Week 4 Narrative Report: Synesthetic Learning Project

Chris Jallaine Mugot

Department of Data Science - College of Information Technology and Computing
University of Science and Technology of Southern Philippines
Lapasan, Cagayan de Oro City 9000, Philippines

This week's work centered on evaluating multiple candidate architectures for the audio-emotion classification task and systematically refining the best-performing model through targeted hyperparameter optimization. Building on the preprocessing and baseline modeling setup completed earlier, the focus for this phase was to rigorously compare architectures, identify the most promising model, and tune it to reach competitive performance suitable for downstream cross-modal retrieval and GAN-based generation.

The first major milestone this week involved conducting a structured model comparison across seven architectures: a baseline CNN, CNN-LSTM, CNN-LSTM with Attention, CNN-GRU, a ResNet feature extractor, an ANN baseline, and a lightweight Transformer. Each model was trained under consistent conditions to ensure a fair comparison, allowing us to isolate architectural strengths rather than noise from training configurations. The results revealed a clear trend: models that relied primarily on convolutional feature extraction outperformed hybrid or fully sequential approaches. The standalone CNN achieved the highest validation accuracy at 0.8254, despite having fewer parameters than many of the more complex architectures. In contrast, the Transformer obtained the lowest accuracy at 0.6731, suggesting that our dataset size and representation format favor localized convolutional processing over global self-attention.

Several insights emerged from this comparison. First, the CNN struck a favorable balance between model capacity and training stability. While architectures like ResNet were more expressive, their significantly longer training times (over two hours compared to ~28 minutes for the CNN) made them less practical for extensive hyperparameter sweeps. Meanwhile, hybrid models such as CNN-LSTM-Attention initially appeared attractive due to their ability to capture temporal dependencies, but in practice they offered only marginal improvements over simpler models and occasionally suffered from overfitting or unstable convergence. These patterns aligned with findings in prior literature showing that Mel-spectrograms often benefit most from spatial feature extraction rather than recurrent temporal modeling. Overall, the results established the CNN as the most efficient and accurate foundation for the next stage of model refinement.

With the CNN identified as the best candidate, the next major task involved performing automated hyperparameter optimization using Optuna. The goal was to fine-tune learning rate, weight decay, batch size, and embedding dimension to push the model beyond its earlier baseline performance. Over the course of 30 trials, Optuna evaluated a broad search space, pruning unpromising configurations early to focus compute resources on high-potential settings. Initial trials exhibited significant variability, with validation accuracies

ranging from the low 0.71–0.73 range up to the high 0.75–0.77 range depending on the stability of the optimizer settings.

A key discovery during tuning was the role of learning rate. Extremely small learning rates resulted in slow learning and inconsistent val accuracy, while overly large ones caused early plateauing. The strongest performance emerged from a moderately small learning rate around 7.48e-04, coupled with a low weight decay value. The best-performing trial achieved a validation accuracy of 0.7785, using a batch size of 16 and embedding dimension of 128, confirming that smaller batch sizes helped the model generalize better and avoid getting stuck in narrow minima. These results established the optimized configuration that would be used for the final training phase.

After selecting the best hyperparameters from the optimization study, I trained the optimized CNN for 20 epochs to evaluate its full learning trajectory. Early epochs showed a gradual improvement in both training and validation accuracy, with a notable jump beginning around epoch 5 as the model learned stronger emotional cues from the spectrograms. Val accuracy continued to rise steadily, eventually reaching a peak of 0.8238, nearly matching the earlier model comparison maximum and confirming the success of the tuning process.

The confusion matrix and classification report provided deeper insight into the model's behavior. The network demonstrated strong precision and recall across all three emotion classes, particularly excelling in recognizing Sad samples, where it achieved a recall of 0.92. This suggests that the spectrogram patterns linked to slower tempos or darker tonal textures were particularly distinctive and well-captured by the convolutional filters. The Happy class showed the lowest recall at 0.71, which aligns with the tendency of lively or bright audio expressions to vary widely in timbre and rhythm, making them harder to capture consistently. Nonetheless, the overall performance metrics—accuracy of 0.82, balanced class performance, and stable loss trends—indicate that the optimized CNN is robust and well-suited for downstream integration.

Finally, the week concluded with preparation steps for the conditional GAN component of the project. With the classifier now stabilized, I began constructing the conditional GAN architecture that will eventually use the predicted emotion embeddings to generate abstract art aligned with emotional cues. Model parameter summaries showed 4.97 million parameters for the Generator and 2.82 million for the Discriminator, marking a reasonable computational footprint for iterative training. Initial training logs showed progressively decreasing loss over the first 3000 steps, indicating that the adversarial training loop was functioning correctly and setting the stage for more comprehensive experiments in upcoming weeks.

Overall, this week solidified the technical foundation of the system's audio-emotion pipeline. Through structured model comparison, targeted hyperparameter tuning, and rigorous experimental iteration, we arrived at a high-performing CNN that balances accuracy, interpretability, and computational efficiency. These advancements position the project strongly for the next stages, where the classifier will be integrated with the cross-modal retrieval system and later the conditional GAN to enable emotion-guided abstract art synthesis.