Project Report

Bacteria Classification Problem

Vatsal Vador, Rohan Hulsure, Jingsheng Li

1. Problem Statement:

The goal of this project is to automate the process of classifying bacteria from images acquired by the microscopes through image classification and deep learning which will help reduce the time required to analyze and classify the bacteria and increase the accuracy of the diagnostic process.

2. Motivation:

Recognition and classification of bacterial genera and species is crucial as there are many important results which can be useful in medicine, biochemistry, food industry or farming. Recognition of such microbiological samples is preceded by culturing process which require much dedicated tools, equipment and chemical agents to complete whole process. The classical laboratory methods of bacteria recognition require an expert knowledge and experience. It is a time-consuming process based on comparative analysis of the obtained samples with referential ones. Recognizing bacteria based solely on the shape would be a difficult one because many bacteria share very similar shapes. Second most differentiating feature is the shape and the size of the colonies formed by the bacteria. Moreover, there are several procedures and safety protocols associated with such methods. If we sum up it is time-consuming process. Therefore, automizing this process with deep learning can help reduce the time and increase the accuracy of the pathologist.

3. Methods:

To achieve the above-mentioned goal, multiple deep learning models were experimented. Convolutional Neural Networks were mainly used as the data consisted of images acquired by the microscopes.

The dataset – Digital Images of Bacteria Species dataset [DIBaS] consisted of 33 bacteria species with 20 images of each species, 660 images in total. These images were collected by the Chair of Microbiology of the Jagiellonian University in Krakow, Poland. All the samples were stained using the Gramm's method. The images were taken with Olympus CX31 Upright Biological Microscope equipped with a SC30 camera. DIBaS dataset is publicly available to other researchers (http://misztal.edu.pl/software/databases/dibas/). As proof of concept for this project we will be only working with 4 classes/species of bacteria mentioned below:

a) Candida albicans
b) Clostridium perfringens
c) Lactobacillus paracasei
d) Staphylococcus saprophiticus

3.1 Approach 1 – CNN with Image Datagenerator

The images in the dataset are of the highest resolution, 2048x1532. They were reshaped to 512x383 and were trained as color images based on the idea of saving enough original information. An image data generator was used, a part of Keras library for real time data augmentation. The image data generator accepts a batch of training images, apply a series of transition to each image in the batch and replace the original image with new randomly transformed batch.

The architecture of CNN to train these sets of images consists of a 2D convolutional layer having filter size 40 with kernel size 7x7 and rectified linear unit as the activation function. The second is the Max Pooling layer with pool size 7x7, strides of 5x5. The output of these layers is flattened and then sent to the following two dense layers, first with 100 neurons and rectified linear unit as the activation and the second dense layer as a final output with total classification groups and softmax activation layer. (Refer Fig 3.1.1 below)

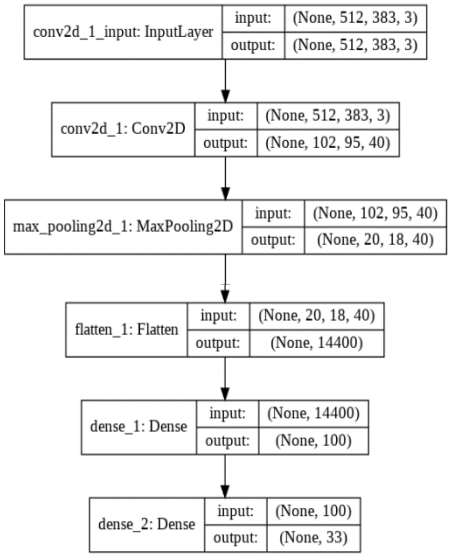

Fig 3.1.1: Architecture of Approach 1

This model used Adam optimizers with default learning rate and compiled as categorial cross entropy

3.2 Approach 2 – CNN with Image breakups

The second method is based on the results of approach 1, analyzed to improve accuracy. As the images in the data set are of high resolution (2048x1532) resizing them to shorter size than that of approach 1 would have led to loosing of important features such as edges and clusters may have been lost. These high-resolution images were therefore broken down into blocks of 224x224 with a custom script. Breaking down of images into blocks also led to having increased training and validation data (~32000 images). They were later transformed into grayscale images, resulting to clear features such as edges shapes and size of the bacteria (Refer Fig 3.2.1 below)

