Diagnose Pneumonia with Convolution Neural Network

Yingru Feng, Liangyu Zhao

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

Acknowledgement

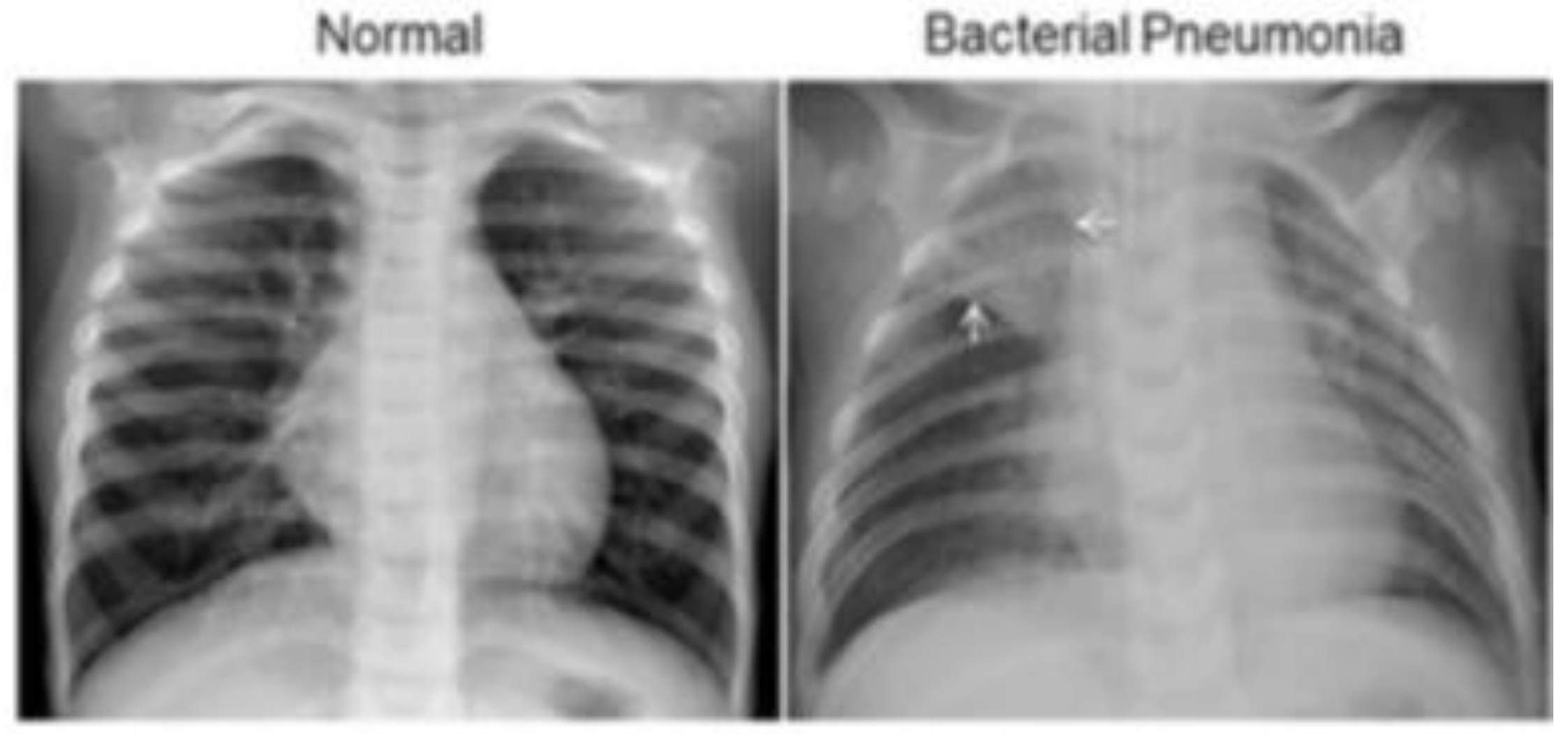
1. Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification, https://data.mendeley.com/datasets/rscbjbr9sj/2
2. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Le arning, http://www.cell.com/cell/fulltext/S0092-8674(18)30154-5
3. Rethinking the Inception Architecture for Computer Vision, https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.pdf

I. Background

Inspired by the paper we found named "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning" published earlier this year on Cell about the breakthroughs of utilizing Machine Learning on certain medical field, our team decided to do classification task on the chest X-ray to be and most interesting and most relevant to what we've learn in the class since will have the opportunity to practice optimization of training convolutional network on real life dataset.

