Multi-stage Glaucoma Detection Using Transfer Learning: An Ablation Study

Natasha Saki Ong
natashao@stanford.edu

Kyung-Tae Kim
kkim801@stanford.edu

Abstract

The prevalence of restricted datasets in medical settings form a bottleneck in training new networks. However, with emerging improved Deep Learning technologies in various fields including diagnosis, Convolutional Neural Networks (CNNs) could provide an insight into improving diagnoses, as well as potentially accompanying or replacing the human element. Therefore, we explore the idea of a transfer learning system to identify the viability in using larger datasets to train on a smaller, multi-class glaucoma dataset. More specifically, we perform an ablation study, looking at how pre-training on a dataset dealing with a similar vs drastically different task impacts the ultimate performance on the target task. We perform our experiments using a Kaggle Diabetic Retinopathy (DR) Dataset [1] and MINC-2500 Dataset [4].

We find that pre-training in general improves overall performance of accuracy and that pre-training on the DR dataset (similar task) especially improves the detection of early stages of glaucoma.

1. Introduction

Deep learning algorithms, especially convolutional networks (CNNs), have increasingly become popular methods for analysing medical images and, consequently, diagnosing medical conditions. In fact, at times, the use of CNNs has achieved expert-level performances in many distinct medical fields including radiology[15], opthamology [9], and dermatology [6]. With these advancements, many researchers have taken greater interest in classifying and diagnosing medical conditions. However, due to privacy issues, differing equipment for collecting data, extensive time required to annotate datasets, and possible unavailability of multiple qualified experts, medical imaging tasks struggles from a dataset-size and availability issue [16]. Therefore, transfer learning becomes an effective tool to rectify this issue.

In the spectrum of classification problems in the medical field, effective diagnoses are more useful when early diagnoses can help greatly reduce, and even partially halt the onset of the condition. One such condition is glaucoma.

Glaucoma is a group of eye conditions that progressively damages the optic nerve leading to various optic disc and cup changes. Moreover, glaucoma is a leading cause of blindness [18]. However, early detection and treatment can slow progression of the disease and prevent vision loss, making early diagnosis integral [8].

The input to our networks are two networks in parallel: A CNN network using

1.1. Problem

Due to privacy issues and extensive resources required to annotate datasets, medical imaging tasks often struggle from a lack of large public datasets on which to train a model [16]. Additionally, due to variance in the output format of medical devices, there exist variance in datasets of the same target problem. Therefore, the effectiveness of transfer learning in order to augment the training in the scope of a limited training dataset needs to be evaluated, in order to provide a baseline for future methodologies.

Glaucoma is a popular problem for evaluating classification problems as well as past experiments on the effectiveness of transfer learning [14]. The problem then presents itself in two parts:

- The efficacy of training on a dataset to augment the limited nature of medical datasets
- The comparative efficacy of different datasets. i.e. The comparative effectiveness of using a dataset of retinal images, versus a similarily sized dataset intended to target a more generalized problem

Therefore, we perform a kind of ablation study to examine to efficacy of the datasets used to pre-train models in our target task of performing multi-stage glaucoma detection.

2. Related Work

2.1. CNNs for classification of Medical Images

CNNs are a popular choice for various classification tasks of medical images, as medical images form datasets

that present features that are classifiable based on image-based features. Therefore, CNNs with applications in the medical domain has become a popular method for tackling the task. However, data collection remains a persistent issue, which has lead to various methods for improving performance on limited data sets, including data augmentation, GANs, aas well as transfer learning. [20]. Other related works in classification problems, especially in Glaucoma detection has attempted to use a CNN-RNN network in order to create a more robust network to target the above issues, by extracting spatial and temporal features. [7]. Other applications of CNN in different medical settings like radiology attempt to define a more complex computer architecture in order to capture a more comprehensive set of features[21].

2.2. Transfer Learning on Medical Imaging

Many existing works explore the comparative effectiveness of conducting transfer learning for the purpose of classifying medical images; Imagenet is a popular dataset for the base training dataset, as it provides a wide enough dataset, while being easily standardizable. This allows it to be suitable for purposes of early transfer learning works. However, more recent works explore the possibilities of different datsets for baseline training, as well as different architectures.[19]
Additionally, early works noted that issues in transfer learning may derive from a growing gap between the task for the source data and the target data(ie. Using Imagenet vs Diabet Retinopathy for the purpose of transfer learning for glaucoma), indicating that highly generalizable datasets like Imagenet may be less optimal than domain specific datasets, such as medical imaging datasets [22]. However, overcoming the specific set of issue presented by training on a generalizable model like Imagenet is an area under exploration. A proposed solution includes using a widely available set of unlabeled medical images for feature extraction(a dataset common in many medical settings), and applying transfer learning to train a small labeled dataset. [2]. A simple comparison on the relative accuracies achieved when training from Imagenet vs a random initialization yielded poor results[14]. On the converse, a modality-bridge transfer learning method, where an intermediate dataset obtained using the same acquisition modality yielded significant results compared to random initializations and datasets of a different acquisition modality.[11].
[17] explores the effects of using different architectures for classification(CifarNet, AlexNet, GoogLeNet) on random initialization or pre-trained models. The results indicate that GoogLeNet was able to obtain significant results compared to previous literature, although it may partly be in part due to the datasets being used in this particualr study being composed of images of a larger FOV, lending itself to extraction of more spatial features.

