Article

Hierarchical Residual Attention Network for Musical Instrument Recognition Using Scaled Multi-Spectrogram

Rujia Chen * , Akbar Ghobakhlou , Ajit Narayanan *

Posted Date: 23 September 2024

doi: 10.20944/preprints202409.1632.v1

Keywords: Spectrograms; Musical Instrument Classification; Audio classification; Audio feature extraction; Music information retrieval; Spectrogram transformation; Residual attention networks; Attention mechanisms; Deep learning for audio

Article

Hierarchical Residual Attention Network for Musical Instrument Recognition Using Scaled Multi-Spectrogram

Rujia Chen *, Akbar Ghobakhlou and Ajit Narayanan

Computer Science and Software Engineering Department, Auckland University of Technology, Auckland, New Zealand
* Correspondence: rujia.chen@autuni.ac.nz

Featured Application: The proposed method could potentially be applied to musical instrument classification tasks, contributing to the organization and analysis of audio data where identifying instruments is required. This may be useful in research, music information retrieval systems, or other related applications.

Abstract: Musical instrument recognition is a relatively unexplored area of machine learning due to the need to analyze complex spatial-temporal audio features. Traditional methods using individual spectrograms, like STFT, Log-Mel, and MFCC, often miss the full range of features. We propose a hierarchical residual attention network using a scaled combination of multiple spectrograms, including STFT, Log-Mel, MFCC, and CST features (chroma, spectral contrast, and Tonnetz), to create a comprehensive sound representation. This model enhances focus on relevant spectrogram parts through attention mechanisms. Experimental results with the OpenMIC-2018 dataset show significant improvement in classification accuracy, especially with the "Magnified 1/4 Size" configuration. Future work will optimize CST feature scaling, explore advanced attention mechanisms, and apply the model to other audio tasks to assess its generalizability.

Keywords: spectrograms; musical instrument classification; audio classification; audio feature extraction; music information retrieval; spectrogram transformation; residual attention networks; attention mechanisms; deep learning for audio

1. Introduction

Musical instrument recognition in recorded music is complex due to the diverse and intricate features in audio signals. Traditional approaches using individual spectrograms like Short-time Fourier Transform (STFT), logarithm mel frequency (log-mel), or Mel-frequency cepstral coefficients (MFCC) capture only specific aspects of the audio signal, such as amplitude and short-term power spectrum. This often fails to capture the full richness of audio features necessary for accurate instrument classification.

To address this, combining multiple spectrograms—such as STFT, Log-Mel, MFCC, along with CST features (chroma, spectral contrast, and Tonnetz)—into a single compact input image has been proposed for various audio tasks [1–5]. This approach leverages the strengths of each spectrogram type, providing a more comprehensive understanding of the audio signal and leading to more accurate classification accuracy.

However, the increased number of spectrogram features can introduce challenges, making it difficult for the model to determine which part of the input image are most important. Attention mechanisms help the model focus on the most relevant spectrogram features, addressing this issue.

doi:10.20944/preprints202409.1632.v

Some studies, like those on bird sound recognition [3] and music annotation [6], integrating attention mechanisms [7,8] to enhance the model.

Building on this foundation, our work introduces a hierarchical attentive convolutional neural network (CNN) for musical instrument classification using scaled spectrograms. Also integrating attention mechanisms within the hierarchical Residual Networks (ResNets) [9], our model can prioritize the most relevant features from the combined spectrograms, enhancing classification accuracy. The scale assigned to each spectrogram type are determined during preprocessing, magnifying or minimizing their representation in the input images. This approach allows the model to focus on the most informative parts of the spectrograms, leading to enhanced classification accuracy.

2. Literature Review

Deep learning for music informatics has demonstrated that convolutional neural networks (CNNs) can effectively learn features directly from audio data, bypassing the need for manual feature design and significantly advancing automatic feature extraction and music classification performance [10]. The integration of Deep Learning Networks [11] and spectrogram features has proven to be a robust approach for audio classification even in noise conditions [12]. Spectrograms, which represent audio signals visually, allow CNNs to learn intricate patterns and features from the time-frequency domain [13–15].

Also, other research [16] proposed a multi-spectrogram encoder-decoder framework that utilizes different types of spectrograms to improve acoustic scene classification. Their approach highlights how integrating diverse spectral features can enhance the robustness of audio classification models.

Various studies [17,18] have explored different spectrogram types, such as STFT, Log-Mel, and MFCC, to capture diverse audio characteristics. Specifically, for musical instrument classification in recorded polyphonic music, the work [19] accurately classifies the NSynth dataset [20] with good performance. The IRMAS [21] dataset's work [22] that achieved 0.79 precision. The Open-Mic [23], dataset has been used more , with one work [24] achieving 0.843 mean average precision (mAP), another work [25] achieving 0.852 mAP, and a benchmark work [26] of 0.855 mAP.

3. Method

3.1. Hypothesis

In this study, we explore the impact of scaled multi-spectrogram inputs on the performance of a hierarchical residual attention network for musical instrument recognition. The Universal Approximation Theorem [27] suggests that a neural network can approximate any continuous function, as expressed in (1).

$$Y = w \cdot x + b \tag{1}$$

where Y is the output, w is the weight, x is the input, and b is the bias.

Building on the Universal Approximation Theorem [27], we preprocess the spectrograms by applying fixed scales to different spectrogram components to enhance classification performance, as illustrated in (2). This approach optimizes the representation of each spectrogram type before they are fed into the model, potentially leading to improved accuracy.

$$Y = w(S_1 * LogMel(x) + S_2 * Chroma(x) + S_3 * SpectralContrast(x) + S_4 * Tonnetz(x)) \tag{2}$$

where Y is the output, w is the overall weight, $(S_1, S_2, S_3,)$and $(S_4)$are the scale for the Log-Mel, Chroma, Spectral Contrast, and Tonnetz spectrogram features respectively, and (x) represents the input audio signal.

