#5039

Deep Convolutional and Recurrent Networks for Polyphonic Instrument Classification from Monophonic Raw Audio Waveforms

Kleanthis Avramidis*, Agelos Kratimenos*, Christos Garoufis, Athanasia Zlatintsi, Petros Maragos
School of ECE, National Technical University of Athens / Robot Perception and Interaction Unit, Athena Research Center

1. Outline

- Audio classification tasks traditionally discard direct waveform modeling for expensive time-frequency feature representations.
- We propose a lightweight end-to-end classifier for Instrument Classification by parameterizing RNN and CNN networks to model raw audio waveforms with comparable performance.

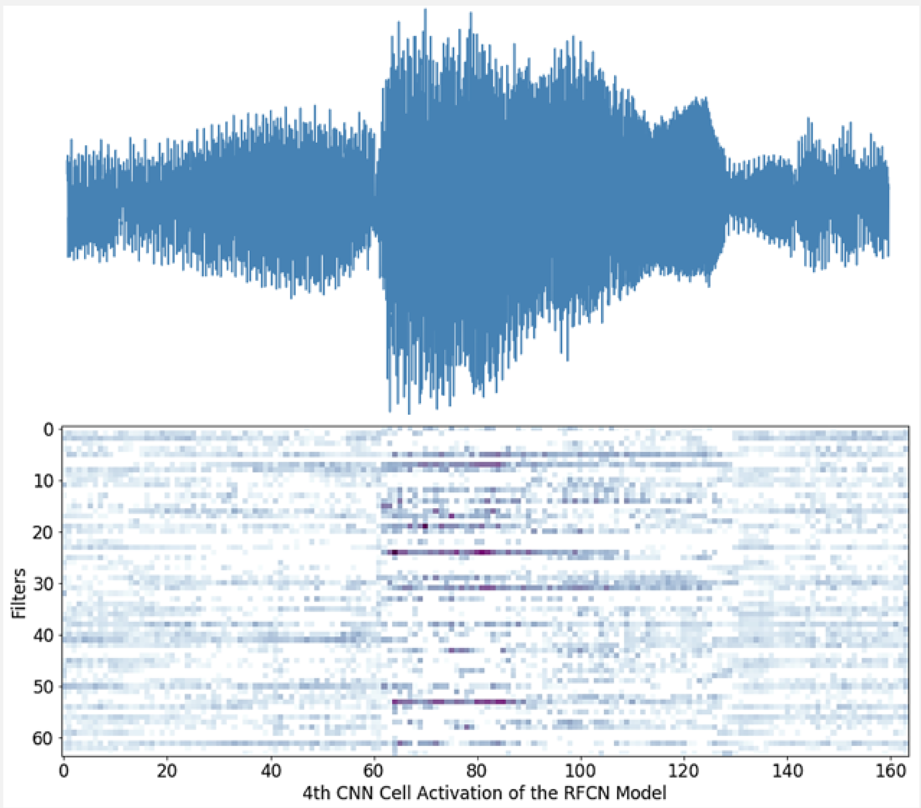

Figure: **Intermediate activation of the RFCN for piano**

2. Experimental Setup

- **IRMAS** [2] is used to train and test our models. Separate splits with 11 annotated instruments.
- 5-fold cross-validation, batch size 64
- BCE Loss for multi-label classification, Adam
- Metrics: LRAP ranking and F1 Score

3a. BiGRU Architectures

Number of Layers	Number of Units
1	128 or 256
2	128, 64
Dropout (0.5)	
Output Dense	

3b. CNN & Combined Architectures

- Architecture based on [1] that yielded strong results on the IRMAS Dataset. CNN cell: 2 stacked identical 1D convolutional layers, Batch Normalization, Leaky ReLU activation and a max pooling layer.
- This module is followed by 2 fully connected layers (DCNN) ⟶ **increases substantially the number of its trainable parameters** ⟶ we experiment by **removing dense layers** (FCN).
- Residual FCN: embed **skip connections** to the previous model, to propagate low-level features.

