu

Jason Eissayou
UC Davis
jdeissayou@ucdavis.edu

Abstract

Background: Under-utilization of machine learning methods in medical diagnostics is costing medical professionals' time and worsening patient outcomes. Few-shot learning algorithms can help translate advanced machine learning models into the low data availability medical diagnostics space. However, few-shot models currently lack the accuracy and robustness for direct application. Recent methods, such as applied vector quantization and ensembles of embedding spaces, have shown improvements in few-shot learning accuracy. Another method, image selection using a sample choosing policy, can make few-shot models more robust in medical landmarking tasks. Combining these recent advances in few-shot learning accuracy and robustness is paramount to developing systems that can swiftly transfer into medical diagnostics. **Methods:** Using ResNet50 and a large lung x-ray binary classification dataset, we trained five models: a fully connected neural network "gold-standard" (GS)[1] classifier trained on the full dataset and a 32-image subset (GS-small)[2]; a few-shot model based on the ensemble of embedding spaces (EE)[3]; a few-shot model based on vector quantization strategies (VQ)[4]; and a high-accuracy (HA)[5] few-shot model developed on both vector quantization and ensemble frameworks. A sample choosing policy was applied to the dataset, and all models were trained on the selected samples and compared to the performance of their 32 random image-trained counterparts. **Results:** The GS model outperformed all other models. There was a significant increase in the classification performance of the EE model compared to the GS-small model. **Conclusion:** Few-shot learning is showing efficacy when compared to the gold standard trained on small datasets. More experiments need to be run to draw further conclusions. Future work will incorporate code for other few-shot methods and sample choosing policies into our experimental design.

1 Introduction

Medical imaging plays a pivotal role in patient disease diagnosis. However, due to the expense of physician time and lengthy analysis processes, patients can experience long waiting periods and may choose to forgo imaging altogether (Bjørn Hofmann, 2023). Aiding physicians in classifying diseases based on medical images can improve patient outcomes by reducing waiting times, scheduling challenges, and procedural costs. Large data machine learning models can successfully classify common images (Dong Su, 2018). However, because of the scarcity of labeled medical imaging data, large data models fall short in medical contexts. Few-shot learning (FSL) is a method designed to utilize little labeled data to achieve high classification accuracy (Quan Quan, 2022). FSL's strategy involves developing class-specific 'prototypes' based on averaged encodings from pre-trained large models such as convolutional neural networks and transformers. During classification, input - 'query' - images are encoded by pre-trained models and labeled based on the prototype with the greatest similarity to the input image. FSL models are well suited for classifying low-labeled data medical images. However, reliable applications in medical diagnostics and decision-making involve rigorous validation and high prediction accuracy.

One notable strategy to improve FSL prediction accuracy is to recruit multiple FSL subspaces to a classification task as a form of ensemble voting (Kshitiz, 2023a). In this strategy, each FSL subspace acts as a voting member in the classification task and affects the final prediction made by the model. Since FSL already performs well on its own, the addition of ensemble subspaces will further increase prediction accuracy. Creating disparity between the subspaces is a proposed step to improving prediction accuracy (Kshitiz, 2023a). Intuitively, greater discrimination between the sub-

spaces will make each individual vote from a subspace more valuable to the overall prediction. Minimizing the sum of the projections of each vector from one subspace to another is the way to do this (Kshitiz, 2023a). The ensemble of subspaces with a discriminative loss function has shown improvements in the performance of FSL models (Kshitiz, 2023a). However, a limitation of this work is the use of a large number of randomly organized prototypes per class, which is not optimal for few-shot prediction performance (Shiqi Huang, 2023).

To further understand the sub-optimal performance of FSL with randomly organized prototype spaces, FSL must be viewed from a vector quantization (VQ) perspective (Shiqi Huang, 2023). Prototypes occupy an embedding space with defined embedding vectors acting as borders between different prototypes. When a new 'query' vector enters the space, it will be labeled corresponding to which prototype's border it lands in. This can best be depicted with Voronoi tessellations. A random organization of prototypes will saturate the embedding space and overfit to the training data (Shiqi Huang, 2023). Furthermore, the bordering vectors will become 'fuzzy' due to the inherent variability of data and cause the model to make mistakes near border cases (Shiqi Huang, 2023). Medical images will have a lot of variability and therefore a lot of border cases. This requires the inclusion of VQ optimization strategies for FSL to achieve the best possible performance.

