UO MS Project Report

Adib Mosharrof

March 2021

Contents

1 Introduction **3**

2 Data **3**

3 Related Work **3**

4 Scoring **4**
- 4.1 Matthews Correlation Coefficient 4
- 4.2 Mediscore 4
- 4.3 Dilation 5
- 4.4 Erosion 5
- 4.5 No Score Region 5
- 4.6 Thresholded MCC 6

5 Methodology **6**
- 5.1 Data Pre processing 6
  - 5.1.1 Pixel 6
  - 5.1.2 Image Resizing 7
  - 5.1.3 Image Patches 7
- 5.2 Architectures 7
  - 5.2.1 Logistic Regression 7
  - 5.2.2 Neural Networks 8
  - 5.2.3 Weighted Ensemble 9

6 Experiments **9**

7 Results **9**

8 Discussion **9**

9 Future Work **9**

Model	Self	Mediscore
LR	0.095	0.235
LGB	0.22	0.23
MLP	0.007	0.004
XGB	0.195	0.238
LR (tuned - log loss)	0.075	0.239
LGB (tuned)	0.19	0.242

Figure 1: Previous Results

1 Introduction

2 Data

3 Related Work

The initial work on this data focused on using traditional machine learning algorithms to perform image localization. In these models one row of data was considered to be a single pixel in an image. A data preprocessing step was applied to convert each image into a 1 dimensional array, thus enabling the use of a single pixel as a row of data. The number of features in a model is equal to the number of indicators that have been selected to be used. As an example of a single row of data, we consider the top left pixel of an image, where we take the value of that pixel from each of the indicators and pass that into the model as input.

One of the major challenges in the previous work was to get a consistent scoring for the algorithms. As mentioned in the previous section, Mediscore was used to evaluate the performance of the models. Since it is a very time consuming step, a local scoring system was created which would provide a score for the models. However, there was no direct correlation between Mediscore and the local scoring method, which proved to be a major drawback when fine tuning models to increase performance. Tweaks in hyperparameters that would increase the score in the local scoring, would not produce the same behavior in from the results returned from Mediscore. A major step in our current work was to create a new local scoring system that would be fast and also be consistent to what Mediscore would produce, thus enabling us to perform experiments without having a tight coupling with Mediscore.

Previous work mainly focused on models with decision trees, regression and boosting. The scores produced from the different models are shown in figure 1

Initial work started with a vanilla Logistic Regression model that produced a decent score in Mediscore.

		Prediction	
Actual		P	N
	P	TP	FN
	N	FP	TN

Figure 2: The Standard Confusion Matrix

Positives (P) and Negatives (N). True Positives(TP) and True Negatives (TN) are the correct predictions, while False Negatives (FN) and False Positive (FP) are the incorrect predictions

4 Scoring

As mentioned in the Data section, in the dataset, the proportion of manipulated pixels when compared to non-manipulated pixels is very small. Unblanced data generally causes algorithms to be biased towards the majority class, thus popular metrics like accuracy, which is the ratio between the number of correctly classified samples and the overall number of samples (for example [8]), will no longer be an accurate metric to use here. Accuracy would provide an overoptimistic estimation of the classifier's ability on the majority class [2]. Consider a data set that has 10% positive class and 90% negative class. A naive classifier that always outputs the majority class label will have a high accuracy of 0.90.

The problem caused due to this imbalance can be addressed by using the Matthews correlation coefficient (MCC), a special case of the ϕ phi coefficient [4]. MCC has been originally developed by B.W. Matthews for comparison of chemical structures [5]. However, it was re-proposed by Baldi and colleagues [1] in 2000 as a standard performance metric for machine learning with a natural extension to the multiclass scenario [3].

4.1 Matthews Correlation Coefficient

Matthews correlation coefficient is defined in terms of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). These terms can be calculated from the confusion matrix as defined in Figure 2. MCC is a contingency matrix method of calculating the *Pearson product-moment correlation coefficient* [6] and can be calculated using Equation 1.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{1}$$

4.2 Mediscore

The scoring mechanism in the Mediscore project has a concept of no score region that was not considered in the previous work. The idea behind the no

score region is to ignore the regions on the boundary of the manipulated and non-manipulated regions. If an algorithm can identify the manipulated region in the image, it could still mislabel a lot of pixels in the boundary region, which could incurr a high penalty to the algorithm and give it a score that does not accurately reflect its performance. To get a more fair score, the score is calculated by considering the pixels in the image excluding the pixels that are present in the no score region. In order to produce the no score region, two of the core morphologic operations; erosion and dilasion, are applied to the image. A difference operation is performed on the images produced from erosion and dilation, which is followed by a binary inverse thresholding to produce the set of pixels that constitute the no score region.