II. Dataset

The dataset contains around 5000 Chest X-ray images of pediatric patients from one to five years old. The images are classified into two classes: the ones with Pneumonia and the normal. They were pre-selected for quality control and all the images are human readable (for doctors) and variable sizes.



VI. Future Work: Balance Classes

In our dataset, the sizes of two classes are imbalanced with pneumonia having 3875 images and normal having only 1341 images. In our exploration, we have only tried simple oversampling the normal data to balance the classes; however, the result is not ideal. The network overfits quickly, which is a clear disadvantage of oversampling technique. There are many other more advanced balancing techniques existing, such as anomaly detection algorithms. Here, we do not have enough time to do experiment on them, but in our opinions, balancing classes has a great potential in further improving the performance of our network.

VII. Conclusion

We used data augmentation and unfrozen higher layers of the CNN to obtain an accuracy over 90% on our test dataset. Because we used pretrained network, the network easily overfit in our training. Our exploration is a continuous combat versus overfitting. If better techniques to prevent overfitting are in use, we believe the network has great potential to have better performance.

III. Choosing Pretrained CNN for Feature Extraction

Our first step was finding a CNN model that suits the task. For every CNN, there are two groups of layers, the feature extraction layers and the linear classification layers based on the extracted features. The feature extraction layers are often the more complex ones in a CNN compared with the linear classification layers. At the same time, the feature extraction layers are much more reusable than the linear classification layers since most of the image recognition tasks depends on similar lower level features. Therefore, we decided to start with using a pretrained version of renowned convolutional network for feature extraction. There were four popular CNNs in our consideration: AlexNet, DenseNet, Inception V3, and ResNet. These networks all have strong feature extraction capabilities. Since we need to choose one from them we need to do some testing to make comparison of their performance on the dataset. However, one issue we encountered is that these network have very different linear classification layers. In order to do a fair comparison of the networks on feature extraction, we decided to replaced the original classifier layers of all four CNNs with our own classifier layers consists of linear and dropout layers. We freezed the convolution layers of these networks in training so that they stay on their pretrained values and only trained the classifier layers. With learning rate 1e-5 and 30 epochs, we got the following results:

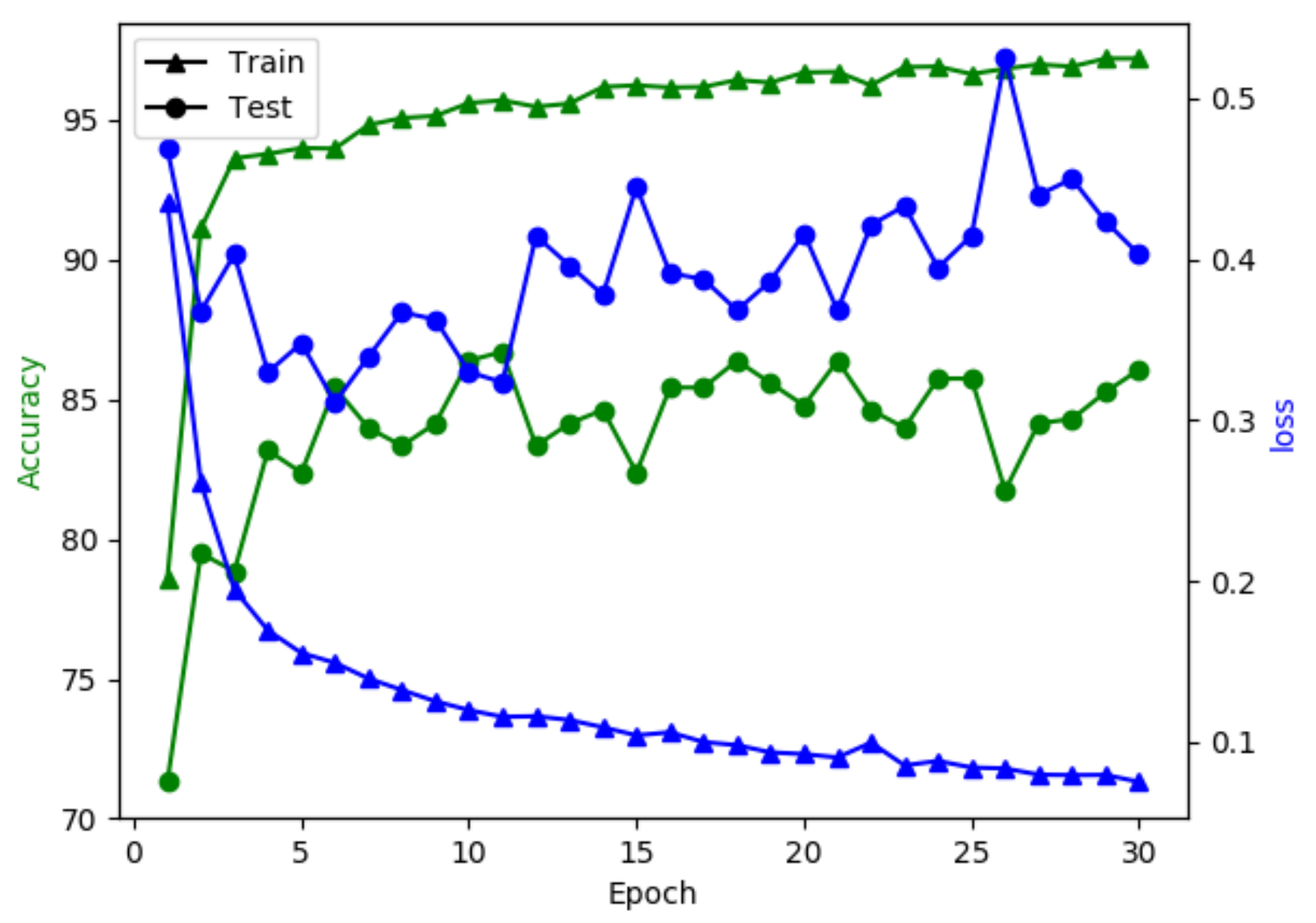


Figure 1.DenseNet

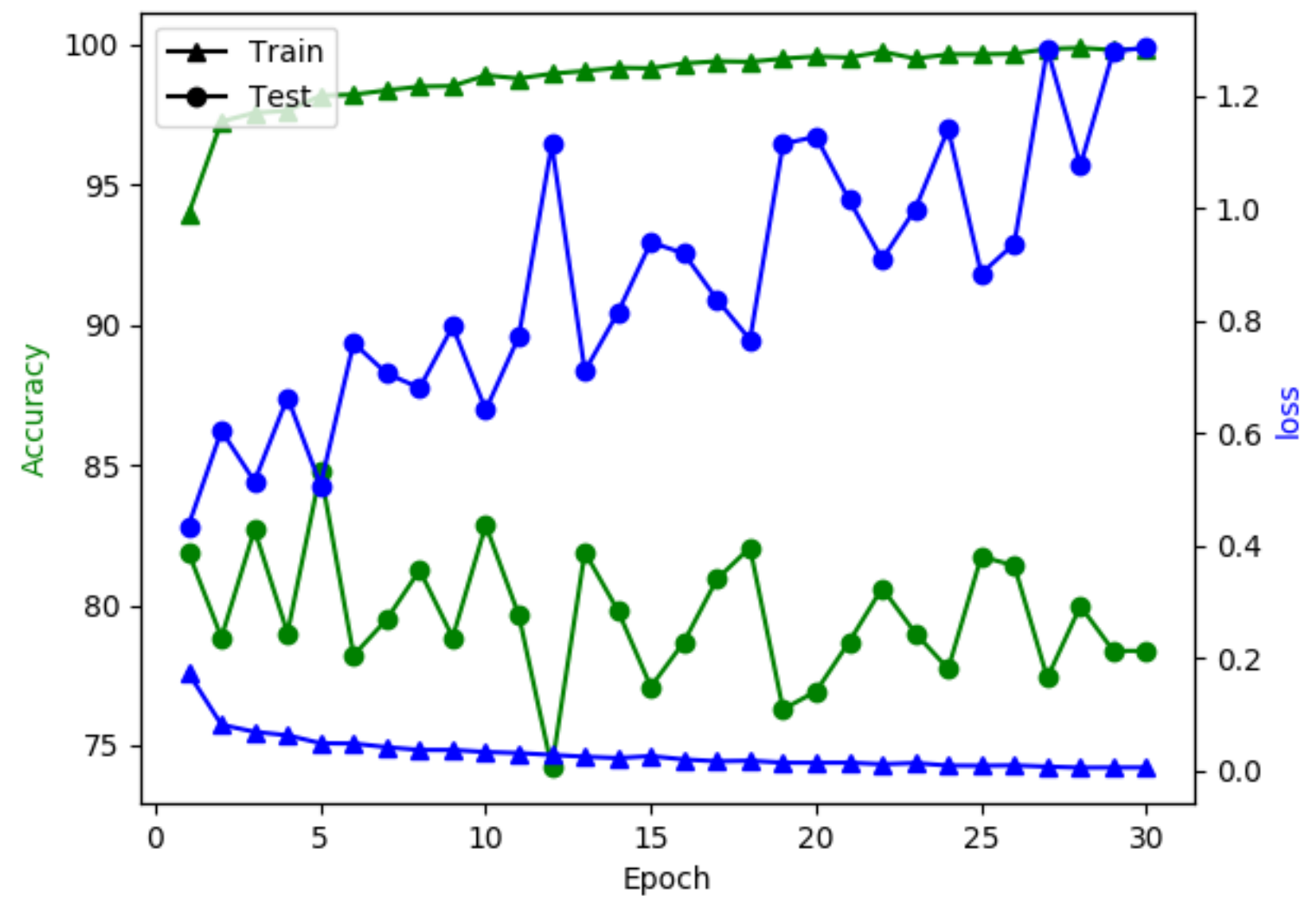


Figure 2. AlexNet

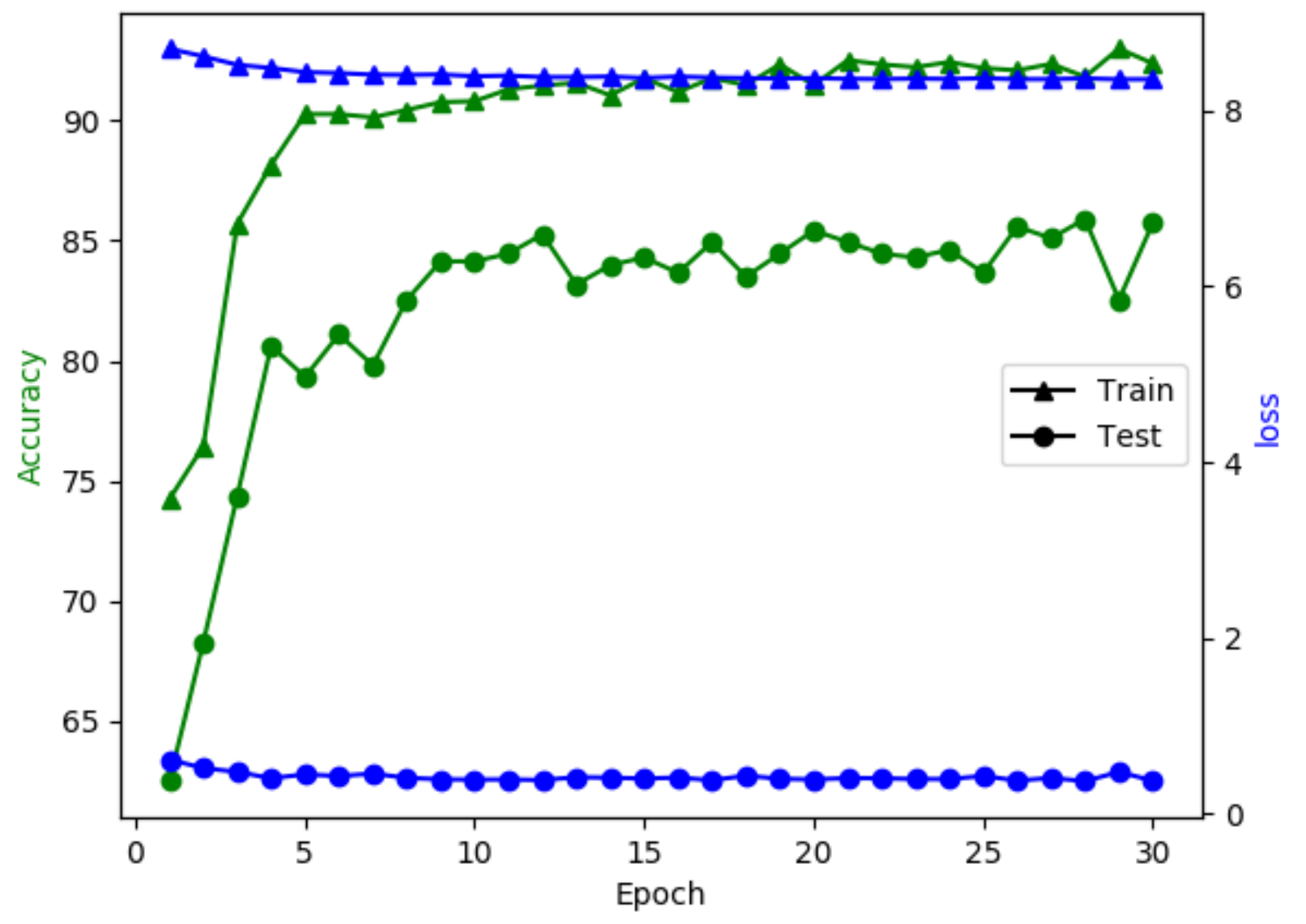


Figure 3. Inception V3

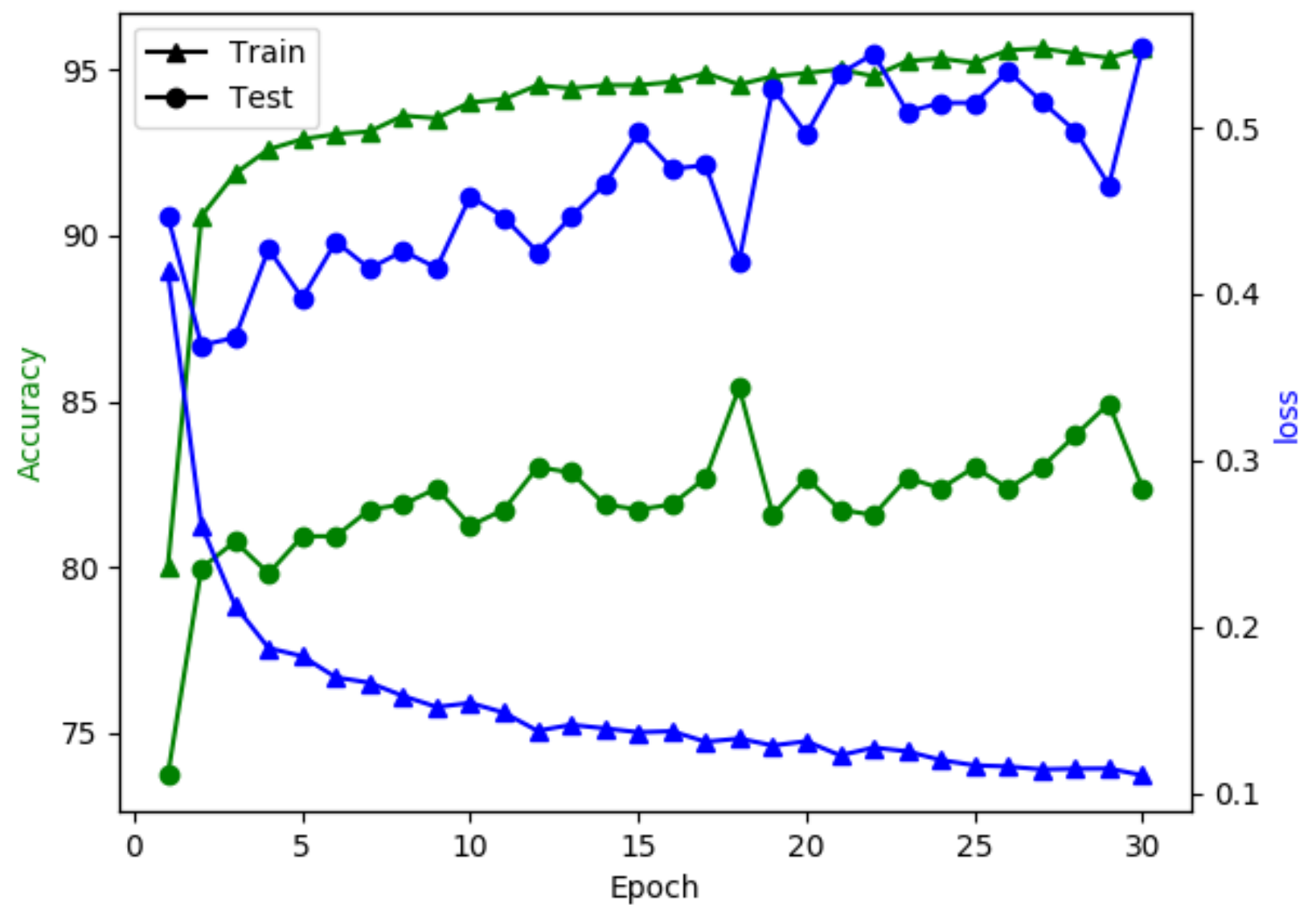


Figure 4. ResNet

As we can see from the results, all networks have similar performance in terms of test accuracy.

However, the test loss of Inception V3 is very stable compared with the other models and keeps on decrease continuously as we ran through the Epochs. The increment of test losses of all the other models is a clear sign that these networks were overfitting. This made Inception v3 standout as we saw its best potential in avoiding overfitting. After these results, we decide to continue our optimization of CNN with Inception V3 network.