Three vector quantization methods that can improve the performance of FSL are self-organizing VQ, grid-form VQ, and residual-oriented VQ. Each can improve prediction performance on its own, but together they can optimize performance (Shiqi Huang, 2023). Self-organized VQ acts to organize similar prototypes by grouping them together in the embedding space. This can bring structure to a usually random organization of prototypes. Self-organizing VQ further acts to distill large groups of prototypes into single best-representing prototypes from each group - essentially pruning the prototype number. Grid-form VQ will act to smooth out the differences between neighboring prototypes via inverse distance weighting. This smoothing process will result in greater similarity within groups and greater disparity outside of the groups, resulting in stronger defined borders between different prototype groupings. Finally, residual-oriented VQ adds the consideration of class labels into the organization process. Residual-oriented VQ will encourage grouping based on labels and result in strong class separation and final prediction accuracy (Shiqi Huang, 2023). When these VQ strategies were applied to a medical landmark detection task, they outperformed other few-shot models (Shiqi Huang, 2023), suggesting that they should be considered strongly in FSL. However, there can be challenges with data availability in medical contexts.

It is common in medicine to encounter rare diseases with inaccessible labeled data to train on. In these cases, the only strategy FSL models can utilize is self-supervised selection of the best representative images for annotation by medical professionals. A Sample Choosing Policy (SCP) is proposed to help with this process (Quan Quan, 2022). The SPC uses a rating score of images generated from two different feature extractors: SIFT and any large pre-trained model. It can be best summarized into three parts: first, radiological images are fed through the large pre-trained feature extractor using self-supervised learning for clustering; the second model performs Key Point Proposal for localizing important image features within groups; and finally, a score is generated using Representative Score Estimation to select the most informative images in each group (Quan Quan, 2022). Using SCP to select training images for FSL models resulted in reduced mean radial error in the task of landmarking medical images when compared to the random selection of images (Quan Quan, 2022).

All of the above studies have shown great contributions to FSL accuracy and robustness. However, to our knowledge, there have been no studies testing the use of an ensemble of FSL subspaces, with a subspace discriminative loss function, and vector quantization strategies for high-performing medical image classification FSL models. To generalize the performance of high-accuracy FSL models to unlabeled data medical contexts, we plan to compare the performance of the full-data trained FSL to an FSL model trained on single class images selected from a sample choosing policy. Improvements made to FSL classifiers are crucial for timely applications of machine learning tools in medicine since FSL models will precede large data models - due to their low data requirements - and will span a broader set of low data applications.

2 Methods

2.1 Preliminaries

Problem Definition: There is a lack of machine learning utilization in medical image classification because of a lack of high quality labeled datasets. Because of this, the accuracy of machine medical image classification is quite low, even with all of the recent advancements in the field of image classification as a whole.

System Overview: We plan to combine the ensemble of prototype subspaces with a cross entropy loss function and vector quantization methods to create high-accuracy few-shot learning models which can address the problem above. We also plan to test performance against a few-shot learning model trained on a smaller dataset selected by a sample choosing policy to generalize our results to unlabeled medical data scenarios. The dataset we are using contains 5.8 thousand images of chest X-rays, each labeled as either "pneumonia" or "normal" (i.e., we have two classes).

2.2 Techniques

The techniques we will employ to solve the above problem is to first take a pre-trained ResNet model (ResNet50 model) and use it as an encoder for the image data. This ResNet model will be able to take in images and distill them down into tensors containing relevant information for image classification. We will then stack a soft-max layer on top of that to do class prediction. Finally, we will train using Few-Shot Learning strategy which uses just a few images from the training set to generate an average tensor representation, called a "prototype", of each class in the data set and then compare each prototype against a "query" or test image to generate an error that is later used to choose the best prototypes. During inference, the model would see which prototype the test image is most closely aligned to and assigns a class based on that. Training this type of model can be done using tens of images rather than hundreds or even thousands.