4.3 Dilation

Applying dilation $\oplus$ to an image expands the shape of the image. Let X be a gray scale shape and B be a symmetric structuring element. The dilation operation on these two elements, $X \oplus B$ can be defined as

$$X \oplus B = X + b = \{x + b : x \in X \& b \in B\} \tag{2}$$

The output of dilation is the set of translated points such that translate of the reflected structuring element has a non-empty intersection with X. Equation 2 is based on obtaining the refelction of B about its origin and shifting this reflection by b. The dilation of X by B is the set of all the displacements, b, such that x and b overlap by atleast one element [7].

4.4 Erosion

Erosion Θ is generally applied to images for eliminating irrelevent details in terms of size.

$$X \Theta B = X - b = \{z : (B + z) \in X\} \tag{3}$$

The output of erosion is the set of translation points such that the translated structuring element is contained in the input set X. Equation 3 indicates that the erosion of X by B is the set of all points b such that B, translated by b, contain X [7]. The general output of erosion is that it shrinks the shape of the image.

4.5 No Score Region

The boundary no score region is produce by performing two operations sequentially, which are defined in Equation 4 and 5 Equation 5 is basically a binaray inverse thresholding.

$$\psi = (X \Theta B) - (X \oplus B) \tag{4}$$

$$\text{NoScore}(x, y) = \begin{cases} 0, & \text{if } \psi(x, y) > 0 \\ 1, & \text{otherwise} \end{cases} \tag{5}$$

4.6 Thresholded MCC

The output of the algorithms is a grayscale image where pixel values are in the range $[0 - 255]$ and the ground truth is an image which has pixel values of 0 or 1, which represent manipulated and non-manipulated pixels respectively. The output image is converted from a 2 dimensional array into a 1 dimensional array and the pixel indices that are in the no score region are filtered out to produce a list of pixels that are to be used in scoring. The indices in this list are called the scoring indices which will be used to calculate the MCC score.

The pixel values in the scoring indices list need to be converted to have a value of either 0 or 1. This conversion is done using a binary thresholding on the scoring indices list, where the threshold values are in the range $[0 - 255]$. Thresholding is applied using every value in the range and the output produced is used with the ground truth to produce an MCC score for that particular threholded value. The maximum MCC value obtained is selected as the MCC score for the image.

5 Methodology

5.1 Data Pre processing

The format of data expected as the input layer varies depending on the type of model being used. The raw image files cannot be served directly as input, thus we need to process these images into the format expected by the appropriate model. The highest level of distinction is whether the model expects 1 dimensional data, or it expects a 2 dimensional data, an image of a certain size. The data preprocessing step exists as an individual module, in which the input is the raw images and the output is either an image or a csv file depending on the configuration specified.

5.1.1 Pixel

Models like Logistic Regresssion, Random Forests and other traditional machine learning algorithms, use a single pixel as a row of data. Images are read in transformed from a 2 dimensional array into a 1 dimensional array, so that it can be fed into the model as input. The output of the model will be a 1

dimensional array of the same size as the input array. A very important step has to be taken next, which is converting this 1 dimensional array back to a 2 dimensional array, so that we can generate an image of the output of the model. The dimension reduction can be done by stacking rows or columns together, so when we are generating the output image from the model prediction, we have to be consistent with the method used previously.

5.1.2 Image Resizing

Neural Network models expect input images to be of a certain dimension, and so in this data pre-processing step, we resize all the input images to a certain size. Since there is a big variation in the image sizes, depending on the resize resolution, some images might be scaled up whereas some might be scaled down.

5.1.3 Image Patches

When resizing images to a fixed size, a lot of information in images are lost. When a high resolution image is shrunken down considerably, there is a lot of blur in the image and fine grained details are lost in the process. To mitigate this problem, we break down a whole image into patches of a fixed size. One main challenge here is that not all images break out into even patches. There are numerous pixels around the boundary of an image that are not enough to fulfil the size of the image patch. To remedy this, we add padding to the image both horizontally and vertically, so that the image is of a size that can be broken down evenly according to the patch size specified.

