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.

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

<table>
<thead>
<tr>
<th>Number of Layers</th>
<th>Number of Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>128 or 256</td>
</tr>
<tr>
<td>2</td>
<td>128, 64</td>
</tr>
</tbody>
</table>

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.

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.

3. 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).

<table>
<thead>
<tr>
<th>Models</th>
<th>F1-micro %</th>
<th>F1-macro %</th>
<th>LRAP %</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRU_2</td>
<td>49.28 ± 2.45</td>
<td>43.18 ± 3.11</td>
<td>67.07 ± 1.81</td>
</tr>
<tr>
<td>DCNN</td>
<td>55.32 ± 0.55</td>
<td>48.30 ± 0.31</td>
<td>73.48 ± 0.38</td>
</tr>
<tr>
<td>FCN</td>
<td>58.45 ± 0.36</td>
<td>49.96 ± 0.29</td>
<td>75.13 ± 0.32</td>
</tr>
<tr>
<td>RFCN</td>
<td>58.55 ± 0.22</td>
<td>50.22 ± 0.35</td>
<td>75.14 ± 0.23</td>
</tr>
<tr>
<td>CRNN_2</td>
<td>60.77 ± 0.26</td>
<td>54.31 ± 0.35</td>
<td>74.74 ± 0.39</td>
</tr>
</tbody>
</table>

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

5. References