ATTENTION-BASED PREDOMINANT INSTRUMENTS RECOGNITION IN POLYPHONIC MUSIC

Lekshmi C. REGHUNATH[1] and **Rajeev RAJAN**[1]

[1]*College of Engineering Trivandrum*, **APJ Abdul Kalam Technological University**, Trivandrum, Kerala, India

ABSTRACT

Predominant instrument recognition in polyphonic music is addressed using the score-level fusion of two visual representations, namely, Mel-spectrogram and modgdgram. Modgdgram, a visual representation is obtained by stacking modified group delay functions of consecutive frames successively. Convolutional neural networks (CNN) with an attention mechanism, learn the distinctive local characteristics and classify the instrument to the group where it belongs. The proposed system is systematically evaluated using the IRMAS dataset with eleven classes. We train the network using fixed-length single-labeled audio excerpts and estimate the predominant instruments from variable-length audio recordings. A wave generative adversarial network (WaveGAN) architecture is also employed to generate audio files for data augmentation. The proposed system reports a micro and macro F1 score of 0.65 and 0.60, respectively, which is 20.37% and 27.66% higher than those obtained by the state-of-the-art Han model. The experiments demonstrate the potential of CNN with attention mechanism on Mel-spectro/modgdgram fusion framework for the task of predominant instrument recognition.

1. INTRODUCTION

Predominant instrument recognition refers to the problem where the prominent instrument is identified from a mixture of instruments being played together [1]. The auditory scene produced by a musical composition can be regarded as a multi-source environment, where different sound sources are played at various pitches and loudness, and even the spatial position of a given sound source may vary with respect to time [2]. In polyphonic music, the interference of simultaneously occurring sounds makes instrument recognition harder. Automatic identification of lead instrument is important since it helps to enhance fundamental music information retrieval (MIR) tasks like source separation [2] auto-tagging [3], and automatic music transcription [4].

Han *et al.* [1] employed the Mel-spectrogram-CNN approach for predominant instrument recognition in polyphonic music using an aggregation strategy over sliding windows. Pons *et al.* [5] analyzed the architecture of Han *et al.* in order to formulate an efficient design strategy to capture the relevant information about timbre. Both approaches were trained and validated by the IRMAS dataset of polyphonic music excerpts. Detecting the activity of music instruments using a deep neural network (DNN) through a temporal max-pooling aggregation is addressed in [6]. The paper [7] employed an attention mechanism and multiple-instance learning (MIL) framework to address the challenge of weakly labeled instrument recognition in the OpenMIC dataset. Dongyan Yu *et al.* [8] employed a network with an auxiliary classification scheme to learn the instrument categories through multitasking learning. Gomez *et al.* [9] investigated the role of two source separation algorithms as pre-processing steps to improve the performance in the context of predominant instrument detection tasks. It was found that both source separation and transfer learning could significantly improve the recognition performance, especially for a small dataset composed of highly similar musical instruments. In [10], the Hilbert-Huang transform (HHT) is employed to map one-dimensional audio data into two-dimensional matrix format, followed by CNN to learn the affluent and effective features for the task. In [11] an ensemble of VGG-like CNN classifiers, trained on non-augmented, pitch-synchronized, tempo-synchronized, and genre-similar excerpts of IRMAS for the proposed task. The modified group delay feature (MODGDF) has already been proposed for pitched musical instrument recognition in an isolated environment [12] and polyphonic predominant instrument recognition [13]. Bosch *et al.* improved the algorithm proposed in [2] with source separation in a pre-processing step [14]. While the commonly applied mel frequency cepstral coefficients (MFCC) feature is capable of modeling the resonances introduced by the filter of the instrument body, it neglects the spectral characteristics of the vibrating source, which also play its role in human perception of musical sounds and genre classification [15]. Incorporating phase information is an effective attempt to preserve this neglected component. It has already been established in the literature that the modified group delay function emphasizes peaks in spectra well [16]. The idea of including modgdgram, GAN-based data augmentation strategy, and CNN with multi-head attention are the main contributions of the proposed scheme.