A major step in this preprocessing is reconstructing the source image from the individual patches. We have to align the patches in the right order and also remove the padding we added. In order to create patches and reconstruct patches from an image, we used a library, patchify, which provides functions to produce patches from an image given a patch shape, and also function to reconstruct images, given a list of patches.

5.2 Architectures

We have used mixture of Machine Learning models and architectures, ranging from traditonal models like Logistic Regression, to simple Neural Network and even a bit more complicated neural architecture, U-NET. Models require data in certain formats, so each model is tied to a certain data preprocessing step, so that it can obtain the data in the format it wants.

5.2.1 Logistic Regression

Logistic Regression is a classical linear Machine Learning model that has proved to be a really simple but efficient model. We have included this to create a baseline score with which we can compare other models. This model has been implemented using the scikit-learn library. One of the major drawbacks of logistic regression is that it has to load the whole data for training, thus we

are limited by the amount of memory we have. It does not have the ability to perform batch learning, thus creating this challenge. Due to this challenge, we had to perform Logistic Regression on data that had been scaled down by a factor of 32 or more so that we can avoid the memory challenges.

5.2.2 Neural Networks

We have explored with different types of Neural Network architectures to fit the data, ranging from simple single layer networks to multi layers and also the complex U-NET architecture. The neural architectures have the ability to work with both 1 and 2 dimensional data. Data preprocessors arecoupled to the type of architecture to feed the data in the required format. The main advantage of neural networks is that it can perform training in batches, thus allowing to train on a large dataset and also on images of larger resolution. For images with large resolutions, we had to minimize the batch size in order to fit the memory constraint.

The networks have been implemented using the Keras 2.0 library. Sigmoid and ReLU activation functions have been used when building the networks. To fine tune the models, different L1 regularization and learning rate values has been used.

5.2.2.1 Multi-layer Neural Networks

A sequential linear multi-layer network has been applied to the 1 dimensional data. The number of neurons in the input layer is equal to the number of indicators in the data. Each dense layer is followed by an activation layer. The final output layer has only one neuron, which predicts whether the current pixel is manipulated or not. The input layer has the maximum number of neurons, and the number of neurons in successive dense layers are halved every time. For example, if we start with an input layer with a dimension of $64x64$, and follow it with 3 dense layers, then the dimensions of these 3 layers would be $32x32$, $16x16$ and $8x8$. These dense layer would then be augmented with the output layer, which would have only 1 node.

Add an image of the architecture here

5.2.2.2 UNET

UNET is a special type of Convolutional Neural Network, that has been designed specifically to solve image segmentation and localization.

MORE DETAILS ABOUT UNET HERE

U-NET works with 2 dimensional data, specifially on the data produced after the preprocessing steps of image patches and image resizing. A single layer in the network is designed to have a convolutional layer, followed by a max pooling layer, which is subsequently followed by a dropout layer. The various hyperparameters have been tinkered with to find the set of values that

produce the best results. More regarding hyperparamter tuning are included in the experiments section.

5.2.3 Weighted Ensemble

The different machine learning algorithms pick up different patterns from the input data, some of which might be very important, while others might be less. There is also the possibility that some patterns are picked up by certain algorithms, whereas other algorithms completely fail to pick them up. In order to tackle these cases, we have introduced a weighted ensemble of the algorithms. Each algorithm runs independently and produces a prediction. Next, a weighted average is performed on the predictions based on the weights provided to produce a final output.

6 Experiments

7 Results

8 Discussion

9 Future Work

References

[1] Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., and Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics 16*, 5 (2000), 412–424.

[2] Chicco, D., and Jurman, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics 21*, 1 (2020), 1–13.

[3] Gorodkin, J. Comparing two k-category assignments by a k-category correlation coefficient. *Computational biology and chemistry 28*, 5-6 (2004), 367–374.

[4] Guilford, J. P. Psychometric methods.

[5] Matthews, B. W. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure 405*, 2 (1975), 442–451.

[6] Powers, D. M. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061* (2020).

[7] Tambe, S. B., Kulhare, D., Nirmal, M., and Prajapati, G. Image processing (ip) through erosion and dilation methods.

[8] Wang, L., Chu, F., and Xie, W. Accurate cancer classification using expressions of very few genes. *IEEE/ACM Transactions on computational biology and bioinformatics 4*, 1 (2007), 40–53.