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:

Harry Averkiadis (2123340)

March 2024

1. Problem Definition

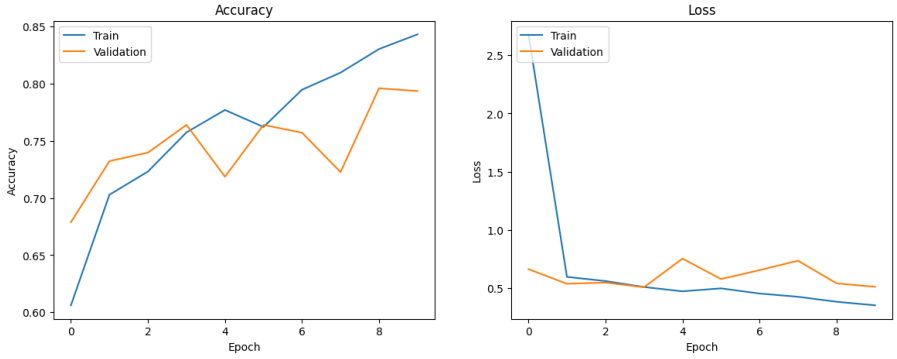
The goal of this project is to evaluate the effectiveness of Convolutional Neural Networks (CNNs) in identifying and diagnosing various lung and colon cancers from images, framing this as an image classification problem. We have a dataset of 25,000 histopathological images, divided into five categories: Lung Benign Tissue, Lung Adenocarcinoma, Lung Squamous Cell Carcinoma, Colon Adenocarcinoma, and Colon Benign Tissue. Our approach starts with developing a basic CNN baseline model. We will then focus on improving its accuracy through fine-tuning and explore the benefits of transfer learning by using pre-trained models to enhance our results.

2. Exploratory Data Analysis

During the Exploratory Data Analysis (EDA) phase, we undertook key steps to prepare and scrutinize the dataset for model training. The process began with data preprocessing, where we resized the dataset images to a consistent 120x120 pixels using the PIL library, enhancing both consistency and computational efficiency. Each image was then converted into a numpy array, with labels assigned based on directory names to categorize them into five distinct groups. These processed images and labels were saved as numpy arrays for convenient future access. To facilitate effective classification, we employed one-hot encoding on the categorical labels, transforming them into binary vectors using LabelEncoder and one-hot encoding, thus allowing the model to recognize each label as a separate category without inferring any inherent order. We then divided the dataset into training, validation, and testing sets, following a 60%, 20%, and 20% distribution, respectively. This stratified split was critical to ensure a representative class distribution across each set, which is pivotal for an unbiased model assessment. Additionally, we performed a visual analysis of sample images and their class distributions to gain a deeper understanding of the dataset. This step was instrumental in identifying class characteristics and verifying the absence of significant imbalances, thereby guiding the modeling strategy.

3. Results of the Baseline Model

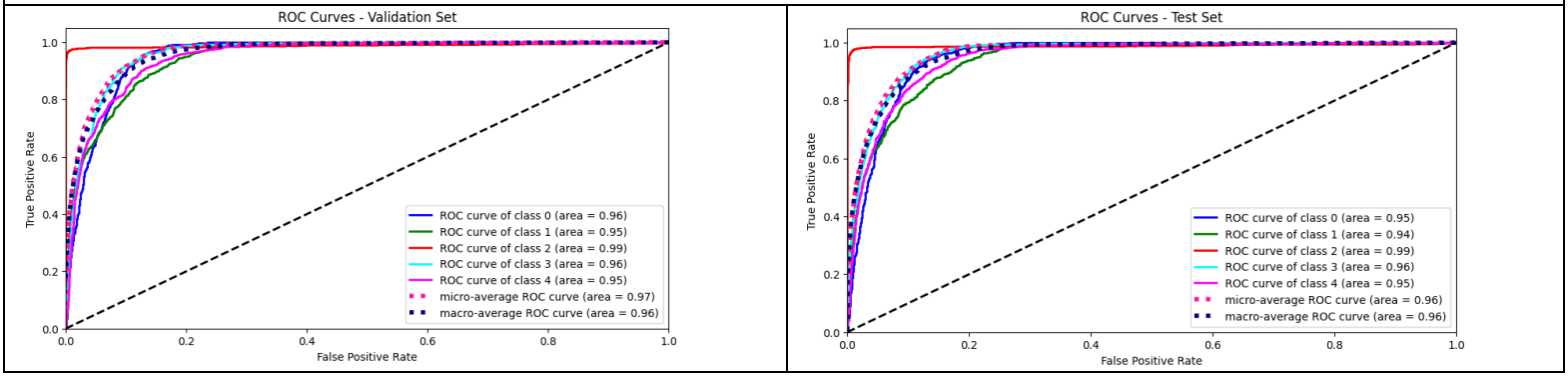
Given that the assignment provides a comprehensive summary of the baseline model, we will not delve into its detailed description. Instead, we will focus on presenting the essential information associated with it, specifically the model's performance metrics from the evaluation phase. To begin, we will illustrate the model's dynamics through graphs depicting its accuracy and loss:



Validation set	Accuracy=0.793	Precision=0.797	Recall=0.793	F1 Score =0.793
Test Set	Accuracy=0.782	Precision= 0.788	Recall= 0.784	F1 Score=0.783

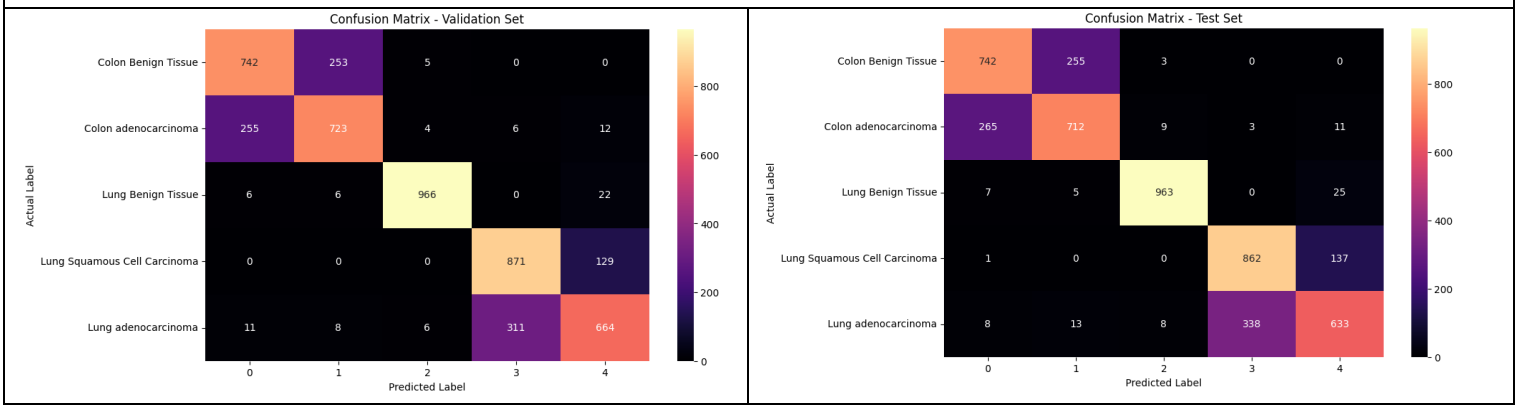
We should also note that for the loss function we used softmax, as this function excel on multiclass classification tasks. Additionally, our assessment of the validation and test dataset highlighted areas for improvement in our baseline model, notably the moderate precision score, crucial for medical diagnostics where false positives carry significant consequences. Notably, signs of overfitting become evident after approximately the 3rd epoch, marked by an increasing evaluation loss alongside a decreasing training loss, suggesting the model's improved performance on the training data does not generalize as well to unseen data. Also there is a divergence between training and validation lines after the 3rd epoch in accuracy, suggesting overfitting. We continue our analysis by presenting the ROC curves for both validation and test sets which can show us the model's performance across the different classes.

