Utilizing Liquid Neural Network for Efficient Audio Classification

Department of Electronic and Computer Engineering
Thapathali Campus

Team Members
Khemraj Shrestha (THA077BCT020)
Niyoj Oli (THA077BCT029)
Om Prakash Sharma (THA077BCT030)
Punam Shrestha (THA077BCT038)

Supervised By
Er. Kshetraphal Bohara

Co-Supervised By
Er. Rojesh Man Shikhrakar

June 2024

Presentation Outline

- Motivation
- Objectives
- Scope of Project
- Proposed Methodology
- Expected Results
- Project Application
- Tentative Timeline
- Estimated Project Expenses
- References

Motivation

- Contemporary models uses millions to billions parameters,
- 19 neurons enough for autonomous driving - Liquid Time Constant (LTC) Neural Network,
- Papers suggest LNN to be efficient for temporal data like Audio.

Objectives

- To develop LTC neural network model for audio classification and benchmark against contemporary models,
- To achieve comparable accuracy while using less computational power.

Scope of Project & Limitation

- Focus on Audio Classification of a small duration (~10 seconds),
- Prioritizes benchmarking LNN for Audio Classification tasks,
- Unable to predict long length videos inherently.
- May not adequately replicate real-world conditions
- Performance limited to chosen datasets.

Methodology-[1]
System Block Diagram

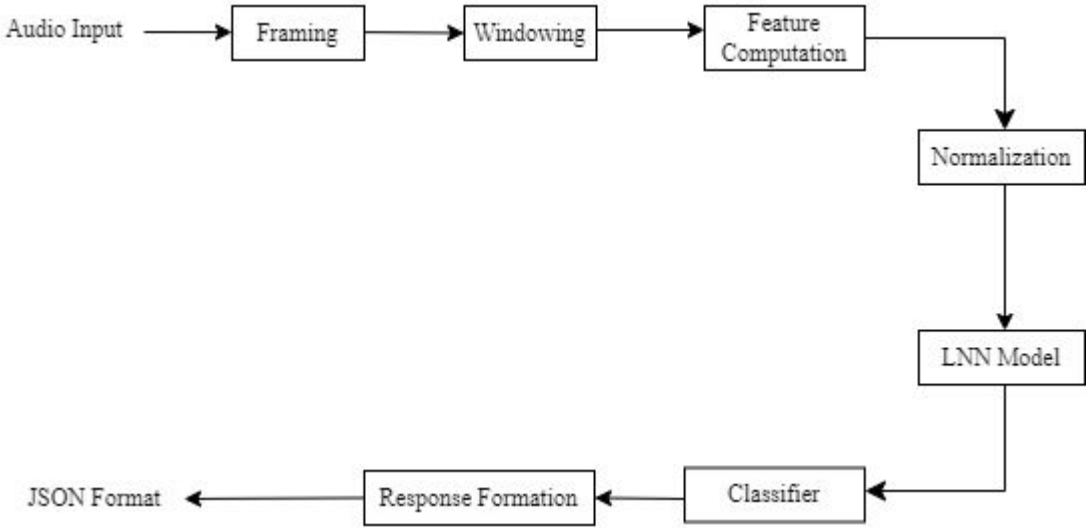

Fig: System block diagram

Methodology-[2]
Dataset Exploration

1. **VGG (Visual Geometry Group)**

- Audio-visual dataset with 210,000 data with 310 audio classes,
- Classes like wind noise, sliding door, car, train, etc.
- 10-second audio clips,
- Roughly 200 audio per each class,
- Used by Mirasol3B has 69.8% accuracy on this dataset.