2.3. Ablation Studies on Transfer Learning

There have been a few studies that have performed ablation studies for the context of Transfer Learning, though usually for different applications. On study performs ablation studies relating to domain adaptation, low-shot learning, size of pre-training corpus, and parameter updating methods for Learning-Based Sentiment Analysis in Japanese [3]. For instance, for the domain adaptaion ablation study, which is most similar to the ablation study we conduct, they found that fine-tuning ULMFiT [10] improved the performance on all datasets while ELMo [13] and BERT [5] showed varied results. Meanwhile, another study focuses on transfer learning for person re-identification, and perform an ablation study evaluating the contribution of a proposed two-stepped fine-tuning with proxy classifier learning strategy against standard one-stepped fine-tuning strategy used by most previous deep learning. They find that their proposed two-stepped fine-tuning is more effective for knowledge transfer in deep re-identification applications. Both papers look at drastically different applications from our target medical application, suggesting that the conclusions found in the study may not generalize to our medical application. The sentiment analysis paper further differs in the types of models tested (language models vs CNNs) as it is more NLP/NLU-focused.

3. Dataset

For our pre-training, we have two different datasets: MINC-2500 Dataset [4] and Kaggle Diabetic Retinopathy (DR) dataset [1]. For our fine-tuning, we use a Glaucoma dataset [12] released by Kim's Eye Hospital in Seoul, Korea.

3.1. Kim's Eye Hospital Glaucoma Dataset

This dataset contains 1,544 RGB fundus photographs that have been pre-processed by scaling to have fixed size of 800 pixels and then cropping at the region of Optic nerve to a final size of 240x240 pixels. As with the DR dataset, these images are labelled by a clinician on the presence and progression of glaucoma using the following scale and distribution:

- early glaucoma 289 images
- advanced glaucoma 467 images
- normal control 788 images

Figure 1 includes a sample image for each class. Since the dataset only includes the images separated by label, we split the dataset into training, validation, and test set using an 80:10:10 ratio, respectively, and ensured equal distributions as shown in Figure 2.

Figure 1: Glaucoma Dataset Sample Images

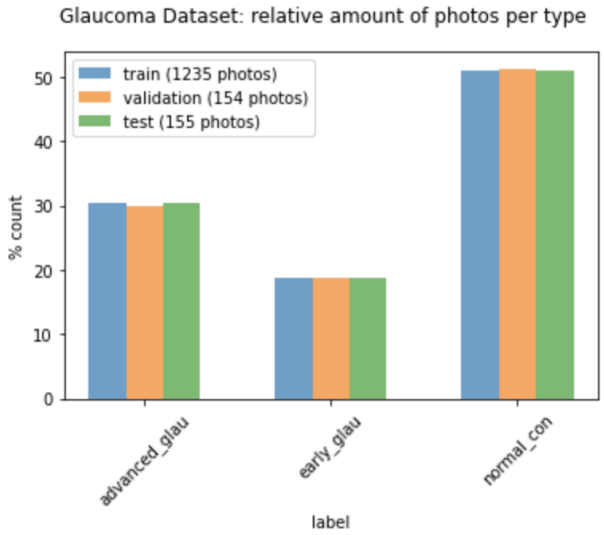

Figure 2: Glaucoma Dataset Label Distribution

3.2. Kaggle Diabetic Retinopathy Dataset

This dataset contains 35,108 high-resolution labelled retina images, where a clinician has rated the presence of diabetic retinopathy in each image on a scale of 0 to 4. The scale and distribution of labels is as follows:

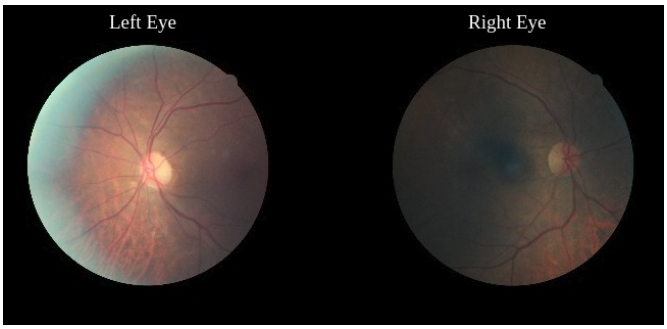

Figure 3: Kaggle Diabetic Retinopathy Dataset

- 0 — No DR 25802 images
- 1 — Mild 2438 images
- 2 — Moderate 5288 images
- 3 — Severe 872 images
- 4 — Proliferative 708 images

Each image in the original dataset is of resolution 1024x1024. To match the glaucoma dataset, we pre-process each image to be 240x240.