IV. Data Augmentation

In order to avoid the overfitting problem, we decided to do some data augmentation on our dataset. Because our training data and target data are all chest X-ray images, it is impossible for target images to be too different from training images in terms of certain characteristics. Therefore, we need to be careful when choosing data augmentation techniques. Some techniques like random vertical flip and color Jitter are not suitable in our case. We ended up with a transform with random horizontal flip, random affine and random crop. Since human lungs are generally symmetric, we use horizontal flip in the data augmentation. Besides the horizontal flip, we have also used random affine with relatively small parameters, because we do not want to change the shape and direction of the X-ray images too much. With data augmentation, the result ends up with a slight improvement on test accuracy. Without data augmentation, test accuracy becomes stable around 83%; with data augmentation, test accuracy becomes stable around 85%.

V. Unfreeze Pretrained Layers

With data augmentation, our network becomes more robust against overfitting. That leaves us room to decide if we want to unfreeze some convolutional layers in training. Lower layers of convolution networks are recognizing edges and shapes of the image, and only the higher layers of the networks are recognizing the overall contents of the image. Because our dataset has several key differences from the ImageNet dataset which the pretrained networks were trained to classify, we may need to retrain higher layers. Note that the lower layers of the pretrained network are still useful, because the edge and shape recognition should be the same across different images. Therefore, we now try to gradually unfreeze higher layers of the Inception network and compare the results. Based on our training result before, we found that the networks generally converge very fast within 5 epochs. This is also what we expected with pretrained network. When we unfroze 9 modules from the pretrained layers, the test accuracy is above 90%. The improvement is significant.