ROC Curves – Baseline Model



The ROC curves indicate a good performance with consistent results across validation and test sets. Specifically, our high value of the area suggests a good class separability. Lastly, we provide the confusion matrices of validation and test set:

Confusion Matrices – Baseline Model



The matrices highlight a significant number of misclassifications. This indicates that our current model struggles to accurately differentiate between categories, suggesting it may lack the complexity needed to capture more intricate patterns and information.

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

To enhance the baseline model, our initial strategy targeted the overfitting of our baseline model. We experimented with integrating dropout layers and implementing

L2 weight regularization, which are popular methods for coping with overfitting. Among the regularization factors tested—0.001, 0.0001, and 0.00001—the factor of 0.0001 proved most effective. Dropout rates of 0.4 and 0.3 were also evaluated but did not outperform the L2 regularization approach.

Further optimization involved extending the number of epochs from 10 to 20, coupled with the implementation of early stopping, employing a patience of 3 epochs. This allowed us to monitor performance on the validation set over a greater number of epochs, halting training when no improvement was observed.

Adjustments to the Adam optimizer's learning rate were also explored, with 0.0001 emerging as the optimal value. The intuition behind adjusting the learning rate was that although a smaller learning rate might make it more likely to get stuck in local minima, such occurrences are infrequent in the high-dimensional landscapes characteristic of deep learning models. Instead, saddle points are more prevalent (Dube, 2018). The chosen small learning rate thus provides a more precise navigation through these saddle points.

With a robust foundation formed by L2 regularization, an optimal learning rate, and early stopping, we sought to capture more information by increasing the model's depth. More layers in a CNN allow for the extraction of increasingly abstract and complex features from the input data (Alzubaidi, 2021). In our case the addition of just one convolutional layer before each max pooling step significantly boosted test accuracy from 0.89 to 0.92.

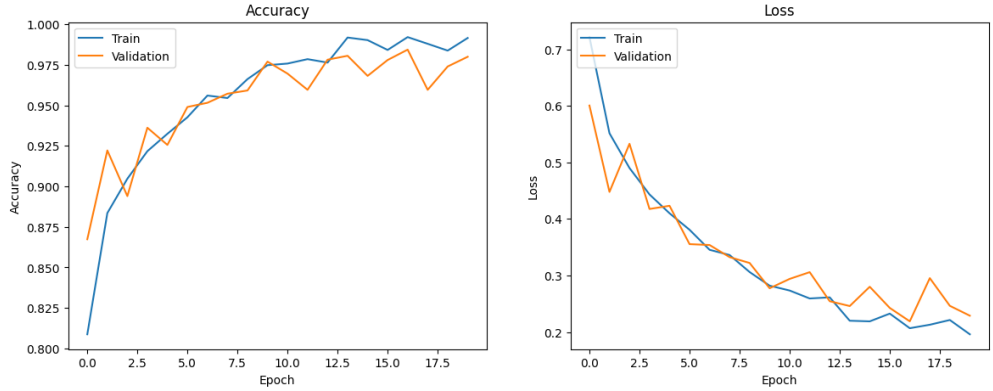
Building on this success, we introduced two more sequences of Conv2D layers followed by max pooling, leading to further accuracy gains on the test data. We then fine-tuned the model's complexity by varying the number of neurons in each layer, adopting a pyramid-like architecture—starting with fewer neurons, increasing them through the network, and then decreasing again. This design leveraged the capacity of deep learning models to extract increasingly complex features at deeper levels, resulting in improved test accuracy (Yu, 2018).

Despite the substantial progress made, achieving a test accuracy of 0.98, we continued refining the model. We tested larger batch sizes and alternative activation functions. While increasing the batch size to 40 did not yield better results, replacing the ReLU activation function with the tanh function expedited convergence and further enhanced accuracy by little.

