Performance and Memory when Diagnosing Retinal Diseases using Classification

Rose Rodrigues - rodr0875@umn.edu

University of Minnesota - Twin Cities

Abstract

Classifications of images are a very important problem in the machine learning field, and a very common sub-problem is analyzing labeled medical images and grouping them based on their diagnoses. There are a multitude of ways to do that and each of them has their own advantages and disadvantages. That is, the more advanced and flexible a model you desire, the longer it takes to train and run them, in addition to the memory and computational effort. As such, an analysis was made of three separate models - a KNN PCA-based model, a custom-made DCNN, and a pre-made model, *MobileNetV3* - and their complexity, accuracy, training time, and predicting time. Each obtained accuracies of 72.6%, 86.4%, and 89.4%, while each model took 1 second, 2 hours and 20 minutes to train. Pre-made models are thus more appropriate for the quick training and evaluation sought when prioritizing for accuracy, while KNNs are more recommended for prioritizing speed and efficiency.

Contents

1 Introduction

The diseases of cataracts, diabetic retinopathy, and glaucoma are diseases that affect close to 190 million people around the world [1][2][3]. Their diagnoses can prove vital for a wide variety of people, but there are moments when doctors may not notice specific details, and it is trivial to imagine that it is in a patient's best interest to obtain a proper diagnosis.

As such, an alert that comes up upon the scan is obtained to help the medical professional be more alert for that disease could be a great addition, but to help with adoption, it needs to be fast, efficient, and compatible with a large array of devices, possibly ones with low specifications. Thus, a question arises as to which technique would be better suited for that task.

1.1 Applications

One of the best ways to have this possible classification model implemented is by integrating it with already existing infrastructures to ease accessibility. For example, misdiagnoses comprise 10% of adverse effects observed in hospital visits, and one of the most impactful factors are cognitive biases, which can be slightly mitigated by the use of analytical reasoning [4].

Thus, the idea came of creating a model to help diagnose images, especially to flag the medical scan for a possible disease was thought of. However, for this to be easily implementable, factors that would be impactful are how fast and how accurately a model classifies an image.

1.2 Question

What would be the best model to choose when it comes to trying to find a balance between accuracy of predictions and performance and memory usage of specific algorithms?

Three models will be used to explore the possibilities. Firstly, a K-Nearest Neighbour model, which can be easily and quickly implemented without the use of complicated libraries or cumbersome dependencies, alongside having a quicker processing time, at a trade-off for higher accuracy. Secondly, a custom-made Deep Convolutional Neural Network, which could be propriety developed by the medical company in question to take advantage of possible hardware benefits or other factors. Finally, a literature-based model called *MobileNetV3-Large* [5], which was developed with low-performance devices in mind, trying to maximize efficiency and accuracy.

1.3 Hypothesis

The pre-made model will have the highest accuracy, but will take the longest to train, given the possible complexity compared to a custom-made DCNN. It, in question, will have an average accuracy and training time, while the KNN will have a very low accuracy of predictions.

2 Methodology

Steps to develop and train models

The three models will be compared using the testing accuracy and a confusion matrix based on the validation set. In addition, all models will be multiclass classifiers, classifying retinal images between **normal**, **cataracts**, **diabetic retinopathy**, and **glaucoma**.

2.1 Database and Preprocessing

The chosen database is **eye_diseases_ classification**, which contains around 4100 different images and scans. However, many of them have different ratios or resolutions, the latter varying from 256×256 to 512×512. The chosen method was to filter all images with a non-square ratio, and scale the remaining ones to 128×128. An exception, however, was made for the KNN, whose images were scaled to 256×256, since it uses up less memory than the other models, thus allowing for higher resolutions and more details. Then, the input images were separated into a train-test-validation set with a 0.7 - 0.15 - 0.15 distribution. The randomization, to preserve results across different runs, was determined with a seed.

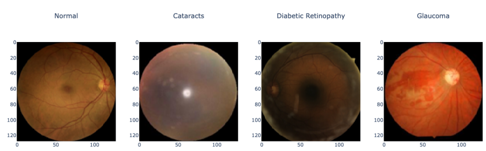

Figure 1: Randomly chosen images of a normal retina followed by ones with cataracts, diabetic retinopathy, and glaucoma, respectively.

2.2 KNN

A PCA will be done to reduce the dimensionality in the model, and the dimensions that yield a variance close to 98% will be chosen. Afterwards, the model will be tested for every K to find the one with the highest test accuracy.

2.3 DCNN + MNV3L

The custom DCNN's specific structure will be determined by tuning the number of layers and non-linearity-inducing methods, alongside the number of kernels. *MobileNetV3-Large* will be trained in the normalized dataset. For both, their training process was designed to keep track of the lowest test losses, and if there are 7 instances of increased loss, early stopping will happen, reverting to the model with the lowest test loss. Every time these increased test losses are detected, the learning rate will be halved to avoid overshooting and redundant epochs.