This approach hypothesizes (2) that by assigning appropriate scale to these components, the model can better capture the distinct characteristics of each feature type, leading to a more accurate classification of musical instruments. The rationale behind this hypothesis is that different spectrogram features highlight unique aspects of the audio signal, and a scaled combination can provide a more comprehensive representation.

3.2. Data Pre-Processing

In this experiment, we use the OpenMIC-2018 dataset [23], an open-source, multilabel music instrument annotated database. The OpenMIC-2018 dataset contains over 20,000 ten-second audio clips, each annotated with multiple instrument labels, providing a diverse and extensive collection of musical instrument samples for analysis. The dataset includes 20 different musical instruments, and all recordings are real audio recordings capturing the authentic characteristics of various instruments in natural settings.

Figure 1 illustrates various audio feature representations extracted from a musical instrument sample and the effects of combining these features at different scales. The log-mel spectrogram in (a) has 128 Mel frequency bins, the chroma features in (b) map to 12 pitch classes, the spectral contrast in (c) has seven bins, and the Tonnetz in (d) uses six bins to represent harmonic intervals. Combining these features at their original sizes, as shown in (b), results in an overwhelming dominance of the Log-Mel spectrogram due to its larger size. To address this, different scaling approaches were generated: (c) shows the features combined at 1/4 of the Log-Mel's size (32), (d) at half size (64), (e) at 3/4 size (96), and (f) with all features scaled to the same size (128 bins). OpenCV [28] was used to stretch the spectral contrast, chroma, and Tonnetz features to match the Log-Mel's 128 bins or different scaled size settings, creating balanced and coherent representations.

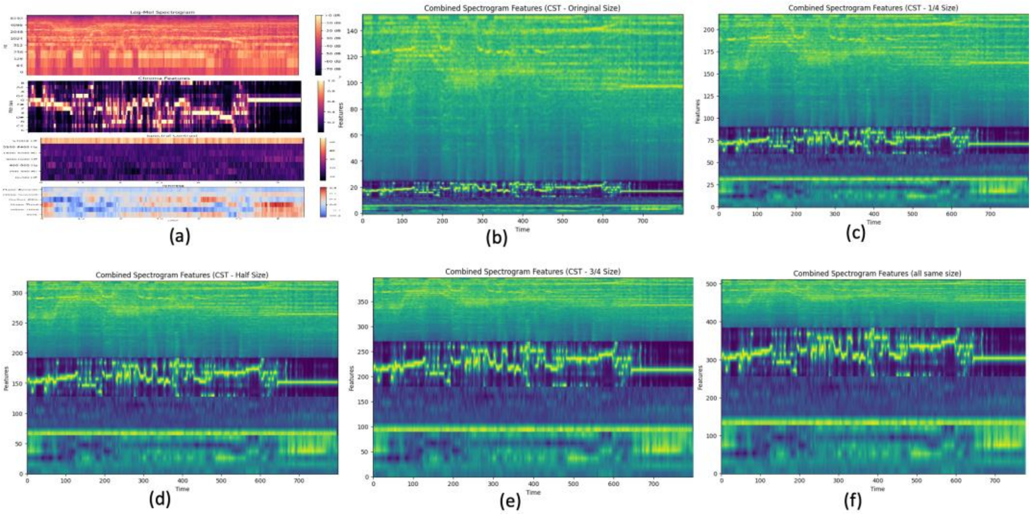

Figure 1. Audio feature representations for a musical instrument sample: (a) Log-Mel Spectrogram, Chroma Features, Spectral Contrast, and Tonnetz combined at original sizes, (b) Combined Spectrogram Features (CST - Original Size), (c) Combined Spectrogram Features (CST - 1/4 Size), (d) Combined Spectrogram Features (CST - Half Size), (e) Combined Spectrogram Features (CST - 3/4 Size), and (f) Combined Spectrogram Features (all same size).

3.3. Convolutional Neural Network Structure

The neural network model used here (Figure 2) classifies musical instruments using spectrogram inputs. It consists of three residual blocks with 32, 64, and 128 filters, respectively, each followed by MaxPooling to reduce dimensions. Attention mechanisms are applied at multiple stages: Early Attention after the first block, Mid Attention after the second, and Late Attention after the third, along with Channel and Coordinate Attention to emphasize critical features. The output is flattened, passed through a dense layer with 512 units and ReLU activation, followed by a 40% dropout layer, and finally a sigmoid activation layer for multi-label classification. The model uses binary cross-entropy loss, Adam optimizer, and accuracy as the metric.

