Analysis of Bimodal Distribution Removal for Facial Expression Identification

Alex Lawrence

Research School of Computer Science,
Australian National University

U5801889@anu.edu.au

Abstract. Bimodal distribution removal in neural networks is a useful tool for noisy datasets. This paper identifies the degree of benefit of this method for facial expression identification and whether the benefit present from previous analysis holds under a more complex structure. This is achieved through implementing bimodal distribution removal in a convolutional neural network and then evaluating the results against the previous findings of applying the algorithm in a two layer neural network. It was found that the variance of the error distribution rose in relation to the decrease in average error. This is crucial for the bimodal distribution removal algorithm, hence a partial implementations of the algorithm was tested to lacklustre results. This result is meaningful as it adds a wider scope to the evaluation performed in the previous paper on the benefit of bimodal distribution removal.

1 Introduction

This paper provides further analysis of the Bimodal Distribution Removal method for the removal of outliers from noisy datasets. It extends upon the previous analysis of BDR[1] within a simple two layer neural network to view the effects of BDR when implemented within a more complex convolutional neural network for a facial expression identification task. The performance of the BDR algorithm is examined through the comparison against a normal back propagation model and then analysed against the results from the two layer neural network model. It was found that the implementation of BDR in the convolutional network did not display the same benefits that were present within the two layer neural network implementation. In order to add context to the results it's important to first summarise the established history of both the dataset used, the BDR method and the previous analysis conducted.

1.1 Background

1.1.1 Dataset

The dataset used in this experiment was first introduced in the paper "*Static Facial Expression Analysis in Tough Conditions: Data, Evaluation Protocol and Benchmark"* [1]. It's labelled as the Static Facial Expressions in the Wild (SFEW) database. The paper [1] uses this dataset along with several others to develop a standard protocol for analysing the accuracy of expression recognition models. The standard that is proposed is called the SPI Baseline which calculates accuracy based on the number of true positives, true negatives, false positives and false negatives. The accuracy calculated for the SFEW database is cited as 43.71%. The explanation for the low accuracy is that due to the real world conditions of the images in the dataset and differing resolutions of the faces depicted, lots of complexity is added to the classification problem. The accuracy for images with a neutral background and high resolution faces experienced much better results. It's for this reason that an outlier removal algorithm like BDR could prove useful, as removal of images that aren't contributing to the network learning has the potential to improve accuracy.

The dataset consists of 675 images taken from movies and television. All the images depict a person displaying a facial emotion. Each image is labelled as one of seven emotions which indicates the facial emotion displayed in the image. The possible emotions displayed in the images are anger, disgust, fear, happiness, neutral, sad and surprise.

1.1.2 Bimodal Distribution Removal

Bimodal distribution removal was introduced by researchers Slade and Gedeon [2] in their paper titled *Bimodal Distribution Removal.* BDR is a method proposed to remove outliers from datasets processed through neural networks. In particular to remove noise generated from real world datasets. This was important as other proposed methods performed substantially less well on noise generated from real world data as opposed to artificial noise. The paper [2] concluded that

[1] Bimodal Distribution Removal

the method performs as well as other outlier detection/removal methods without the disadvantages of them, such as slowing down learning and increasing the chance of overfitting.

How bimodal distribution removal works is detailed in the paper by Slade and Gedeon [2]. They state that a measure of variance of the error distribution provides a metric to determine when patterns are starting to form that don't provide expected values provided the input. It's stated that when there is a low enough variance value in the error set (which they state is approximately ≤ 0.1) this indicates outliers are distinctive enough to start the removal process. Then the patterns with error greater than the mean of the error over the training set, forming a subset. Then the mean *(m*) and standard deviation (*s*) is calculated for this subset, and removes patterns from the training set that have an error $\geq m + \alpha s$ (where α is a value between 0 and 1) as these are judged by the method as outliers. The network is then trained for 50 epochs before starting the process again to learn the features from the altered set of patterns. BDR stops when the variance of the error set is so low that beyond that continuing will potentially eliminate useful data from the training set. In the paper [2] this value is stated as 0.01. This provides a natural termination for the neural network which is why is avoids issues like overfitting and slow learning.

1.1.3 BDR Analysis for Mark Prediction

This paper is an extension of a previous study on the effectiveness of the BDR algorithm when implemented in a simple two layer neural network designed to predict the final marks of students without the final exam mark. The dataset for this study originally came from the paper by Gedeon & Turner [3] which looked at examining the decision process of neural networks. The dataset was chosen for the analysis of the BDR algorithm due to the noise created by removing the final exam mark which accounted for 60% of the student's grade [3]. This creates noise in the dataset which makes it a prime candidate for testing the BDR algorithm. This is the same reason that the facial recognition task on the SFEW database [1] was selected for this further analysis. As all the images are in real world backgrounds instead of a lab environment as well as the fact that the faces in the images are of different resolutions [1], thus a lot of noise is generated. Therefore providing an environment for which the BDR algorithm could provide benefit.