3.3. MINC-2500 Dataset

Materials in Context Database (MINC) [4] is a large scale material in the wild dataset. In this work, we use a publicly available subset of the original called MINC-2500 (section 5.4 of original MINC paper above). MINC-2500 is a patch classification dataset with 2500 samples per category and 23 categories in total. Each image in the dataset is of size 362 x 362 and each category is sampled evenly. To match out final/target glaucoma dataset, we further pre-process the images to 240x240 to feed into our model.

4. Methods

We perform an ablation study on transfer learning, specifically investigating the effectiveness of different datasets used for pre-training under the application of multi-stage glaucoma detection.

4.1. Model

We create a relatively simple convolutional neural network (CNN) for our model on which to run the baseline as well as perform the other experiments. The model consists of four convolutional layers, followed by a fully connected layer. Each convolutional layer is followed by a ReLU non-linearity, max-pooling(2), and dropout (0.5). Moreover, since we use ReLU non-linearity, we initialize our weights using He initialization. The He initialization draws samples from a truncated normal distribution centered on $\mu = 0$ with standard deviation of $\sqrt{\frac{2}{fan_{in}}}$ where fan_{in} is the number of input units in the weight tensor. For backpropagation, we use a cross entropy loss function since we are dealing with multi-class classification and its equation is given by:

$$\text{loss}(x, class) = -\log\left(\frac{\exp(x[class])}{\sum_j \exp(x[j])}\right) = -x[class] + \log\left(\sum_j \exp(x[j])\right)$$

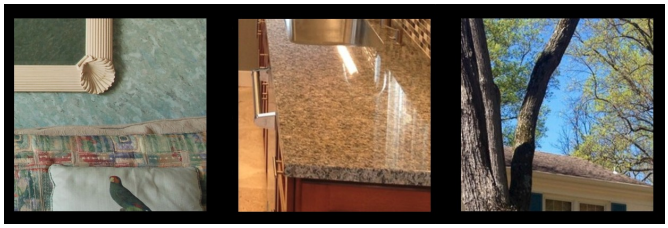

Figure 4: Minc-2500 dataset

Moreover, we use Adam optimiser with $\beta_1 = 0.9, \beta_2 = 0.999$, and learning rate of $1e-4$ since these are Pytorch's default parameter for Adam Optimizer and seemed reasonable for a model that we were going to keep constant across experiments. Figure 5 shows the architecture of our model.

4.2. Metrics

To evaluate our models, we look at overall accuracy and loss curves, as well as recall and precision for each class. We investigate the performance for each class to get a better understanding of what our model does well with and struggles with as the overall accuracy collapses these factors into a single number. For a given class c, its recall can be calculated as:

$$Recall_c = \frac{TP_c}{True_c}$$

Here, TP_c is the number of true positive (ie correctly classified) labels for class C, and $True_c$ is the total number of actual (true) labels for class c. More simply put, this is the number of correctly predicted divided by number actual/true labels for the class. Similarly, the precision can be calculated as:

$$Precision_c = \frac{TP_c}{Pred_c}$$

Here, TP_c is as defined above and $Pred_c$ is the number of samples predicted to be class c. More simply put, precision is the ratio between the True Positives and all the Positives.

We use these metrics to get a better sense of performance, and also because in medical applications, recall is very important (likely more so than precision) since we want to prevent having False-negatives (ex. concluding someone has no glaucoma when they do).

4.3. Experiments

We conduct two experiments using the Diabetic Retinopathy and the MINC-2500 datasets.

```
Sequential(
  (0): Conv2d(3, 16, kernel_size=(7, 7), stride=(1, 1), padding=(2, 2))
  (1): ReLU()
  (2): MaxPool2d(kernel_size=2, stride=2, padding=0, dilation=1, ceil_mode=False)
  (3): Dropout(p=0.5, inplace=False)
  (4): Conv2d(16, 32, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
  (5): ReLU()
  (6): MaxPool2d(kernel_size=2, stride=2, padding=0, dilation=1, ceil_mode=False)
  (7): Dropout(p=0.5, inplace=False)
  (8): Conv2d(32, 64, kernel_size=(5, 5), stride=(1, 1), padding=(2, 2))
  (9): ReLU()
  (10): MaxPool2d(kernel_size=2, stride=2, padding=0, dilation=1, ceil_mode=False)
  (11): Dropout(p=0.5, inplace=False)
  (12): Conv2d(64, 32, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
  (13): ReLU()
  (14): MaxPool2d(kernel_size=2, stride=2, padding=0, dilation=1, ceil_mode=False)
  (15): Dropout(p=0.5, inplace=False)
  (16): Flatten()
  (17): Linear(in_features=6272, out_features=3, bias=True)
)
```