The rest of the paper is organized as follows. Section 2 gives an overview of the proposed model. The model architecture is described in Section 3. The performance evaluation is described in Section 4, followed by results in

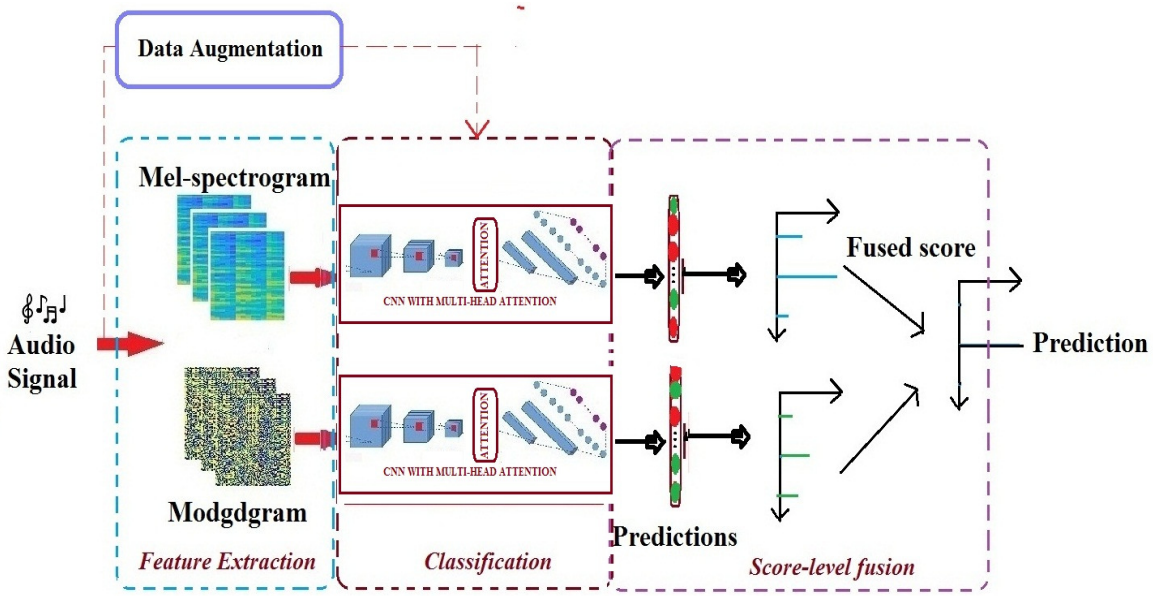

Figure 1. Block diagram of the proposed method of predominant instrument recognition.

Section 5. The paper is concluded in Section 6.

2. SYSTEM DESCRIPTION

The block diagram of the proposed method of predominant instrument recognition is illustrated in Figure 1. In the proposed model, CNN with multi-head attention is used to learn the distinctive characteristics from Mel-spectro/modgd-gram to identify the leading instrument in polyphonic context. As a part of data augmentation, additional training files are generated using WaveGAN. During the testing phase, the probability value at the output nodes of the trained model is treated as the score corresponding to the input test file. The input audio file is classified to the node which gives the maximum score during testing. In the fusion framework, the individual scores of Mel-spectro/modgd-gram experiments are fused at the score-level to make a decision. The fusion score S_f, is obtained by,

$$S_f = \beta S_{spectro} + (1-\beta) S_{modgd} \quad (1)$$

where $S_{spectro}$, S_{modgd}, β are the Mel-spectrogram score, modgdgram score and weighting constant, respectively. β = 0.5 is chosen empirically in our experiment. The performance of the proposed system is compared with that of Han's model and a DNN framework. Feature extraction is described in the following subsections.

2.1 Mel-spectrogram

Mel-spectrogram is widely used in recent music processing applications [17, 18]. Mel-spectrogram approximates how the human auditory system works and can be seen as the spectrogram smoothed, with high precision in the low frequencies and low precision in the high frequencies [19]. It is computed with a frame size of 50 ms and a hop size of 10 ms with 128 bins for the given task.

2.2 Modified group delay functions and Modgdgram

Group delay features are being employed in numerous speech and music processing applications [16, 20–22]. The group delay function is defined as the negative derivative of the unwrapped Fourier transform phase with respect to frequency. Modified group delay functions (MODGD), $\tau_m(e^{j\omega})$ are obtained by,

$$\tau_m(e^{j\omega}) = \left(\frac{\tau_c(e^{j\omega})}{|\tau_c(e^{j\omega})|}\right)(|\tau_c(e^{j\omega})|)^a, \quad (2)$$

where,

$$\tau_c(e^{j\omega}) = \frac{X_R(e^{j\omega})Y_R(e^{j\omega}) + Y_I(e^{j\omega})X_I(e^{j\omega})}{|S(e^{j\omega})|^{2b}}. \quad (3)$$

The subscripts R and I denote the real and imaginary parts, respectively. $X(e^{j\omega})$, $Y(e^{j\omega})$ and $S(e^{j\omega})$ are the Fourier transforms of signal, $x[n]$, n.$x[n]$ (signal multiplied with index), and the cepstrally smoothed version of $X(e^{j\omega})$, respectively. a and b ($0 < a,\ b \leq 1$) are introduced to control the dynamic range of MODGD [16]. Modgdgram is the visual representation of MODGD with time and frequency in the horizontal and vertical axis, respectively. The amplitude of the group delay function at a particular time is represented by the intensity or color in the third dimension. Modgdgrams are computed with a frame size of 50 ms and hop size of 10 ms using a and b values of 0.9 and 0.5 respectively.