doi:10.20944/preprints202409.1632.v

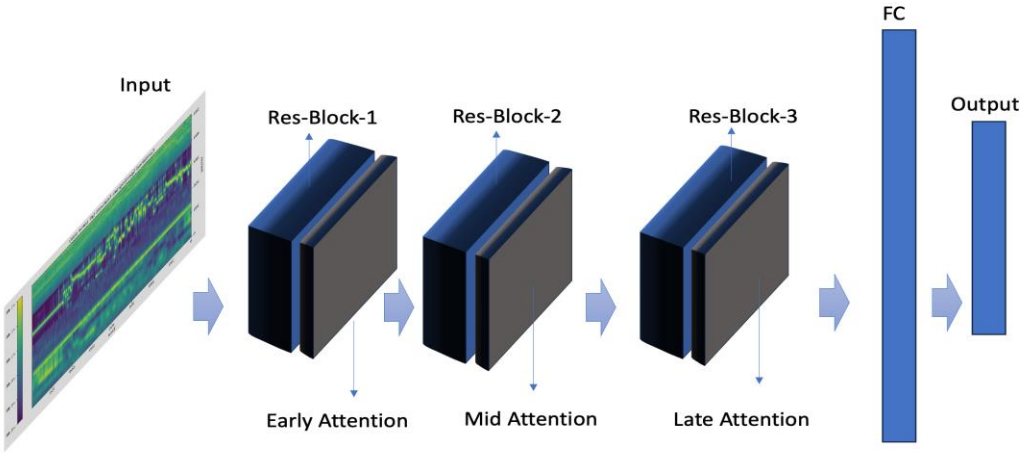

Figure 2. Overview of the neural network architecture used for musical instrument classification. The model includes three Residual Blocks, each followed by an attention layer (Early Attention, Mid Attention, Late Attention), and a final fully connected (FC) layer for classification).

This architecture effectively combines residual connections, attention mechanisms, and dense layers to achieve robust feature extraction and classification performance for musical instrument recognition tasks.

4. Results

4.1. Benchmark Comparison

Figure 3 presents a comparison of mean average precision (mAP) achieved by different methods on the Open-MIC dataset. The x-axis lists the methods, while the y-axis indicates the mAP values. The graph includes benchmark methods, our proposed methods, and highlights our best model. The benchmark methods serve as a reference point, showing the progression in performance over the years, including the Baseline [23] , PaSST [24], EAsT-KD + PaSST [25], and DyMN-L [26]. Our methods include various configurations of combining spectrogram features and attention mechanisms. Specifically, the "Single Log-Mel" approach, "Log-Mel CST Combined Spectrogram," and "Log-Mel CST with Attention Layer" configurations explore the impact of different spectrogram features on model performance. Additionally, we experimented with magnifying the spectrogram features to different extents: "Magnified 1/4 Size," "Magnified 1/2 Size," "Magnified 3/4 Size," and "Magnified Full Size."

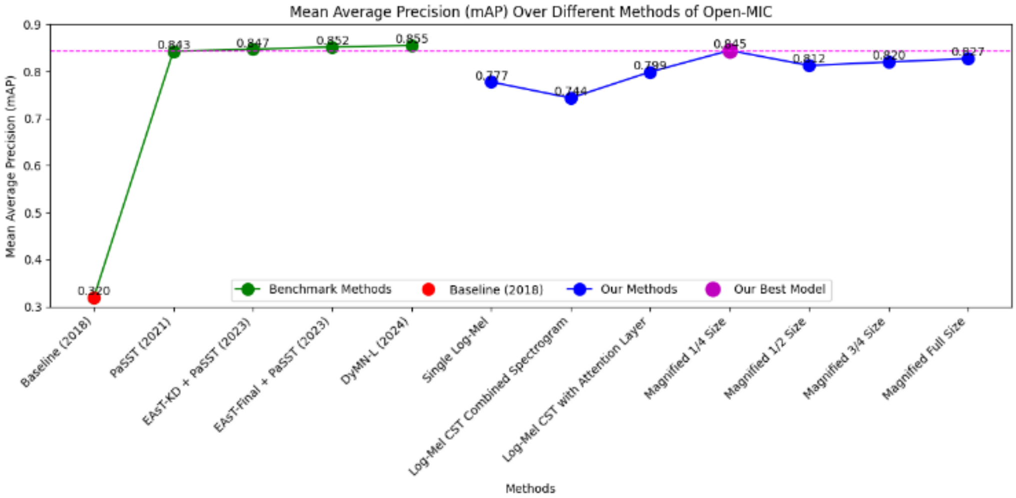

Figure 3. Mean Average Precision (mAP) Comparison Across Various Methods on the Open-MIC Dataset.

Among these, the "Magnified 1/4 Size" model achieved a mAP of 0.8445, demonstrating a performance close to the leading benchmark methods. This model's success highlights the importance of carefully scaling the spectrogram features and incorporating attention mechanisms to enhance the model's focus on the most informative parts of the input data. The magenta marker and horizontal magenta line on the graph emphasize the noteworthy performance of this model, illustrating the potential effectiveness of our approach in musical instrument recognition tasks.

4.2. Evaluation Metrics Comparison among Each Scaled Multi-Spectrogram Settings

Figure 4 presents a comprehensive comparison of the precision, recall, and F1-score metrics for various instrument recognition models using different configurations of spectrogram features. The x-axis represents the different musical instruments, while the y-axis indicates the metric values.