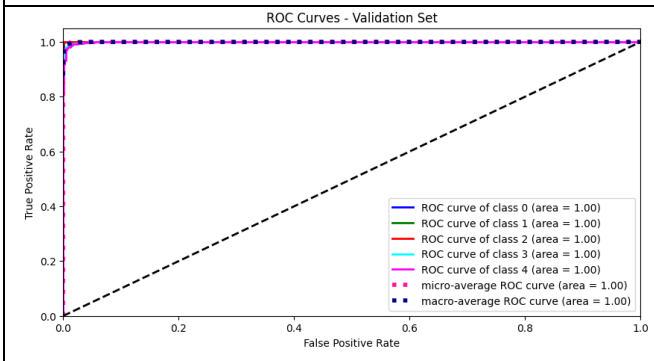
Lastly, we contemplated the addition of another fully connected dense layer, hypothesizing that increased model complexity could translate to better performance, as it has done so far. However, this change did not result in further improvements. Considering all that has been mentioned, our final model summary is the following:

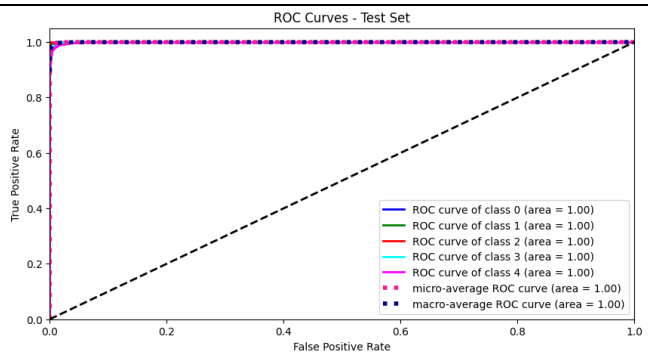
```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d (Conv2D)             (None, 120, 120, 64)      1792

 conv2d_1 (Conv2D)           (None, 120, 120, 64)      36928

 max_pooling2d (MaxPooling2  (None, 60, 60, 64)        0
 D)

 conv2d_2 (Conv2D)           (None, 60, 60, 128)       73856

 conv2d_3 (Conv2D)           (None, 60, 60, 128)       147584

 max_pooling2d_1 (MaxPoolin  (None, 30, 30, 128)       0
 g2D)

 conv2d_4 (Conv2D)           (None, 30, 30, 256)       295168

 conv2d_5 (Conv2D)           (None, 30, 30, 256)       590080

 max_pooling2d_2 (MaxPoolin  (None, 15, 15, 256)       0
 g2D)

 conv2d_6 (Conv2D)           (None, 15, 15, 512)       1180160

 conv2d_7 (Conv2D)           (None, 15, 15, 512)       2359808

 max_pooling2d_3 (MaxPoolin  (None, 7, 7, 512)         0
 g2D)

 conv2d_8 (Conv2D)           (None, 7, 7, 256)         1179904

 conv2d_9 (Conv2D)           (None, 7, 7, 256)         590080

 max_pooling2d_4 (MaxPoolin  (None, 3, 3, 256)         0
 g2D)

 flatten (Flatten)           (None, 2304)              0

 dense (Dense)               (None, 128)               295040

 dense_1 (Dense)             (None, 64)                8256

 dense_2 (Dense)             (None, 5)                 325

=================================================================
Total params: 6758981 (25.78 MB)
Trainable params: 6758981 (25.78 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Moreover, the model's performance metrics from the evaluation phase are described below:

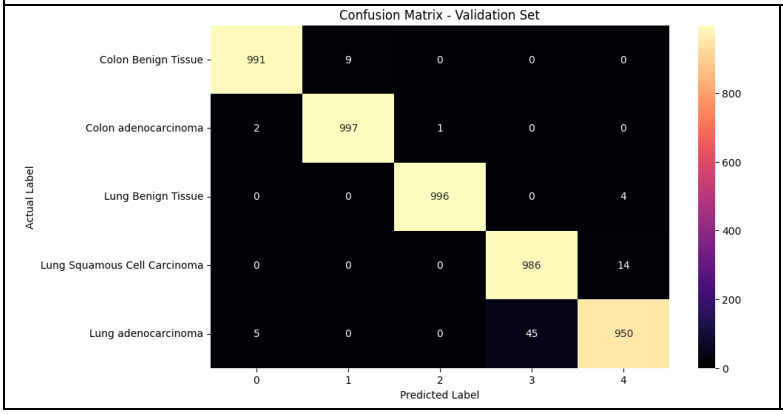


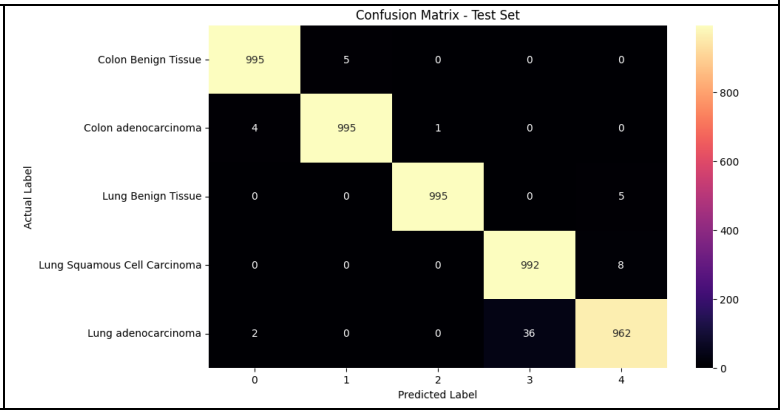
ROC Curves – Enhanced Model





Confusion Matrices – Enhanced Model





Validation set	Accuracy=0.984	Precision=0.985	Recall=0.984	F1 Score =0.984
Test Set	Accuracy=0.988	Precision= 0.988	Recall= 0.988	F1 Score=0.988

The hyperparameter tuning proved to be highly effective, elevating the test accuracy from 0.78 to 0.98. The improvement in precision suggests a substantial reduction in false positives, critical in areas like medical diagnostics where such errors are costly. The boost in recall points to the model's enhanced ability to identify all relevant instances, minimizing missed detections. Lastly, the rise in the F1 score reflects a balanced improvement in precision and recall, indicating the model's effectiveness in maintaining a trade-off between minimizing false positives and negatives.

