Computers Love Music

Can Computers Mimic Music?
Music Translation

• The goal: translating music across instruments, genres and styles
• The method: neural networks - multi-domain wavenet autoencoder
• The challenge: no data!
Technical Background

In order to understand the research, we’ll discuss some concepts and terms first:

- Neural networks
- Domain transfer
Neural Networks
Neural Networks - Types

- Convolutional (CNN): in our case - a classifier that receives an input and determines which class it belongs to
  - Can provide a clear-cut or a probable answer
Neural Networks - Types

• Auto-regressive (AR): creates the next frame in time, adds it to history, thus lengthening the history and building the “future” upon it.
Domain Transfer

- The challenge of translating input from one domain to another
- Can be unsupervised
The method

LET’S HEAR SOME MUSIC 🎶
Data

- 6 input musical domains: Mozart - symphonies, Bach - orchestra and choir, Bach - organ, Bach - harpsichord, Beethoven - piano
- Data separated to train and test sets
- Each musical piece split to 1-second segments
Encoding

- NNs work on numbers, not music
- Need to encode the music to numbers
- Can’t do notes - too specific, too complicated, existing results for simpler tasks are not good enough
- One encoder to rule them all
Encoding

• Based on WaveNet

• Input music is encoded to latent space

• In order to prevent the encoder from memorizing music - noise was added to the data

  • In each 1-sec file, the pitch of a randomly chosen segment length of between 0.25-0.5 seconds gets modulated by a -0.5 to 0.5 half-tone
Data Augmentation

- The goal: prevent the system from encoding data that is domain-specific
- The means: confusion network - another network, used only during training, which is responsible for minimizing the classification loss

**Diagram:**
- **Common Encoder**
- **Domain Confusion Network**

**Text:**
- *Generalize* an instrument in latent space
- *Differentiate* an instrument in latent space

**Icons:**
- (closer)
- (far apart)
Training

\[ \sum_j \sum_{s^j} E[ L(D^j(E(O(s^j, r))), s^j)] \]

\[ - \lambda \sum_j \sum_{s^j} E[ L(C(E(O(s^j, r))), j)] \]
Loss Function, Explained

In red - the decoder is given an encoded sample, outputs a “cover” in the same style.

In blue - the domain confusion network is given an encoded sample, and outputs which domain it belonged to.
Evaluating the New Music

• How do you give a score to a cover version?

• Compare the network’s results to the same task performed by human musicians
  • The task - convert 60 segments of 1 second each, to piano

• Comparison done by both human listeners and automatic score
Results

• The human scoring was done using CrowdMOS (mean opinion score), an open source tool for Mechanical Turk that helps detect and discard inaccurate scores

• The users were asked 2 questions: on a scale of 1 to 5 -
  • what's the quality of the audio?
  • How well does the converted version match the original?

<table>
<thead>
<tr>
<th>Converter</th>
<th>Harpsichord→ Piano</th>
<th>Orchestra→ Piano</th>
<th>New domains→ Piano</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Audio quality</td>
<td>Translation success</td>
<td>Audio quality</td>
</tr>
<tr>
<td>E</td>
<td>3.89 ± 1.06</td>
<td>4.10 ± 0.94</td>
<td>4.02 ± 0.81</td>
</tr>
<tr>
<td>M</td>
<td>3.82 ± 1.18</td>
<td>3.75 ± 1.17</td>
<td>4.13 ± 0.89</td>
</tr>
<tr>
<td>A</td>
<td>3.69 ± 1.08</td>
<td>3.91 ± 1.16</td>
<td>4.06 ± 0.86</td>
</tr>
<tr>
<td>Our</td>
<td>2.95 ± 1.18</td>
<td><strong>3.07 ± 1.30</strong></td>
<td>2.56 ± 1.04</td>
</tr>
</tbody>
</table>
Results

- The automatic scoring was done by pitch matching
- The system was more true-to-source than the pianists

Table 2: Automatic quality scores for the conversion task.

<table>
<thead>
<tr>
<th>Converter</th>
<th>Harpsichord→Piano</th>
<th>Orchestra→Piano</th>
<th>New domains→Piano</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NCC</td>
<td>DTW</td>
<td>NCC</td>
</tr>
<tr>
<td>E</td>
<td>0.82</td>
<td>0.98</td>
<td>0.78</td>
</tr>
<tr>
<td>M</td>
<td>0.69</td>
<td>0.96</td>
<td>0.65</td>
</tr>
<tr>
<td>A</td>
<td>0.76</td>
<td>0.97</td>
<td>0.73</td>
</tr>
<tr>
<td>Our</td>
<td><strong>0.84</strong></td>
<td><strong>0.98</strong></td>
<td><strong>0.82</strong></td>
</tr>
</tbody>
</table>
Significance of This Research

- Superior results compared to existing methods
- Breaking ground in the field of musical AI
- Democratization of music
- Changing what was considered possible