Methodology-[3]
Dataset Exploration

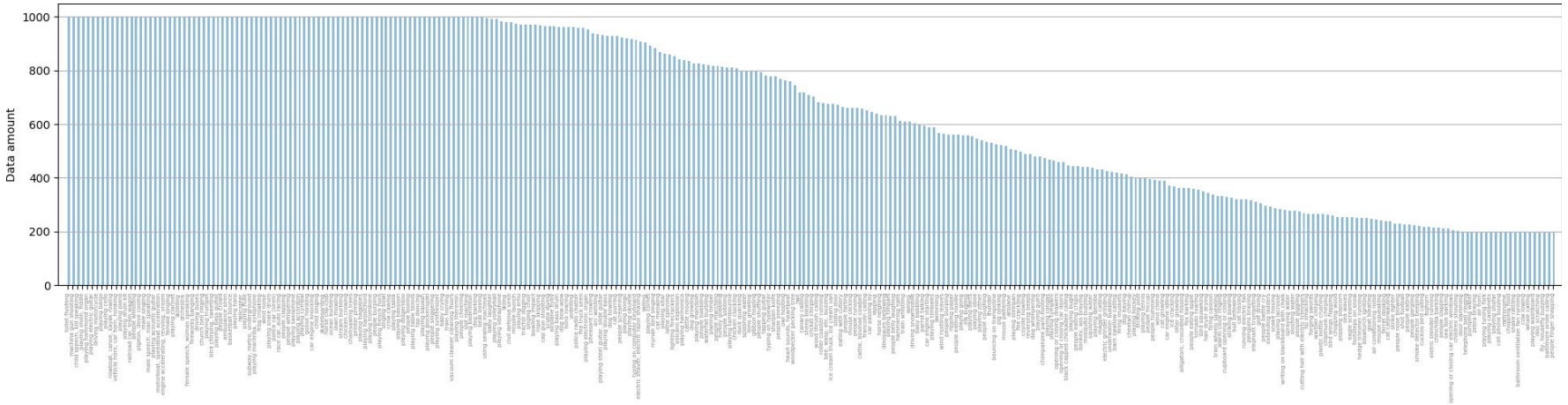

Fig: Dataset distribution for VGG Dataset

Methodology-[4]
Dataset Exploration

2. ESC-50

- 2000 labelled environmental audio recordings,
- Each clip of 5 seconds, covering 50 distinct classes,
- Includes classes like animals, water sound, natural soundscapes, etc.
- Pre-arranged in 5-folds,
- Used by OmniVec-2 Model with 99.1% accuracy.

Methodology-[5]
Dataset Exploration

3. UrbanSound8K

- Comprising 8,732 labelled sound excerpts,
- Each clip of 4 seconds, with total 27 hours of audio,
- Includes classes like air conditioner, car horn, children playing, dog bark, etc,
- Used by ASM-RH-I with 97.96% accuracy (10-fold).

Methodology-[6]
Dataset Exploration

4. AudioSet

- 2,084,320 YouTube videos containing 527 labels,
- 10-second sound clips sourced from YouTube videos and labelled by humans,
- Includes classes like music, speech, vehicle, car, etc.,
- Used by OmniVec with 0.548 mAP.

