COMP6208 Advanced Machine Learning

Group Coursework - Diabetic Retinopathy Detection

Michelle Abela, mna1u18, 30557313
Joseph Early, je5g15, 27309061
Sandeep Mistry, sbm1g18, 30628768
Aran Smith, abs2n18, 30538289

September 16, 2019

Abstract

Diabetic retinopathy (DR) is a medical condition that causes blindness due to diabetes. We present a study into DR detection using various machine learning techniques, including convolutional neural networks (CNNs), support vector machines (SVMs) and random forest (RF). We employed the state of the art, publicly available CNN architectures AlexNet, ResNet-18 and VGG-19 with batch normalisation. We re-balanced the dataset to avoid over-fitting and implemented different data augmentation techniques. The best classification method we found placed us in the top ten percent of teams in the online Kaggle competition.

1 Introduction

Diabetic retinopathy (DR) is a medical condition that causes blindness due to diabetes. Owing to the large number of cases of diabetes (422 million worldwide in 2014 [1]), DR is a leading cause of blindness, particularly in working adults, where it is the most frequent cause of blindness [2]. The progression of DR is heavily dependent on the duration of diabetes, however it has few symptoms until it reaches a more progressed state at which is starts affecting vision, and once it becomes more severe it is difficult to reverse any loss of vision [3]. One effective detection technique is to examine fundus photographs of retinas, allowing classification of the severity of DR due to lesions on the retina [4]. However, this is currently a slow process, requiring analysis of each retinal scan by specially trained clinicians. An automated method for DR screening that is effective at detecting DR (including in its early stages when vision loss can still be halted) is highly desirable in streamlining the process and taking the growing burden of DR away from clinicians. To this end, a Kaggle competition[1] was launched in 2015 to drive development of automated DR detection systems. This report summarises our work in attempting to classify the severity of DR in a large selection of fundus images using a range of techniques.

2 Dataset and preprocessing

The Kaggle competition provides a training and test set consisting of 35126 and 53576 images respectively, where each image has been classified into one of five categories, from zero (no DR) to

[1] `https://www.kaggle.com/c/diabetic-retinopathy-detection/`

Figure 1: A selection of retinal scans demonstrating the noise and variation in the images. The scans increase in DR severity from left to right, from class zero on the far left to the class four on the far right.

Class	DR Severity	Number of Training Images	Number of Test Images
0	None	25810	39533
1	Mild	2443	3762
2	Moderate	5292	7861
3	Severe	873	1214
4	Proliferative	708	1206

Table 1: A breakdown of the train and test DR datasets.

four (proliferative DR). The images come in pairs of left and right per patient, and while each pair should have a similar level of DR, there is no guarantee that both images fall into exactly the same category. The images are subject to noise, which originates from the use of different retinal scanning devices under different conditions. The noise includes orientation of the retina, colour differences, and positional differences, meaning the detection methods developed must be very robust. Figure 1 shows a selection of images from the training set with varying levels of DR.

The datasets provided are unbalanced; they have a vastly different number of images for each class. As can be seen in Table 1, there are far more images in class zero, meaning it possible to get a train and test accuracy of $\sim$0.73 by just classifying every image as the majority class, class zero. A straightforward way of dealing with an unbalanced dataset is to re-sample the classes such that there are an equal number of images in each class, i.e. by oversampling the minority classes and undersampling the majority classes [5]. In this work, the oversampling was done at random with replacement and the undersampling was done at random without replacement, such that there were 7000 images for each class, giving a new training set of 35000 images, which is approximately the same as the number of images in the original training set. This means the computational cost of training is the same as with the original dataset, and it is assumed that undersampling class zero does not lead to a loss of information as the DR scans for class zero do not have features to detect, i.e. the distinguishing DR features such as lesions only appear in the other classes.

Due to the re-sampling, the new training set now contains duplicate images for the minority classes. In the worse case (class four), each image is duplicated on average almost 10 times, as the class is scaled from 708 images to 7000. This can artificially inflate learning, as the model effectively gets a 10x boost for correctly classifying a class four image. Therefore, to negate the duplication effect, noise is added to each of the images in the training set to ensure that there are no identical images, which is also useful in ensuring the robustness of the developed model. This was achieved by aug-

menting images before classification through random rotation and centre cropping.

We attempted a number of preprocessing approaches on the data. Our first technique involved appending the left and the right eyes next to each other into a 512x512 image so that the network would classify the highest level of DR for a patient from either of the two eyes. This complicates data augmentation, and requires a smarter network. Classifying single eyes removes the need for complicated data augmentation techniques and presents our classifiers with a simpler problem.