3. MODEL ARCHITECTURE

3.1 Fusion without attention

First, we designed a three-layer CNN to encode the Mel-spectrogram or modgdgram images. Filters of sizes 32, 64, and 192 are used in the convolutional layers. Each convolutional layer is followed by 2x2 max-pooling. ReLU is used as the activation in the hidden layer and the same padding is employed to maintain the spatial resolution. We used filters with a very small 2×2 receptive field, with a fixed stride size of 1. For implementing fusion without attention, we flattened the last convolutional layer followed by fully connected layers. 20% of training data is used for tuning the hyperparameters during training. Softmax is used as the activation function for the output layer with 11 outputs. The number of parameters learned by the baseline model is 1830315 parameters.

3.2 Attention model

Multi-head attention is employed after three convolutional layers. It expands the model's ability to focus on different positions of Mel-spectrogram/modgdgram. For constructing the attention the principle behind is to create smaller linear representations of the same block by splitting the content of a block into query vectors(q), key vectors(k), and value vectors(v). Using key and query vectors we can create weights for the value vectors that will be used to create the output vector. In the paper [23], such attention blocks are used to encode and decode the sentences. The various parameters chosen are shown in Table 1.

N	l	d	h	dv	$dout$	dk
6	6*6 = 36	64*3 = 192	8	8*3 = 24	32	36

Table 1. Various hyperparameters (top row) and its values (bottom row) selected for attention model.

where N represents the number of encoders or decoders present in the self-attention layer, l is the number of blocks in the feature map the convolutional network made, d is the dimension of block, dv is the dimension of linear space the input to be projected, $dout$ is the output of the block after applying attention and h is the number of heads or number of projections for each block. dk is the dimension of the query vector. $q1, k1, v1$ are the query, key and value vectors of input and $q2, k2, v2$ are the h projections with size dv of the $q1, k1, v1$ vectors. Then each query vector is multiplied with each key vector to get the softmax predictions over h value vectors. The outputs from all 8 attention heads are concatenated to form a single output vector before passing it through the feed-forward network. After the attention layer, a normalization layer is also added to increase the speed of convergence. It makes the tensor have a standard normal distribution, at the same time it acts as another smaller attention by deleting some dimensions of the vector that are not important. The norm layer is followed by flattened layers. The network is trained using adam optimizer with a learning rate of 0.001. The network learns the model with 473163 parameters which are approximately 4 times smaller than the baseline model without attention. The model summary of the proposed method of CNN with multi-head attention is shown in Table 2.

Input size	Description
1x28x28	Modgdgram /Mel-spectrogram
32x28x28	2x2 Convolution, 32filters
32x14x14	2x2 Max-pooling
64x13x13	2x2 Convolution, 64 filters
64x7x7	2x2 Max-pooling
192x6x6	2x2 Convolution, 192 filters
192x36	Reshape
32x36	Multi-head attention
32x6x6	Reshape
32x6x6	Normalization layer
1152	Flattened and fully connected
256	Dense
11	Softmax

Table 2. Proposed CNN architecture with multi-head attention.

4. PERFORMANCE EVALUATION

4.1 Dataset

IRMAS dataset [2], comprising eleven classes, is used for the evaluation. The classes include cello (Cel), clarinet (Cla), flute (Flu), acoustic guitar (Gac), electric guitar (Gel), organ (Org), piano (Pia), saxophone (Sax), trumpet (Tru), violin (Vio) and human singing voice (Voice). The training data consists of 6705 audio files with excerpts of 3 s from more than 2000 distinct recordings. Since the dataset consists of testing audio samples with multiple predominant instruments as labels, we have considered all the audio files with a single predominant instrument (single label) during the testing phase.