Methodology-[7]
Pre-processing Pipeline

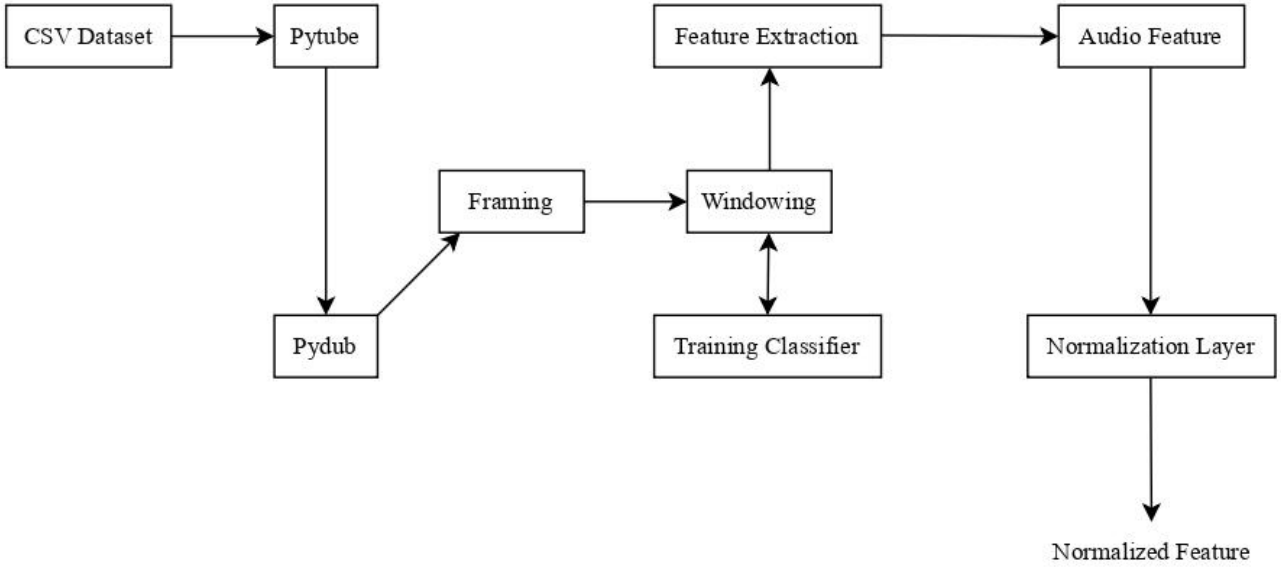

Fig: Pre-processing pipeline for training

Methodology-[8]
Evaluation Metrics

- **F1-Score**

- Used when the class distribution is imbalanced,
- Provides a single measure that balances both the false positives and false negatives.
- **Precision** is the ratio of true positive detections to the total number of positive detections,
- **Recall** is the ratio of true positive detections to the total number of actual positives.

Methodology-[9] Evaluation Metrics

- **F1-Score**
  - Harmonic mean of precision and recall.

$$F1Score = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}$$

Methodology-[10]
Evaluation Metrics

- **Mean Average Precision (mAP)**
  - **Average Precision** is the area under the precision-recall curve for a single query or class.

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i$$

Methodology-[11]
Evaluation Metrics

- **Accuracy**

- measures the proportion of correct predictions made by the model out of all predictions.

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{All Predictions}}$$

Methodology-[12] Instrumentation

1. **Kaggle Notebook**
   - Kaggle Notebooks are essentially Jupyter Notebooks hosted on the cloud
   - Provides 4 CPU cores, 20GB of RAM, and 1 x Nvidia Tesla P100 GPU with 4 cores and 29 GB of RAM,
   - GPU can be used for 30 hours a week and 9 hours per session.

Methodology-[13] Instrumentation

2. Microsoft Azure Notebook

- Provides a 16-core CPU, 110GB of RAM, and 1x NVIDIA Tesla T4 16GB vRAM, which costs $1.32 per hour,
- Charges on a pay-as-you-use basis,
- Will be used for during the final training and benchmarking phase.

Methodology-[14] Instrumentation

3. Librosa

- Python package for music and audio analysis,
- Calculation of time domain features like Zero-crossing rate,
- Calculation of frequency domain features.

Methodology-[15] Instrumentation

4. Pytorch

- Open-source deep learning framework developed by Facebook's AI Research lab,
- Uses Dynamic Computation Graph,
- Rich ecosystem and community support.

