Speech Understanding

Programming Assignment 3

Prateek Singhal (M22AIE215)

Experiment Setup:

1. Obtained SSL_W2V model trained on LA and DF tracks of the ASVSpoof dataset.
2. The data utilised consists of a custom dataset and a foreign dataset. Where "for" was utilised for fine-tuning, validating, and testing. Custom is just utilised for testing purposes.
3. Utilised Python 3.9's most recent version and established a virtual environment using conda.

AUC, EER and ROC on Custom Dataset:

Settings:

1. To anticipate using the model, use the score from index 1 as shown in the original git repository.
2. Developed a recursive programme to locate all .wav and .mp3 files within a directory and save their paths in a data frame. The type is designated as train/test and the labels as real/fake based on the folder names.

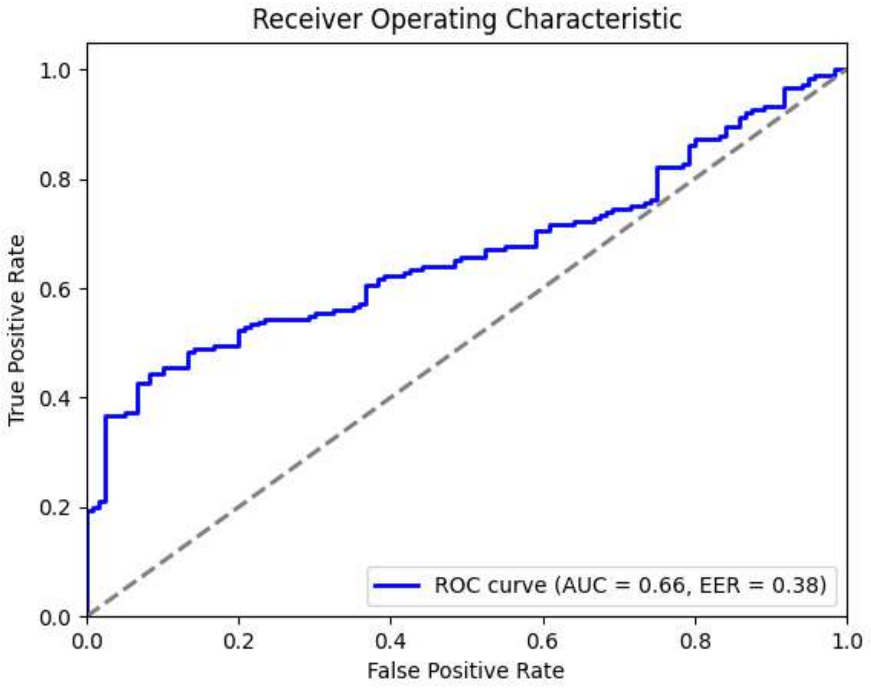

Fig.1

Analysis (Refer Fig.1):

1. AUC: 0.66
2. EER: 0.38
3. This model demonstrates only moderate discriminative ability as indicated by an AUC of 0.66, which suggests that the model is somewhat better than random guessing but far from perfect.
4. The EER of 0.38 implies that the model exhibits a significant number of classification errors at the threshold where the false positive rate equals the false negative rate. This higher error rate can be a concern if the application requires high precision or recall.
5. The curve is consistently below the diagonal line of no discrimination until very high false positive rates, indicating that the model struggles to achieve high true positive rates without substantially increasing the false positives.

Overall, the model exhibits moderate predictive capacity but lacks robust accuracy, suggesting potential for enhancements. Improvements could involve training on a broader or more representative dataset, optimizing model parameters, or adopting advanced modelling techniques. Additionally, the model's performance challenges may stem from inadequate generalization to new data, a common issue in machine learning.

Fine tune the model on FOR dataset:

Settings

1. We will be adjusting the weights and biases of the first five and last five layers of the model due to its large size.
2. The rationale for selecting these layers is based on the model's significant impact on feature extraction and final output.
3. The cross-entropy weight of [0.1, 0.9] was utilised as specified in the original git repository.
4. Separated the train set into training and validation sets with fractions of 0.8 and 0.2, respectively.
5. Implemented a hyper-parameter configuration in the main function that can be adjusted to manage tuning parameters such as data subset, number of trainable layers, batch size optimised for GPU memory, epochs, and learning rate.
6. Logged all tuning on wandb.ai and posted on GitHub too. wandb github.
7. Conducted several repetitions with various configurations, and here are the results from a few of them.
8. The subsets of data used for finetuning are 10%, 50%, and 100%.