Figure 5: Our Model Architecture

4.3.1 Pre-training on the Diabetic Retinopathy Dataset

The DR dataset more closely resembles the final task than the MINC-2500 dataset. There are 5 labels/classes for this dataset that represent the progression of DR, while our final task and dataset has 3 labels representing the progression of glaucoma. The final class numbers/labels are therefore very similar in number and magnitude. Moreover, both datasets use retinal images, specifically fundal photographs, so the model will likely learn more relevant features through pre-training on this dataset. As noticeable from the dataset section, the DR dataset is highly imbalanced, with most of the samples being from class 0, ie No DR case. This is typical of medical datasets, but to tackle this issue, we alter our loss function to accept a weights parameter. More specifically, we give greater weight to classes that are less represented, with a class's weight calculated by:

$$w_{class} = \frac{N_0}{N_{class}} = \frac{25802}{N_{class}}$$

4.3.2 Pre-training on the MINC-2500 Dataset

The MINC-2500 dataset contains images that are less similar or aligned to our final task. As mentioned above, the dataset has 23 classes consisting of 2500 each, of various textures that may be commonly found, and may be efficiently differentiable(glass, wood, metal, etc). In contrast to the Diabetic Retinopathy Dataset, the MINC-2500 dataset has a much larger number of classes, which may present another variable in evaluating the efficacy of transfer learning. Since this dataset is completely balanced, there we can simply use the original loss function aforementioned.

4.3.3 Fine-tuning on the Glaucoma Dataset

Finally, we freeze most of the layers except the final fully connected layer, whose output we change to have 4096 features. Moreover, we add a ReLU and another fully connected layer, with output of 3, since we have 3 classes – no glaucoma, early glaucoma, advanced glaucoma. We added a ReLU layer and another FC/linear layer since our dataset is much smaller and we want to prevent overfitting (ie we don't fine-tune the entire model). We train for a total of 20 epochs.

4.3.4 Ensuring Robust Generalization

To ensure robust generalization such that we can draw conclusions, we carefully consider how we conduct these experiments. Some considerations are as follows:

1. **Using the same base model**
   We use the same model on both experiments, keeping the hyperparameters and architecture the same.

This enables us to conclude that performance variances comes from the dataset itself rather than changes in the model. To use the same model without potentially facing dimension issues, we pre-processed all input images to 240 x 240. We then feed these to the model to train on.

2. **Dataset size**
   Generally, training on more data helps the model learn and perform better, assuming regularization methods are applied to reduce issues like overfitting. As such, though it is very difficult to find datasets of the exact same size/examples, we attempt to control the experiment as much as possible by using datasets for pre-training that are of similar size in magnitude.

4.4. Baseline

As a baseline, we evaluate the model's performance on the target glaucoma dataset without transfer learning, i.e. without pre-training on a either of the larger datasets previously mentioned. In particular, using the model from the above section, we directly train on our target glaucoma dataset for 60 epochs.

5. Results and Discussion

5.1. Baseline Results

To evaluate the baseline, we produce loss and accuracy curves to see performance over the epochs. We also produce a confusion matrix to evaluate the classification performance for each of the three classes. The findings are summarized in Figure 6, 7, and Figure 8. Our accuracy on the test set came out to be 75.48%. Instead of training on the glaucoma dataset for 20 epochs as we did for the other models, here we trained for 60 epochs, to provide the baseline a better chance to perform better and likely converge.