3 Methodology

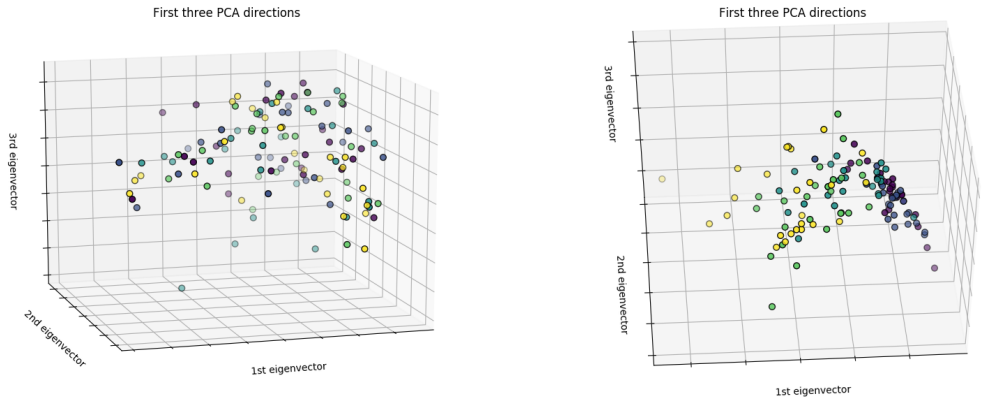
The main aim of this work was to experiment with a number of different techniques for classifying DR from fundus images. Our main focus was on Convolutional Neural Networks (CNNs) as they have shown strong performance in image related tasks, including medical imaging domains [6, 7]. Custom CNN architectures were developed based on the literature and trained from scratch, but pre-trained CNNs were also used with minor alterations for our work. The deep CNNs were used as feature extractors, either with fully connected classifiers on the end or using alternative classification techniques such as Random Forest (RF) [8] and Support Vector Machines (SVMs) [9]. Transfer learning was used with VGG-19-BN [10], ResNet-18 [11], and AlexNet [6], where the final dense layer of the network was replaced with the correct number of class outputs. The pre-trained models were fine tuned in two different ways: freezing their weights and only training a final dense layer for classification, or fine tuning every network layer.

3.1 CNNs

The first architecture we implemented (RetCNN) came from Pratt et al. [12]. This included 10 convolutional layers with ReLu, BN, and Maxpool layers between each, and three dense layers with ReLu, Dropout (0.5), and Softmax on the final layer. The idea behind a network with many CNN layers is that it can learn higher level features that might hint at a certain level of DR: lower levels learn basic lines and edges, while higher levels can respond well to vein patterns and exudates that can indicate DR. The dense layers are then able to discriminate the level of DR based on the features produced by the convolutional layers.

The problem with training a custom architecture from scratch is that is has to learn all of the low level feature representations such as lines and edges, which can take a significant amount of time. To negate this, it is possible to use networks that have already been pre-trained against another large dataset (not necessarily against similar images). By removing the final classification layer of the pre-trained network and replacing it with a fresh dense layer that has the correct number of outputs, it is then possible to use the pre-trained network for DR detection. The difference here is that the pre-trained network already responds well to certain patterns in an image, meaning it may be able to provide a set of feature vectors that are easily separable, either linearly or non-linearly.

Feature vectors produced by these pre-trained networks are not necessarily separable, and the only way found to perform well in training here was to over-fit. Even a dense network capable of defining complex non-linear patterns would have a hard time not over-fitting due to entangled feature vectors (Fig. 2a). A remedy here is to allow the entire pre-trained network to be modified during training. This is known as unfreezing the weights, and fine-tuning the network. This approach is advantageous as it allows us to utilise pre-existing knowledge that is useful, and ignore knowledge that is not useful



(a) Frozen ResNet (b) Unfrozen ResNet

Figure 2: Principle Component Analysis of the feature vectors extracted from frozen and unfrozen ResNet networks. In the unfrozen network, the feature vectors cluster, allowing better generalisation and easier training.

in predicting DR. Feature vectors produced after the training process are found to be easier to discriminate with fully connected layers (Fig. 2b).

3.2 SVM

The extracted feature vectors were classified using an SVM, with varying parameters. Since SVMs perform better on small datasets [13], a new dataset was constructed from the existing balanced one, with 200 images from each class in the training set and 100 images from each class in the test set. This was increased to 400 images per class for the training set, giving the standard train/test split ratio of 80/20, however no increase in test accuracy was observed so the size of the training set was not increased any further. Exploratory testing was undertaken, where the C Parameter, degree of the polynomial kernel function and the decision function shape (one-vs-rest or one-vs-one) were altered, but only negligible differences were observed in both overall test accuracy and per class test accuracies. As little change was observed, more exhaustive testing was not undertaken, and the parameters were kept at their default values of $C = 1.0$, degree $= 3$ and a one-vs-rest decision function shape.