The structure of the network used was a two layer network that utilised MSE[2] for the loss function and SGD[3] as the optimiser. The activation function for the hidden layer neurons was the ReLU (rectified linear unit) function. This differs significantly from the structure of the network for this paper and thus the comparison will provide useful information.

The experiment conducted consisted of examining the difference between the normal back propagation model and the model with BDR implemented under varying network conditions. There were four network variables that were tweaked which include:

- Changing the k value in k-fold cross validation
- Changing the value of α for the bimodal distribution removal algorithm
- Changing the number of hidden neurons
- The number of epochs taken to train each model.

The k-value showed the difference in performance of the BDR algorithm in relation to the size of the training dataset. It was found that the model performs better with BDR when the dataset is smaller. The value of α controls how many patterns are deleted when the BDR algorithm is utilised. It was found that the average error for the model lowered as α increased. Altering the number of hidden neurons allowed the performance of the algorithm to be analysed in different network structures. It was found that the model reached the halt condition of BDR faster the more hidden neurons the network contained.

The main finding for the study was that the training speed for bimodal distribution removal was much faster for a slightly less accurate model, while also providing a natural stopping point. This finding is meaningful for models which value a decrease in training time over optimised prediction accuracy. Therefore study found that for a simple 2-layer neural network, the BDR algorithm provides benefit. This paper seeks to further explore the usefulness of the BDR algorithm by applying it to a larger convolutional network to see if these benefits hold.

1.2 Aim

This study extends the previous paper and examines the positive and negative effects of implementing Bimodal Distribution Removal on a convolutional neural network that's attempting to perform an image classification task of identifying emotions displayed within an image of a person's face. Furthermore the degree of benefit for implementing the Bimodal Distribution Removal algorithm in this model will be compared to the results that were found within the first study of the two layer neural network. The reasoning behind this is that it will provide a larger picture surrounding the

[2] Mean squared error
[3] Stochastic gradient descent

usefulness of the BDR algorithm, and will show how the two network structures differ in performance when the algorithm is applied.

2 Method

2.1 Network Structure

The network structure utilised for the task of facial expression identification is a convolutional neural network (CNN). Convolutional neural networks work exceptionally well for image classification tasks. An example and basis for the neural net constructed for this task is the convolutional neural network designed by Alex Krizhevsky [4]. The CNN used in this study consists of four convolutional layers, three max pooling layers and then two fully connected layers. The convolutional layers extract features out of the images which the fully connected layers can then categorise. The pictures are relatively high resolution with 720x576 pixels. Therefore the first layer has quite a large kernel and a max pooling layer following it in order to obtain a workable amount of filters. The activation function used for every layer excluding the output layer was the ReLU[4] function. This is the same function that was used for the previous study. The cross-entropy loss function was utilised for the model as it has been shown to be very effective within convolutional neural networks [5]. The Adams optimiser was used as it performed the best out of the candidate optimisers. In particular the function used in the previous study was attempted (SGD), however it failed to effectively modify the network as the training and test results were not changing, despite modifying the learning rate and momentum value. The batch size selected was 64. The model was trained for 100 epochs before testing occurred.

2.2 Input Handling

The file structure of the SFEW database consists of folders named by the facial expression shown in the images within that folder. In order to encode the images into a format in which the convolutional neural network would recognise them the Torchvision ImageFolder function was utilised [6]. This function takes as input the folder containing the database and encodes the pixels of the images in tensors, then attaches a number to indicate the folder in which the image was found. These numbers can then be used as targets to train/verify the network. The images are shuffled so that the split of partitions for the training and testing set are randomised.

2.3 Bimodal Distribution Removal

After the normal back propagation network was established, the Bimodal Distribution Removal method was then implemented, as per the description in the original paper [2]. Which in summary consists of training the network until the variance over the errors of the training set falls below 0.1, at which point a subset of the patterns with an error greater than the mean of the errors is taken. Then the mean (m) and standard deviation (s) is calculated for this subset and finally all the patterns in the subset with an error $\geq m + \alpha s$ where $0 \leq \alpha \leq 1$ are removed from the training set. Then after 50 epochs, (to allow the network to learn the new training set) the function runs again and this continues until the variance of the errors is less than 0.01. At this point the training process is terminated.