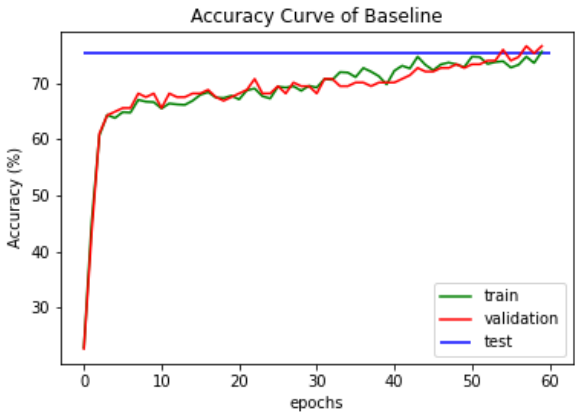

Figure 6: Training and Validation Accuracy of Baseline

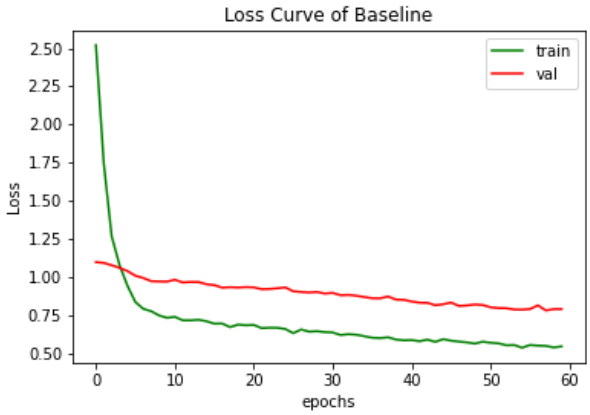

Figure 7: Training and Validation Loss of Baseline

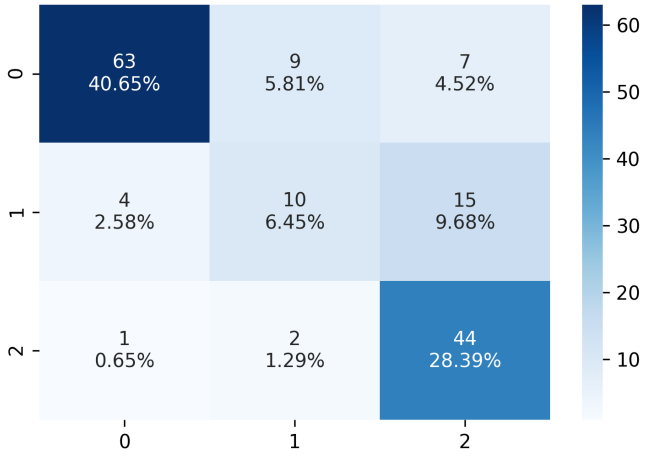

Figure 8: Confusion Matrix for Baseline
0: normal eyes, 1: early glaucoma, 2: advanced glaucoma

Thus, we treat the final accuracy of 75.48% as a kind of high/upper-bound performance of a baseline.

5.2. Model Pre-trained on DR Dataset

The model pre-trained on the DR Dataset achieved a final accuracy of around 88.4%. Figure 9 shows the accuracy and loss curves and Figure 12 includes a confusion matrix which allows for classification comparisons per class.

From the accuracy and loss curves, we notice that in fine-tuning, the model actually overfits to the training set a bit, though the validation curves still follow the training curves quite well. The overfitting may be attributed to a number of factors including hyperparameter and not enough regularization but also to the similarity of the pre-training dataset to the target dataset. Moreover, since the target glaucoma dataset is quite small, we may not need to run as many iterations (but we kept the number of epochs to be 20 so as

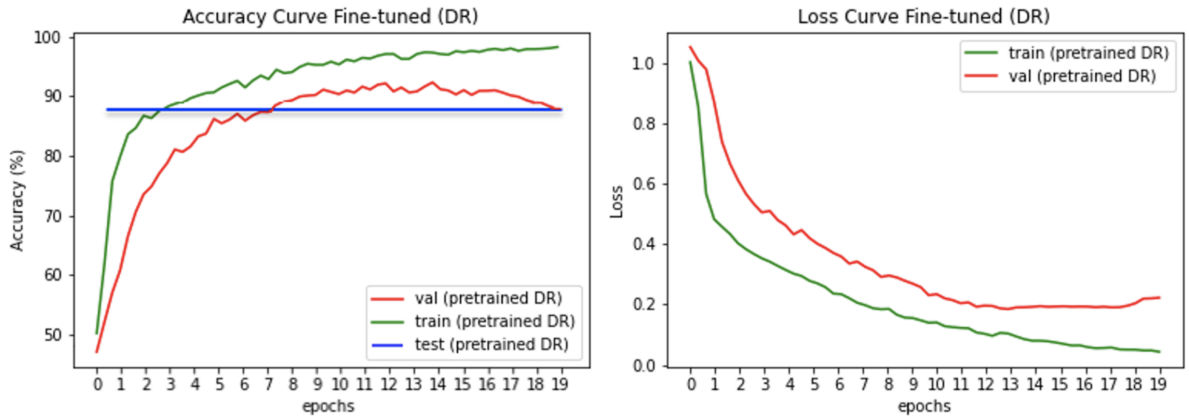

Figure 9: Accuracy and Loss Curves during Fine-tuning for Model Pre-trained on DR Dataset

to not to change too many variables between the models). Nevertheless, even with the overfitting the model outperforms the baseline by over 10%. Looking specifically at the classification for each of the labels (normal, early glaucoma, advnaced glaucoma) through the confusion matrix, we see that our model does a better job predicting the correct labels for each class. For both the no glaucoma class and the advanced glaucoma class, the model predicts the next similar label much more frequently that it does the third class (ie for advanced glaucoma, we only predict normal condition for one image vs 7 for the baseline). We see that though the model performs much better for the early glaucoma case, it is still the class that the model has most difficulty predicting correctly. We also see this pattern for the baseline.

5.3. Model Pre-trained on MINC Dataset

The model pre-trained on the MINC Dataset achieved a final accuracy of around 83.9%, better than the baseline but worse than the model pre-trained on DR. Figure 10 shows the accuracy and loss curves and Figure 13 includes class-by-class visualization through a confusion matrix.

From the accuracy and loss curves, we notice that in fine-tuning, the model actually overfits to the training set and plateus fairly quickly, which was quite surprising. Since we kept the model the same (since we wanted to directly see the effects of pre-training on specific datasets), we did not alter or add extra layers to prevent overfitting (ex. adding dropout layers or adding penalization to loss function). We also did not touch the learning rate for the same reasons above, which may alter the shape of the curve, rate of convergence, etc. Moreover, since the target glaucoma dataset is quite small, we may not need to run as many iterations (but kept the number of epochs to be 20 for the same reasons as above).

