Adaptive Model Optimization for Audio Classification on Edge Devices

Alice Hu, Emily Shao, Yash Kumar Johar
Final Project Presentation
ECE 661, Fall 2024

Project Description

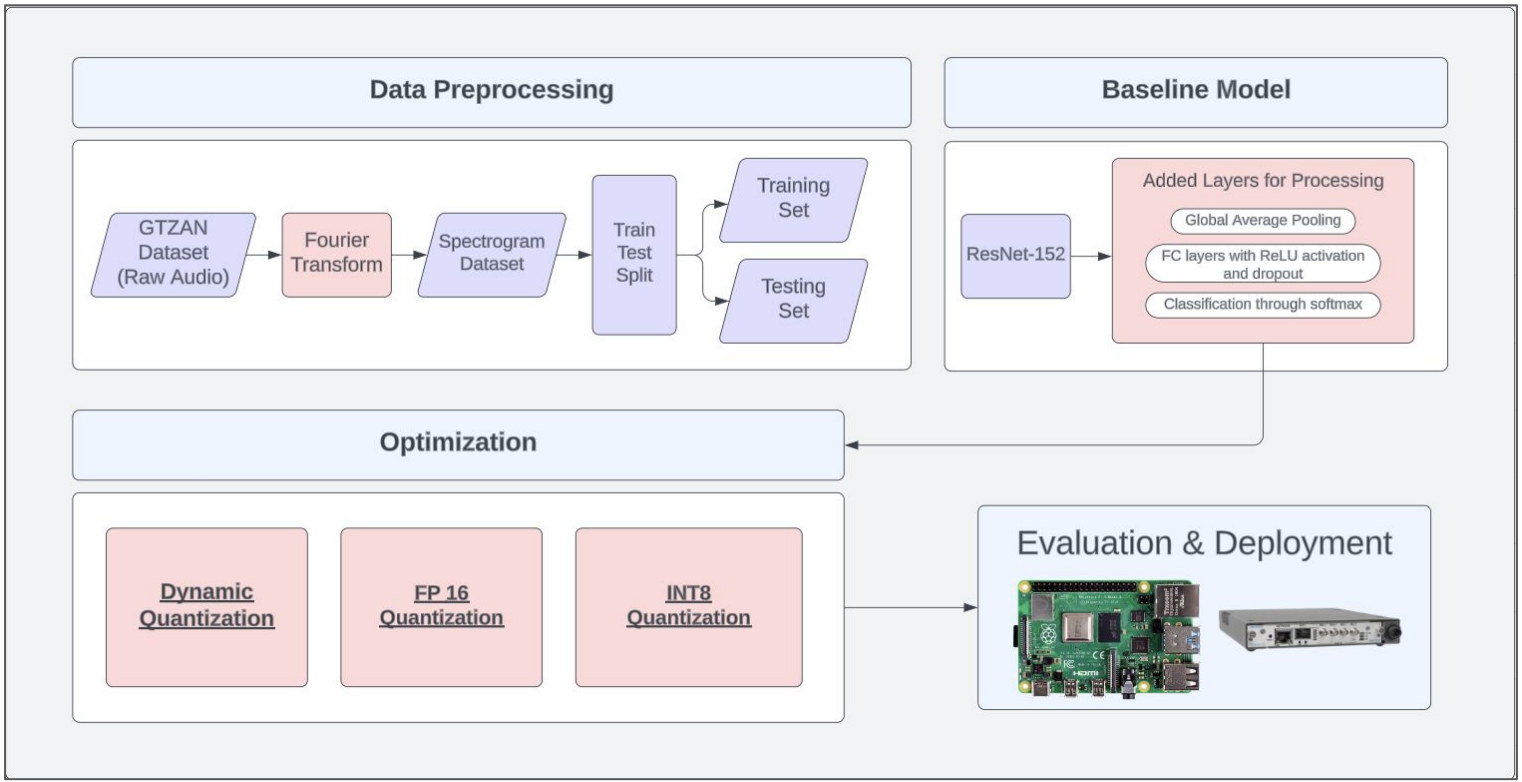
Problem: Deploying deep learning models on resource-constrained devices, such as Raspberry Pi or ESP32, can be challenging due to their limited memory and computational power.

This Project will optimize a ResNet-152 model trained on the GTZAN dataset for audio classification into 10 music genres by applying adaptive optimization techniques.