3.3 Random Forest

Feature vectors from the balanced dataset were extracted by the AlexNet and ResNet-18 pre-trained networks and then used for classification by Random Forest (RF), an ensemble method that incorporates a large number of decision trees and classifies the data according to the majority voting [8]. The RF method was run against different variations of the data (unbalanced vs balanced), with different pre-processing techniques applied. It was found that simple normalisation on the balanced dataset gave the best performance. Different parameters were tested, including class weight, which did not improve the model, and increasing the number of estimators, which granted a slight improvement in accuracy.

4 Results

The Kaggle competition provides a large test set of images that can be used to evaluate the various techniques outlined in §3. While the test labels were previously hidden when the Kaggle competition was live, they have since been publicly released to allow further research. Three metrics were used to evaluate the efficacy of our methods against the test set: accuracy, f1-score, and quadratic weighted kappa (QWK [14]). QWK was included as it compensates for slight misclassification; for example a model that classifies a class four image as a class three image will get a better score for that particular image than a model that classifies it as class zero. It is also the metric used by the Kaggle leaderboard for this competition, so allows comparison of our techniques with the ones that were previously entered. Table 2 shows a comparison of the techniques used as part of this work.

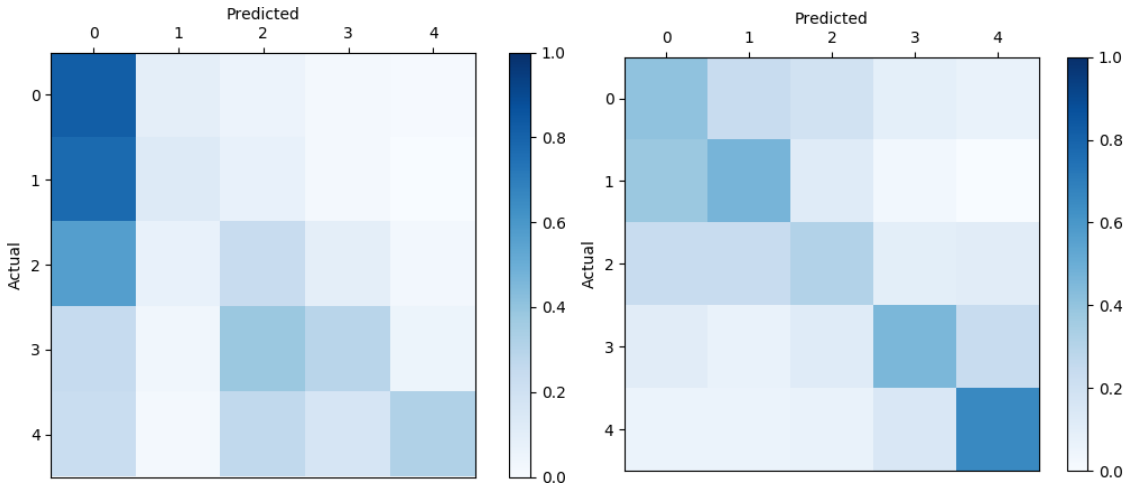
The RetCNN architecture required arduous training - it contained a total of 6818533 parameters and was trained from a random initialisation, meaning the network requires a large amount of time to train. As a result, for the performance we record here it completely over-fits to class zero, not returning any other classes.

Transfer learning showed more promise, especially when all of the network's weights were fine-tuned, not just the output layer. The unfrozen experiments achieved strong performance on our re-balanced dataset, achieving classification accuracies of over 0.85 during training. However, the unfrozen networks were subject to a large amount of over-fitting, giving much lower performance on the test dataset. This could be alleviated by applying more random transforms to the duplicate images, and using methods such as dropout, however implementing these methods was beyond the scope of this project.

The SVM classifiers give the best QWK scores, despite having worse test accuracy than some of the other methods. The RF methods give the best accuracy and also do well on the QWK metric, but are still subject to some over-fitting in the same manner as the deep CNN methods. This is due to the unbalanced test set - by over-fitting to class zero, it is possible to achieve a high level of accuracy, but this then gives a low QWK score. The SVM ResNet classifier is ultimately the best classifier we developed as it has the best QWK score, placing us in the top ten percent of teams in the Kaggle competition. The confusion matrix for the SVM is compared to the unfrozen AlexNet confusion matrix in Figure 3, clearly demonstrating its better generalisation and thus better QWK score.