4.2 Data augmentation using WaveGAN

GAN has been successfully applied to a variety of problems in image generation [24] and style transfer [25]. WaveGAN v2 is used here to generate polyphonic files with the leading instrument required for training. WaveGAN is similar to DCGAN, which is used for Mel-spectrogram generation, in various music processing applications. The transposed convolution operation of DCGAN is modified to widen its receptive field in WaveGAN. Specifically, longer one-dimensional filters of length 25 are used instead of two-dimensional filters of size 5x5 and are upsampled by a factor of 4 instead of 2 at each layer. The discriminator is also modified similarly, using length 25 filters in one dimension [26]. The output dimensionality of WaveGAN v2 is 65536 samples (corresponding to 4.01 s of audio at 16 kHz). For the generator, the input is a random noise uniformly distributed between -1 and 1. For training, the WaveGAN optimizes WGAN-GP using Adam for both generator and discriminator. A constant learning rate of

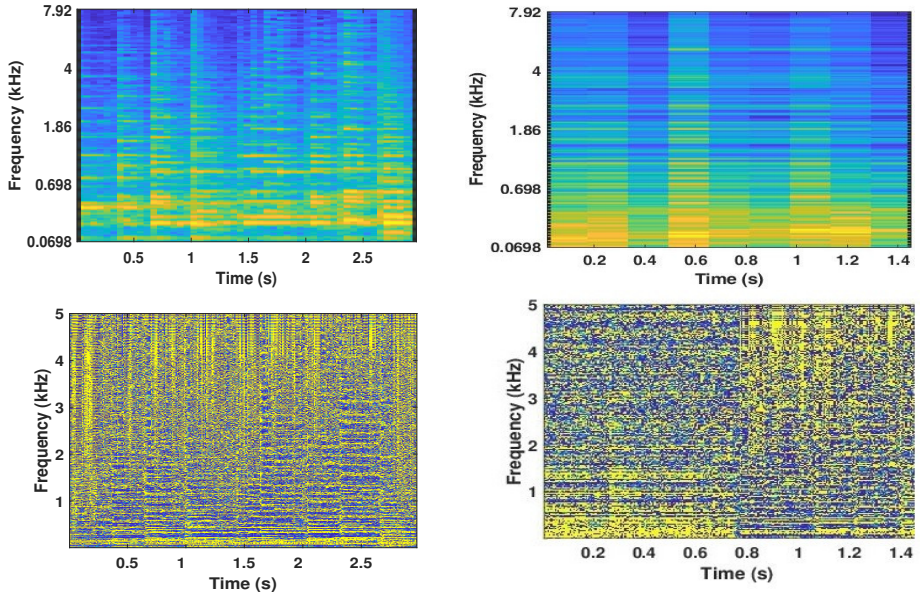

Figure 2. Visual representation of an audio excerpt with acoustic guitar as leading, Mel-spectrogram of original and WaveGAN-generated (Upper pane left and right). Modgdgram of original and WaveGAN-generated (Lower pane left and right).

0.0001 is used with $\beta_1 = 0.5$ and $\beta_2 = 0.9$. WaveGAN is trained for 2000 epochs on the three sec audio files of each class to generate similar audio files based on a similarity metric (s) [27] with a criteria $s > 0.1$. A total of 6585 audio files with cello (625), clarinet (482), flute (433), acoustic guitar (594), electric guitar (732), organ (657), piano (698), saxophone (597), trumpet (521), violin (526) and voice (720) are generated.

The quality of generated files is evaluated using a perception test. It is conducted with ten listeners to assess the quality of generated files for 275 files covering all classes. Listeners are asked to grade the quality by choosing one among the five opinion grades varying from poor to excellent quality (scores, 1 to 5). A mean opinion score of 3.64 is obtained. This value is comparable to the mos score obtained in [26] and [28] using WaveGAN. The generated files are denoted by $Train_g$ and training files available in the corpus are denoted by $Train_d$. Mel-spectrogram and modgdgram of natural and generated audio files for acoustic guitar are shown in Figure 2. The experiment details and a few audio files can be accessed at https://sites.google.com/view/audiosamples-2020/home/instrument

4.3 Experimental set-up

The experiment is progressed in three phases namely Mel-spectrogram-based, modgdgram-based, and score-level fusion-based. 1305 polyphonic files comprising eleven classes with a single label are used for the testing phase. The performance of the proposed method is compared with that of Han's model [1]. As different from our approach, they used a sliding window to perform short-time analysis, and sigmoid outputs were aggregated by taking class-wise average. After normalization, the candidate with maximum probability is assumed to be the most predominant instrument. Han's baseline model is implemented for the given experiment with 1 s slice length for performance comparison [1]

A DNN framework on musical texture features (MTF) is also experimented with to examine the performance of deep learning methodology on handcrafted features. MTF includes MFCC (13 dim), spectral centroid, spectral bandwidth, root mean square energy, spectral roll-off, and chroma STFT. The features are computed with a frame size of 40 ms and a hop size of 10 ms using Librosa framework [2]. DNN consists of seven layers, with increasing units from 8 to 512. ReLU has been chosen for hidden layers and softmax for the output layer. The network is trained for 500 epochs using Adam optimizer with a learning rate of 0.001.

