Topic: Data-Centric AI Alice Tang and Yinda Chen

TF Lite Breast Cancer Detection Week 10: Data-Centric AI

1. Introduction

As part of this week's focus on data-centric AI, we worked towards refining the data preprocessing pipeline to improve model performance by improving the quality of the input data. Our intent was not to alter the model architecture itself, but to create a new preprocessing function to better prepare the data. These steps included background removal to ensure that only important features are highlighted, Contrast Limited Adaptive Histogram Equalization (CLAHE) for improving image contrast and enhancing key features, and pixel normalization to standardize the input across images. Furthermore, we have ensured the images were resized to the target dimensions and adjusted to have three color channels. Ultimately, the intent was to take a data-centric approach in optimizing the data fed into the model, thereby improving its ability to learn from the data and optimize performance.

2. Three Data-Centric Improvements for Model Enhancement

Based on the initial performance results of our chosen model, specifically the issues with precision and recall, there were multiple crucial improvements to be made to the data preprocessing pipeline to improve the model's ability to generalize and accurately classify mammography images.

2.1 Background Removal

One of the major challenges in mammography images is dealing with irrelevant background issues that divert the model from focusing on the actual breast tissue and potential malignancies. Through incorporating background removal in the preprocessing function, we have ensured that the model's attention was focused toward the most critical areas of the image— therefore potentially improving the precision and recall. Naturally, reducing noise from irrelevant features is a safe and beneficial step, particularly in medical imaging, as it enhances clarity and highlights the relevant diagnostic areas without compromising the integrity of data, ultimately improving model performance,

2.2 Contrast Limited Adaptive Histogram Equalization (CLAHE)

Utilizing the application of CLAHE highly improves the contrast of the mammography images, In turn, subtle features like small tumors or calcifications, can become more apparent. Completing this step is essential to improving the quality of the data, as well as boosting the model's ability to detect elusive abnormalities, especially when there are varying lighting conditions or image quality. We chose to complete this step as we also believe enhancing the contrast may address some of the false negatives we observed, where small or low-contrast anomalies may have been overlooked in the original data before training.

2.3 Normalization and Resizing

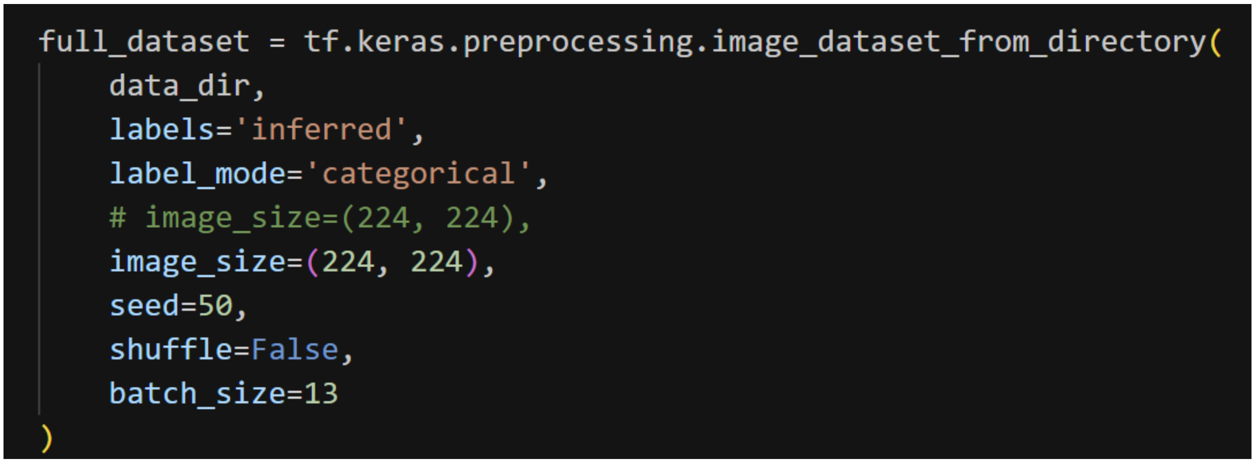
Normalizing the images is done to make sure there are consistent intensity values across all of the input images. Doing so, aims to standardize the data to ensure better model convergence. Our method consisted of combining with resizing images to a uniform target size of 224x224 pixels, certifying the model receives inputs with a consistent scale and format— thus improving the training efficiency and generalization. Additionally, by verifying the correct number of color channels (3 channels), we are able to ensure the model can work with images in a standard RGB format and improves its ability to learn patterns in the image data effectively.