The ResNet50 model is a pre-trained model that is used primarily in image encoding / feature extracting. It allows us to condense a large image down into just a single vector. The model has 25.6 million parameters, a depth of 107, and a total size of 90 MB (source: https://keras.io/api/applications/).

3 Experiments

3.1 Set Up

The set up for this project is to develop and train three different models to compare their effectiveness at the task of medical image classification.

1. We first trained a large classification model that uses ResNet50 for feature extraction, feeding encoded images into a dense, fully connected neural network with 256 nodes and utilizing a final two-node softmax layer for classification. This model is trained on the full dataset and serves as a benchmark "gold standard" model that works best given large amounts of labeled data.

2. The second model has the same structure as the first, the only difference is that it will be trained using only a small subset of images from the training set, similar to the number of images that will be used to train the few-shot model.

3. The third model is our high accuracy few-shot model, trained using 32 images broken up into a support set (22 images) and a query set (10 images). It uses a renset encoder with only a two-node softmax layer and no dense, fully connected layer. The support set is used to generate our class prototypes, and our query set is used to test how good the prototypes in this training step are. There is an ensemble of prototype embedding spaces generated from the support set. The ensemble methodology is directly imported from the Kshitiz et.al code (Kshitiz, 2023a). We maximize the differences between the embedding spaces by minimizing the following discriminative loss function during training:

$$\mathcal{L}_{\text{dis}} = \sum_{\forall x,y} \frac{\theta_x \cdot \theta_y}{\|\theta_x\| \|\theta_y\|}$$

Here θ_x and θ_y represent embedding spaces and $\mathcal{L}_{\text{dis}}$ represents our total discriminative loss. The cross entropy loss function for this model comes from the distance between the prototype for a class and an image of the same class type from the query set.

We will train and compare these three models against each other. Our goal is to fine-tune the few-shot model to be as close in performance to the gold-started model as possible.

3.2 Datasets

A wide array of different datasets from Kaggle could have been selected to demonstrate the superior accuracy of our proposed few-shot model.

To ensure a fair comparison with a conventional ResNet model, choosing a dataset with sufficient data to train the ResNet model effectively was essential. After extensively evaluating various datasets, including both multiclass and binary options, we ultimately decided on the chest-xray-classification dataset on Kaggle provided by user Keremberke.

This dataset comprises 5,824 labeled chest X-ray images, each annotated with a 0, indicating a normal chest X-ray, or a 1, indicating the presence of pneumonia. To standardize these X-rays, the dataset's author employed preprocessing techniques that included auto-orienting the pixel data (with EXIF-orientation stripping) and resizing all images to a uniform 640x640 resolution. Notably, no image augmentation techniques were applied, ensuring that each labeled image remained unique without duplication through transformations such as flipping, rotating, or masking. Following normalization and labeling, the dataset was divided into training, validation, and test sets. The training set comprises the majority of the images, approximately 4,000, while the validation set includes around 1,100 images, and the test set contains about 600 images. This amount of data allows us to train the ResNet model without worrying about the lack of labeled data, ensuring a fair comparison to showcase the accuracy of our proposed few-shot model.

4 Results and Analysis

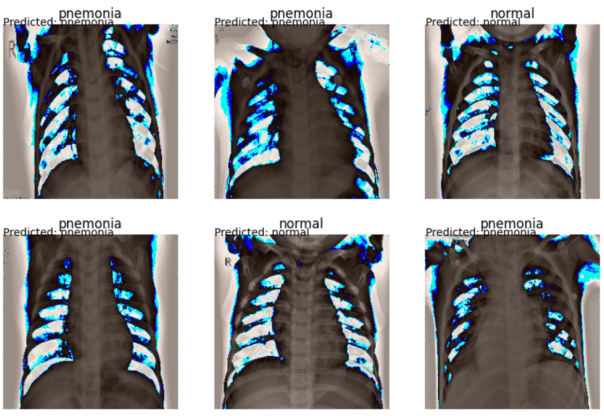

Figure 1: Example Image Predictions made by our large model

Our experiments evaluated the performance of three different models on the chest X-ray classification dataset: the baseline ResNet model, a reduced data ResNet model, and our few-shot learning model. The baseline ResNet model, trained on the full dataset of 5,824 labeled chest X-ray images using the default ResNet 50 parameters, achieved an accuracy of 90% - 94% between multiple training tests. This model served as a benchmark for comparing the performance of the few-shot learning model.