Since the number of annotations for each class was not equal, we computed precision, recall, and F1 measures for both the micro and the macro averages. For the micro averages, we calculated the metrics globally, thus giving more weight to the instrument with a higher number of appearances. On the other hand, we calculated the metrics for each label and found their unweighted average for the macro averages. Overall accuracy is also used as a metric for performance evaluation. ‘

5. RESULTS AND ANALYSIS

The overall performance of different phases of the experiment is tabulated in Table 3. Fusion with Attention (Fusion-Attn) network achieved micro and macro F1 mea-

[1] https://github.com/Veleslavia/EUSIPCO2017
[2] https://librosa.org/doc/latest/tutorial.html

SL.No	Class	MTF-DNN			Han's Model			Mel-spectrogram-Attn			Modgdgram-Attn			Fusion-Attn		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	Cello	0.54	0.59	0.56	0.55	0.44	0.49	0.48	0.70	0.57	0.09	0.15	0.12	0.79	0.67	**0.73**
2	Clarinet	0.15	0.40	0.22	0.23	0.64	0.33	0.74	0.68	0.71	0.10	0.28	0.15	0.54	0.80	**0.65**
3	Flute	0.17	0.21	0.19	0.54	0.54	0.54	0.66	0.60	0.63	0.21	0.16	0.18	0.68	0.72	**0.70**
4	Acoustic guitar	0.59	0.39	0.47	0.62	0.51	0.56	0.61	0.44	0.51	0.61	0.38	0.47	0.66	0.57	**0.61**
5	Electric guitar	0.56	0.46	0.51	0.57	0.51	0.54	0.53	0.64	0.58	0.49	0.47	0.48	0.66	0.68	**0.67**
6	Organ	0.22	0.45	0.29	0.20	0.42	0.27	0.26	0.69	0.38	0.12	0.18	0.14	0.30	0.62	**0.40**
7	Piano	0.70	0.36	0.47	0.72	0.58	0.64	0.68	0.66	0.67	0.64	0.55	0.59	0.73	0.71	**0.72**
8	Saxophone	0.03	0.40	0.06	0.13	0.60	0.21	0.10	0.70	0.18	0.03	0.20	0.05	0.18	0.70	**0.29**
9	Trumpet	0.14	0.57	0.23	0.30	0.86	0.44	0.86	0.43	0.57	0.19	0.36	0.24	0.59	0.71	**0.65**
10	Violin	0.24	0.53	0.33	0.43	0.58	0.49	0.85	0.31	0.45	0.38	0.45	0.42	0.56	0.51	**0.53**
11	Voice	0.55	0.34	0.43	0.69	0.55	0.61	0.74	0.47	0.57	0.68	0.54	0.60	0.78	0.61	**0.68**
	Macro	0.35	0.43	**0.34**	0.45	0.57	**0.47**	0.59	0.57	**0.53**	0.32	0.34	**0.31**	0.59	0.66	**0.60**
	Micro	0.39	0.39	**0.39**	0.54	0.54	**0.54**	0.56	0.56	**0.56**	0.43	0.43	**0.43**	0.65	0.65	**0.65**

Table 3. Precision (P), recall (R), and F1 score for the experiments with data augmentation.