Each of these improvements to the data preprocessing pipeline were created in alignment with addressing the initial model’s limitations. Through removing background noise, improving image contrast, and ensuring data consistency, we have created a more robust dataset that will be better suited to training the model for more accurate and reliable predictions. Ultimately, our hope is that these steps will mitigate the issues with precision and recall and create a more reliable tool for breast cancer detection.

3. Performance Metrics Results

3.1 Issue with Data Shuffling and Its Impact on Model Performance

An important point of realization during this week’s task was regarding the shuffling behavior in our dataset loading process. Initially, we were using this following code to load the dataset:



```
full_dataset = tf.keras.preprocessing.image_dataset_from_directory(
    data_dir,
    labels='inferred',
    label_mode='categorical',
    # image_size=(224, 224),
    image_size=(224, 224),
    seed=50,
    shuffle=False,
    batch_size=13
)
```

In the code above, we've set the **shuffle = False** parameter for both training, validation, and testing datasets. However, we now understand that this choice significantly negatively affects the performance of our model, especially in terms of its ability to generalize and learn meaningful features.

During training, shuffling the dataset is essential because it discourages the model from learning patterns based on the order of the data. When **shuffle = True** is set for training, the model is then allowed to see the data in a random order during each epoch, ensuring that the learned features are generalized across different subsets of the data. Without the shuffling, the model may inadvertently learn the patterns based on the order of images versus the content itself— leading to poor generalization and overfitting.

On the other hand, for validation and testing, we realized that shuffling must be set to False. This is because the validation and test sets are meant to provide an unbiased evaluation of the model's performance on unseen data. Keeping the order fixed allows us to ensure that all evaluation metrics (accuracy, precision, recall, etc;) are consistent across each training epoch, and are not affected by any randomness in the order of the data. Taking this step showcases the performance results without any sort of artificial variations caused by shuffling.

Impact on Model Performance

Ultimately, we believe that the failure to shuffle the dataset during training most likely contributed to the poor performance observed last week, particularly with precision and recall. By not shuffling the data, the model may have been exposed to the same sequences of images in each epoch— thereby resulting in overfitting and poor generalization to new data. Not to mention, the order of the samples in the validation and test sets may have skewed the evaluation, perhaps making it seem as though the model was performing better or worse than it actually was.

Through correcting this issue and ensuring that **shuffle = True** during training and **shuffle = False** during validation and testing, we are expecting to see more reliable and consistent results which will improve the model's ability to generalize and produce more accurate predictions for our task.

3.2 Comparison of Performance Metrics

Model	Final Validation Loss	Final Validation Accuracy
Week 9 Baseline Model	0.1032	0.972
Data-Centric Model	0.0986	0.9735

In our evaluation, our Week 9 Best Model and the Data-centric model had differing performance on key metrics, specifically in validation loss and validation accuracy:

Validation Loss:

- **Baseline Model:** 0.1032
- **Data-Centric Model:** 0.0986

The Data-Centric model achieved a lower validation loss than the Baseline Model, showing that it may fit the validation data more closely. In this case, having a lower validation loss shows better predictive accuracy and a smaller gap between the predicted and actual values.

Validation Accuracy:

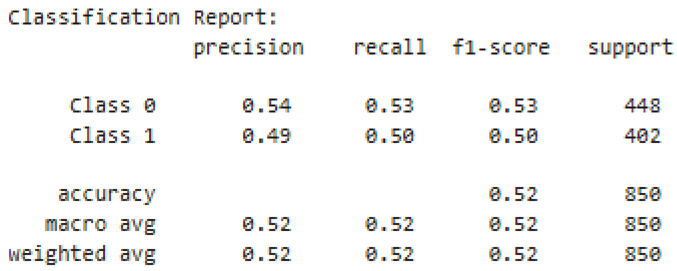
- **Baseline Model:** 0.972
- **Data-Centric Model:** 0.9735

The Data-Centric Model also demonstrated a slightly higher validation accuracy (97.35%) compared to the Baseline model (97.2%). This metric shows that the Data-Centric Model still retained a minor edge in accurately classifying validation samples, perhaps due the data-centered enhancements.