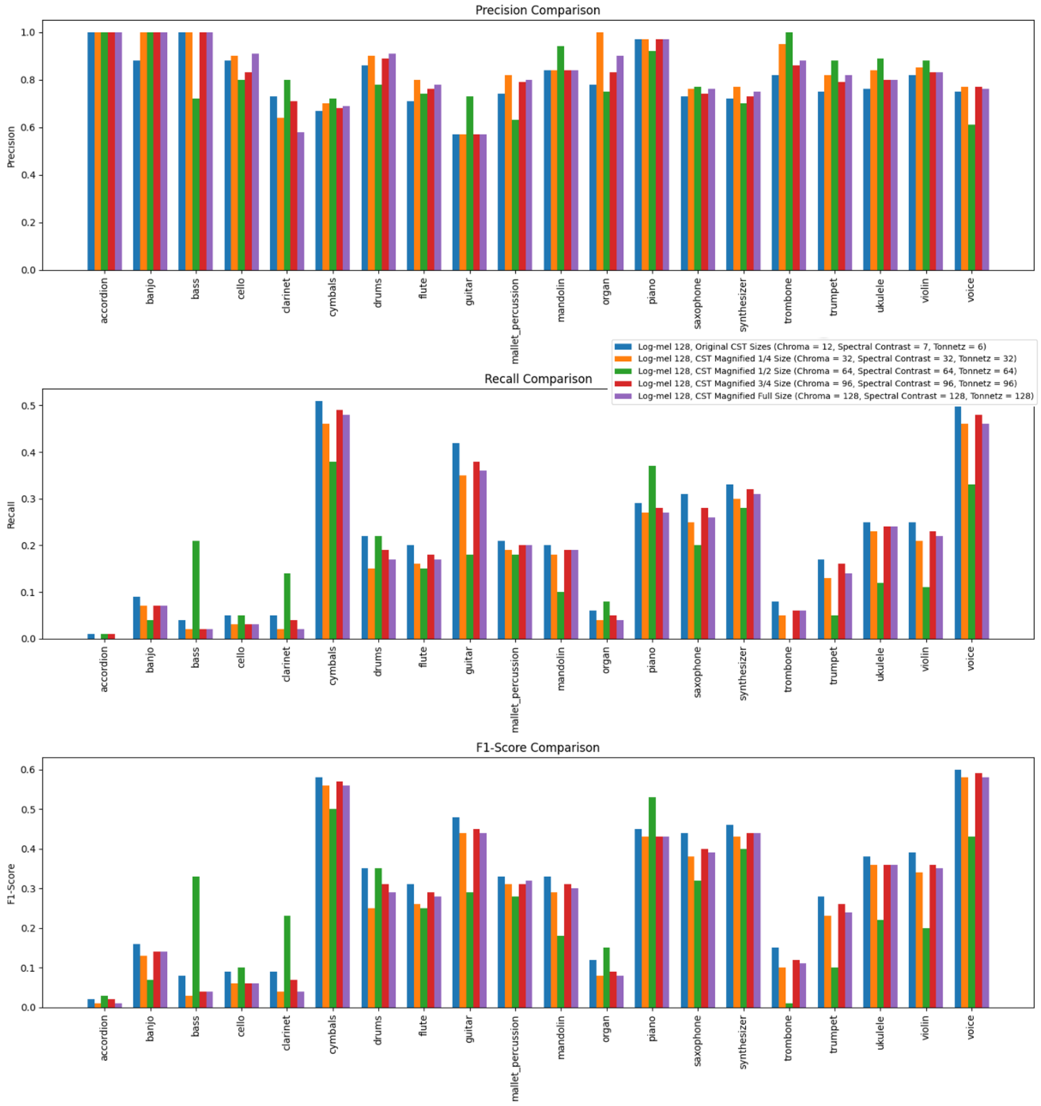

Figure 4. Precision, Recall, and F1-Score Comparisons for Different Spectrogram scaled size.

The models compared include configurations such as Log-Mel 128 with Original CST Sizes (Chroma = 12, Spectral Contrast = 7, Tonnetz = 6), Log-Mel 128 with CST Magnified to 1/4 Size (Chroma = 32, Spectral Contrast = 32, Tonnetz = 32), Log-Mel 128 with CST Magnified to 1/2 Size (Chroma = 64, Spectral Contrast = 64, Tonnetz = 64), Log-Mel 128 with CST Magnified to 3/4 Size (Chroma = 96, Spectral Contrast = 96, Tonnetz = 96), and Log-Mel 128 with CST Magnified to Full Size (Chroma = 128, Spectral Contrast = 128, Tonnetz = 128). Each sub-plot within the figure illustrates a

doi:10.20944/preprints202409.1632.v

specific metric comparison, with the top plot showing Precision Comparison, the middle plot showing Recall Comparison, and the bottom plot showing F1-Score Comparison.

According to Figure 4, for instruments like accordion, banjo, bass, drums, guitar, marimba, piano, synthesizer, and trumpet, high precision is consistently maintained across all configurations, indicating effective differentiation with minimal false positives. However, for cello, clarinet, flute, mandolin, violin, and voice, precision varies, suggesting certain spectrogram features better reduce false positives. Notably, the "Log-Mel 128 CST Magnified 1/4 Size" configuration generally provides a balanced performance, capturing essential characteristics effectively. Instruments like cymbals, organ, saxophone, and trombone exhibit significant precision fluctuations, indicating overlapping features with other instruments, making accurate differentiation more challenging.

Recall metrics reveal consistently high detection rates for accordion, banjo, bass, piano, synthesizer, trumpet, ukulele, violin, and voice, showcasing the model's effectiveness. Variability in recall for cello, clarinet, cymbals, flute, guitar, marimba, mandolin, and saxophone, with the "Log-mel 128 CST Magnified Full Size" often achieving higher recall, suggests larger feature sizes capture more relevant characteristics. Lower and more variable recall rates for drums, organ, and trombone indicate these instruments are less distinct or more challenging to detect accurately. High and consistent F1-scores for instruments like accordion, banjo, bass, drums, guitar, piano, synthesizer, and trumpet reflect a good balance between precision and recall. In contrast, cello, clarinet, cymbals, flute, mandolin, organ, saxophone, and trombone show fluctuating F1-scores, with the "Log-mel 128 CST Magnified 1/4 Size" and "Log-mel 128 CST Magnified Full Size" configurations often performing better. This indicates these configurations provide a better trade-off between detecting instruments and minimizing false predictions, while voice, violin, marimba, and ukulele show variability in F1-scores, suggesting room for optimization in feature size and attention mechanisms.