We will:

- Process Baseline Model to achieve High Accuracy on GTZAN dataset
- Implement Model Quantization approaches (INT8, FP16, Activation, Dequantization)
- Deploy Baseline and Fine-Tuned Models to Edge Device

Project Workflow



Data Preprocessing

Audio Data Processing Workflow

1. Audio File Resampling

All files were resampled at **22,050 Hz** for consistency

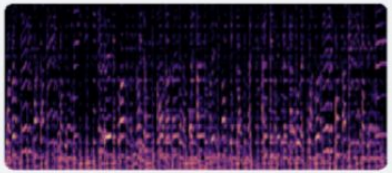
2. Mel Spectrogram Computation

(a) Short-Time Fourier Transform

$$X(m,k) = \sum_{n=0}^{N-1} x(n)w(n-mR)e^{-j2\pi kn/N}$$

(b) Mel Filter Bank

$$S(m,k) = \log\left(\sum_{j=0}^{M-1} H(k,j)|X(m,j)|^2\right)$$



Spectrogram of Blues .wav file

3. Normalization

Spectrogram values were normalized to **dB scale**

4. Image Resizing

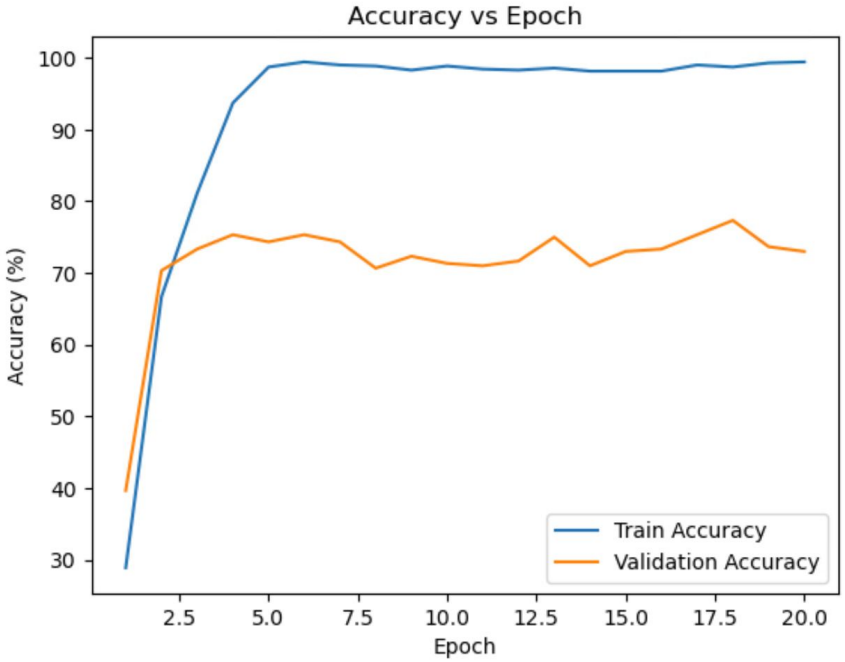
Spectrograms were resized to **299x299 pixels** to match the ResNet-152 input requirements.

5. Dataset Splitting

The dataset was divided into **80% training** and **20% testing**

Baseline Model

- Model: ResNet 152
  - Testing initial model baseline with no changes
  - Modifications
    - 20% test split and 20 epochs
    - 1e-4 learning rate
    - Mel Spectrogram input



2. Mel Spectrogram Computation

(a) STFT was applied

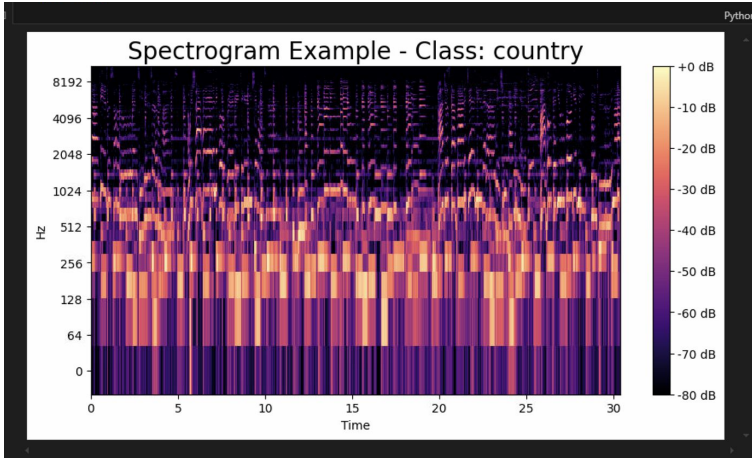
$$X(m,k) = \sum_{n=0}^{N-1} x(n)w(n-mR)e^{-j2\pi kn/N}$$