However, it's essential to recognize that accuracy should not be the sole metric of evaluation in this context, as the dataset has more non-cancer cases than cancer cases, naturally skewing accuracy results. In this scenario, precision and recall are more critical metrics. Both are particularly important due to the serious implications of false positives and false negatives— misclassifications that could lead to significant medical consequences for patients. As reflected in the classification report we will discuss below, prioritizing metrics such as precision and recall will offer a more meaningful assessment of our model's ability to minimize diagnostic errors.

3.3 Final Model Performance Metrics

Week 9 Baseline Model Performance Metrics on Test Set

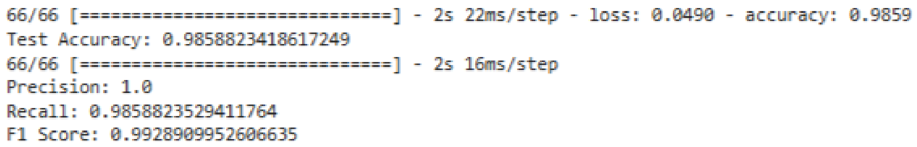


```
Classification Report:
              precision    recall  f1-score   support

     Class 0       0.54      0.53      0.53       448
     Class 1       0.49      0.50      0.50       402

    accuracy                           0.52       850
   macro avg       0.52      0.52      0.52       850
weighted avg       0.52      0.52      0.52       850
```

Week 10 Data-Centric AI Performance Metrics on Test Set



```
66/66 [==============================] - 2s 22ms/step - loss: 0.0490 - accuracy: 0.9859
Test Accuracy: 0.9858823418617249
66/66 [==============================] - 2s 16ms/step
Precision: 1.0
Recall: 0.9858823529411764
F1 Score: 0.9928909952606635
```

When comparing the performance of the data-centric AI model with the baseline model, the differences are apparent.

For the **baseline model**, the test accuracy shows as 52%. The precision, recall, and F1 score for both classes are all around 0.5, meaning that the model is only barely better than random guessing. Once again, these results suggest that the baseline model struggles to distinguish between the classes— making it unsuitable in a sensitive field like medical diagnosis where false negatives or positives can be highly detrimental.

On the other hand, the **data-centric AI model** has a test accuracy of 98.59%, with a loss of 0.0490. The precision is perfect at 1.0, and recall is very high at 98.59%. The F1 score, which balances precision and recall, is incredibly high at 0.993, which indicates this model performs wonderfully in identifying both classes. This model shows a promising performance, especially in a medical context where misdiagnosis could lead to serious consequences.

Overall, we see that the data-centric model significantly outperforms the baseline on the test set, showcasing the power of improved data preprocessing and a focus on enhancing the dataset rather than only optimizing the model.

4. Final Insights

Based on the evaluation of the test, validation, andt training errors, we can draw several key insights. Firstly, the **Data-Centric Model** outperforms the **Baseline Model** in multiple ways. While both models achieved similar validation accuracy, the Data-Centric Model exhibited a slightly lower validation loss, suggesting that it better fits the validation data and has a smaller gap between predicted and actual values. However, most importantly, the Data-Centric Model shows superior performance in precision, recall, and F1 score, all of which are crucial for minimizing diagnostic errors in a medical setting. The baseline model with a 52% accuracy, only has minimal ability to distinguish between classes, which makes it highly unfit for deployment in a context where false positives and false negatives can have severe consequences. However, we do acknowledge that the shuffle configuration may have also contributed to the low accuracy score on the test set with the baseline model.

In any case, based on these results, the **Data-Centric Model** is clearly the better performing model and will be selected as the final model for deployment. Its ability to accurately classify both cancerous and non-cancerous cases, with a high level of precision and recall, makes it an exceedingly more reliable choice for real-world applications, especially where patient safety must be considered first and foremost. This model's improvements, driven by enhanced data preprocessing, demonstrate the significant impact that a data-centric approach can have on model performance.