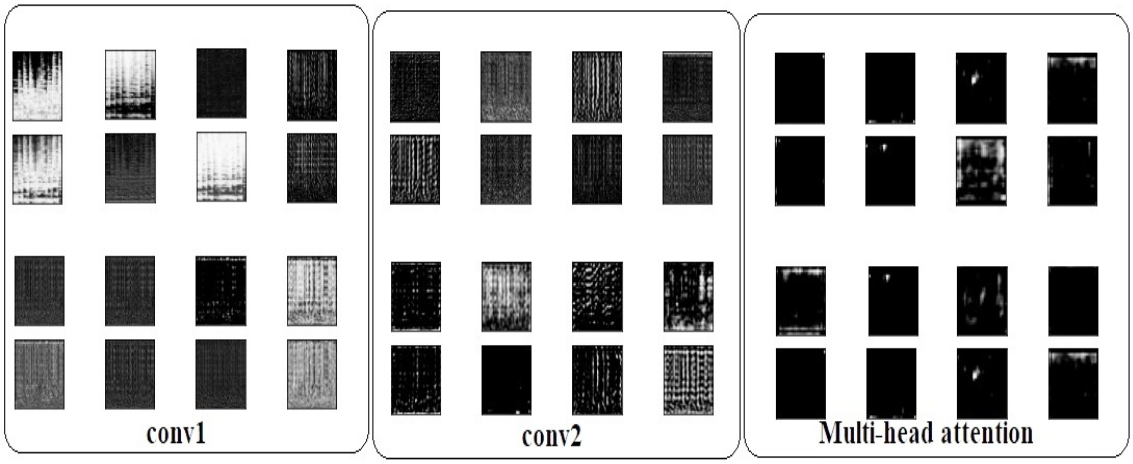

Figure 3. Visualisation of feature maps of convolutional layers and attention. The upper pane represents the feature maps for Mel-spectrogram inputs and the lower pane represents the feature maps for modgdgram inputs.

sures of 0.65 and 0.60, respectively. The State-of-the-art Han model reports micro F1 and macro F1 scores of 0.54 and 0.47, respectively. Micro F1 and macro F1 are 20.37% and 27.66% higher than those obtained for the baseline model. Modgdgram added complementary information to the spectrogram approach. Conventionally, the spectrum-related features used in instrument recognition take into account merely the magnitude information. However, there is often additional information concealed in the phase, which could be beneficial for recognition [12]. Han's model and the proposed Mel-spectrogram approach show similar performance with better performance for the proposed architectural choice. It is worth noting that modgdgram itself outperforms the MTF-DNN methodology. It reveals the importance of phase information in musical processing tasks.

5.1 Effect of attention

Attention mechanisms have become an integral part of compelling sequence modeling and transduction models since these models show superior quality while being more parallelizable and requiring significantly less time to train. [23]. It is applied to a variety of speech and music processing applications like speech emotion recognition [29], music instrument recognition [7], music generation [30] etc. For polytimbral music instrument recognition attention model focus on specific time segments in the audio relevant to each instrument label. The ability of the attention model to weigh relevant and suppress irrelevant predictions for each instrument leads to better classification accuracy [7]. Compared to self-attention the multi-head attention gives the attention layer multiple representation subspaces, and as the image passes through different heads

Sl.No	Model	Micro			Macro		
		P	R	F1	P	R	F1
1	Fusion-without Attn ($Train_d + Train_g$)	0.54	0.54	0.54	0.46	0.58	0.48
2	Fusion-Attn ($Train_d$)	0.57	0.57	0.57	0.53	0.63	0.54
3	Fusion-Attn ($Train_d + Train_g$).	0.65	0.65	0.65	0.59	0.66	0.60

Table 4. Performance comparison of the models with and without data augmentation.

Sl.No	Model	Micro			Macro		
		P	R	F1	P	R	F1
1	Bosch et al. [14]	0.50	0.50	0.50	0.41	0.45	0.43
2	Han et al./1446k [1]	**0.65**	0.56	0.60	**0.54**	0.51	0.50
3	Single-layer/62k [5]	0.61	0.52	0.56	0.52	0.48	0.48
4	Multi-layer/743k [5]	0.65	0.54	0.59	0.55	0.52	0.52
5	**Fusion-Attn** ($Train_d + Train_g$)**/473k**	0.63	**0.63**	**0.63**	0.51	**0.55**	**0.52**

Table 5. Performance comparison for IRMAS dataset.

predictions about the predominant instruments are more refined than employing single head self-attention. Another important point is that it requires very few trainable parameters to learn the model, which helps to reach convergence faster than the models employing CNN alone. The significance of attention in the proposed model can be analyzed from Table 4. Fusion without Attention reports micro and macro F1 scores of 0.54 and 0.48 respectively. Fusion with Attention reports micro and macro F1 scores of 0.65 and 0.60, respectively, with an improvement of 20.37% and 25% higher than that obtained by Fusion without Attention.

Visualization of the feature maps extracted from the first two convolutional layers and attention layer is shown in Figure 3. It is created with 8 feature maps as subplots. The feature maps close to the input detect small or fine-grained detail, whereas attention layer feature maps capture more general and refined features for predominant instrument recognition.

5.2 Effect of data augmentation

For deep learning, the number of training examples is critical for the performance compared to the case of using hand-crafted features because it aims to learn a feature from the low-level input data [1]. The significance of data augmentation in the proposed model can be analyzed from Table 4. Fusion-Attn without data augmentation ($Train_d$) reports micro and macro F1 score of 0.57 and 0.54 respectively. Fusion-Attn ($Train_d + Train_g$) reports micro and macro F1 score of 0.65 and 0.60, respectively, with improvement of 14.03% and 11.11% higher than that obtained by Fusion ($Train_d$).