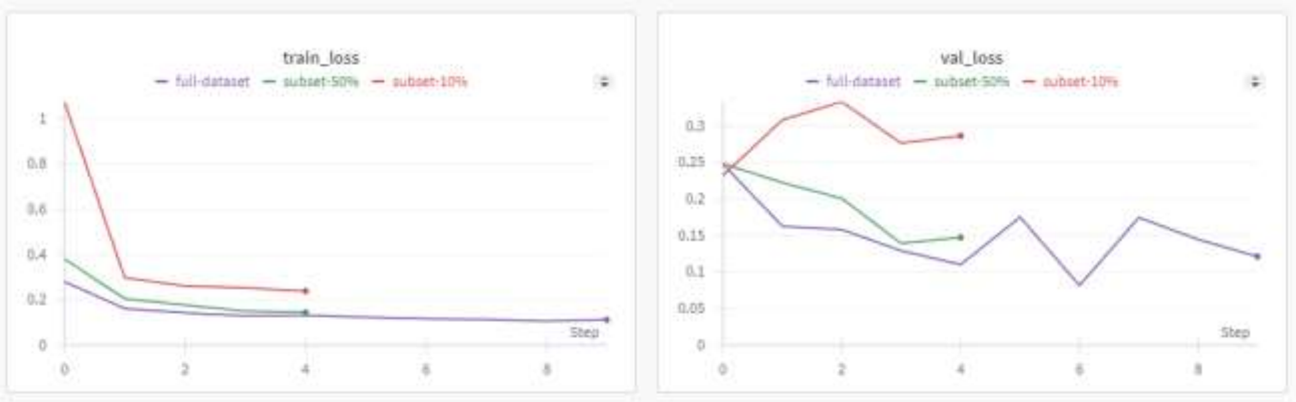

Fig.2

Analysis (Refer Fig.2):

1. The charts illustrate the training and validation losses for a model fine-tuned on different subsets of a dataset containing 2-second audio clips. The model's training involved fine-tuning only the first 5 and last 5 layers.
2. Training Loss Insights:
    a. **Full Dataset:** Exhibited a sharp initial drop and rapid stabilization, indicating efficient learning from the full dataset.
    b. **Subset 50%:** Displayed a significant, though less steep, decrease, suggesting effective learning with reduced data.
    c. **Subset 10%:** Showed the smallest decline, pointing to limited learning capabilities due to a significantly smaller dataset.
3. Validation Loss Insights:
    a. **Full Dataset:** Started low but showed slight increases, potentially indicating overfitting.
    b. **Subset 50%:** Demonstrated moderate stability and less overfitting compared to the full dataset.
    c. **Subset 10%:** Began with the highest losses but improved, suggesting initial underfitting yet decent generalization as training progressed.

Conclusions:

Fine-tuning selective layers proved effective across all data subsets. Utilizing the full dataset maximizes learning but risks overfitting, whereas smaller subsets, especially 50%, balance between learning depth and generalization. This indicates that data volume and diversity critically influence model performance and training dynamics.

Evaluate Finetuned Model on FOR2 dataset:

Settings:

1. Used finetuned model on 50% dataset as it is most balanced model.

Analysis (Refer Fig.3):

1. AUC: 0.98
2. EER: 0.06
3. The AUC of 0.98 is indicative of excellent model performance, showing that the model can almost perfectly discriminate between the two classes with very high confidence.
4. An EER of 0.06 suggests that the model achieves a very balanced trade-off between sensitivity and specificity at the optimal threshold, with minimal error rates, making it highly reliable for practical applications.
5. The ROC curve rises sharply towards the top-left corner of the plot, demonstrating that the model achieves high true positive rates at very low false positive rates, which is ideal for most applications requiring high accuracy.

