Deep Learning

880663-M-6

Assignment

Using Deep Learning to Perform Multi-Class Classification on the Lung and Colon Cancer Histopathological Image Dataset (LC25000)

Report by:

Javier Torralba (2042878)

Note: all codes, visualizations, models and results can be found under "assignment_DL_Javier_Torralba", submitted on canvas

March 2024

1. Problem Definition

Histopathology is the microscopic examination of tissue to detect diseases, especially cancer. These images are studies of diseased cells and tissues with a microscope. With the advancements in deep neural networks, it is now possible for computers to analyze these images and detect different types of tumors. In this report, I focus on using a Convolutional Neural Network (CNN) to evaluate these images and classify different types of lungs and colon cancer.

2. Exploratory Data Analysis

The data used comes from a skin cancer database (Skin Cancer, 2024). The original dataset consists of 250 benign lung tissue, 250 lung adenocarcinomas, and 250 lungs squamous along with 250 benign colon tissue and 250 colon adenocarcinomas cell carcinomas. The images have been augmented to a total of 25,000 images of five types of colons and lung cancer. The images were resized into 120x120 pixels for computational workload. In figure 1 you can find 15 random images from the dataset,

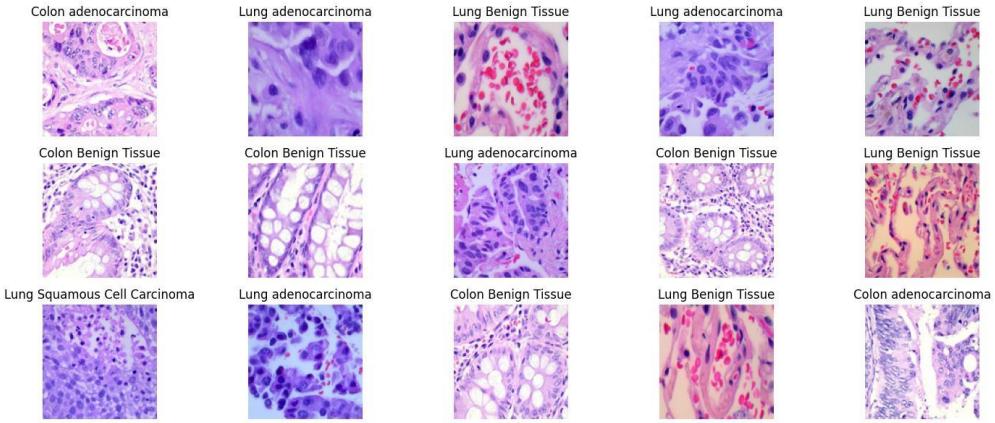

Figure 1

A distribution of classes was also performed to determine how many images per category are present. In figure 2 you can see the visualization with these results.

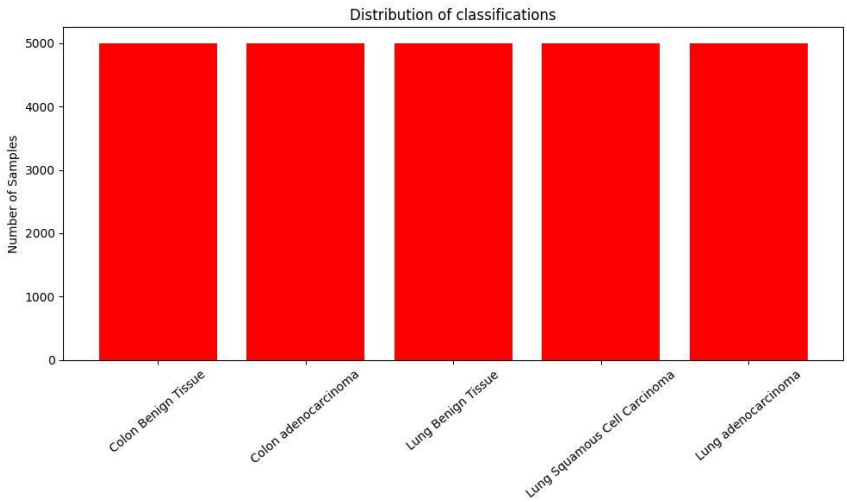

Figure 2

After analyzing the distribution of the classes, one hot encoding was performed on the image labels. There is an even distribution of all five categories on the dataset.