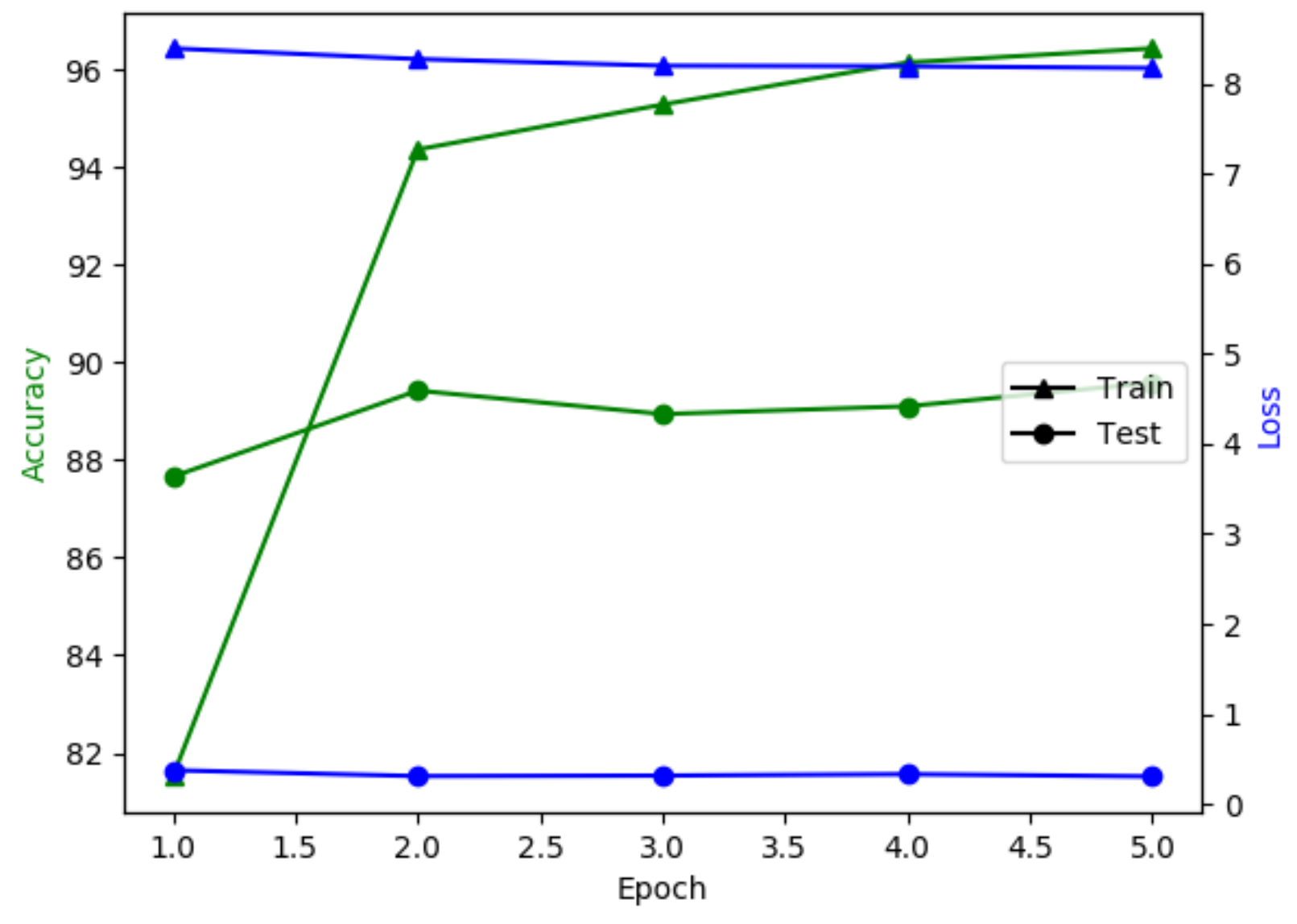


Figure 1. Unfreeze 3 modules with data augmentation

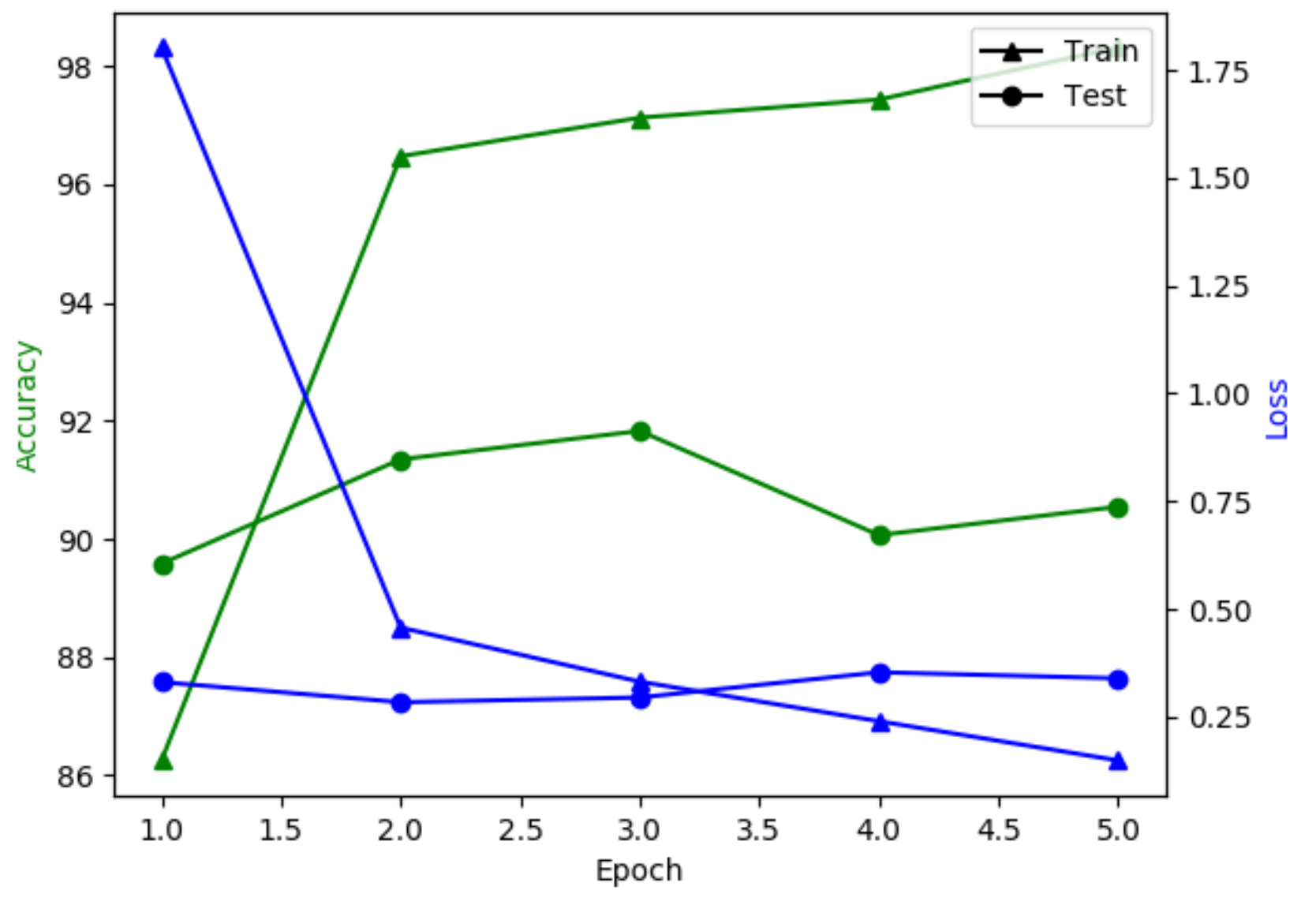