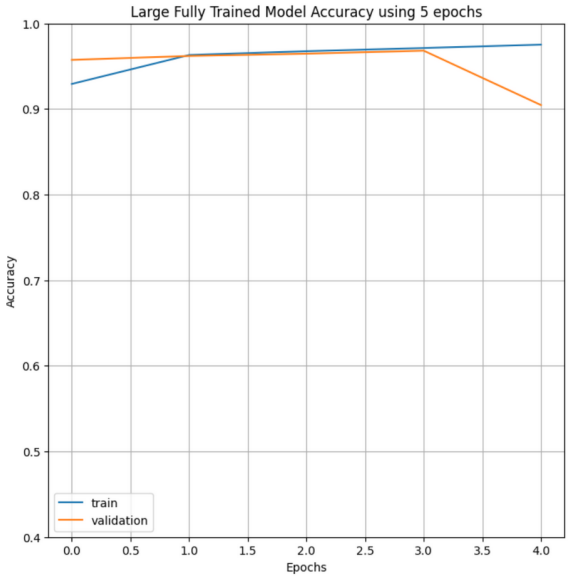

Figure 2: Large Model Accuracy over the training process

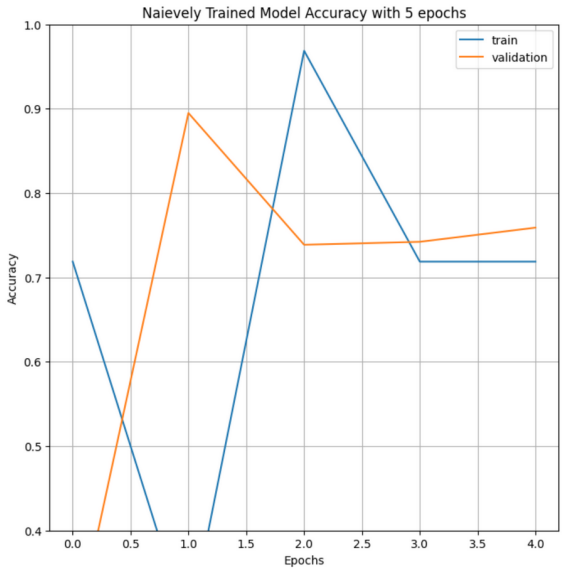

Figure 3: Small Model Accuracy over the training process

Next, we trained the same ResNet model but with a significantly reduced subset of the dataset. The number of images used for this model corresponded to those that were used in the few-shot

learning model. As we expected, the accuracy of the reduced data ResNet model dropped significantly to consistently less than 75% between all tests. This substantial decrease in accuracy highlights the limitations when training a model with insufficient labeled data, making it unsuitable for medical applications where high accuracy is of the utmost importance.

Finally, our few-shot learning model was trained using a support set of 22 images and a query set of 10 images. Despite the limited data this few-shot model was trained on, it achieved an impressive accuracy of 86-90% between all tests, nearly matching the performance of the baseline ResNet model that was trained on the full dataset. This result shows the effectiveness of the few-shot learning approach in addressing the issue of limited labeled data in medical image classification.