Looking at the accuracy and loss graphs, the enhanced model demonstrates a higher overall accuracy on both the training and validation sets, and most importantly, the validation accuracy remains consistent with the training accuracy, indicating better generalization. The loss graphs also show a more stable decrease for both sets with less discrepancy between training and validation loss, suggesting that overfitting has been significantly reduced.

The ROC curves for the enhanced model display perfect scores (area = 1.00) for all classes on both validation and test sets, an improvement compared to our first model.

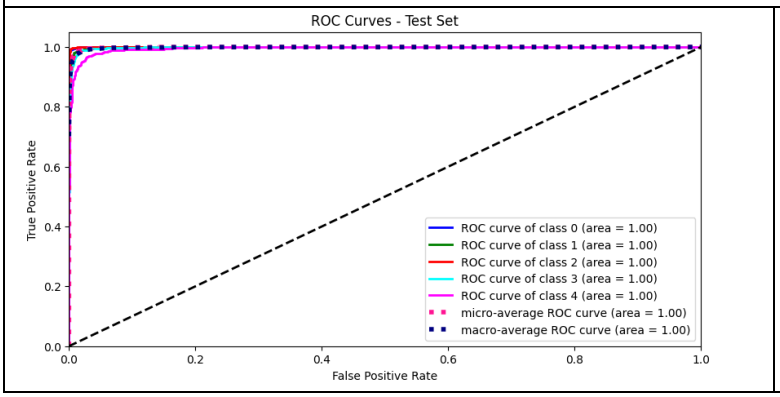
When comparing the confusion matrices, the enhanced model has greatly improved its predictive accuracy, with a substantial decrease in false positives and false negatives across all classes.

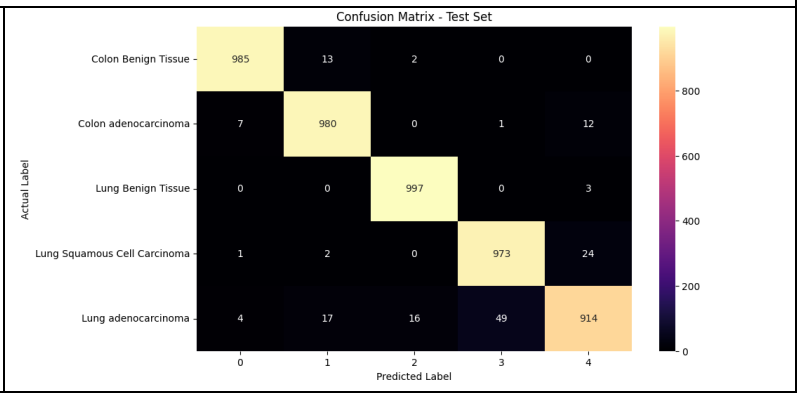
Overall, the enhanced model not only shows better performance in classification accuracy but also suggests it has better learned the underlying patterns in the data, providing a more reliable and robust solution for the problem at hand.

5. Transfer Learning Model and Its Results

The transfer learning model, utilizing the VGG16 architecture pre-trained on ImageNet, demonstrated notable predictive performance with an accuracy of 0.97 on the test set. VGG16's convolutional base was frozen to preserve its learned features, and the 3 new fully connected layers like the ones we had in our enhanced model were added (with same number of neurons and activation function). Also, we applied early stopping, the same way that we did in our improved model. The results for the test set are shown below:

ROC & Confusion Graphs of Transfer Learning Model





Test Set	Accuracy=0.970	Precision= 0.970	Recall= 0.970	F1 Score=0.970

While the transfer learning model did not surpass the enhanced CNN model's accuracy of 0.98, it still achieved a commendable accuracy of 0.97. Moreover, it showcased an exceptional ROC curve, with an area under the curve (AUC) score of 1.00 for all classes, highlighting its excellent classification capabilities. The confusion matrix further confirmed its efficiency, displaying high true positive rates with minimal misclassifications. Compared to the baseline model, the transfer learning model marked a significant enhancement in all evaluated metrics. Furthermore, there was a significant decrease in the misclassification rates across all categories, underscoring its robust performance

6. Discussion