3. Results of the Baseline Model

Ahead you can find results for the baseline model for accuracy and loss on the validation set. The validation accuracy peaks at about 0.75 on the 7th epoch and the best loss is of around 0.5 with a validation loss of 0.5. Figure 3 shows the results. This model is likely underfitting, as it does not have a very high performance on either the validation accuracy or loss.

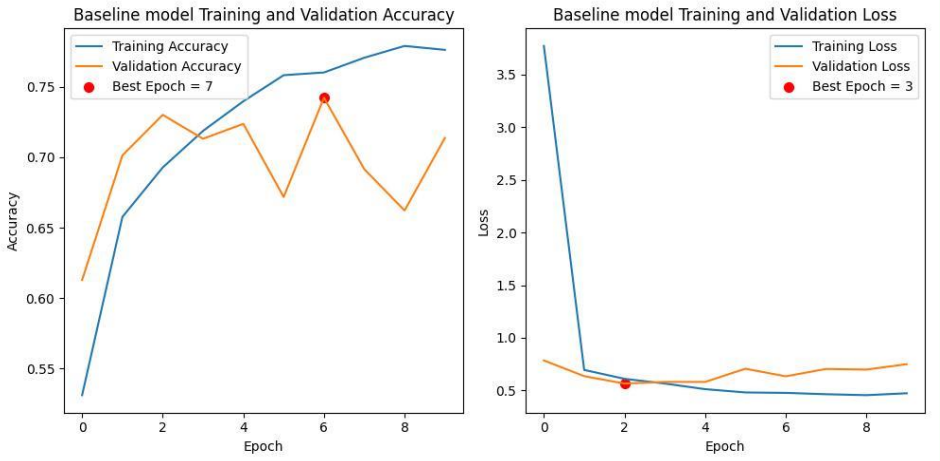

Figure 3

Similarly, you can find the confusion matrix and ROC curves with the relevant AUC for the test set. The model seems to have a high accuracy classifying lung benign tissue, but struggles differentiating between the two types of lung carcinomas. Similarly, the model seems to have trouble differentiating between carcinomas and benign tissues on the colon, as one can see on figures 4 and 5. Figure 4 shows the performance on the validation set and figure 6 shows the performance on the test set.