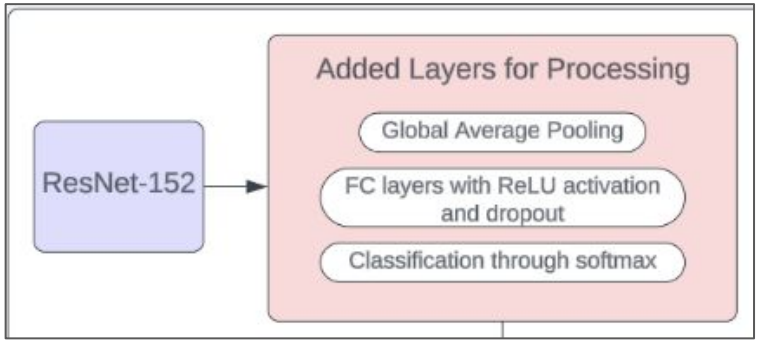
(b) Mel filter bank

$$S(m,k) = \log\left(\sum_{j=0}^{M-1} H(k,j)|X(m,j)|^2\right)$$

Modified Baseline

- Model: ResNet 152
- Dataset: GTZAN dataset
  - Collection of musical samples (3sec + 30sec) that are categorized into 10 genres of music
- Initial enhancements
  - Conversion to spectrograms or better processing results
  - Added Layers for Processing
- Validation accuracy: 74%





Pruning & Quantization Results – MB

Baseline	Pruning (0.5%)	Quantization (FP16)
Kind: Document Size: 268,263,408 bytes (268.3 MB on disk)	Kind: Document Size: 246,208,760 bytes (246.2 MB on disk)	Kind: Document Size: 121,797,220 bytes (121.8 MB on disk)

Quantization Approaches Comparison

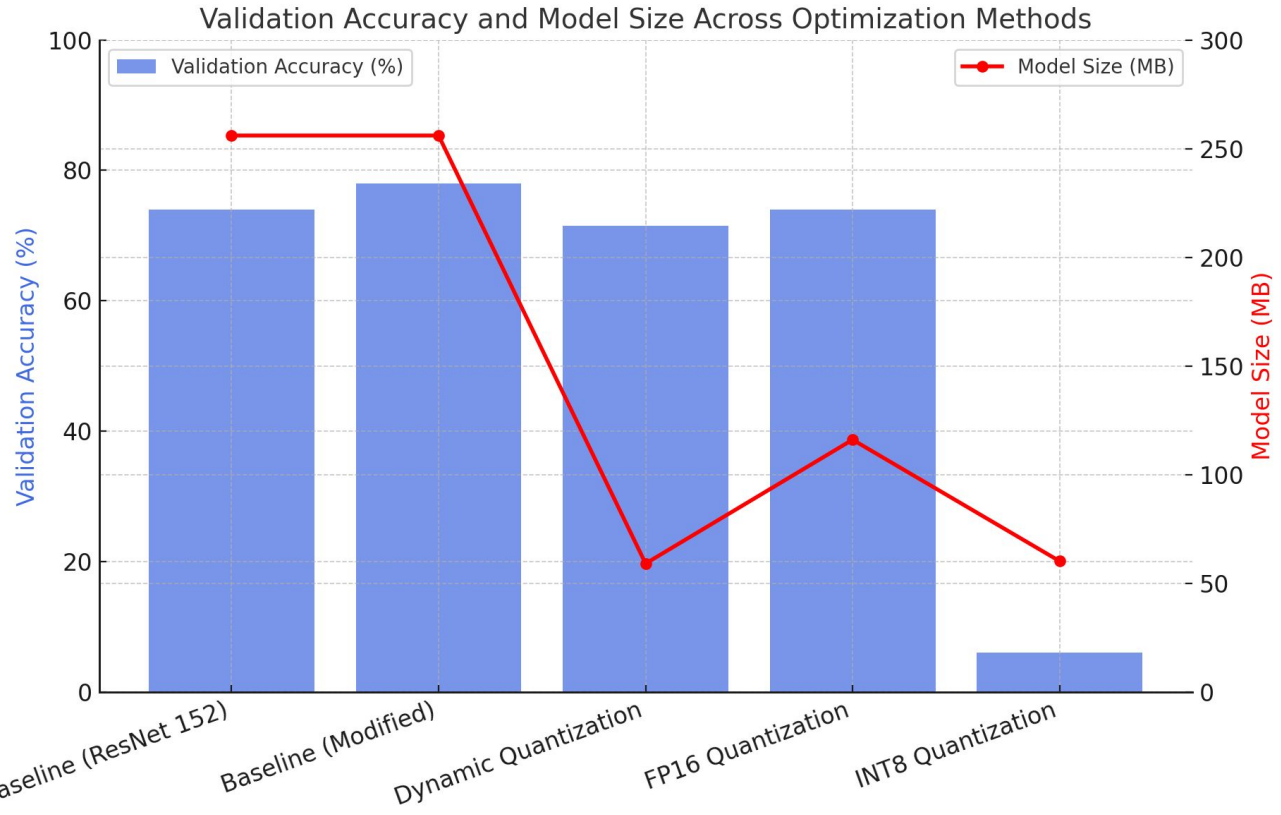
Approach	Definition	Key Characteristics
Dynamic Quantization	Converts only **weights** to INT8 at inference time while keeping activations in FP32.	- Reduces storage and improves inference speed. - Minimal accuracy loss. - Best suited for transformer and LSTM models.
FP16 Quantization	Converts both **weights and activations** to 16-bit floating point (FP16).	- Preserves floating-point operations for better numerical stability. - Balances precision and efficiency. - Suitable for deep networks sensitive to precision changes.
INT8 Quantization	Converts both **weights and activations** to INT8, requiring dequantization for computations.	- Highest compression and computational efficiency. - Can cause significant accuracy degradation. - Works best with **per-channel quantization** to reduce precision loss.