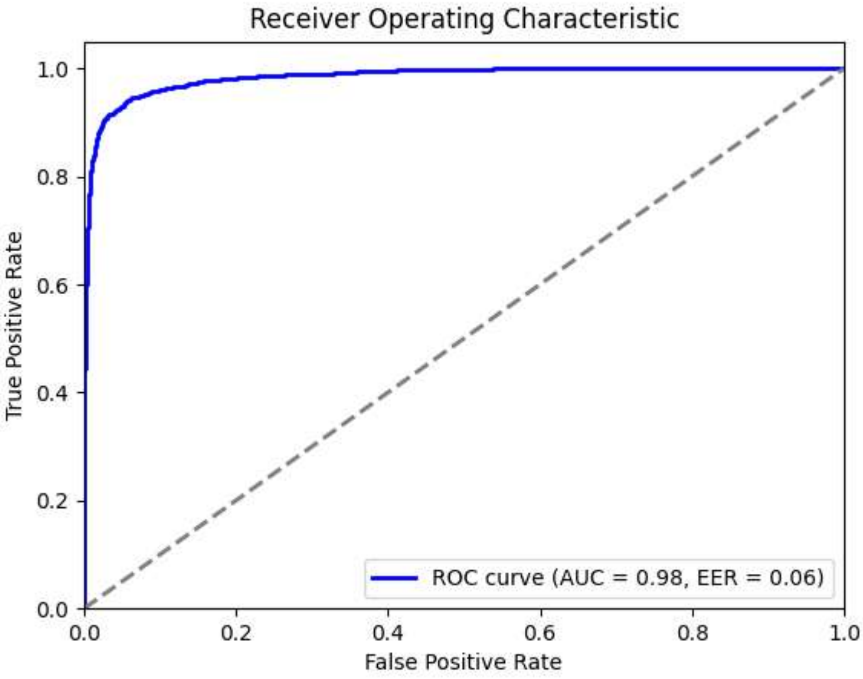

Fig.3

Evaluate Finetuned Model on Custom dataset:

Settings:

1. Used finetuned model on 50% dataset as it is most balanced model.

Analysis (Refer Fig.4):

2. AUC: 0.72
3. EER: 0.37
4. Extended Analysis:
5. The AUC of 0.72 shows an improvement over the original model's performance on the same custom dataset, indicating that finetuning has provided some beneficial effects in terms of overall discriminatory power.
6. Despite the improvement in AUC, the EER remains high at 0.37, close to that of the original model. This consistency in EER suggests that while the model may distinguish between classes better overall, it still fails to optimally balance false positives and false negatives at the best threshold.
7. The curve shows a more gradual rise compared to the finetuned model on the FOR2 dataset, indicating that achieving high true positive rates without increasing false positives is more challenging on this dataset.

Comparative Analysis on all Evaluations:

1. The finetuning process significantly improved the model’s performance on the FOR2 testing data, suggesting that the adjustments made are highly effective for that dataset. This is evident from the high AUC and low EER values.
2. When the finetuned model is applied back to the custom dataset, there is an improvement in AUC compared to the original model's performance on the same dataset, which indicates beneficial transfer of learning. However, the minimal change in EER suggests persistent challenges in balancing sensitivity and specificity.

3. The differing performances highlight the importance of model generalization. The finetuned model performs excellently on the dataset similar to its training data but does not reach the same level of performance on a different dataset, indicating potential overfitting or dataset-specific tuning.
4. These analyses suggest that while finetuning can dramatically improve performance for similar datasets, ensuring robustness and generalizability across different types of data remains a critical challenge.

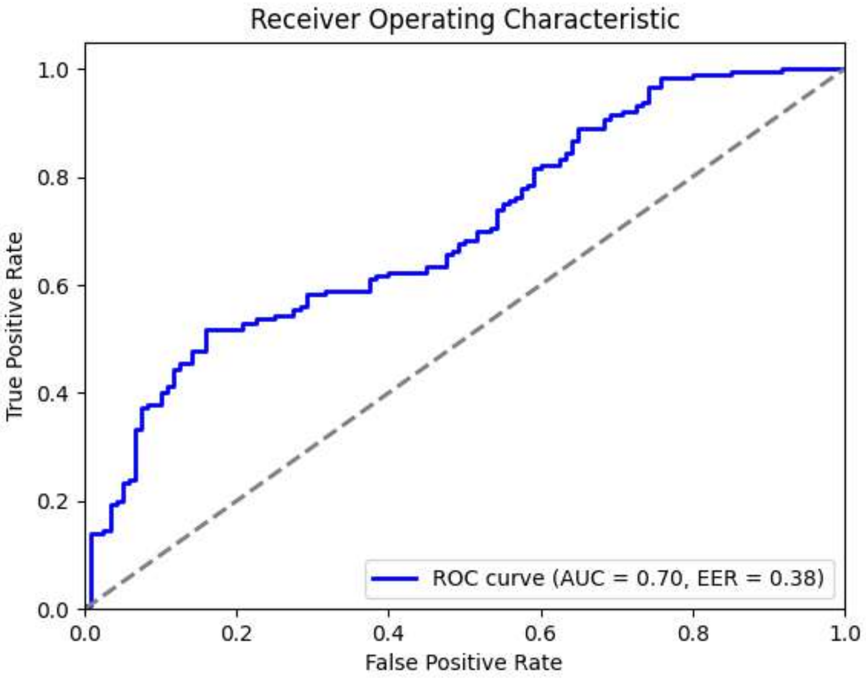

Fig.4