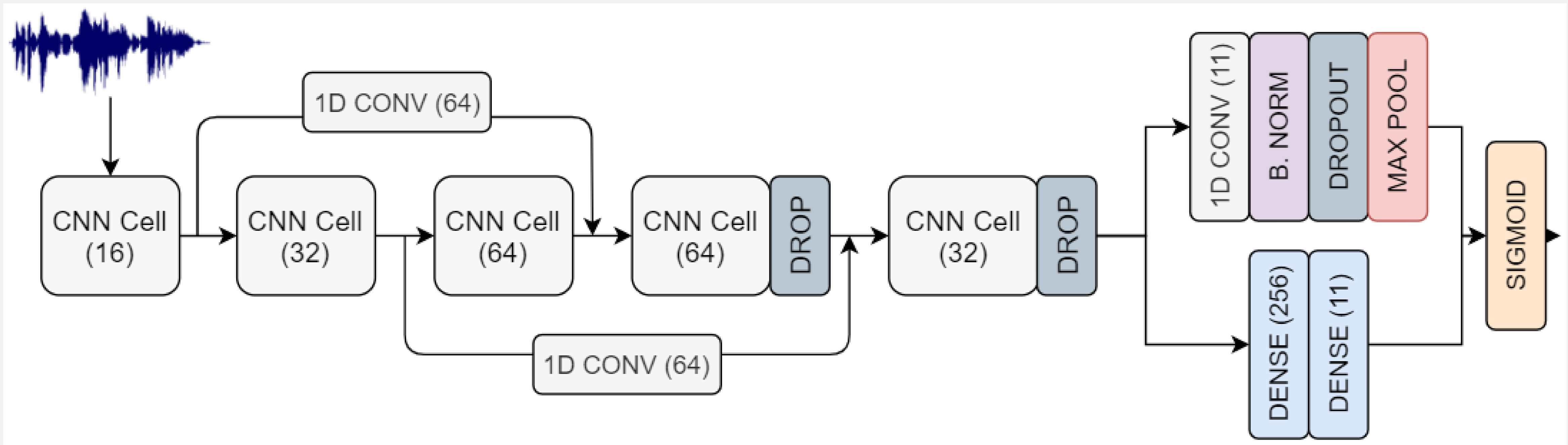

Figure: **The DCNN, FCN and RFCN architectures used in the experimental evaluation**

- CNNs concentrate on spatial features and, in the context of waveforms, on temporally **local correlations**, while recurrent ones are useful in modeling longer-term **temporal structure**.
- Combined RCNN: We attach the best performing BiGRU model into our RFCN.

4b. Instrument-wise Analysis

- We examine the class-wise performance in terms of the F1 metric. The results are visualized along with the corresponding results obtained from CQT spectrogram modeling from our previous work [1].
- **Brass** instruments (clarinet, flute, saxophone) are recognized much better using raw waveforms.
- **Predominant** instruments, i.e. guitars, piano or voice, are distinguished better through CQT models.

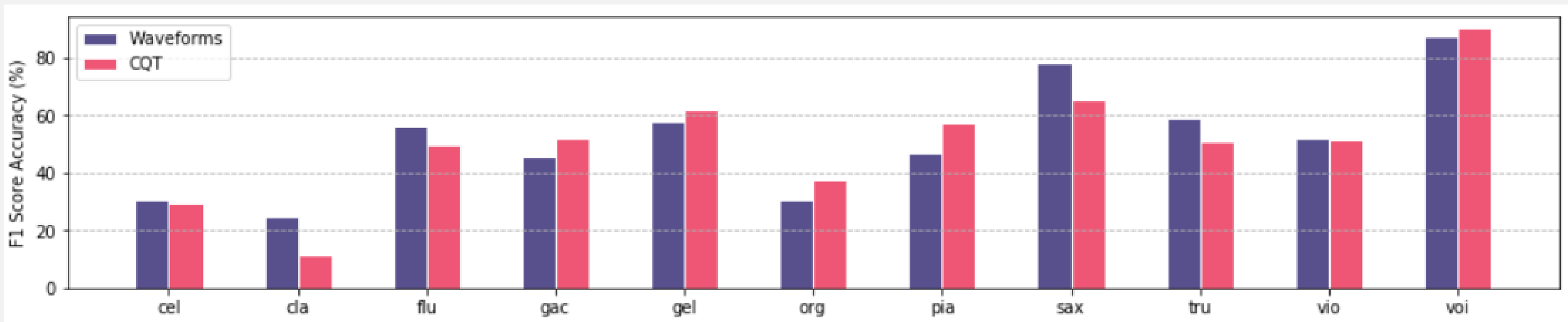

Figure: **Instrument-wise performance of the proposed model and the monophonic [1] in terms of F1-score**

4a. Results

- A simple RNN cannot sufficiently decode the information, while 1D CNNs are performing almost as well as 2D CNNs on spectrograms.
- Removing the dense layers reduces the number of trainable parameters and increases accuracy substantially (spatial correlations).

Models	F1-micro %	F1-macro %	LRAP %
GRU_2	49.28 ±2.45	43.18 ±3.11	67.07 ±1.81
DCNN	55.32 ±0.55	48.30 ±0.31	73.48 ±0.38
FCN	58.45 ±0.36	49.96 ±0.29	75.13 ±0.32
RFCN	58.55 ±0.22	50.22 ±0.35	**75.14 ± 0.23**
$CRNN_3$	**60.77 ± 0.26**	**54.31 ± 0.35**	74.74 ±0.39

- Only certain residual connections and RNN placements work well in enhancing scores.
- Comparable results to literature with reduced number of model trainable parameters.

Models	F1-micro	F1-macro	LRAP
Bosch et al. [2]	0.503	0.432	–
Pons et al. [3]	0.589	0.516	–
Han et al. [4]	0.602	0.503	–
Kratimenos et al. [1]	**0.616**	0.506	**0.767**
[1] Reduced	0.520	0.458	0.689
Proposed	0.608	**0.543**	0.747

5. References

[1] A. Kratimenos et al. Augmentation Methods on Monophonic Audio for Instrument Classification in Polyphonic Music. In *Proc. EUSIPCO*, 2020.

[2] Bosch et al. A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals. In *Proc. ISMIR*, 2012.

[3] J. Pons et al. Timbre analysis of music audio signals with convolutional neural networks. *Proc. EUSIPCO*, 2017.

[4] Y. Han et al. Deep convolutional neural networks for predominant instrument recognition in polyphonic music. *IEEE/ACM Tr. on Audio, Speech, and Language Processing*, 2017.