3 Results

Observed results from the three used models

3.1 KNN

For the K-Nearest-Neighbour model, Principal Component Analysis was used to process the images, since calculating distances of a $256 \times 256 = 65536$- dimensional vector would be too computationally intensive. After creating the analysis, a graph was plotted of the chosen factors by variance preserved. It was observed that the first 128 dimensions were sufficient, preserving 98.16% of the variance.

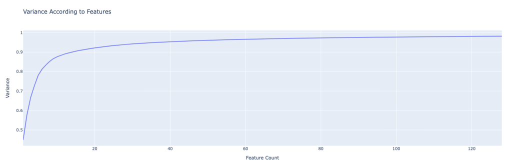

Figure 2: Graph of chosen PCA-d dimensions (x) by preserved variance (y).

Afterwards, Ks from 1 to 50 were chosen, and for each K, a model was trained, and its training and testing accuracies were plotted in a secondary graph, to confirm which K provided a higher accuracy. This led to the conclusion that the best value for K is $K = 1$.

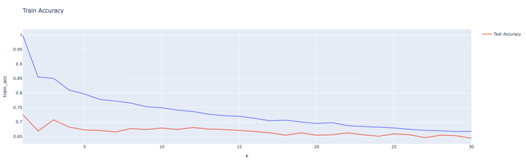

Figure 3: Graph of training (blue) and testing (red) accuracies (y) according to K (x)

It was found that on the testing dataset, the 1NN obtained a **72.6%** accuracy, and on the validation dataset, the following confusion matrix was generated:

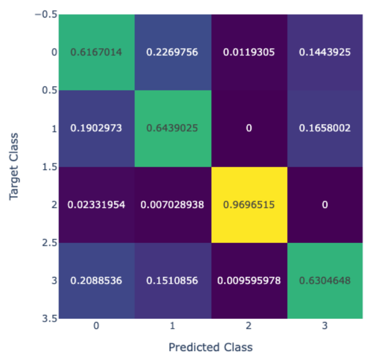

Figure 4: Confusion matrix of 1NN

The total training time was less than **1 second**.

3.2 DCNN

For the Deep Convolutional Neural Network, the following training parameters were chosen:

- **Maximum Epochs:** 50
- **Batch Size:** 35
- **Loss Function:** CrossEntropyLoss
- **Optimizer:** Adam

For the layer structure and activation functions, the chosen tuning methods were trial and error to see which structures yielded the highest testing accuracies. In the end, the following structure was settled upon:

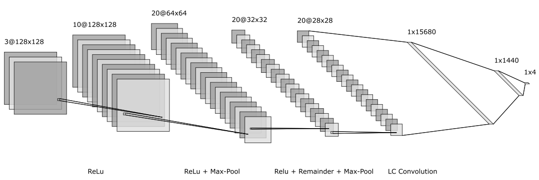

Figure 5: Graphic representation of DCNN.

The DCNN includes 6 hidden layers, each with the following transformations and activation functions:

1. 10 5×5 kernels with 2 pixels of zero-padding, followed by a ReLu
2. 20 3×3 kernels with 1 pixels of zero-padding, followed by a ReLu and a Max-Pool
3. 20 3×3 kernels with 1 pixels of zero-padding, followed by a ReLu, a residual sum from the previous layer, and a Max-Pool
4. 20 5×5 kernels on a locally-connected convolution
5. Flat linearization
6. Feed-forward layer of 15680 neurons to 1440

This network has a total number of 30,450,794 parameters. Despite the maximum number of epochs being set at 50, only 29 epochs ran before early stopping went into action, returning to a model with 1.6194 testing loss at epoch 18. It resulted in a model with an **86.4%** testing accuracy, and the following confusion matrix based on the validation dataset:

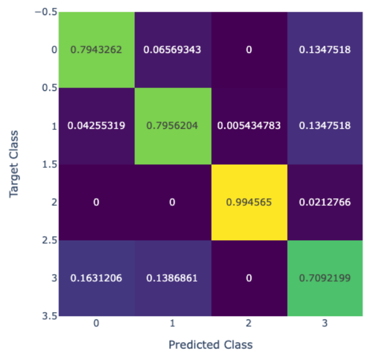

Figure 6: Confusion matrix of DCNN

The training time was around **2 hours**.

3.3 MNV3L

For *MobileNetV3-Large*, the same training parameters were used, with a final layer being added for the output layer. However, the images were normalized with ImageNet in mind, reshaping each channel from $[0, 255] \to [0, 1]$, and for each image $\iota = [r, g, b]$, $\iota_{\text{normalized}} = [\frac{r-\bar{r}}{\sigma_r}, \frac{g-\bar{g}}{\sigma_g}, \frac{b-\bar{b}}{\sigma_b}]$, with the means and standard deviations coming from the ImageNet dataset. It has a total of 4,207,156 parameters.