5. Discussion

The different configurations of CST feature sizes (Original, 1/4, 1/2, 3/4, Full) highlight how scaling affects model performance. Smaller sizes generally provide a compact representation, leading to higher precision but potentially lower recall. Conversely, larger sizes capture more details, improving recall but possibly introducing more false positives. The "Magnified 1/4 Size" configuration often achieves a good balance, making it a preferred choice for general purposes. The variability in performance across different instruments suggests that certain instruments benefit more from specific feature configurations. For instance, instruments like drums and organ might require more sophisticated feature combinations or additional attention mechanisms to enhance detection accuracy. Incorporating attention mechanisms at various stages helps the model focus on the most relevant parts of the spectrogram, improving overall performance. The results indicate that these mechanisms are crucial, particularly for instruments with overlapping or subtle features. Further research could explore optimizing CST feature sizes and experimenting with other attention mechanisms. Additionally, applying the hierarchical residual attention network to other audio classification tasks could test its generalizability and effectiveness beyond musical instrument recognition

5.1. Early Attention Layer Analysis

In the early layers (Figure 5), the feature maps primarily capture basic structures and low-level features of the spectrogram. As observed in the first and second row of images, filters from the early convolutional layers (Conv Layer 1 and Conv Layer 2) identify fundamental frequency components and broad harmonic structures. The filters show a high level of activity across the entire spectrogram, indicating that the network is learning to detect general patterns and frequencies.

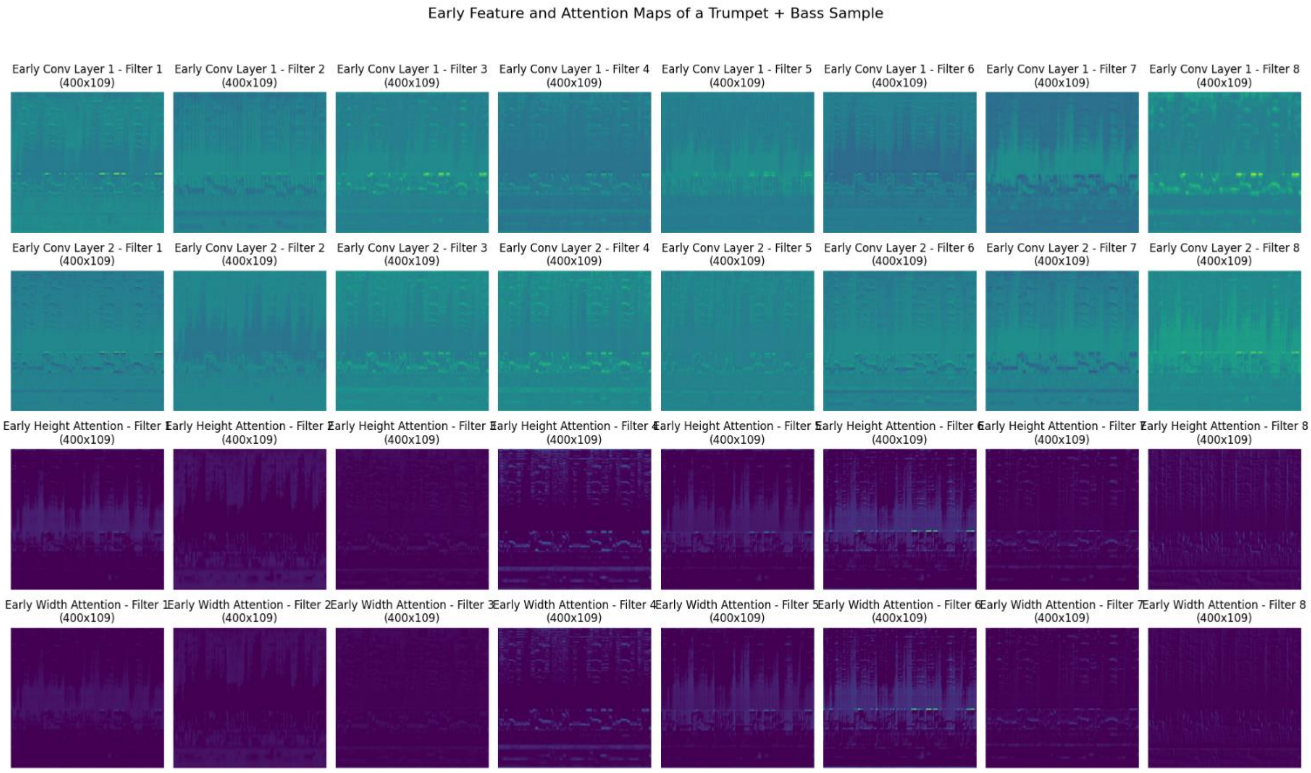

Figure 5. Early Conv Layer 1 and Early Conv Layer 2 feature maps of a trumpet and bass sample, showing the network's initial extraction of basic structures and low-level features. Height and width attention maps illustrate the focus on specific frequency bands and temporal patterns, respectively.

The attention maps on second and third row highlight the importance of vertical segments of the spectrogram, which correlate to specific frequency bands. The height attention maps reveal that the network is focusing on certain frequency ranges more than others, possibly identifying critical harmonic regions of the trumpet and bass sounds. The width-based attention maps suggest that the network is also learning to identify temporal patterns and consistency across time. The early width attention maps show how the model emphasizes different time segments, helping to capture rhythmic and temporal features of the input sounds.