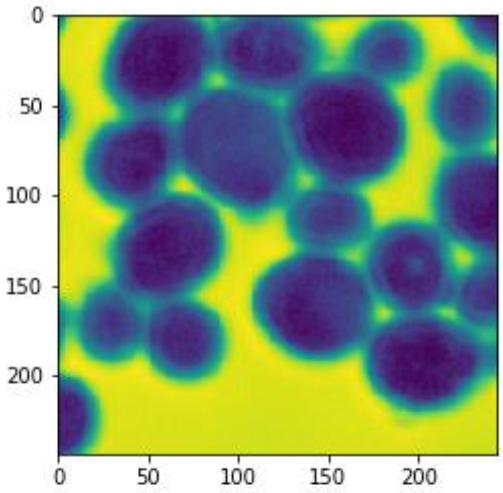

Fig 3.2.1: Transformed grayscale Image block of size 224x224

These transformed images were labelled according to their specific class names and then trained of a more tuned CNN architecture than the previous approach. The architecture consists of two different 2D Convolutional and Max Pooling layers having filter size 32 and 64, kernel size of 3x3, activation function as Rectified Linear Unit and pooling size of 2x2 respectively. The outputs of these two layers were flatten and sent to the Dense layers. First dense layer having 64 neurons with Rectified Linear Unit as the activation function, and second dense layer as the output layer with total classification groups and softmax activation layer. (Refer Fig 3.2.2 below)

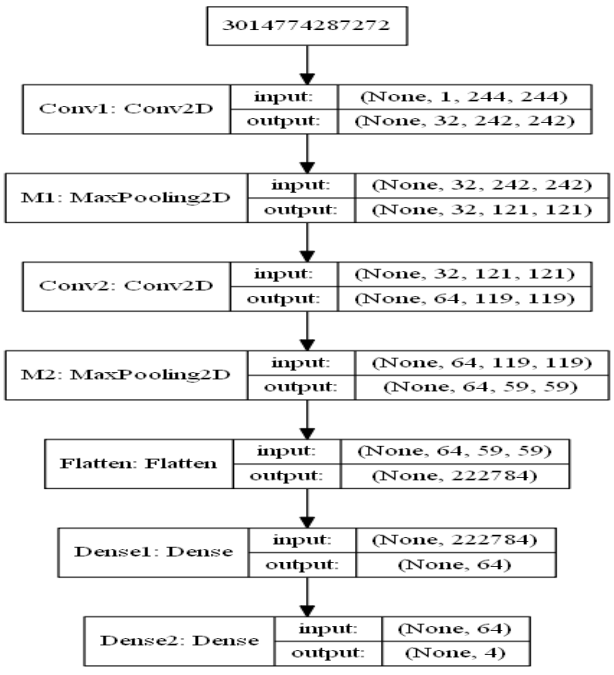

Fig 3.2.2: Architecture of Approach 2

The second model used Adam optimizers and compiled as categorial cross entropy with learning rate of 0.00025 tuned for higher accuracy.

4. Results:

Both the approaches provided varying results. The results were analyzed on the metrics of accuracy vs loss for both training and the validation data, classification report from the sci-kit library showed the precision of models.

4.1 Approach 1 –

The accuracy for approach 1 lies best between 75% - 80%. The first model was trained for 30 epochs with batch size 50. The following plots depict the behavior of the model in approach 1. The model was trained with colored images reshaped to 512x512 augmented using image data generator. As these images were of high resolution the acquired huge main memory space and slowed down the training therefore maximum of 30 epochs and 50 batch size was used.

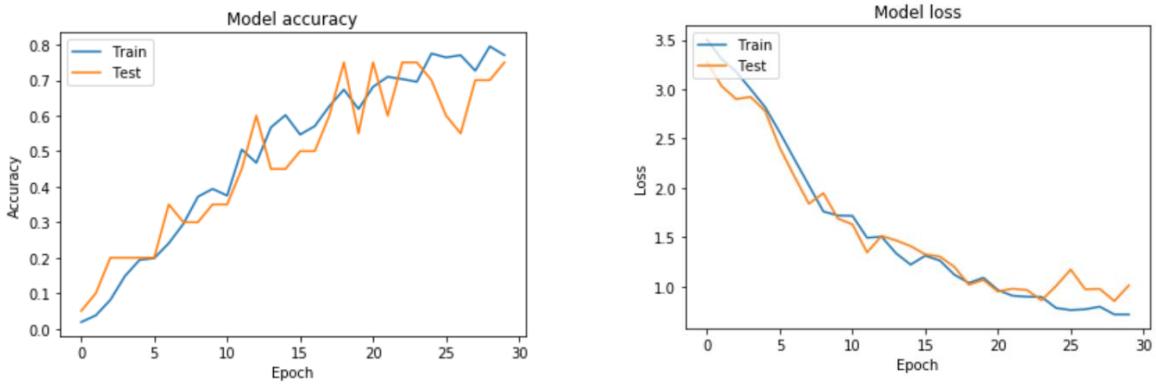

Fig 4.1.1: Accuracy and Loss Plot for approach 1

4.2 Approach 2-

Analyzing the results of approach 1 we tuned the architecture for approach 2 for improved accuracy, result – success. The accuracy for model 2 lies between 92% - 93% The model was trained with grayscale transformed images for 100 epochs and batch size of 90 (exceeding 90 produced resources exhausted error).