After the training, the resulting model had an 89.4% accuracy on the testing set, and the following confusion matrix based on the validation dataset:

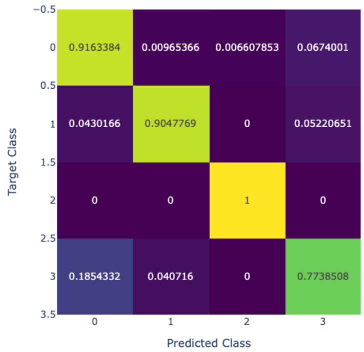

Figure 7: Confusion matrix of MNV3

The training time was around **20 minutes**.

4 Discussion

Discussing obtained results

It's interesting to note that the images for diabetic retinopathy were always consistently classified with extremely high accuracy.

On a separate note, it was impressive to see that a simple 1-NN was able to achieve 72.6% accuracy in less than a second of training.

The custom-made DCNN did fare remarkably well, but its training time was the longest. It also has a large number of parameters, especially when compared with MNV3L's.

Possible ethical limitations of these models would probably be ways to obtain more data points while not violating patient privacy. In addition, it's important to stress that these models should in no way substitute actual doctors, but rather serve as another tool to help them.

A possible limitation of this dataset specifically is the different number of scanning techniques. According to the dataset description: "These images are collected from various [sources] like IDRiD, Oculur recognition, HRF etc." [6]. This could impact the accuracy of certain models diagnosing specific diseases, and could be the reason why diabetic retinopathy is consistently classified correctly.

5 Conclusions

All three models were to correctly classify retinas more often than not, especially considering the lowest random-guessing accuracy would be 25%.

As mentioned in the previous section, across the entire spectrum of possible models, the most certain diagnosis is diabetic retinopathy. One of the possible reasons could be a peculiarity when it comes to its images, either because of how it affects the retina, or because of the manner in which its scans specifically where chosen.

Despite the velocity of 1NN, it does misclassify images a fair number of times, despite being correct about around 18 images for every 25. The custom-made DCNN was close to being the best, but its training time and almost $8\times$ more parameters than MNV3L makes it so that it would not be the priority choice when adjusting for memory, performance, and accuracy.

In the end, MNV3L was able to achieve the highest accuracy while still maintaining a great training speed. This tracks, since it was made with memory restrictions and lower-capacity phone CPUs in mind.

These results were very different from the hypothesized results, as the KNN had a much higher accuracy than what was previously thought. In addition, the pre-made model had a lower training time than the custom DCNN. Also the disparity of training times opens up the possibility for an unthought of conclusion.

If a classifier to help facilitate the diagnoses of these diseases without interfering with the amount of time a medical professional would spend is desired, *two* choices can be made: if you prioritize performance and memory above all, a KNN would be sufficient with a high accuracy still; but if you prioritize accuracy above all, a model like *MobileNetV3-Large* could be worthwhile to implement.

References

[1] Z. L. Teo, Y.-C. Tham, M. C. Yan Yu, *et al.*, "Global prevalence of diabetic retinopathy and projection of burden through 2045: Systematic review and meta-analysis," *Ophthalmology*, vol. 128, May 2021. DOI: 10.1016/j.ophtha.2021.04.027.

[2] H. A. Quigley and A. T. Broman, "The number of people with glaucoma worldwide in 2010 and 2020," *British Journal of Ophthalmology*, vol. 90, pp. 262–267, Mar. 2006. DOI: 10.1136/bjo.2005.081224.

[3] M. Li, W. Jia, J. Song, *et al.*, "Global prevalence and years lived with disability (ylds) of cataract in 204 countries and territories: Findings from the global burden of disease study 2021," *Eye*, Mar. 2025. DOI: 10.1038/s41433-025-03743-z. [Online]. Available: https://www.nature.com/articles/s41433-025-03743-z (visited on 03/14/2025).

[4] G. Neale, H. Hogan, and N. Sevdalis, "Misdiagnosis: Analysis based on case record review with proposals aimed to improve diagnostic processes," *Clinical Medicine*, vol. 11, pp. 317–321, Aug. 2011. DOI: 10.7861/clinmedicine.11-4-317. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5873736/.

[5] *Searching for MobileNetV3*, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Nov. 2019, pp. 1314–1324. DOI: 10.48550/arXiv.1905.02244. [Online]. Available: https://arxiv.org/abs/1905.02244v5.

[6] G. Venkat Doddi. "Eye_diseases_classification," www.kaggle.com. (May 2022), [Online]. Available: https://www.kaggle.com/datasets/gunavenkatdoddi/eye-diseases-classification/data (visited on 05/08/2025).