5.2. Mid Attention Layer Analysis

In the mid layers (Figure 6), the network refines its understanding of the input spectrogram, capturing more detailed and intermediate-level features. The feature maps in the first and second row of the mid layer images show more complex patterns and finer details than those in the early layers. The filters here are responsive to specific harmonic and melodic structures, which are essential for distinguishing between different musical instruments.

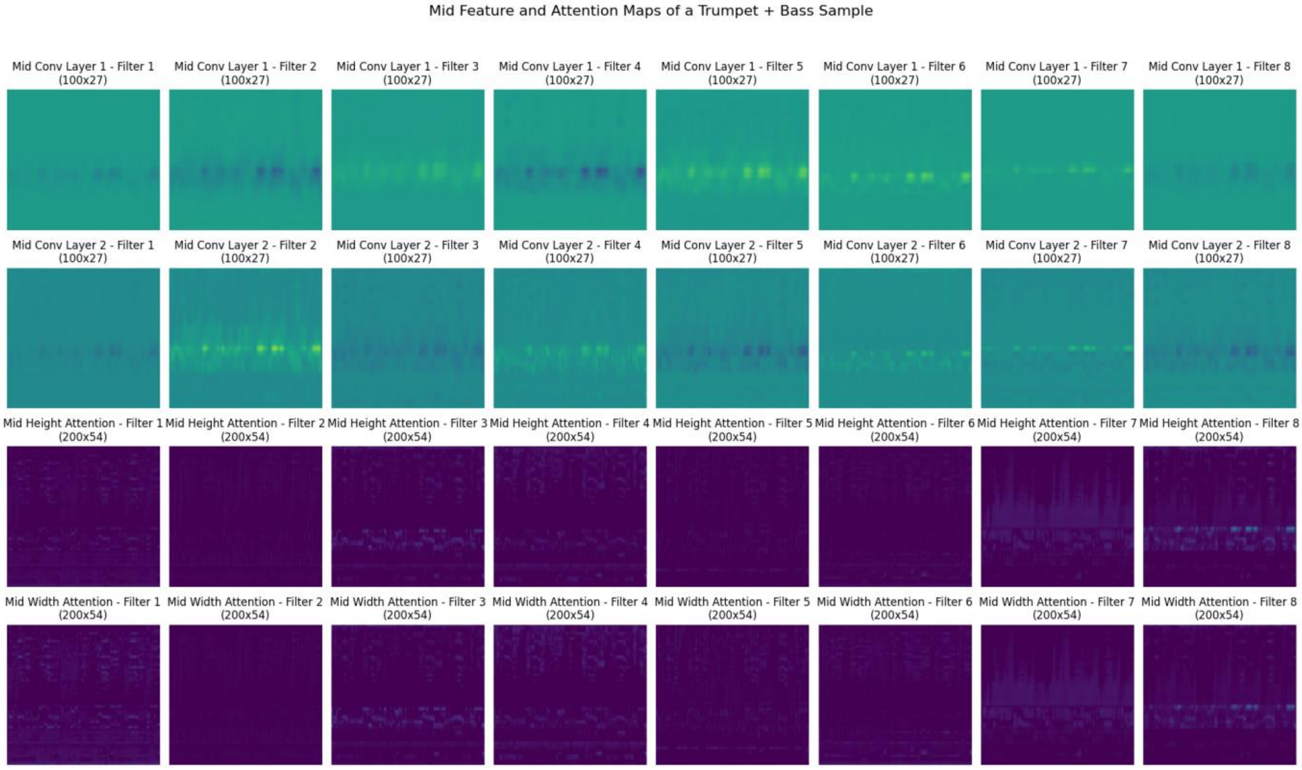

Figure 6. Mid Conv Layer 1 and Mid Conv Layer 2 feature maps for a trumpet and bass sample, highlighting the network's refinement of intermediate-level features. Height and width attention maps show the network's increased focus on specific harmonic and melodic structures, as well as temporal variations.

The height attention maps in the mid layers further narrow down the frequency bands of interest, showing more focused attention on specific harmonics and overtones. This level of attention helps in identifying unique timbral characteristics of the instruments, such as the brightness of the trumpet or the depth of the bass.

Also, width attention maps in the mid layers continue to highlight temporal structures, but with more precision compared to the early layers. These maps demonstrate the network's ability to track temporal variations and dynamics within the spectrogram, crucial for capturing the expressive qualities of musical performances.

5.3. Late Attention Layer Analysis

In the late layers (Figure 7), the network consolidates its feature extraction to capture high-level, abstract features that are directly relevant to the classification task. The convolutional feature maps in the late layers (first row) show highly specific patterns, often isolating key harmonic and rhythmic elements that are distinctive for each instrument.