Baseline & Quantization Results

Table 1: Summary of Model Performance

Model	Train Accuracy	Validation Accuracy	Model Size
Baseline (ResNet 152)	89%	74%	256 MB
Baseline (ResNet 152) with Modification	89%	78%	256 MB
Dynamic Quantization	– %	77.5%	59 MB
FP16 Quantization	– %	78%	116 MB
INT8 Quantization	– %	6%	60.1 MB
Pruned Model (Pending Optimization)	– %	74%	235 MB
FP16 ResNet-152 on Edge Device	– %	77.21%	116 MB
Dynamic ResNet-152 on Edge Device	– %	76.34%	59 MB

Baseline & Quantization Results



Model Deployment & Challenges

- Raspberry Pi Model B
- Datasets
  - GTZAN validation set
  - Stored on SD Card (not internal memory)
- Validation Test
  - Accuracy Dynamic: 76%
  - Accuracy FP16: 77%
  - Accuracy INT8: 6%

Raspberry Pi Results

Model	Validation Accuracy	Model Size
FP16 Quantized ResNet-152	78%	116 MB
Dynamic Quantized ResNet-152	77.5%	59 MB
FP16 ResNet-152 on RPi	77.21%	116 MB
Dynamic ResNet-152 on RPi	76.34%	59 MB

Component	Estimated Memory (MB)
Python Interpreter	5–10 MB
TensorFlow (Installed + Runtime)	1,600–2,200 MB
NumPy	100 MB
Matplotlib	120 MB
SciPy	100 MB
Model (Loaded into RAM)	300–500 MB
TensorFlow Execution Overhead	500–1,500 MB
Total Estimated Memory	**2.7 GB – 4.5 GB**

Table 3: Estimated Memory Usage Breakdown

References

[1] Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, *10*(5), 293-302. https://doi.org/10.1109/TSA.2002.800560

[2] NVIDIA ADLR. (2019, August 13). MegatronLM: Training Billion+ Parameter Language Models using GPU Model Parallelism. https://nv-adlr.github.io/MegatronLM

[3] Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016, February 24). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. *arXiv.org*. https://arxiv.org/abs/1602.07360

[4] Howard, A., Sandler, M., Chu, G., Chen, L., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019, May 6). Searching for MobileNetV3. *arXiv.org*. https://arxiv.org/abs/1905.02244

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770-778). https://doi.org/10.1109/CVPR.2016.90

[6] Luo, X., Liu, D., Kong, H., Huai, S., Chen, H., Xiong, G., & Liu, W. (2024). Efficient Deep Learning Infrastructures for embedded Computing Systems: A Comprehensive Survey and Future Envision. ACM Transactions on Embedded Computing Systems, 24(1), 1–100. https://doi.org/10.1145/3701728

[7] Raspberry Pi 4 specs and benchmarks — The MagPi magazine. (n.d.). The MagPi Magazine. https://magpi.raspberrypi.com/articles/raspberry-pi-4-specs-benchmark

Team Progress Documents

Initial proposal:

https://docs.google.com/document/d/1yOxUrbp8bAvgqmjfXcBR5Sv4PFLzor_InDYUeoIdnHg/edit?tab=t.0

TA Check-In Meeting

https://docs.google.com/document/d/1U3xawos_iKo98e7ye_vS1GMfizV-UWRWg7--K5vUSMA/edit?tab=t.0

Revamped proposal:

https://docs.google.com/document/d/1x7q0mtCAlYSWsA6GRdIh4Ib1jfn1urPA8pIIm06Jjng/edit?tab=t.0