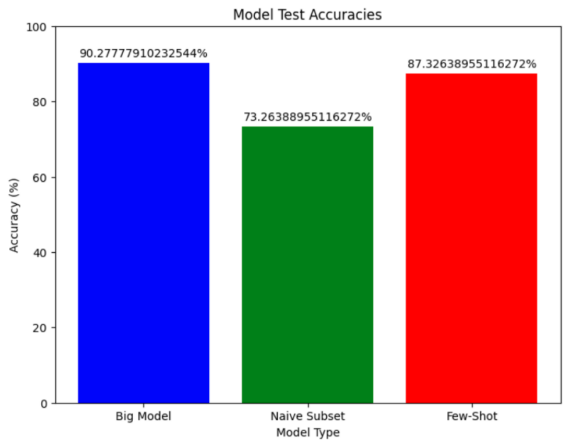

Figure 4: Accuracy differences between models in our final test

Between all of our experiments, we found that the few-shot model was much more sensitive to overfitting than the models trained using more traditional methods, i.e., Large NN with cross-entropy loss. The more traditional models' performance capped out around 5 training epochs, but going beyond that did not negatively impact the performance of these two models very much. The Few-Shot model, on the other hand, needed closer to 30 training epochs to get it to be as accurate as possible, but anything beyond 40 routinely resulted in massive overfitting and huge performance drop-offs.

We hypothesize that implementing a Sample Choosing Policy (SCP) to select the most informative images for training could further enhance the accuracy of the few-shot model. By prioritizing images with distinctive features, the SCP would help define better class prototypes, potentially improving classification performance. This approach could make the few-shot learning model even more robust and reliable for medical diagnostics, where accurate and timely image classification is essential.

5 Related Work

The first paper that we looked at is "Few-Short Diagnosis of Chest X-Rays Using an Ensemble of Random Discriminative Subspaces" by Garvit Garg, Angshuman Paul, published in ICLR (Kshitiz, 2023b). This paper uses the same large data set that we are working with, attempting to train a model using a low number of data points by extracting features from the images, splitting them up into random subspace vectors, and then training a model on this. Its main goal is to learn classification with only a few data points per class (for images). The motivation for this is that there are not a lot of documented medical images out there, and the data sets that do exist are generally pretty sparse. A model like this would make it so this lack of data was less of an issue while also likely being quicker to train than traditional large models.

Our plan is to train a robust model using ResNet and all the data present in the set to see how strong a model we can make. As an extension of this, we will also try to train a model using a small subset of the images we have provided and see how it compares to a more robust model trained using all the data possible. Our goal is to figure out how few images we can use to train a model before it starts getting reasonably close.

The second paper that we looked at is titled "LVM-Med: Learning Large-Scale Self-Supervised Vision Models for Medical Imaging via Second-Order Graph Matching"(Duy Nguyen, 2023). This paper creates a new novel model called LVM Med, which is a self-supervised learning technique using ResNet. This was done because while there are many models that are trained on images, not many of them are trained on specifically medical images. Because of this, there is a significant gap in the performance of any pre-trained model with medical images. This paper compares a few self-supervised algorithms to see which one is best when it comes to learning medical images.

We are also attempting to train a model specifi-

cally using medical images but something we want to do that is slightly different is try to find the optimal number of images needed for these medical image models. Where this paper attempts a few different algorithms, trying to find the best for this task, our goal is to try to find a lower end number of images needed to train a model that is fairly well performing.

The third paper that we looked at is titled "Which images to label for few-shot medical landmark decisions"(Quan Quan, 2022). This paper also deals with the fact that finding good, well-labeled medical images to train a model on is very difficult. Because of this few shot learning, which is a method where we only train a small amount of labeled data on the model, is widely used. This paper discusses how the quality of medical images that a model is trained on can vastly change how well it performs. Thus, they propose a "Sample Choosing Policy," which is basically a way to determine which images are the most worthy of being included in the training data set. This novel policy was found to decrease the error rates (MRE) of models trained with it.

Our project will also be looking into this problem of low amount and quality data sets of medical images to train on. Our goal, though, is to not only train an actual working model but to test out models trained with lower amounts of data. We will compare these to a large robust model that was trained using all of the data (90,000 annotated images). Our goal is to see if we can find a lower bound number for how many images a model needs before it can get performance reasonably close to the large robust model. This paper has opened our eyes to the fact that there are big differences in the quality of images that a model can be trained on and that this could greatly affect our performance, which is something we will need to keep in mind during our experiments. As a quick fix for now, we will make sure that as our models get smaller and smaller, they will use subsets of the set of images originally used to train some of the larger prior models.