Expected Results

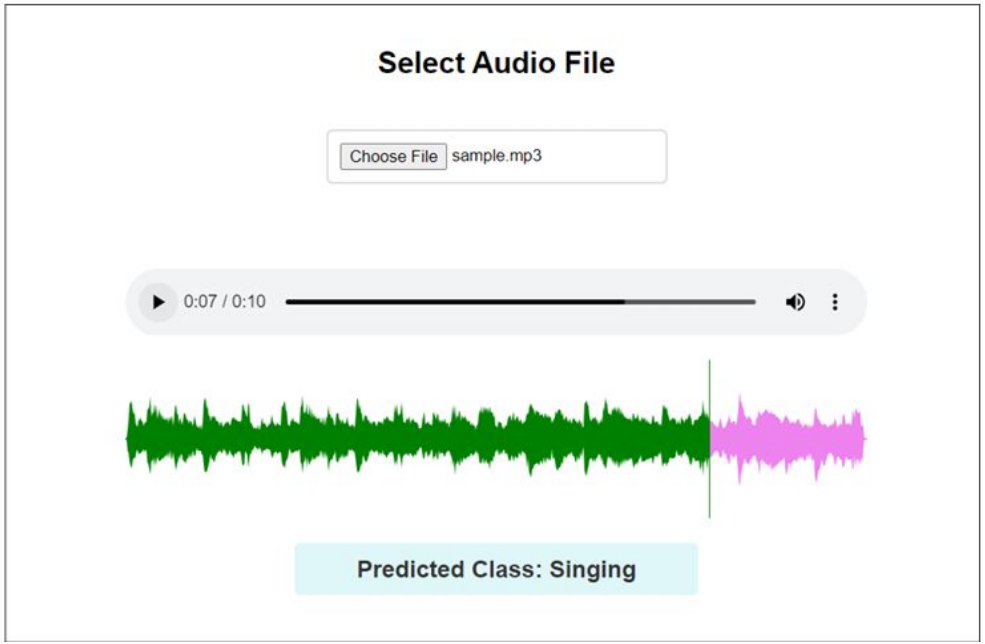

Fig: Web application interface

Project Applications

- Audio Event Detection
- Speech Recognition
- Music Information Retrieval
- Environmental Sound Classification
- Health Monitoring
  - Identifying patterns associated with Respiratory illness, cardiovascular conditions or vocal abnormalities

Tentative Timeline

	Itineary	June				July				August				September				October				November				December				January				February			
		W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4
Part A	Problem Understanding	//////	//////																																		
	Literature Review & Analysis	//////	//////	//////	■	■	■	■	■																												
	Mathematical Understanding			//////	■	■	■	■																													
	Mathematical Model Formulation			//////	■	■	■	■	■	■																											
	Dataset Analysis		//////	//////	■	■	■	■																													
	Baseline Model Preparation			//////	■	■	■																														
	Preparation of Data Pipeline					■	■	■	■																												
	Dataset Preprocessing									■	■	■																									
	Documentation	//////	//////	//////	■	■	■	■	■	■	■	■	■																								
Part B	Model Training													■	■	■	■	■	■	■	■																
	Model Validation & Tuning																	■	■	■	■	■	■	■	■												
	Comparison Model Training																				■	■	■	■	■	■	■	■	■	■	■	■	■				
	Benchmarking																																■	■	■	■	■			
	Documentation													■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■

Legend	
Upcoming	■
Completed	//////

Fig: Gantt chart

Estimated Project Budget

Particulars	Price	Total Cost
Standard NC16as T4 v3 (16 cores, 110GB RAM, 352GB Storage)	$ 1.32 per hour	$ 1900.80
Azure Storage (1000GB HDD)	$ 20 per month	$ 160
Total cost		$ 2060.80

Fig:Project Budget

Reference - [1]

[1] R. Hasani, M. Lechner, A. Amini, D. Rus, and R. Grosu, “Liquid time-constant networks,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 9, 7657–7666, May 2021. DOI: 10.1609/aaai.v35i9.16936. https://ojs.aaai. org/index.php/AAAI/article/view/16936.

[2] S. Srivastava and G. Sharma, Omnivec: Learning robust representations with cross modal sharing, 2023. arXiv: 2311.05709 [cs.CV].

Reference - [2]

[3] M. Chahine, R. Hasani, P. Kao, et al., “Robust flight navigation out of distri bution with liquid neural networks,” Science Robotics, vol. 8, no. 77, eadc8892, 2023, Published online 2023 Apr 19, ISSN: 2470-9476.

[4] H. Ju, J.-X. Xu, and A. M. VanDongen, “Classification of musical styles using liquid state machines,” in The 2010 International Joint Conference on Neural Networks (IJCNN), 2010, 1–7.