5.3 Multiple predominant instrument recognition

The IRMAS dataset contains testing files of variable length and has multiple predominant instruments. For our initial work, we considered only the variable-length polyphonic testing files with a single predominant instrument. The same experiment is repeated for multiple predominant instrument recognition using the entire 2874 testing files available in the corpus. For that, we trained our networks using fixed-length excerpts containing a single predominant instrument and estimated an arbitrary number of instruments from variable-length audio files having multiple predominant instruments.

The standard metrics for various algorithms on the IRMAS corpus are reported in Table 5. The number of trainable parameters is also indicated. Bosch *et al.* [14] algorithm used typical hand-made timbral audio features with their frame-wise mean and variance statistics to train SVMs with source separation technique called flexible audio source separation framework (FASST) in a preprocessing step. The state-of-the-art Han model [1] reports micro and macro F1 score of 0.60 and 0.50 respectively. Han *et al.* [1] developed a deep CNN for instrument recognition based on Mel-spectrogram inputs. Pons *et al.* [5] customized the architecture of Han *et al.* and introduced two models, namely, single-layer and multi-layer approaches. They used the same aggregation strategy as that of Han's model by averaging the softmax predictions and finding the candidates with a threshold of 0.2. As different from the existing approaches, we estimated the predominant instrument using the entire Mel-spectrogram without sliding or aggregation analysis. As our Mel-spectrogram/modgdgram inputs pass through multiple heads the presence of predominant instruments is refined from the simultaneously occurring partials. Our Fusion approach reports a micro and macro F1 score of 0.63 and 0.52, which is a 5 % and 4 % increase from Han's model. Also, our proposed method shows better micro and macro recall than the existing techniques. Our proposed method reports a micro and macro recall of 0.63 and 0.55, which is an 12.5 % and 7.84 % increase from Han's model. It reveals the importance of the attention

mechanism in predicting multiple instruments. A significant improvement is also observed over the method proposed in [14]. Also, the fusion network reports better results with very less trainable parameters, compared to existing techniques. In [7], the usage of an attention layer was shown to improve classification results in the OpenMIC dataset, when applied to a set of Mel-spectrogram features extracted from a pre-trained VGG net. While the work [7] focusses on Mel-spectrogram, we experimented with the effect of phase information along with magnitude information. The experimental results in the paper show the potential of Mel-spectrogram and modgdgram on recognizing predominant instruments in a polyphonic environment with multi-head attention.

6. CONCLUSIONS

We presented an Attention-based predominant instrument recognition system using Mel-spectro/modgd-gram inputs. CNN with multi-head attention is used to capture the instrument-specific characteristics and then do further classification. The proposed method is evaluated using the IRMAS dataset. Data augmentation is also performed using WaveGAN. The fusion framework outperforms the latest model proposed by Han *et al.* The results show the potential of score-level fusion of magnitude and phase-based approaches and the attention mechanism empowers the network to focus on specific regions of Mel-spectrogram/modgdgram in predominant instrument recognition in polyphonic music.

Acknowledgments

The first author would like to acknowledge the CERD of APJ Abdul Kalam Technological University, Trivandrum, Kerala, India for providing a Ph.D. fellowship.

7. REFERENCES

[1] Y. Han, J. Kim, and K. Lee, "Deep convolutional neural networks for predominant instrument recognition in polyphonic music." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 208–221, 2017.

[2] F. Fuhrmann and P. Herrera, "Polyphonic instrument recognition for exploring semantic similarities in music," *in Proc. of 13th International Conference on Digital Audio Effects DAFx10, Graz, Austria*, vol. 14, no. 1, pp. 1–8, 2010.

[3] J.-Y. Liu and Y.-H. Yang, "Event localization in music auto-tagging," *in Proc. of the 24th ACM international conference on Multimedia*, pp. 1048–1057, 2016.

[4] Z. Duan, J. Han, and B. Pardo, "Multi-pitch streaming of harmonic sound mixtures," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 138–150, 2013.

[5] J. Pons, O. Slizovskaia, R. Gong, E. Gómez, and X. Serra, "Timbre analysis of music audio signals with convolutional neural networks," *in Proc. of 25th European Signal Processing Conference (EUSIPCO)*, pp. 2744–2748, 2017.