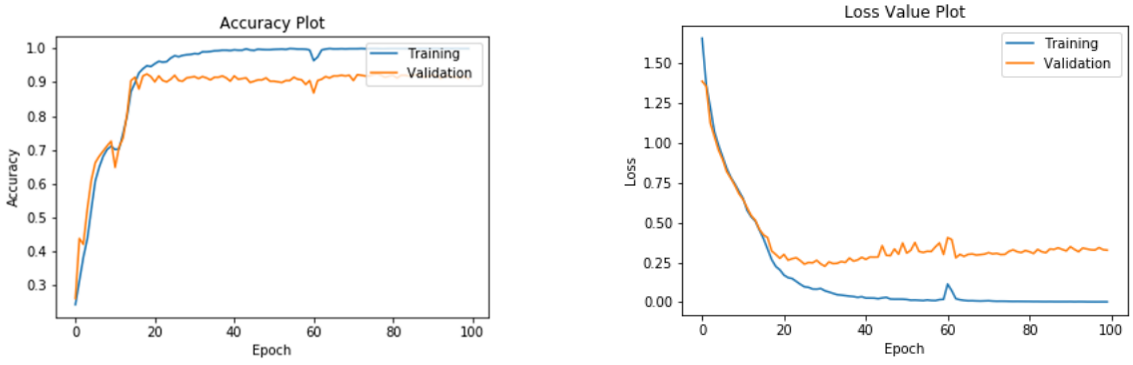

Fig 4.1.2: Accuracy and Loss Plot for approach 2

Analyzing the plots we can see that training accuracy reaches 99% while the validation accuracy stays between 90% – 93%. We tried to fined tune the model by varying the learning rate. Learning rate of 0.00025 gives the best results, reducing or increasing the learning rate drops down the accuracy of the model. The classification report below gives indepth information on the precision of the model for 4 classified species of bacteria.

Classification report

	precision	recall	f1-score	support
0	0.99	0.96	0.97	426
1	0.87	0.78	0.82	282
2	0.99	1.00	1.00	377
3	0.85	0.94	0.89	353
accuracy			0.93	1438
macro avg	0.92	0.92	0.92	1438
weighted avg	0.93	0.93	0.93	1438

Fig: 4.2.3: Classification report for approach 2

The dataset consisted of 2582 training images and 1438 validation images. The classification report shows higher precision value for classes Candida albicans and Lactobacillus paracasei. The precision for classes Clostridium perfringenss and Staphylococcus saprophyticus drops a little as they are similar in shape, but they do differ in size. Below is the confusion matrix to provide how exactly the prediction performed on class level.

Predicted Values / True Values	0	1	2	3
0	409	16	0	1
1	1	221	2	58
2	0	0	377	0
3	4	17	1	331

Table 4.2.1: Confusion Matrix

5. Related Work:

a. This paper provided an approach based on geometric features or cell, they did by segmenting digital results cell images and extracting geometric features for cell classification.(link:https://www.researchgate.net/publication/228845327_Automatic_Classification_of_Bacterial_Cells_in_Digital_Microscopic_Images)

b. This paper proposed a combination of methods which can be used for versatile, efficient processing. (link: https://aem.asm.org/content/aem/64/9/3246.full.pdf)

c. This paper mainly used Hep-2 cells for the identification of antinuclear autoantibodies (ANA). Every entry includes 162 features and used a data mining algorithm to find out the relevant features. (link: https://link.springer.com/chapter/10.1007/3-540-45497-7_33)

6. Limitations:

In our approach we have broken down the high-resolution image into blocks of size 224x224 and then transformed into grayscale, the prediction is solely based on the blocks of images and the model does not notify the prediction on the full-size image (2048x1532). When this model is considered into a real-world application we need to find out a way to break down the images into blocks arriving from the microscope in real time and feed it to the model and generate results for the image in its full size as pathologists needs classification on the image from the microscope but not on their blocks. This limitation leaves us with the opportunity to explore and implement a method to introduce the model in real world application in near future.

7. Closing Remarks:

There are many different approaches to this problem and better chance to make our prediction model more accurate. Current method used stained samples of bacteria colonies and they were observed under oil immersion. We only focused on bacteria classification from sample images of individual bacteria colonies so there was no use case of mixed bacteria sample in single image. We can extend this classification problem to advance level in which we can directly identify bacteria colonies on plate rather than one colony of bacteria. This approach may require different technique in image input as well as data extraction. In terms of data collection, we can use laser beam method to expose colonies and take high resolution pictures. So, data gathering technique may vary and problem can first be to localization of bacteria colonies and on next level we can classify each colony. This research has much brighter scope in detecting human life exitance on Mars or other planet where scientists and researchers believe that life is possible in those environmental conditions.