Still, as with the model pre-trained on the DR dataset, we see this model outperforms the baseline, though not by as big of a margin. Looking specifically at the classification for each of the labels (normal, early glaucoma, advnaced glaucoma) through the confusion matrix, we see that our model does better in correctly predicting the normal eyes class and the early glaucoma class. There was a slight decrease in accuracy predicting the advanced glaucoma class (by one label). As with the other models, this model also has the most difficulty predicting the early glaucoma case, though this makes sense since it is the crux of the problem (glaucoma struggles from early detection though it has a great need for early detection to slow the progression of the condition down significantly).

5.4. Qualitative Results

Here, we provide a sample of wrongly predicted test images, with both the correct label and predicted label attached. These images are provided in Figure 11. The bottom row corresponds to some wrongly classified images by the model pre-trained on the MINC Dataset, while the top row corresponds to that of the model pre-trained on the DR dataset. Due to high pressure, Glaucomic eyes experience damages to the optic nerve and poor vascular supply to the fovea. However, as we do not have expertise in this field, ie we have untrained eyes, it is hard for us to concretely conclude what the model fails to capture, such as occlusions, fine details/features etc that may be more obvious with more common object classification. It is hard to even intuitively know what the correct label is even looking at normal and advanced glaucoma eyes.

Nevertheless, one potential explanation for misclassified images can also be wrongly assigned labels. This can be an issue with medical datasets, where different clinicians would disagree on the label to assign to a particular image. We were unfortunately unable to find detailed information in the process of labelling this dataset (ex. how many clinicians agreed upon the label, confidence of la-

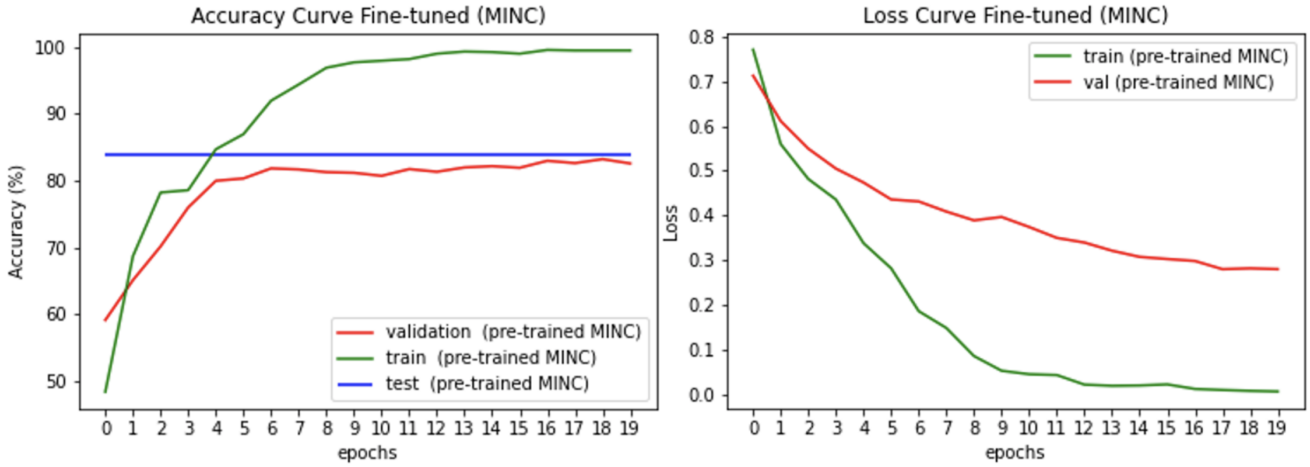

Figure 10: Accuracy and Loss Curves during Fine-tuning for Model Pre-trained on MINC Dataset

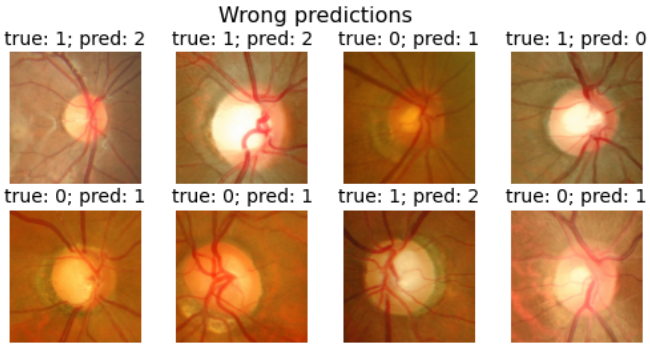

Figure 11: Wrongly Classified Test Images. Bottom Row: Model Pre-trained on DR Dataset ; Top Row: Model Pre-trained on MINC Dataset

belling, etc).

5.5. Overall Results and Discussion

To more closely compare the results of each of the models, we compile the accuracy and calculate recall, precision, and F1-score for each class, ie no glaucoma, early glaucoma, and late glaucoma. The results can be seen in Figure 14.

