Malaria Cell Images Classification

Shaojun Yu, Wenjing Ma, Melanie Zhao

1. Objectives:

This project aims to classify red blood cells as uninfected or parasitized by Malaria based on cell images. We would like to use image classification techniques to detect cells contain Malaria or not.

2. Dataset:

The dataset was taken from https://www.kaggle.com/iarunava/cell-images-for-detecting-malaria. It contains a total of 27,558 cell images with equal instances of parasitized and uninfected cells.

3. Environment & Packages used:

- Python-3.6
- Keras
- Sklearn
- Numpy
- Matplotlib

4. Method:
   (1) Transfer cell images to numpy array for parasitized and uninfected cells and add labels
   (2) Normalize data
   (3) Split the data into Training set and Testing set (Ratio:9:1)
   (4) Build CNN model
   (5) Plot Evaluation metrics

5. Model:

```
The summary of the CNN model is:
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_1 (Conv2D)            (None, 50, 50, 16)        208
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 25, 25, 16)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 25, 25, 32)        2080
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 12, 12, 32)        0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 12, 12, 64)        8256
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 6, 6, 64)          0
_________________________________________________________________
dropout_1 (Dropout)          (None, 6, 6, 64)          0
_________________________________________________________________
flatten_1 (Flatten)          (None, 2304)              0
_________________________________________________________________
dense_1 (Dense)              (None, 500)               1152500
_________________________________________________________________
dropout_2 (Dropout)          (None, 500)               0
_________________________________________________________________
dense_2 (Dense)              (None, 2)                 1002
=================================================================
Total params: 1,164,046
Trainable params: 1,164,046
Non-trainable params: 0
```

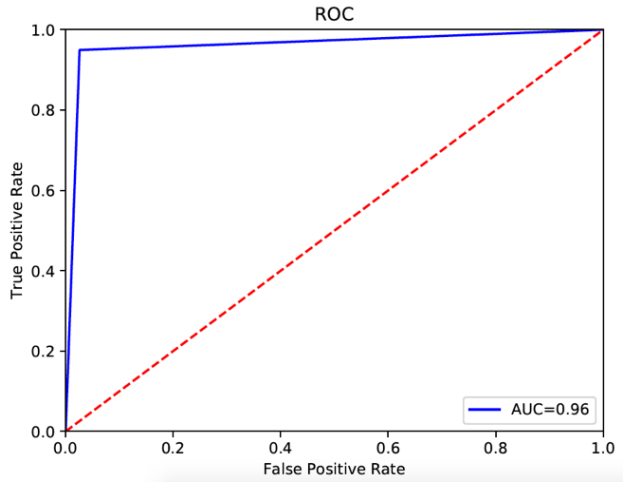
We built a Sequential model which has 3 convolutional layers and 2 dense layers and one 2 dropout layers.

6. Performance:
    1) Accuracy:
       0.9615384615384616
    2) Loss:
       0.2789698895718117
    3) F1 Score:
       0.9618430525557955
    4) Confusion Matrix:
       [[1314 35]
       [71 1336]]

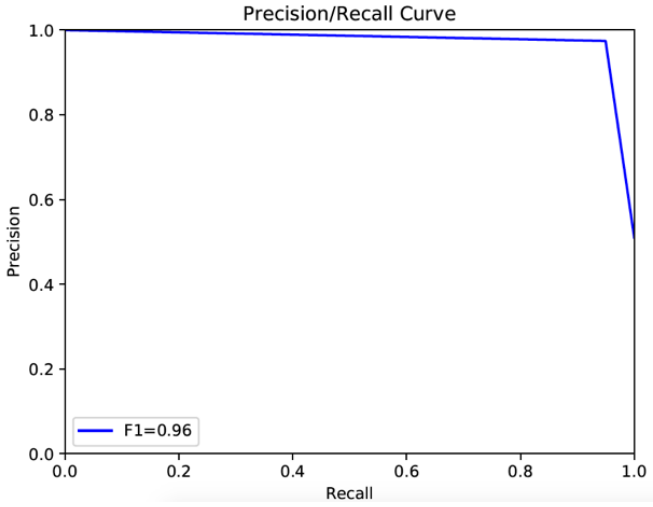
Based on the above data, our model predicts well on the Testing Dataset, which has accuracy of 0.9615 and F1 score of 0.9618.

We also plotted some figures to show the performance.

5) AUC- ROC Curve

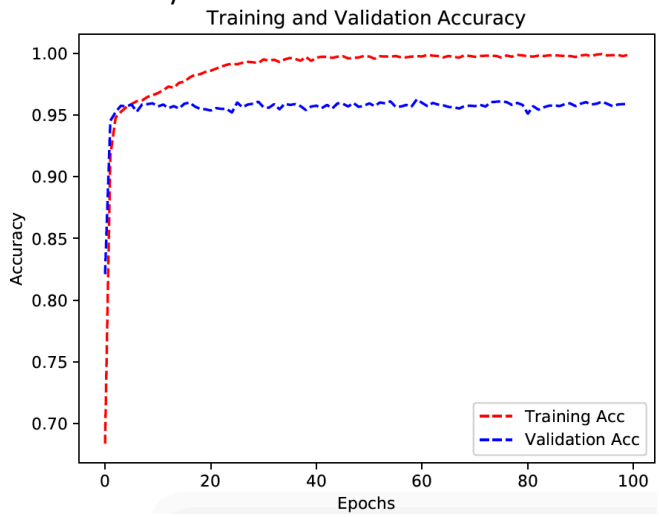


6) PR Curve

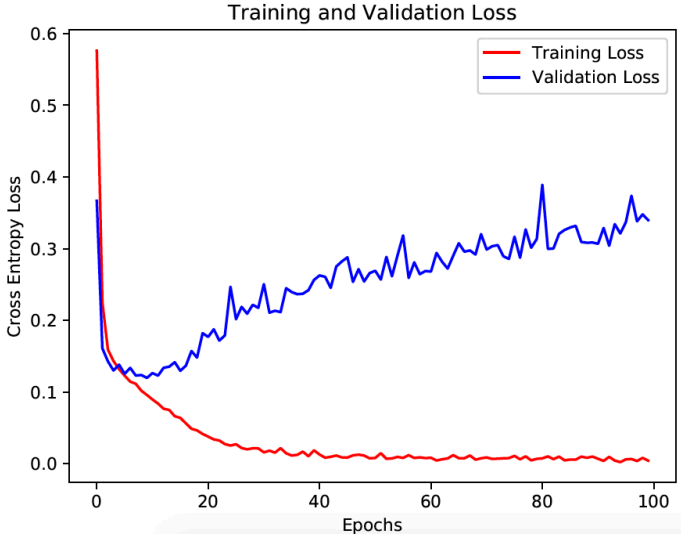


From the ROC and PR Curve, we can see that our model is capable of distinguishing between uninfected cells and parasitized cells.

7) Training vs. Validation Accuracy Plot



8) Training vs. Validation Loss Plot



From the above plot, we can see that after 15-20 epochs, the validation loss starts to increase, which means that our model starts to overfit after 15-20 epochs.

7. Conclusion

From the results above, we can conclude that our model performs well on this classification task, but there are still some overfitting issues need further research. Overall, our model meets the expectation.