XGenomes: Lambda-phage T7-phage Classification Through Supervised Learning

Team 25: Eugene Hsiao, E Chengyuan, Jeff Leong, Jiang Linhui

INTRODUCTION

The goal of our project is to differentiate between two bacteriophages. Figure 1 shows what an example bacteriophage looks like. Specifically, we are to differentiate if a genome sample is a Lambda phage or a T7 phage. The difficulty of this task includes localizing the image and processing it such that phages can be identified clearly.

DATASET

We are provided 5000 sample images of each bacteriophage along with XY coordinates and signal intensities for each image. The processed data directory is divided into the T7 and Lambda directories. Both directories contain 5000 images of the T7 and Lambda phage, respectively. Minimal preprocessing is needed because this is a binary classification problem and we are given two training sets of the two types to be classified.

Our metric of success lies in our accuracy. We are targeting an accuracy of greater than 90% in classifying if an image is a T7 phage or a Lambda phage.

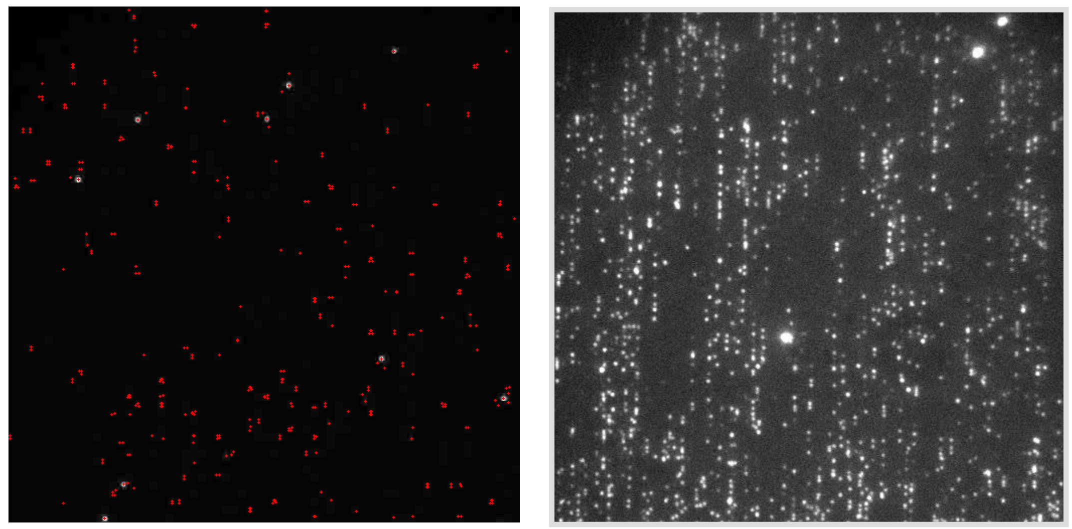


Fig. 1. Sample bacteriophages.The left is a T7 phage labeled to show binding sites. The right image is a Lambda phage unlabeled because intensities are high.

APPROACH

Since our goal is to differentiate between Lambda and T7, we treat this task as a binary classification problem.
It could be described as below: Given x, want to predict $\hat{y} = P(y = 1 | x)$
Input (denoted as x) : images of Lambda and T7. Since each image is 150 pixel * 150 pixel. Each image will be reshaped into a 67500(150*150*3) * 1 column vector, where 3 denotes the RGB channel.

Output (denoted as y) : $\hat{y} = \sigma(w^T x + b)$ $\sigma(z) = \frac{1}{1 + e^{-z}}$. w is column vector which has the same dimension as x and b is constant number. Since both w and x are column vectors, their product will be a number, which could be added by b. Since our output is the probability that x is lambda or T7, it should be in between 0 and 1. Therefore, we use sigmoid function $\sigma(z) = \frac{1}{1 + e^{-z}}$ to keep our output value in between 0 and 1.

Loss function:
We are using binary cross-entropy loss function, which is listed as below. For each individual y, J(w,b) will be smaller when

$$J(w, b) = \sum_{i=1}^{m} y log \hat{y}^i + (1 - y) log(1 - \hat{y}^i)$$

m is the training sample size.

Libraries: Numpy, Matplotlib, PIL and scipy, Tensorflow, Keras

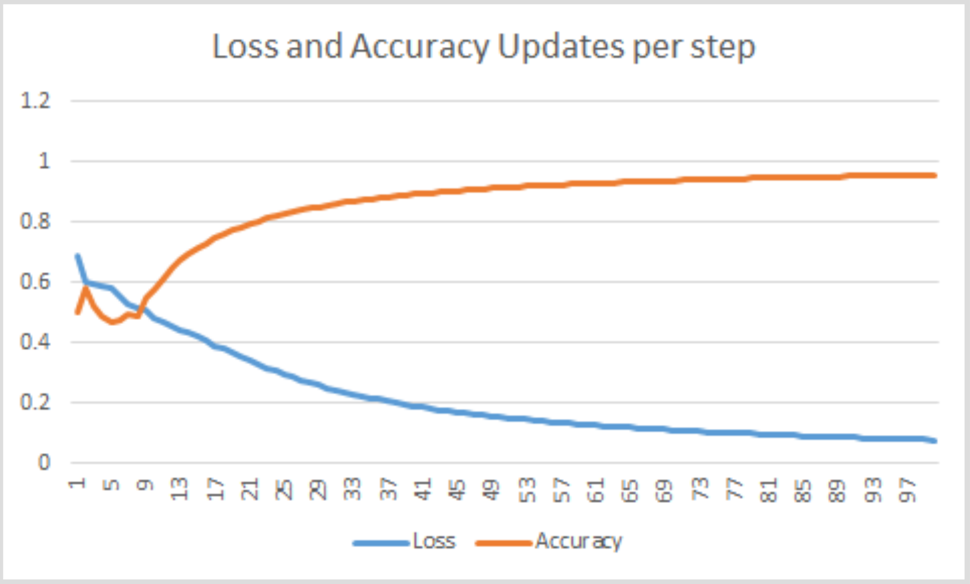
Coding: Import training dataset, Define the model structure, Initialize the model's parameters,
Create a loop for calculating current loss; calculating current gradient and update parameters.
Test with test sample.

Software: Python, Thunderstorm using ImageJ

RESULTS

Steps	Accuracy
1	0.51
20	0.79
50	0.92
90	0.95

Final accuracy 0.9569
Loss 0.0763



METHOD

Activation Function: ReLu
Loss function- Binary-Cross entropy
Optimizer- AdamOptimizer/RMSprop
Metric - Accuracy

We are currently testing the following Hyper Parameters:
Layers: 2,3,4
Learning rate: 1e-4, 1e-3, 1e-5
Activation function: ReLu, Sigmoid
Epochs: 50,100, 1000
Dropout: On vs. Off
Strides: 1,2,4
Padding: On vs. Off

CONCLUSION

As a result of our training, we were able to produce classification a 95.7%, easily beating our metric for success.

This result means that our model is reasonably accurate at distinguishing between the T7 and Lambda phages. Although humans can identify the difference between phages, this model can sort through thousands of images quickly replacing the current slow, manual method.

REFERENCES

Amgarten, Deyvid Emanuel, et al. "MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins." *Frontiers in genetics* 9 (2018): 304.

Wang, Chen, and Yang Xi. "Convolutional Neural Network for Image Classification." *Johns Hopkins University Baltimore, MD* 21218.

Ovesný, Martin et al. "ThunderSTORM: a comprehensive ImageJ plug-in for PALM and STORM data analysis and super-resolution imaging" *Bioinformatics (Oxford, England)* vol. 30,16 (2014): 2389-90.