[6] S. Gururani, C. Summers, and A. Lerch, "Instrument activity detection in polyphonic music using deep neural networks," *in Proc. of International Society for Music Information Retrieval Conference (ISMIR)*, 2018.

[7] S. Gururani, M. Sharma, and A. Lerch, "An attention mechanism for musical instrument recognition," *in Proc. of 20th International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[8] D. Yu, H. Duan, J. Fang, and B. Zeng, "Predominant instrument recognition based on deep neural network with auxiliary classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 852–861, 2020.

[9] J. S. Gómez, J. Abeßer, and E. Cano, "Jazz solo instrument classification with convolutional neural networks, source separation, and transfer learning." *in Proc. of International Society for Music Information Retrieval Conference (ISMIR)*, pp. 577–584, 2018.

[10] X. Li, K. Wang, J. Soraghan, and J. Ren, "Fusion of hilbert-huang transform and deep convolutional neural network for predominant musical instruments recognition," *in Proc. of 9th International conference on Artificial Intelligence in Music, Sound, Art and Design*, 2020.

[11] A. Kratimenos, K. Avramidis, C. Garoufis, A. Zlatintsi, and P. Maragos, "Augmentation methods on monophonic audio for instrument classification in polyphonic music," *in Proc. of 28th European Signal Processing Conference (EUSIPCO)*, pp. 156–160, 2021.

[12] A. Diment, P. Rajan, T. Heittola, and T. Virtanen, "Modified group delay feature for musical instrument recognition," *in Proc. of 10th International Symposium on Computer Music Multidisciplinary Reserach, Marseille, France*, pp. 431–438, May 2013.

[13] R. Ajayakumar and R. Rajan, "Predominant instrument recognition in polyphonic music using gmm-dnn framework," *in Proc. of International Conference on Signal Processing and Communications (SPCOM)*, pp. 1–5, 2020.

[14] J. J. Bosch, J. Janer, F. Fuhrmann, and P. Herrera, "Acomparison of sound segregation techniques for predominant instrument recognition in musical audio signals." *in Proc. of 13th International Society for Music Information Retrieval Conference (ISMIR)*, 2012.

[15] F. Fuhrmann., "Automatic musical instrument recognition from polyphonic music audio signals," *PhD thesis, Universitat Pompeu Fabra*, 2012.

[16] H. A. Murthy and B. Yegnanarayana, "Group delay functions and its application to speech processing," *Sadhana*, vol. 36, no. 5, pp. 745–782, 2011.

[17] M. Sukhavasi and S. Adappa, "Music theme recognition using cnn and self-attention," *arXiv:1911.07041*, 2019.

[18] D. Ghosal and M. H. Kolekar, "Music genre recognition using deep neural networks and transfer learning." *in Proc. of Interspeech*, pp. 2087–2091, 2018.

[19] D. O'shaughnessy, "Speech communication: human and machine," *Universities press*, pp. 1–5, 1987.

[20] R. Rajan and H. A. Murthy, "Two-pitch tracking in co-channel speech using modified group delay functions," *Speech Communication*, vol. 89, pp. 37–46, 2017.

[21] ——, "Music genre classification by fusion of modified group delay and melodic features," *in Proc. of National Conference on Communications*, 2017.

[22] ——, "Melodic pitch extraction from music signals using modified group delay functions," *in Proc. National Conference on of the Communications (NCC)*, pp. 1–5, February 2013.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *in Proc. of Neural Information Processing Systems (NIPS)*, 2017.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Aaron Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[25] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," *in Proc.of International Conference on Machine Learning*, pp. 1857–1865, 2017.

[26] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," *in Proc. of International Conference on Learning Representations (ICLR)*, pp. 1–16, 2019.

[27] A. Madhu and S. Kumaraswamy, "Data augmentation using generative adversarial network for environmental sound classification," *in Proc, of 27th European Signal Processing Conference (EUSIPCO)*, pp. 1–5, 2019.

[28] G. Atkar and P. Jayaraju, "Speech synthesis using generative adversarial network for improving readability of hindi words to recuperate from dyslexia," *Neural Computing and Applications*, pp. 1–10, 2021.

[29] Y. Yu and Y.-J. Kim, "Attention-lstm-attention model for speech emotion recognition and analysis of iemocap database," *Electronics*, vol. 9, no. 5, p. 713, 2020.

[30] G. Keerti, A. Vaishnavi, P. Mukherjee, A. S. Vidya, G. S. Sreenithya, and D. Nayab, "Attentional networks for music generation," *arXiv preprint arXiv:2002.03854*, 2020.