In the section of our baseline model, we noted from its evaluation that there are opportunities for enhancement, particularly concerning some evident overfitting. As a result, we pursued a systematic approach to refining the model, employing hyperparameter adjustments and regularization techniques to optimize and stabilize the performance of the foundational model.

The improved performance of our model is due to several optimizations. L2 regularization reduced overfitting, while early stopping with more epochs prevented unnecessary training and overfitting. A lower learning rate helped the model converge more effectively, and increased depth allowed for capturing complex patterns. The introduction of a pyramid-like layer structure optimized the model's ability to learn hierarchical features without overcomplicating the architecture. Lastly, switching to the tanh activation function from ReLU contributed to faster convergence and increased accuracy, possibly due to its properties that can sometimes prevent the vanishing gradient problem. These combined enhancements are reflected in the model's higher accuracy, as evidenced by the evaluation metrics.

Future enhancements to our approach could include a deeper investigation into the application of data augmentation techniques, particularly in light of the overfitting observed in our model. Data augmentation has the potential to bolster model generalization by infusing greater diversity into the training dataset, thereby mitigating the risk of overfitting (Shorten, 2019). Moreover, leveraging transfer learning with models pre-trained on medical-specific datasets could offer a robust starting point by utilizing feature representations that are more closely aligned with medical imaging tasks, thus potentially enhancing diagnostic accuracy (Shin, 2016). Additionally, hybrid architectures, which combine the strengths of CNNs with other neural networks, such as RNNs or attention mechanisms, present a promising avenue for capturing complex patterns in medical images, effectively improving model performance by integrating spatial and contextual information crucial for accurate cancer diagnosis (Kumar, 2016).

References

Alzubaidi. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of big Data*.

Dube. (2018). High dimensional spaces, deep learning and adversarial examples. *arXiv preprint arXiv:1801.00634*.

Kumar. (2016). An ensemble of fine-tuned convolutional neural networks for medical image classification. *IEEE journal of biomedical and health informatics*.

Shin. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*.

Shorten. (2019). A survey on image data augmentation for deep learning. *Journal of big data*.

Yu, W. (2018). Hierarchical semantic image matching using CNN feature pyramid. *Computer Vision and Image Understanding*.