2.4 Partial Implementation

There were some unforeseen circumstances that prevented the early stopping method within the BDR algorithm. This is discussed further within the results section. Due to this the full implementation for the algorithm was not possible. Therefore the algorithm was altered such that it didn't contain a stopping condition anymore and just ran the pattern removal process every x epochs where x is assigned before running. This allows for some comparison between this study and the last even though the full implementation wasn't possible. The amount of epochs in which the altered BDR algorithm ran was altered and the accuracy of predictions was collected to show the performance difference of running the algorithm more or less for this model.

2.5 Performance Measures

To ensure accurate performance measures k-fold validation was implemented. This method consists of breaking down the dataset into randomised partitions of size k. One partition is selected as the test set and the rest of the partitions form the training set. The partitions are then iterated through so that every partition is selected as the test set exactly once. The error for each of the test sets are summed and divided by k to get the average error for the model. This validation

[4] Rectified linear unit

method was chosen as it doesn't waste very much data while still capturing performance of test sets which contain less common patterns. The *k* value chosen for the model was 5. This is for two reasons. Firstly since the dataset contains 675 images, having *k=5* results in a training sets of 540 and a testing set of 135, an 80/20 split, which is an advantageous ratio. Secondly 5-fold cross validation was utilised in the original paper where this dataset was used [1]. Using the same validation method will make comparisons between results more meaningful

The measure for performance used for the model was the percentage of the total correct predictions for the facial expression featured in the images within the test set.

3 Results & Discussion

3.1 Relationship Between Average Error and Variance

During the training process of the model, an unexpected relationship emerged which brought issues with the implementation of BDR. The average error for the model declined the longer the training process continued, which is what was expected to occur. However what wasn't expected was that the variance of the error distribution rose the more the model was trained. The seemingly inverse relationship between the average error and the variance of the error distribution can be seen in Fig 1 below. The increasing variance breaks the BDR algorithm. Since BDR starts when variance is less than 0.1 the only time where it could be activated is near the beginning of training where the training patterns aren't properly learned and thus outliers aren't properly distinguished from other data. Furthermore the stopping condition of 0.01 is never met. These values were increased to various assignments to see whether the variance would dip after BDR algorithm ran, however it still rose. Therefore an altered implementation of the BDR algorithm was tested where variance wasn't considered and the removal took place every *x* amount of epochs.

Variance declined as training occurred during the mark prediction analysis. This could be due to a few reasons. The first reason could be that originally the BDR algorithm was designed to run on models that train for thousands of epochs such as the previous mark prediction task which ran for 3000 epochs instead of the 100 that this model does. The second possible reason is that the loss function for the mark prediction task (MSE), creates less variance in the error distribution than cross-entropy loss which was used for this task. It could also relate to the large size of the dataset for this task, as was specified in the original BDR paper, BDR doesn't provide much benefit on a large dataset [2]. Despite this not being expected it does highlight some limitations of the BDR algorithm, and shows that it is inapplicable for some models that have variance similar to this example. Further study could be conducted into modifying the BDR algorithm to perform on variance that rises in relation to the training epochs.

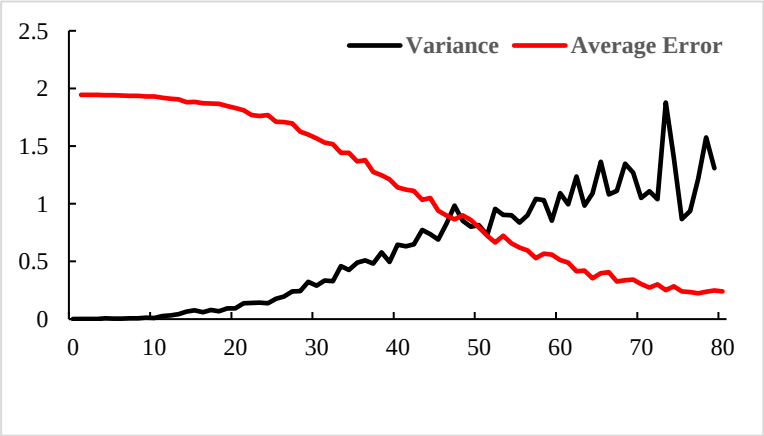

Fig. 1. Relationship between average error and variance of error distribution

3.2 Partial Implementation

Due to the inapplicability of the original BDR algorithm, the results were collected for the model without BDR as well as the altered BDR algorithm operating every 20, 40 and 60 epochs. See Fig 2 below for a visual comparison between the results. These intervals were chosen as they each represent different behaviours. With removal occurring every 20 gives the model time to learn the patterns but still occurs 4 times within 100 epochs. With a 40 epoch interval, removal occurs twice and gives the model longer to learn the initial patterns before removal takes place. With a 60 epoch interval, the removal process happens once and gives more time to the initial learning process.