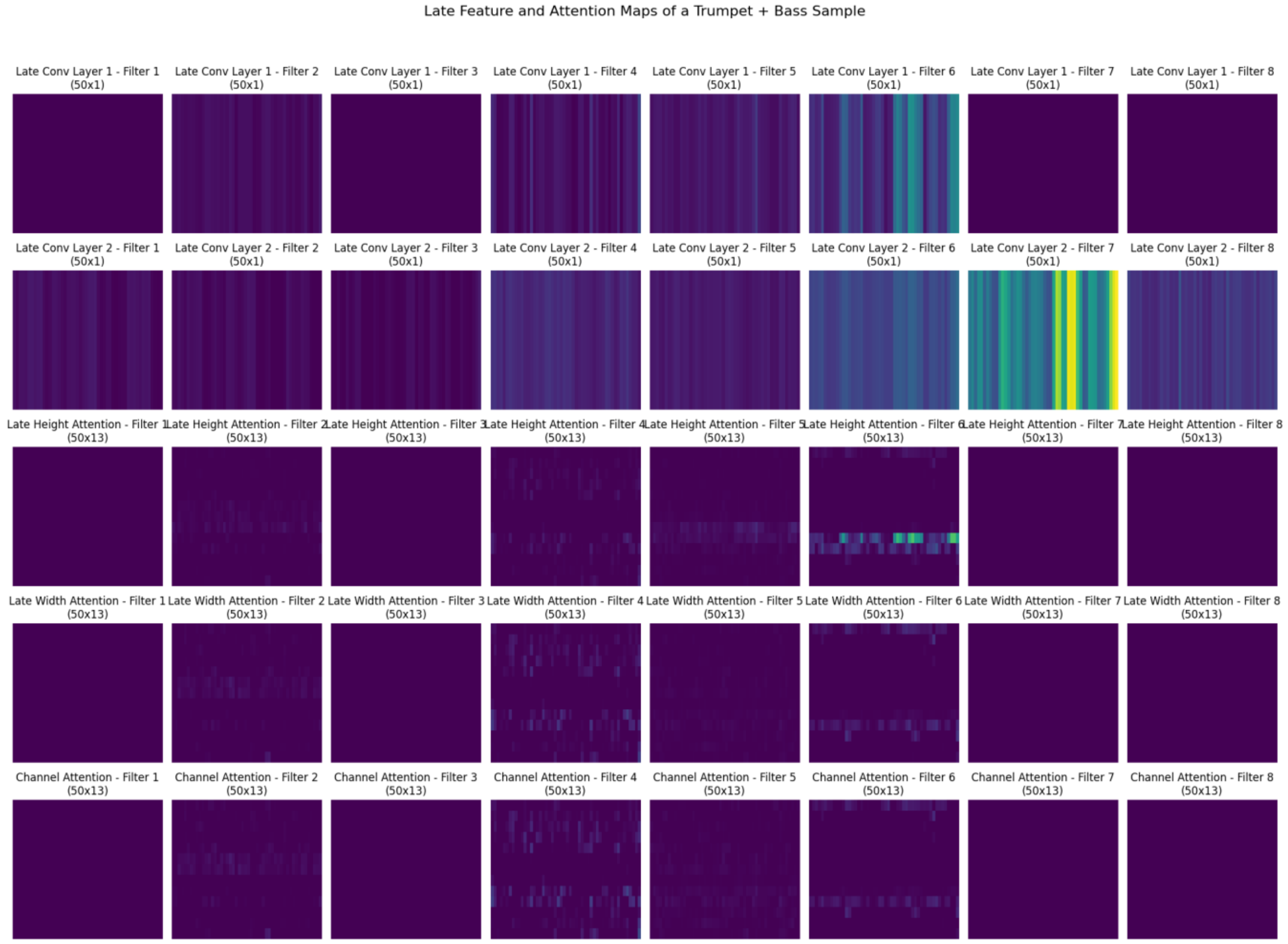

Figure 7. Late Conv Layer 1 and Late Conv Layer 2 feature maps of a trumpet and bass sample, capturing high-level, abstract features. Height, width, and channel attention maps emphasize the most critical frequency bands, temporal segments, and feature maps, respectively.

Height attention maps in the late layers provide a final refinement of frequency focus, isolating the most critical frequency bands that define the instrument's timbre. These attention maps are sparser, indicating that the network has focused on the most informative frequency components.

Width attention maps in the late layers show a similar sparsity, with attention concentrated on the most temporally significant segments of the spectrogram. This focused attention helps the network in making fine-grained distinctions between similar instruments and identifying subtle temporal nuances.

Lastly, the channel attention maps (last row) illustrate the network's emphasis on specific channels or feature maps within the spectrogram. This type of attention helps in weighing the contribution of different feature maps, ensuring that the most relevant features dominate the final classification decision.

6. Conclusion

In this study, we presented a hierarchical residual attention network for musical instrument recognition using scaled multi-spectrogram features. By combining Log-Mel feature with CST features (chroma, spectral contrast, and Tonnetz), our model captures a more comprehensive representation of audio signals. The attention mechanisms, applied at various stages of the network, allow the model to focus on the most relevant parts of the combined spectrogram, improving classification performance.

Our experiments demonstrated that the "Magnified 1/4 Size" configuration achieved the best balance of precision, recall, and F1-score, highlighting the effectiveness of scaling spectrogram features to enhance model performance. The use of hierarchical residual connections and attention mechanisms significantly improved the model's ability to classify instruments accurately, even those with subtle or overlapping features.

7. Future Work

Future research could enhance our model by making the scaling of spectrogram features learnable within the neural network itself. Instead of applying fixed scales during preprocessing, each spectrogram type—Log-Mel, Chroma, Spectral Contrast, and Tonnetz—could have its own adaptive weight in the network's initial layer. This would allow the model to automatically determine the optimal emphasis for each feature during training, potentially improving classification accuracy. This approach eliminates the need for manual scaling in pre-processing stage and enables the model to focus more effectively on the most informative features for instrument recognition.

Additionally, analyzing the learned weights could provide insights into the relative importance of different spectrogram features, guiding future feature selection and extraction methods. Exploring this adaptive scaling in combination with advanced attention mechanisms or architectures could further enhance the model's performance. Applying this strategy to other audio classification tasks would help evaluate its generalizability and effectiveness across various applications in audio processing.