Simply looking at the overall accuracy, we see that pretraining on a larger dataset helps as both the accuracy for DR ($\approx 88.4\%$) and MINC ($\approx 83.9\%$) are higher than that of the baseline ($\approx 75.5\%$). Looking more closely at the precision and recall for each class, we see that the model pre-trained on DR dataset performs better than the other models for all classes. For all 3 models, we notice that the early glaucoma is the hardest to classify, given the significantly lower scores. However, intuitively, this makes sense given the problem since the early glaucoma case falls in between the normal and late glaucoma conditions. Moreover, in practice, detecting glaucoma early is a difficult task even for ophthalmologists since the progression of the disease is mild, and there is a big push to try to detect glaucoma in its earlier stages (one of the incentives for using the glaucoma dataset that we chose).

Nevertheless, for the early glaucoma class, we see drastically higher recall and precision values for the model pretrained on the DR dataset while the baseline model and model pretrained on MINC have tradeoffs between precision and recall. The significant improvement in recall and precision from the baseline in classifying early glaucoma can likely be attributed to the fact that in pre-training on the DR dataset, the model was able to learn many important underlying features that separate different degrees of eye conditions (the DR dataset had the model distinguish between 5 degrees of Diabetic Retinopathy).

For the advanced glaucoma class, we actually notice that the recall for the baseline is quite high, higher than that of the model trained on MINC. But, we see the benefits of pretraining through the significantly increased precision, ie over 15% increase. At the same time, it is important to know that since the dataset is quite small, the test set is even smaller, so a few misclassified images can drastically impact the percentages reported. Lastly, for the no glaucoma or control case, we see that the baseline has marginally higher precision than the model pre-trained on MINC but drastically lower recall than either of the pre-trained models.

6. Conclusion and Future Work

Ultimately, looking at the overall accuracy, DR dataset performed best, followed by the MINC dataset and our baseline. Moreover, we see the general benefits of pre-

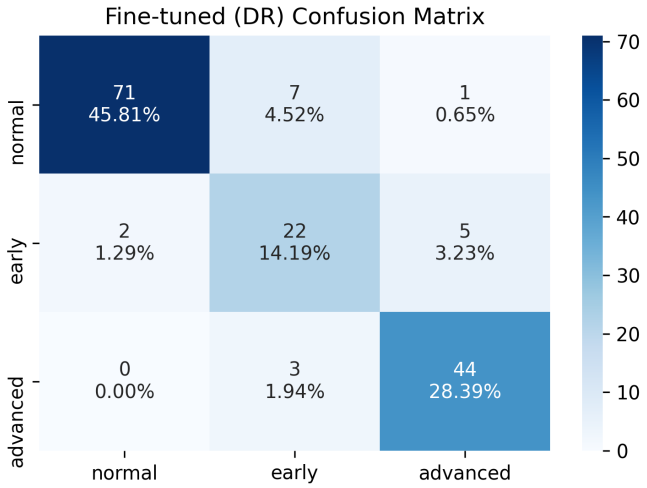

Figure 12: Confusion Matrix for Model Pre-trained on DR Dataset

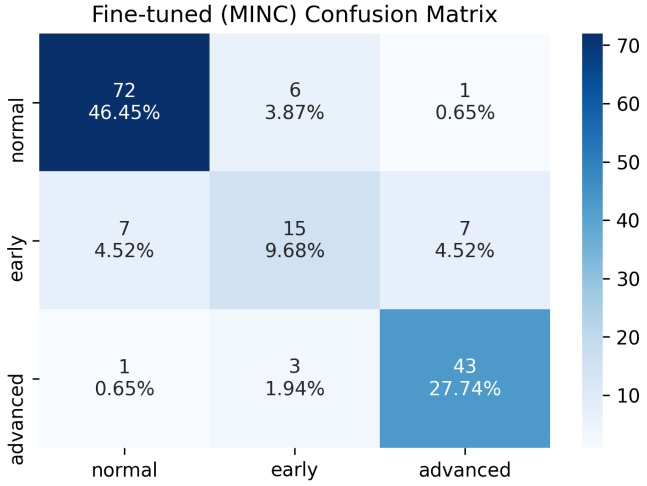

Figure 13: Confusion Matrix for Model Pre-trained on MINC Dataset

Class/Model	Baseline	Pre-trained on DR	Pre-trained on MINC
No Glaucoma (normal eyes)	Recall: 79.7%	Recall: 91.0%	Recall: 91.1%
	Precision: 92.6%	Precision: 97.3%	Precision: 90%
Early Glaucoma	Recall: 52.6%	Recall: 75.9%	Recall: 51.7%
	Precision: 47.6%	Precision: 68.8%	Precision: 62.5%
Advanced Glaucoma	Recall: 93.6%	Recall: 93.6%	Recall: 91.5%
	Precision: 66.6%	Precision: 88.0%	Precision: 84.3%

Figure 14: Comparison of recall and precision per class for each of the models

training through increases in overall accuracy and significant increases in certain class's recall and precision values (in the case of MINC). Comparing the two pre-training models, we can identify clear benefits to pre-training on a similar task (ie other eye condition problems vs general object image classification), especially in the classification of early glaucoma.