The model without BDR implemented performed very similarly to the results in the original paper [1] with an accuracy of 41.63%. While this result could be improved through further optimisation of the network structure it seemed irrelevant for this task of assessing the impact of the BDR algorithm.

With removal occurring every 20 epochs the model achieved an accuracy of 40.44%. This is not a major decrease and comparable to the decrease in performance observed in the mark prediction task. However unlike the mark prediction task the model didn't experience faster training as the natural stopping point was removed.

With removal occurring every 40 epochs the model further decreased in accuracy to 38.96%. This is a surprising result as it was hypothesized that the outliers in the training set would be more visible after 40 epochs of being trained when compared to just 20 epochs. This could be a result of variance due to the 20 epoch model being luckier with their predictions or outliers were learnt enough to be undetected by the removal algorithm.

The removal every 60 epoch model performed the best with an accuracy of 42.37%. This could have performed better than the others due to the fact that it only runs once when the patterns are learnt to an adequate degree, thus removing only the hardest patterns to learn.

Overall the results indicate that for this image classification task, BDR is not very useful. Due to the removal of the stopping condition which is one of the major appeals of the algorithm as well as the statistical based selection of when to apply the removal of patterns, the arbitrary picking of epochs for when to apply the removal procedure yields inconsistent and for the most part worse results than no implementation. The significance of this result is that it provides context to the results gained in the mark prediction study as it shows that there are situations where the BDR algorithm provides no benefit. A direction for further study could include further identification of areas where BDR can be applied beneficially, possibly for evolutionary algorithms.

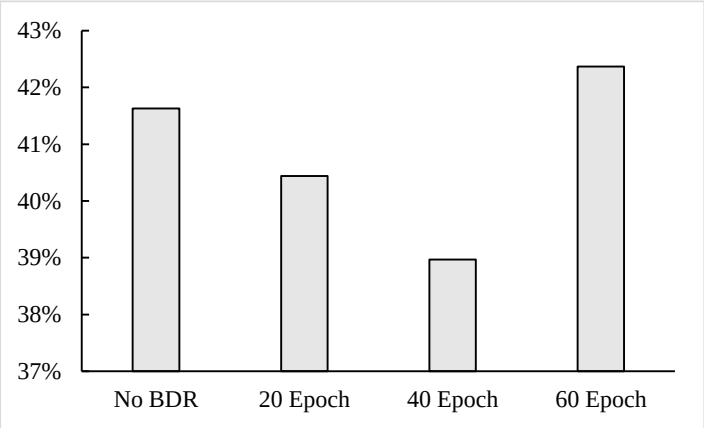

Fig. 2. Comparison of results from applying modified BDR algorithm

4 Conclusion

This paper aimed to provide further analysis on the usefulness of the implementation of the BDR algorithm within the neural network architecture. The study looked at the performance differences between different implementations of an altered version of the BDR algorithm on a convolutional neural network designed to identify facial expressions within images. It was found that the variance acted differently than previously observed in the simpler 2 layer network used for the mark prediction study and thus could not employ the natural stopping condition or the condition that triggers the removal of patterns. Therefore an altered version of the algorithm was used that produced mediocre results. The implication of these findings is that while the BDR algorithm may have had benefits for the mark prediction task, it's not applicable for all neural network structures. This suggests that further study could include further investigations into structures in which there is a benefit of bimodal distribution removal.

5 References

1. A. Dhall, R. Goecke, S. Lucey and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, Barcelona, 2011, pp. 2106-2112, doi: 10.1109/ICCVW.2011.6130508.
2. Slade P., Gedeon T.D. (1993) Bimodal distribution removal. In: Mira J., Cabestany J., Prieto A. (eds) New Trends in Neural Computation. IWANN 1993. Lecture Notes in Computer Science, vol 686. Springer, Berlin, Heidelberg
3. T. D. Gedeon and H. S. Turner, "Explaining student grades predicted by a neural network," *Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan)*, Nagoya, Japan, 1993, pp. 609-612 vol.1.
4. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).
5. T. N. Sainath, A. Mohamed, B. Kingsbury and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, 2013, pp. 8614-8618, doi: 10.1109/ICASSP.2013.6639347.
6. Sébastien Marcel and Yann Rodriguez. 2010. Torchvision the machine-vision package of torch. In Proceedings of the 18th ACM international conference on Multimedia (MM '10). Association for Computing Machinery, New York, NY, USA, 1485–1488.