Also, experimenting with other attention mechanisms or combinations like Vision Transformer [29] reveal further performance enhancements, with a particular focus on each spectrogram individually to refine the model's ability to focus on the most informative parts of the spectrogram.

8. Reproducibility

The data and code that support the findings of this experiment are openly available in our GitHub repository at https://github.com/fireHedgehog/music-intrument-OvA-model/tree/main/open-mic .

References

1. Chi, Z.; Li, Y.; Chen, C. Deep Convolutional Neural Network Combined with Concatenated Spectrogram for Environmental Sound Classification. In Proceedings of the 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT); IEEE, 2019; pp. 251–254.
2. Ghosal, D.; Kolekar, M.H. Music Genre Recognition Using Deep Neural Networks and Transfer Learning. In Proceedings of the Interspeech; 2018; pp. 2087–2091.
3. Xiao, H.; Liu, D.; Chen, K.; Zhu, M. AMResNet: An Automatic Recognition Model of Bird Sounds in Real Environment. *Applied Acoustics* **2022**, *201*, 109121.
4. Xing, Z.; Baik, E.; Jiao, Y.; Kulkarni, N.; Li, C.; Muralidhar, G.; Parandehgheibi, M.; Reed, E.; Singhal, A.; Xiao, F.; et al. Modeling of the Latent Embedding of Music Using Deep Neural Network. *arXiv preprint arXiv:1705.05229* **2017**.
5. Kethireddy, R.; Kadiri, S.R.; Alku, P.; Gangashetty, S.V. Mel-Weighted Single Frequency Filtering Spectrogram for Dialect Identification. *IEEE Access* **2020**, *8*, 174871–174879.
6. Wang, Q.; Su, F.; Wang, Y. A Hierarchical Attentive Deep Neural Network Model for Semantic Music Annotation Integrating Multiple Music Representations. In Proceedings of the Proceedings of the 2019 on International Conference on Multimedia Retrieval; 2019; pp. 150–158.
7. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2021; pp. 13713–13722.
8. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Advances in neural information processing systems* **2017**, *30*.
9. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition; 2016; pp. 770–778.
10. Humphrey, E.J.; Bello, J.P.; LeCun, Y. Moving beyond Feature Design: Deep Architectures and Automatic Feature Learning in Music Informatics. In Proceedings of the ISMIR; 2012; pp. 403–408.

11. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE* **1998**, *86*, 2278–2324.
12. McLoughlin, I.; Zhang, H.; Xie, Z.; Song, Y.; Xiao, W. Robust Sound Event Classification Using Deep Neural Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2015**, *23*, 540–552.
13. Griffin, D.; Lim, J. Signal Estimation from Modified Short-Time Fourier Transform. *IEEE Transactions on acoustics, speech, and signal processing* **1984**, *32*, 236–243.
14. Davis, S.; Mermelstein, P. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE transactions on acoustics, speech, and signal processing* **1980**, *28*, 357–366.
15. Slaney, M. Auditory Toolbox. *Interval Research Corporation, Tech. Rep* **1998**, *10*, 1194.
16. Pham, L.; Phan, H.; Nguyen, T.; Palaniappan, R.; Mertins, A.; McLoughlin, I. Robust Acoustic Scene Classification Using a Multi-Spectrogram Encoder-Decoder Framework. *Digital Signal Processing* **2021**, *110*, 102943.
17. Schmidt, E.M.; Kim, Y.E. Learning Rhythm And Melody Features With Deep Belief Networks. In Proceedings of the ISMIR; 2013; pp. 21–26.
18. Han, K.; Yu, D.; Tashev, I. Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine. In Proceedings of the Interspeech 2014; 2014.
19. Schlüter, J.; Gutenbrunner, G. Efficientleaf: A Faster Learnable Audio Frontend of Questionable Use. In Proceedings of the 2022 30th European Signal Processing Conference (EUSIPCO); IEEE, 2022; pp. 205–208.
20. Engel, J.; Resnick, C.; Roberts, A.; Dieleman, S.; Norouzi, M.; Eck, D.; Simonyan, K. Neural Audio Synthesis of Musical Notes with Wavenet Autoencoders. In Proceedings of the International Conference on Machine Learning; PMLR, 2017; pp. 1068–1077.
21. Bosch, J.J.; Fuhrmann, F.; Herrera, P. IRMAS: A Dataset for Instrument Recognition in Musical Audio Signals 2018.
22. Racharla, K.; Kumar, V.; Jayant, C.B.; Khairkar, A.; Harish, P. Predominant Musical Instrument Classification Based on Spectral Features. In Proceedings of the 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN); IEEE, 2020; pp. 617–622.
23. Humphrey, E.; Durand, S.; McFee, B. OpenMIC-2018: An Open Data-Set for Multiple Instrument Recognition. In Proceedings of the ISMIR; 2018; pp. 438–444.
24. Koutini, K.; Schlüter, J.; Eghbal-Zadeh, H.; Widmer, G. Efficient Training of Audio Transformers with Patchout. *arXiv preprint arXiv:2110.05069* **2021**.
25. Ding, Y.; Lerch, A. Audio Embeddings as Teachers for Music Classification. *arXiv preprint arXiv:2306.17424* **2023**.
26. Schmid, F.; Koutini, K.; Widmer, G. Dynamic Convolutional Neural Networks as Efficient Pre-Trained Audio Models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2024**.
27. Hornik, K.; Stinchcombe, M.; White, H. Multilayer Feedforward Networks Are Universal Approximators. *Neural networks* **1989**, *2*, 359–366.
28. Bradski, G. The Opencv Library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer* **2000**, *25*, 120–123.
29. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision; 2021; pp. 10012–10022.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.