Due to time constraints, we were unable to explore different data augmentation methods for both the glaucoma dataset (small) and DR dataset (very imbalanced). To deal with the imbalanced DR dataset, we altered the cross entropy loss function to take in a weight parameter, which assigned each class with a weight. In the future, we might consider other data augmentation methods like taking random crops, rotating the image, etc to increase the number of samples we have as well as the robustness of our models. At the same time, since we are dealing with medical images, we would need to be careful to ensure that the data augmentation methods do not alter the integrity of the original image itself. In terms of dealing with the imbalanced dataset, oversampling or undersampling (in addition to data augmentation) could be worth exploring to investigate whether there are any considerable differences in performance.

Moreover, in future works, we would like to experiment more with fine-tuning our model, experimenting with aspects such as how many layers to fine-tune as tuning the hyperparameters so as to further increase accuracy and also potentially reduce overfitting, which is evident through some of the loss/accuracy curves.

Furthermore, it would be useful to create a separate model just trained on the glaucoma dataset. We would evaluate this model in the same way as the 3 models investigated in this paper, ie overall accuracy and precision, recall, and F1-score by class. This could serve as another baseline that would help investigate the necessity of transfer learning in general. In other words, if this model performs very well, there may not be as much of a benefit to have transfer learning, which may require more time to pre-train and fine-tune. Similar in vein, we could explore how or whether the findings of this paper extend to other datasets or transfer learning applications.

7. Contributions and Acknowledgements

NO worked on pre-training models on the corresponding datasets, JK worked on pre-processing the dataset, NO and JK worked on fine-tuning the models, and NO and JK contributed to the manuscript. The authors would like to thank the CS231N course staff for instruction during the quarter and would like to especially thank Chris for providing feedback and guiding us through this project.

References

[1] Kaggle diabetic retinopathy detection training dataset (drd). 1, 2

[2] L. Alzubaidi, M. Al-Amidie, A. Al-Asadi, A. J. Humaidi, O. Al-Shamma, M. A. Fadhel, J. Zhang, J. Santamaría, and

Y. Duan. Novel transfer learning approach for medical imaging with limited labeled data. *Cancers*, 13(7), 2021. 2

[3] E. Bataa and J. Wu. An investigation of transfer learning-based sentiment analysis in japanese. *arXiv preprint arXiv:1905.09642*, 2019. 2

[4] S. Bell, P. Upchurch, N. Snavely, and K. Bala. Material recognition in the wild with the materials in context database. *Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2, 3

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[6] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017. 1

[7] S. Gheisari, S. Shariflou, J. Phu, P. J. Kennedy, A. Agar, M. Kalloniatis, and S. M. Golzan. A combined convolutional and recurrent neural network for enhanced glaucoma detection. *Scientific Reports*, 11(1), 2021. 2

[8] M. H. Goldbaum, P. A. Sample, K. Chan, J. Williams, T.-W. Lee, E. Blumenthal, C. A. Girkin, L. M. Zangwill, C. Bowd, T. Sejnowski, et al. Comparing machine learning classifiers for diagnosing glaucoma from standard automated perimetry. *Investigative ophthalmology & visual science*, 43(1):162–169, 2002. 1

[9] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016. 1

[10] J. Howard and S. Ruder. Universal language model fine-tuning for text classification, 2018. 2

[11] H. G. Kim, Y. Choi, and Y. M. Ro. Modality-bridge transfer learning for medical image classification. *CoRR*, abs/1708.03111, 2017. 2

[12] U. Kim. Machine learn for glaucoma, 2018. 2

[13] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018. 2

[14] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. Transfusion: Understanding transfer learning for medical imaging, 2019. 1, 2

[15] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017. 1

[16] M. I. Razzak, S. Naz, and A. Zaib. Deep learning for medical image processing: Overview, challenges and the future. *Classification in BioApps*, pages 323–350, 2018. 1

[17] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, 2016. 2

[18] J. Susanna, Remo, C. G. De Moraes, G. A. Cioffi, and R. Ritch. Why Do People (Still) Go Blind from Glaucoma? *Translational Vision Science Technology*, 4(2):1–1, 03 2015. 1

[19] H. Xie, H. Shan, W. Cong, X. Zhang, S. Liu, R. Ning, and G. Wang. Dual network architecture for few-view ct - trained on imagenet data and transferred for medical imaging. 07 2019. 2

[20] S. S. Yadav and S. M. Jadhav. Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big Data*, 6(1), 2019. 2

[21] R. Yamashita, M. Nishio, R. K. Gian Do, and K. Togashi. Convolutional neural networks: an overview and application in radiology. *Insights IMaging*, page 611–629, Aug 2018. 2

[22] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? *CoRR*, abs/1411.1792, 2014. 2