Method	**Accuracy**	**F1-Score**	**QWK**
RetCNN	0.738*	0.627	0.00
Frozen AlexNet	0.620	0.593	0.137
Unfrozen AlexNet	0.664	0.662	0.422
Unfrozen ResNet	0.634	0.604	0.173
Unfrozen VGG	0.369	0.443	0.149
SVM - AlexNet	0.444	0.429	0.554
SVM - ResNet	0.462	0.461	**0.579**
Random Forest - AlexNet	0.679	0.652	0.320
Random Forest - ResNet	**0.683**	**0.688**	0.508

Table 2: A comparison of the different classification techniques. Each technique was evaluated against the complete set of test images. The best scores are highlighted in bold. *Misleading accuracy as it only classifies as class zero, i.e. it has not learnt anything.



(a) Unfrozen AlexNet

(b) SVM

Figure 3: A comparison of the confusion matrices for the Unfrozen AlexNet classifier and the SVM ResNet classifier. The over-fitting of the Unfrozen AlexNet classifier is very apparent, with lots of images being incorrectly labelled as class zero, whereas the SVM classifier gives better generalisation despite having worse overall accuracy.

5 Future Work

Future work could explore other architectures for pre-trained learning. The networks trained here can be further improved by tuning hyper-parameters such as learning rates, L2 weight decay, and investigating the use of other loss functions. When accuracy saturates during training, learning rate reduction should be utilised as this can increase accuracy. Other gradient methods could also be explored, including the AdaBound and AmsBound.

The main issue we encountered was dealing with the unbalanced dataset, which lead to model over-fitting. Other preprocessing techniques could be explored to deal with the unbalanced dataset (such as SMOTE, a data balancing method that intelligently samples and augments minority classes [15]) and further network techniques such as dropout could be utilised to balance both network over-fitting while still being expressive of signs associated with the different levels of DR. A wider range of ensemble methods could be tested, including AdaBoost, Gradient Boosting, XGBoost and CART. In addition, K-Fold Cross Validation could be implemented on the dataset, and probabilistic classifiers, such as Naive Bayes, could also be explored.

6 Conclusion

DR is one of the major causes of blindness worldwide, meaning early, automated detection methods would have important clinical significance. An automated screening method takes the burden away from trained clinicians, making it highly desirable. Of all the techniques we implemented, random forest (using features extracted from a pre-trained ResNet-18) achieved the highest test accuracy of 68%, however an SVM using the same feature extraction on a smaller dataset achieved a better quadratic weighted kappa score of 0.579, giving the best generalisation performance.

References

[1] World Health Organisation, "Global report on diabetes," 2016.

[2] D. S. Fong, L. Aiello, T. W. Gardner, G. L. King, G. Blankenship, J. D. Cavallerano, F. L. Ferris, and R. Klein, "Retinopathy in diabetes," *Diabetes care*, vol. 27, no. suppl 1, pp. s84–s87, 2004.

[3] A. W. Stitt, T. M. Curtis, M. Chen, R. J. Medina, G. J. McKay, A. Jenkins, T. A. Gardiner, T. J. Lyons, H.-P. Hammes, R. Simo, *et al.*, "The progress in understanding and treatment of diabetic retinopathy," *Progress in retinal and eye research*, vol. 51, pp. 156–186, 2016.

[4] Early Treatment Diabetic Retinopathy Study Research Group and others, "Grading diabetic retinopathy from stereoscopic color fundus photographs—an extension of the modified airlie house classification: Etdrs report number 10," *Ophthalmology*, vol. 98, no. 5, pp. 786–806, 1991.

[5] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 4, pp. 42–47, 2012.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[7] J.-G. Lee, S. Jun, Y.-W. Cho, H. Lee, G. B. Kim, J. B. Seo, and N. Kim, "Deep learning in medical imaging: general overview," *Korean journal of radiology*, vol. 18, no. 4, pp. 570–584, 2017.

[8] A. Liaw, M. Wiener, *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[9] B. Scholkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[12] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional neural networks for diabetic retinopathy," *Procedia Computer Science*, vol. 90, pp. 200–205, 2016.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[14] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.," *Psychological bulletin*, vol. 70, no. 4, p. 213, 1968.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

7 Marks Split

All members indicate that they would like an equal marks split as they all contributed equally to the project.

Michelle Abela

Joseph Early

Sandeep Mistry

Aran Smith