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

Kleanthis Avramidis*, Agelos Kratimenos*, Christos Garoufis, Athanasia Zlatintsi and Petros Maragos

National Technical University of Athens, School of ECE
Computer Vision, Speech Communication and Signal Processing Group

kle.avramidis@gmail.com; ageloskrat@yahoo.gr; cgaroufis@mail.ntua.gr; [nzlat, maragos]@cs.ntua.gr
List of Contents

1. Introduction
2. Architectures
3. Experimental Setup
4. Results & Discussion
5. Contributions
List of Contents

1. Introduction
2. Architectures
3. Experimental Setup
4. Results & Discussion
5. Contributions
Waveforms & End-to-End Models

**Waveform**: Abstract representation of a sound wave

- Complex, non-intuitive structure
- Inherits noise from surroundings / equipment / sound event

Instead: **Time-Frequency Representations** (i.e. CQT, STFT)

**But**: Which should we use? What is their computational cost?
In MIR and Instrument Classification particularly, there is strong intuition into utilizing frequency-related representations, since notes and instruments are densely associated with specific frequency events.

**Remark**: Challenging and computationally expensive to design specialized feature representations for each different recognition task.

**Proposal**: Take advantage of Deep Learning methods to build efficient feature extractors from raw waveforms. Should handle:

- High input dimensionality and noisy structure
- Low-level temporal correlations and features
- Reduced computational cost without performance loss
List of Contents

1 Introduction
2 Architectures
3 Experimental Setup
4 Results & Discussion
5 Contributions
Recurrent Networks (RNN)

Have been widely used in waveform and generally sequence modeling thanks to their ability to handle long-range temporal dependencies.

Bidirectional GRU:

- Lower computational cost compared to LSTM
- Comparable performance to LSTMs for audio sequences
- Considers both past and future features for dependencies

We experiment on the number of layers and utilized GRU units:

<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>

Dropout (0.5)
Output Dense
Convolutional Networks (CNN)

- Traditionally operate on images or time-frequency features.
- Already exhibited results in audio waveform processing [1].

Network based on [2] with alterations:

- DCNN: 2 dense layers to predict - many trainable parameters
- FCN: Dense layers → unit-kernel convolutions and filter pooling
- RFCN: embed skip connections to the previous model

CNNs concentrate on temporally **local correlations** in waveforms, while RNNs are useful in modeling **longer-term** temporal structure.

We expect that by efficiently combining these networks we will combine **different kinds** of discriminative features.

We attach the best-performing RNN model of our experiments to the RFCN model in various positions.

**Connection:** The embedded model takes the output of the corresponding CNN cell and its output is reduced to classes through convolution and Global Average Pooling. The final representation is the **average** of the 2 modules’ outputs.

Empirically search the optimal way of integrating the recurrent model information into a robust classifier.
List of Contents

1. Introduction
2. Architectures
3. Experimental Setup
4. Results & Discussion
5. Contributions
The **IRMAS Dataset** [3]: 11 instruments/classes

- [ cello, clarinet, flute, acoustic/electric guitar, organ, piano, saxophone, trumpet, violin, voice ]

- **Training Set**: A set of 3-sec monophonic audio chunks (music tracks with a predominant instrument) for each class

- **Testing Set**: A set of multilabeled polyphonic tracks

Each training track was:

- cut to 1-sec segments
- downsamplesd and downmixed
- normalised by RMS energy

Training Protocol & Evaluation

- 5-fold Cross-Validation
- Binary Cross-Entropy Loss (Multi-label Task)
- Adam Optimizer ($10^{-3}$ learning rate)
- Learning Rate Reduction & Early Stopping

Utilized evaluation metrics:
- Label Ranking Average Prediction (LRAP): Suitable for multi-label tasks, ranking intuition, threshold independent
- $F_1$ Score: Comparable evaluation, class imbalance

IRMAS Testing Set: Tracks ranging from 5-20 sec. We average the per-sec predictions to obtain a single prediction for each track. Labeled instruments are active throughout the track.
Training Protocol & Evaluation

- 5-fold Cross-Validation
- Binary Cross-Entropy Loss (Multi-label Task)
- Adam Optimizer ($10^{-3}$ learning rate)
- Learning Rate Reduction & Early Stopping

Utilized evaluation metrics:

- **Label Ranking Average Prediction (LRAP)**: Suitable for multi-label tasks, ranking intuition, threshold independent
- **$F_1$ Score**: Comparable evaluation, class imbalance

**IRMAS Testing Set**: Tracks ranging from 5-20 sec. We average the per-sec predictions to obtain a single prediction for each track. Labeled instruments are active throughout the track.
List of Contents

1. Introduction
2. Architectures
3. Experimental Setup
4. Results & Discussion
5. Contributions
A simple recurrent network cannot sufficiently decode the information included in a waveform.

<table>
<thead>
<tr>
<th>BiGRU</th>
<th>F1-micro %</th>
<th>F1-macro %</th>
<th>LRAP %</th>
<th>#Params</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (128)</td>
<td>43.76 ± 1.95</td>
<td>37.37 ± 1.90</td>
<td>57.26 ± 3.28</td>
<td>103.4K</td>
</tr>
<tr>
<td>1 (256)</td>
<td>43.51 ± 2.46</td>
<td>39.19 ± 2.23</td>
<td>58.47 ± 2.73</td>
<td>403.4K</td>
</tr>
<tr>
<td>2</td>
<td>49.28 ± 2.45</td>
<td>43.18 ± 3.11</td>
<td>67.07 ± 1.81</td>
<td>225.6K</td>
</tr>
</tbody>
</table>

1D CNNs are capable of extracting the most discriminative features from raw waveforms, almost as well as 2D models on spectrograms.