Figure 2. Unfreeze 9 modules with data augmentation

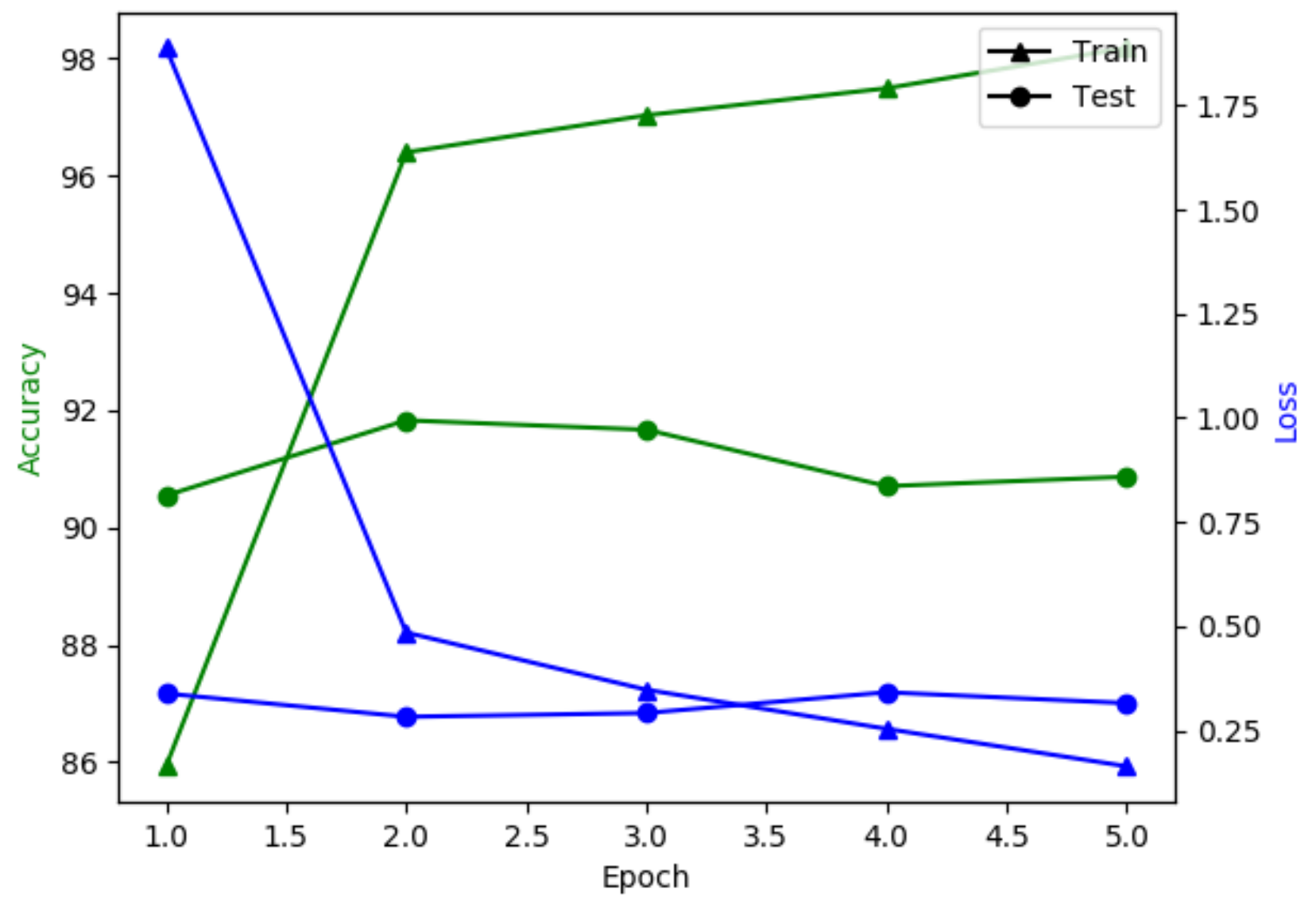


Figure 3. Unfreeze 12 modules with data augmentation