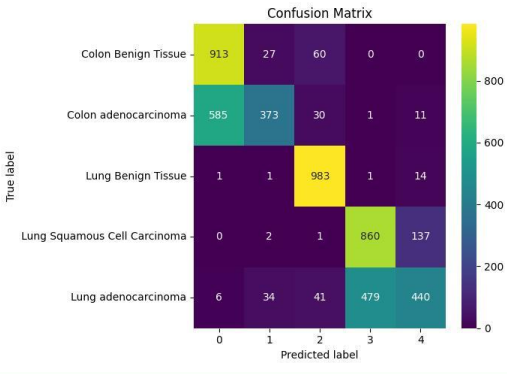

Figure 4

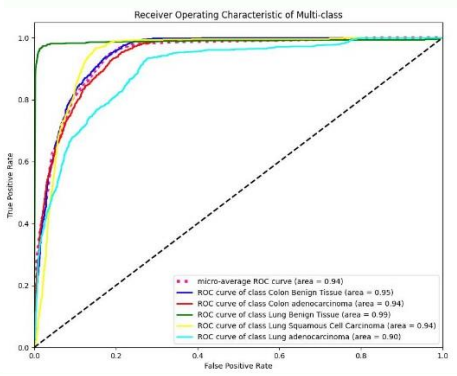

Figure 5

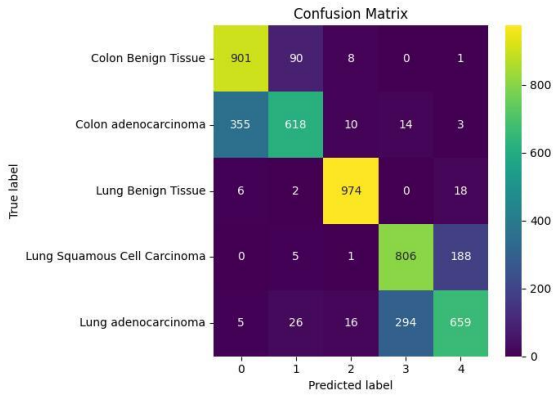

Figure 6

4. Improved (Fine-tuned) Model and Its Results

I explored several finetuning of hyperparameters to improve the baseline model. I took inspiration from several sources from deep learning and computer vision studies to make decisions. Since the baseline model seemed to be underfitting with a low validation and training accuracy, at a high level, I aimed to make the model train as best as possible and then adding counter measures to control overfitting. All the steps taken below can be found under the notebook "Assignment_DL_Javier_Torralba", with each hyperparameter change as "iteration n" where n is the number of iteration performed.

First, the paper by Krizhevsky et al. (2017) uses a CNN with five neural convolutional layers and three fully connected layers. Similarly, Ben Hamida et al. (2021) evaluate different models for histopathological image classification, and all the models they evaluate have at least 5 convolutional layers. In contrast to the baseline which had two convolutional layers, I decided to add three additional convolutional layers (and further into the iterations another additional convolutional layer) with the first two layers being followed by Max Pooling layers as well as the last one. Adding these layers aims to counter the underfitting in the baseline model. This helps better capture the complexities of the images and it proved to be effective, bringing up the validation accuracy from 0.75 to 0.94 and a validation loss of 0.15.

Second, taking inspiration from the Krizhevsky et al. (2017), I replicated the order of the pooling layers and followed their logic on overlapping pooling. The baseline model uses traditional local pooling, while their model uses overlapping pooling, which means the pool size (3x3) is bigger than the strides it takes (2x2). The overlapping pooling helps preserve fine-grained details in the feature map. Considering we are using very detailed histopathological images this was a sensible hyperparameter to experiment with. However, the performance compared to the pervious iteration of the model was very similar, but with more variability on the accuracy. Using the Ocan Razor principle of simplicity, I decided to not use the overlapping pooling.

Third, given that the baseline model and other iterations had 10 epochs, and the model seemed to still be improving around that number. Also, Ben Hamida et al. (2021) repeatedly mentions that some of the networks being trained had 20 or 30 epochs, signaling that more epochs could enhance the performance of the model. I decided to increase the number of epochs to 100 and include early stopping with a patience of 5. This helped with further improving on the training data, but not on the validation data, with similar results to the first iteration, showing signs of overfitting. Keeping the early stopping allows to gauge until what

point is it relevant to keep training the model, hence I kept this hyperparameter change for future iterations.

Fourth, noticing that the model was overfitting with the training data, I decided to include dropout with a probability of 0.5 in the first two fully connected layers, taking inspiration from Krizhevsky et al. (2012). Dropout does not allow some hidden neurons from participating in back propagation, forcing the model to learn more robust features that are useful all together. This proves useful in fighting overfitting, with similar results in the loss and accuracy of validation and training data. These showed to be helpful, as the training and validation losses and accuracies started returning similar results, hinting at a control of overfitting.