FCN: in the absence of a dense layer, the network generalizes better upon the information from spatial processing + less parameters.
A simple recurrent network cannot sufficiently decode the information included in a waveform.

### BiGRU

<table>
<thead>
<tr>
<th></th>
<th>F1-micro %</th>
<th>F1-macro %</th>
<th>LRAP %</th>
<th>#Params</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (128)</td>
<td>43.76 ± 1.95</td>
<td>37.37 ± 1.90</td>
<td>57.26 ± 3.28</td>
<td>103.4K</td>
</tr>
<tr>
<td>1 (256)</td>
<td>43.51 ± 2.46</td>
<td>39.19 ± 2.23</td>
<td>58.47 ± 2.73</td>
<td>403.4K</td>
</tr>
<tr>
<td>2</td>
<td><strong>49.28 ± 2.45</strong></td>
<td><strong>43.18 ± 3.11</strong></td>
<td><strong>67.07 ± 1.81</strong></td>
<td>225.6K</td>
</tr>
</tbody>
</table>

1D CNNs are capable of extracting the most discriminative features from raw waveforms, almost as well as 2D models on spectrograms.

FCN: in the absence of a dense layer, the network generalizes better upon the information from spatial processing + less parameters.

### Models

<table>
<thead>
<tr>
<th>Models</th>
<th>F1-micro %</th>
<th>F1-macro %</th>
<th>LRAP %</th>
<th>#Params</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCNN</td>
<td>55.32 ± 0.55</td>
<td>48.30 ± 0.31</td>
<td>73.48 ± 0.38</td>
<td>1.14M</td>
</tr>
<tr>
<td>FCN</td>
<td>58.45 ± 0.36</td>
<td>49.96 ± 0.29</td>
<td>75.13 ± 0.32</td>
<td>81.8K</td>
</tr>
<tr>
<td>RFCN</td>
<td><strong>58.55 ± 0.22</strong></td>
<td><strong>50.22 ± 0.35</strong></td>
<td><strong>75.14 ± 0.23</strong></td>
<td>85K</td>
</tr>
</tbody>
</table>
Simply averaging the RNN and CNN model outputs lowers accuracy → inadequate standalone performance of the BiGRU

We thus inserted the BiGRU in various locations in the RFCN model:

<table>
<thead>
<tr>
<th>Models</th>
<th>F1-micro %</th>
<th>F1-macro %</th>
<th>LRAP %</th>
<th>#Params</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRNN₂</td>
<td>59.80 ± 0.66</td>
<td>53.20 ± 0.52</td>
<td>74.16 ± 0.66</td>
<td>1.03M</td>
</tr>
<tr>
<td>CRNN₃</td>
<td><strong>60.77 ± 0.26</strong></td>
<td><strong>54.31 ± 0.35</strong></td>
<td><strong>74.74 ± 0.39</strong></td>
<td>1.07M</td>
</tr>
<tr>
<td>CRNN₄</td>
<td>60.07 ± 0.67</td>
<td>53.73 ± 0.59</td>
<td>74.11 ± 0.50</td>
<td>1.08M</td>
</tr>
<tr>
<td>CRNN₅</td>
<td>59.21 ± 0.56</td>
<td>52.18 ± 0.46</td>
<td>74.32 ± 0.65</td>
<td>1.03M</td>
</tr>
</tbody>
</table>

Table: The subscript denotes the CNN layer in which the RNN was connected.

No observed improvement in performance for the LRAP metric, steady increase however for F1 scores

The combined models consist of significantly more parameters
### Literature Comparison

<table>
<thead>
<tr>
<th>Models</th>
<th>F1-micro</th>
<th>F1-macro</th>
<th>LRAP</th>
<th>#Params</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bosch et al. [3]</td>
<td>0.503</td>
<td>0.432</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Pons et al. [5]</td>
<td>0.589</td>
<td>0.516</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Han et al. [4]</td>
<td>0.602</td>
<td>0.503</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Kratimenos et al. [2]</td>
<td><strong>0.616</strong></td>
<td>0.506</td>
<td><strong>0.767</strong></td>
<td>24.3M</td>
</tr>
<tr>
<td>Reduced [2]</td>
<td>0.520</td>
<td>0.458</td>
<td>0.689</td>
<td>1.20M</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.608</td>
<td><strong>0.543</strong></td>
<td>0.747</td>
<td><strong>1.07M</strong></td>
</tr>
</tbody>
</table>

**Table:** Comparison of our work with previous studies on the IRMAS Dataset

- F1 micro surpasses most studies on the task, while we observe dominant performance at the more competitive F1 macro score.
- Results obtained with a significantly reduced number of trainable parameters, low training - testing time and minimal pre-processing.

Instrument-wise Analysis

- We use the per-class $F_1$ score for this experiment
- We examine how each instrument can be discriminative in either waveform or time-frequency representation.

- Brass instruments (ex. clarinet, flute, saxophone) → Waveforms
- Predominant and leading instruments (ex. guitars, piano, voice) → Constant Q Transform Spectrograms
Contributions & Future Work

- Experiments with various architectures that are favourable towards waveform modeling, like Fully Convolutional and Residual Nets and information fusion.

- A residual FCN-BiGRU model (1M parameters) outperforms the state-of-the-art with CQT spectrograms (24M parameters).

- Brass instruments are being identified easier through waveforms, while leading instruments benefit more from time-frequency features.

- Future work: alternate methods to exploit RNNs / enhance performance of predominant instruments / ways to deal with inherent noise.
Thank you