6 Conclusion

Through all of our experiments we have concluded that Few-Shot Learning is a viable and best way to train an image model given limited amounts of training data. Since it is fairly uncommon to have large amounts of labeled medical images, this makes it an ideal approach to building models with a focus on medical image classification. Our Few-Shot model was able to achieve an accuracy of 90% on the high end between the multiple test runs that we did, whereas the traditionally trained model using a similar amount of training data was unable to get past an accuracy of 75%. This is a massive difference as the data set we worked with had only two classes, meaning that random guessing would have already achieved roughly 50% accuracy.

We found that while the Few-Shot model does very well with sparse training data, it also requires much more fine-tuning in its meta parameters than the other models. The Few-Shot model requires significantly more epochs to train with in order to achieve the best performance possible, but it is very easy to overshoot this ideal number and quickly overfit the model. Because of this training, a Few-Shot model might require a bit more trial and error in order to find the batch sizes and epoch amounts that will work best for your specific training set.

A final thing to take into account is that all ResNet-based models that are trained on sparse data can benefit from a Sample Choosing Policy. This Sample Choosing Policy can be used to determine the "best" images to train on or, especially in the medical image field, the best images to give to professionals to label. Some images are better than others when it comes to training models, so being able to determine what is best can help save healthcare professionals hours of work by reducing the number of images they need to label down to only a handful, allowing them to do their actual jobs.

7 Future Work

While we are happy with the results we got from our work on the few-shot model there are a few things that we would like to still work on. The first is extending our work to a larger dataset with 15 classes. We believe that our few shot model will demonstrate even better performance in this case than the traditional model trained using low amounts of data, as few-shot models are able to build reasonably accurate prototypes even with low amounts of training data per class.

In addition, we would like to work on employing a Sample Choosing Policy for this larger data set. Our goal in the project was to adapt the landmarking sample choosing policy from Quan et al., which turned out to be significantly more difficult than ex-

pected. Given more time, we would like to develop our own sample choosing model that can inform us what images are the best to use in training. We think that this could increase the performance of our few-shot model by a wide margin.

We did not get to explore any vector quantization strategies for few-shot learning in our current works. Applying VQ methods to the embedding spaces of the ensemble model is something we plan to explore further in this project. We found a great paper that was able to apply vector quantization to basic few-shot learning models, but their code will not be publicly available until later this year (Shiqi Huang, 2023). We believe that the ensemble method with VQ will create the most state-of-the-art few-shot model based on current literature.

The last thing that we would like to do in the future is fine tune the meta parameters of our few shot model. As of right now, most of the work went into finding the best number of epochs to train the model using 32 images, but there is still more work to be done in determining the best number of images to use in the support and query sets of the few shot model. On top of that, we didn't enforce any specific class distributions between the two sets, which is something else we would like to do in future work.

References

Elin Kjelle Bjørn Hofmann, Ingrid Øfsti Brandsaeter. 2023. Variations in wait times for imaging services: a register-based study of self-reported wait times for specific examinations in norway.

Hongge Chen Jinfeng Yi Pin-Yu Chen Yupeng Gao Dong Su, Huan Zhang. 2018. Is robustness the cost of accuracy? – a comprehensive study on the robustness of 18 deep image classification models.

Nghiem T. Diep Duy Nguyen, Hoang Nguyen. 2023. Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching.

Angshuman Paul Kshitiz, Garvit Garg. 2023a. Few-shot diagnosis of chest x-rays using an ensamble of random discriminative subspaces.

Angshuman Paul Kshitiz, Garvit Garg. 2023b. Few-shot diagnosis of chest x-rays using an ensemble of random discriminative subspaces.

Jun Li S.Kevin Zhou Quan Quan, Qingsong Yao. 2022. Which images to label for few-shot medical landmark detection?

Ning Shen Feng Mu Jianan Li Shiqi Huang, Tingfa Xu. 2023. Rethinking few-shot medical segmentation: A vector quantization view.