Fifth, seeing that the model could still see improvements, I decided to increase and restructure the number of filters in the model. Taking inspiration from the paper Zankoya Dihuk et al. (2020) in which they study the ideal number of filter numbers and order of them in the convolutional layers, I changed the number of filters to 128 and 256 with less filters in the early layers and more in the later ones. The advantages of including less filters early on and more later are that you can capture low-level features early and then increase the attention to detail in the image, which helped the model achieve a test accuracy of 95.8% and a loss of 0.13.

Sixth, (Levi & Hassner, 2020) argue for using larger filter sizes on the first convolutional layers and having smaller filters when you go deeper into the layers. This allows the model to focus on more general shapes and textures early and then focus on finer details deeper into the model. However, the model did not perform well and dropped accuracy by 10%. This could've happened because the model did not learn the relevant features early on, leading to the finer details being missed later in the model. Because of the drop in performance, I decided not to include this change in the final enhanced model.

Altogether, the final model achieved an accuracy of 96.80% on the test set with a loss of 0.1. Figure 7 shows the visualizations of the loss and accuracy performance on validation and training and figures 8 and 9 show the confusion and ROC curve accordingly.

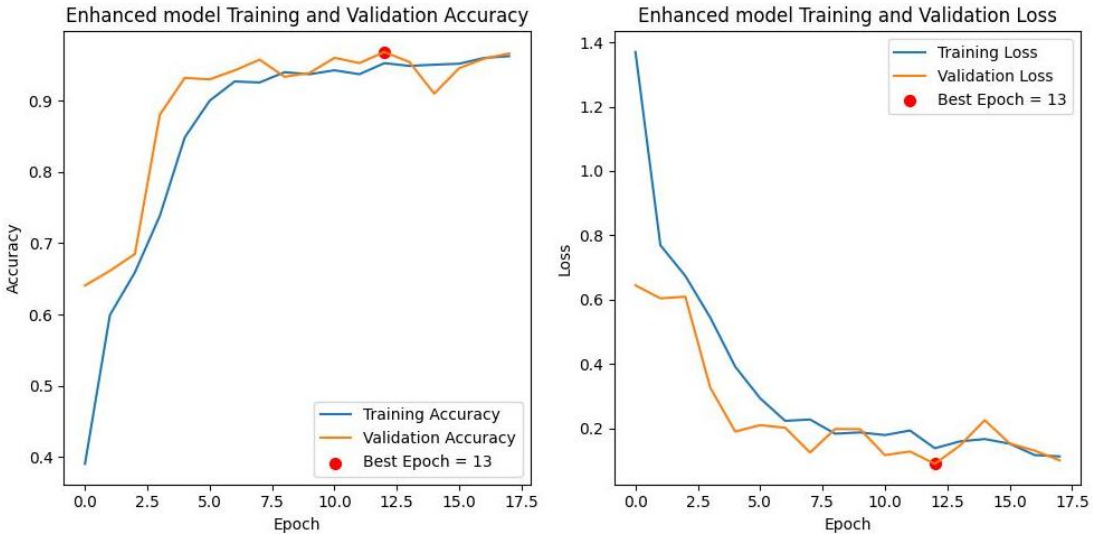

Figure 7

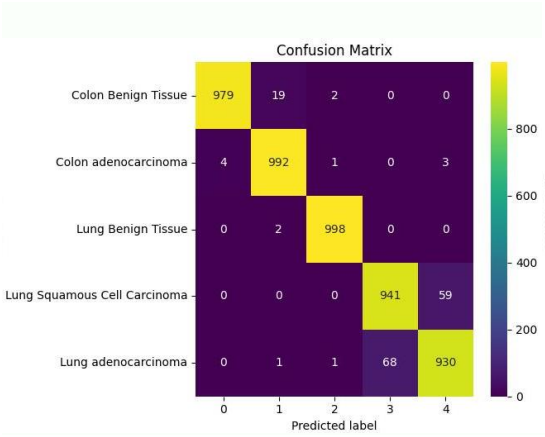

Figure 8

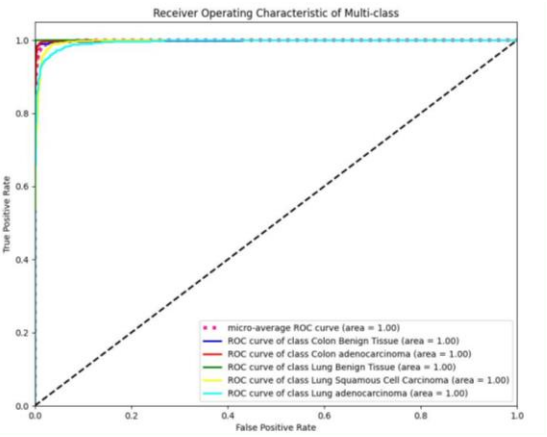

Figure 9

5. Transfer Learning Model and Its Results

The transfer learning model I used is the ResNet50, with the same number of dense layers and sizes as the baseline model. Ben Hamida et al. (2021) used pretrained models on histopathological images and concluded that a different ResNet model (ResNet18) performed the best among other pre-trained models (96% after fine tuning), which inspired me to use this model. Given the knowledge from the literature and the enhanced model performing better with deeper neural networks, choosing to use a ResNet model with 50 layers seemed like a logical choice, enabling to pick up complex patterns and test whether a deeper neural network would contribute to the accuracy and validation loss. The best validation accuracy achieved a 98%, and the best validation loss achieved 0.06 as shown in figure 10.

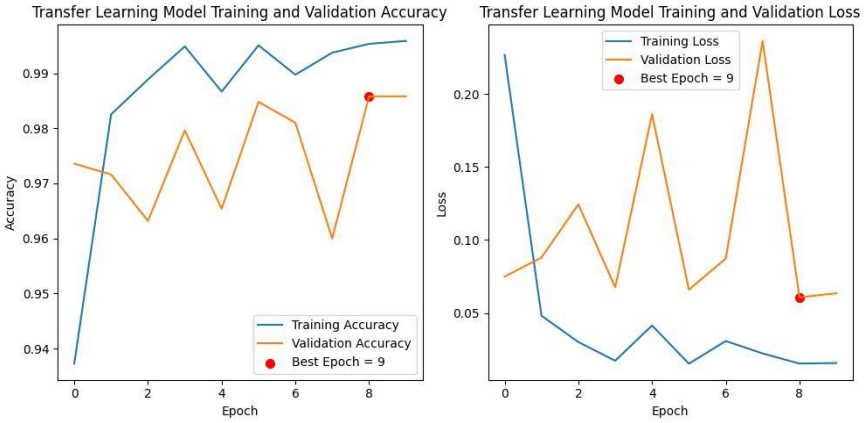

Figure 10

Similar to the baseline model, the transfer learning model is relatively accurate on lung benign tissue and colon classification. Like the baseline model, the transfer learning model seems to perform relatively worse on the classification of lung carcinomas. Yet, the transfer learning model seems to perform best out of all the models we have presented. Figures 11 and 12 show the ROC curve and confusion matrix respectively.

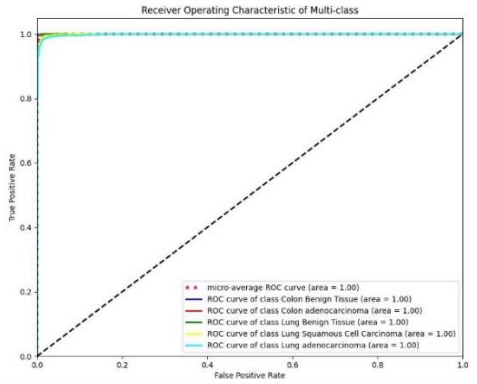

Figure 11

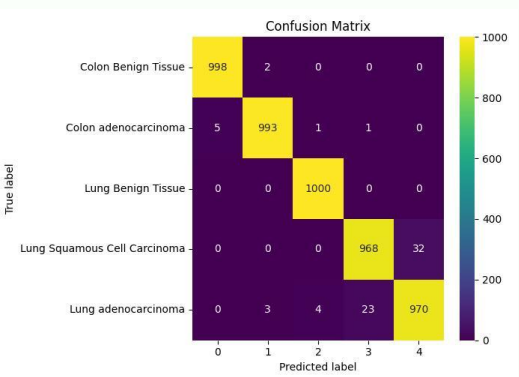

Figure 12

5. Discussion

Analyzing the three models performed, we can summarize and compare their performance. Firstly, the baseline model clearly underfitted the data. Several hyperparameters could be tuned to increase the performance of the model, as it achieved 75% accuracy on the validation set and a loss of 0.74. Several ideas on how to increase the performance of this model came from existing literature on CNNs. Secondly, I explored these changes by tunning them on the enhanced model. The most effective changes were adding layers to the neural network to fight underfitting, changing the number of filters to guide the model on what to focus on, use measures against overfitting using dropout on the dense layers, and increasing the number of epochs all contributed to a higher accuracy test score of 95.8% and a loss of 0.1. Thirdly, I used transfer learning to evaluate a pretrained model in the form of a ResNet50 to evaluate its performance. This model replaced the convolutional layers of the baseline model and performed best, with a test set accuracy of 98.52% and a loss of 0.06. Seeing this high performance without any hyperparameter tunning suggests that changes could be made to this model to achieve an even better performance.

While the enhanced model proved to be an improvement over the baseline, there could still be even more hyperparameter tunning to improve the performance of the network. Even though taking inspiration from different sources leads to the hyperparameter tunning, such as increasing the size of filters (Levi & Hassner, 2020), number of filters (Zankoya Dihuk et al, 2020), the number of layers, including a dropout rate, altering the Max Pooling overlap (Krizhevsky et al., 2017), and increasing the number of epochs all contributed to a more robust model, the magnitude and combination of this hyperparameters could be further tested to empower the performance of the enhanced model even more. Likewise, further changes could be made based on the domain of histopathological images. For example, following models like the SegNet model (Ben Hamida et al., 2021) could further increase the performance of classifying this type of images.

When analyzing the performance of the best model, the transfer learning with a ResNet50, there are several improvements that could be done to this model to have an even better performance and generalization. Given the high training accuracy (99%), measures against overfitting could be tested. For instance, using a dropout rate like Krizhevsky et al. (2017), or exploring other measures like L2 regularization, could prevent the model from overfitting and give even better generalization. Also, increasing the number of epochs and including early stopping could contribute to creating a robust and generalizable model. In future research, the combination of these tools against overfitting and better fitting of the data in more epochs could contribute to an even better performance on the classification of histopathological images.

6. References

Ben Hamida, A., Devanne, M., Weber, J., Truntzer, C., Derangère, V., Ghiringhelli, F., Forestier, G., & Wemmert, C. (2021). Deep learning for colon cancer histopathological images analysis. *Computers in Biology and Medicine*, *136*. https://doi.org/10.1016/j.compbiomed.2021.104730

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). *ImageNet Classification with Deep Convolutional Neural Networks*. http://code.google.com/p/cuda-convnet/

Levi, G., & Hassner, T. (2020). *Age and Gender Classification using Convolutional Neural Networks*. www.openu.ac.

Skin Cancer Dataset: https://academictorrents.com/details/7a638ed187a6180fd6e464b3666a6ea0499af4af, accessed on 8/03/2024

Zankoya Dihuk, Institute of Electrical and Electronics Engineers, & Institute of Electrical and Electronics Engineers. Iraq Section. (2020). *Proceeding of the 2020 International Conference on Computer Science and Software Engineering : 16-18 April 2020, University of Duhok, Duhok, Kurdistan Region, Iraq*.