Theme Transformer
Symbolic Music Generation with Theme-Conditioned Transformer

*IEEE Transactions on Multimedia*

Yi-Jen Shih\textsuperscript{1,2}, Shih-Lun Wu\textsuperscript{1,2}, Frank Zalkow\textsuperscript{3}, Meinard Müller\textsuperscript{3}, Yi-Hsuan Yang\textsuperscript{2}

\textsuperscript{1}National Taiwan University, Taiwan
\textsuperscript{2}Academia Sinica, Taiwan
\textsuperscript{3}International Audio Laboratories Erlangen, Germany
About me

• Ian Shih
• B.S. in Electrical Engineering at National Taiwan University
• Part-time Research Assistant at Music and AI Lab
• Love playing some improvisation on piano (SoundCloud)
• Research Interest:
  • Music Generation (Prof. Yi-Hsuan Yang)
  • Visual Grounded Speech Models (Prof. Hung-Yi Lee)
• Website: atosystem.github.io
Outline

• Overview
• Technical Background
• Results
• Conclusion & Contribution
Outline

• Overview
• Technical Background
• Results
• Conclusion & Contribution
Excerpt from Perfect – Ed Sheeran

slow Your heart is all I own And in your eyes you hold mine Baby

I’m Dancing in the dark You between my arms Barefoot on the grass

Listening to our favourite song When you said you looked a mess I whispered

underneath my breathe You heard it darling you look perfect tonight

Am G F G C C C G

I found a woman Stronger than any one I know

Shes shares my dreams I hope that some day I’ll share her home I found a
Overview - Theme

• Themes
• Sequentia
• Motivic Development
• Music Expectancy
Theme is crucial in music composition

But how do recent model generate music?
Overview

Prompt Conditioned Music Generation

Music Transformer

Given a prompt (Input)

Generate the continuation (Output)
Prompt Conditioned Music Generation

The generated music has no knowledge of theme variation and repetitions
How to teach models to compose music base on a given Theme?
Overview

Theme Conditioned Music Generation

Theme (Input)

Dancing in the dark

Entire Song (Output)
Overview - Difficulties

• Definition of Musical Theme is quite *ambiguous*
• *Lack of Dataset* for Musical Theme Annotations
• Recent Music Generation Models have problems recognizing “Theme”, not to mention *variations* and *repetitions*
Overview

Theme Retrieval

Dancing in the dark

slow Your heart is all I own And in your eyes your holding mine Baby
I'm Dancing in the dark You between my arms Bare-foot on the grass
Listening to our favourite song When you said you looked a mess I whispered
underneath my breath You heard it darling you look perfect tonight
I found a woman Stronger than any one I know

She shares my dreams I hope that someday I'll share her home I found a
Overview

Theme Conditioned Music Generation

Theme (Input)

Entire Song (Output)

Dancing in the dark
Outline

• Overview

• Technical Background
  • Theme Retrieval
  • Theme based Music Generation

• Results

• Conclusion & Contribution
Theme Retrieval

• Previous Works
  • String-based
    • Correlative matrix (Hsu et al., 2001)
  • Geometric-based
    • COSIATEC (Meredith et al., 2010)
    • RECURSIA-RRT (Meredith, 2019)

• Requires hyperparameters tuning and prone to noise in data
Theme Retrieval

Encode Melody into Vector Space

Music Segment

Extract Theme

Density based clustering

Themes should be in the largest cluster
Theme Retrieval

• Adopt idea from SimCLR (Chen et al., 2020)

Data Augmentations

Pitch shift on scale

Last note duration variation

Note Splitting & Combination

Music Segment 1

Dancing in the dark

Music Segment 2 (from other song)

I found a woman
Theme Retrieval

• Adopt idea from SimCLR (Chen et al., 2020)

Data Augmentations

- Pitch shift on scale
- Last note duration variation
- Note Splitting & Combination

Music Segment 1

Music Segment 2 (from other song)

Transformer

Transformers

Pull Together

Pull Together
Theme Retrieval

• Contrastive loss
  \[ -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\alpha)}{\sum_k 1_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\alpha)} \]

• Apply DBSCAN to cluster music segments
  \[ D(S_i, S_j) = \|\text{Emb}(S_i) - \text{Emb}(S_j)\|_2 \]

• Regard the largest segment as “Theme”

• Results: (F1 retrieval with human annotators)

<table>
<thead>
<tr>
<th></th>
<th>CL (proposed)</th>
<th>CL w/o Note Duration Augmentation</th>
<th>CL w/o Pitch Shift Augmentation</th>
<th>CM</th>
<th>COSIATEC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average F1</td>
<td><strong>.378</strong></td>
<td>.220</td>
<td>.336</td>
<td>.345</td>
<td>.297</td>
</tr>
</tbody>
</table>
Outline

• Overview

• Technical Background
  • Theme Retrieval
  • Theme based Music Generation

• Results

• Conclusion & Contribution
Theme-based Music Generation

• Background – Representation **REMI** (Hung et al., 2018)

![Musical notation and time representation](image)

We added additional **Theme-Start, Theme-End** tokens to represent Theme Regions

Figure from Chou et al., 2021
Theme-based Music Generation

• Background – Autoregressive Model
  \[ p(x_t | x_{<t}) \]

• Recently works employ Transformer as main backbone
  • Music Transformer (Huang et al., 2018)
  • Pop Music Transformer (Hung and Yang, 2020)
  • Compound Words Transformer (Hsiao et al., 2021)

• Train by minimizing Negative log-likelihood
  • \[ - \sum_{t=1}^{T} \log p(x_t | x_{<t}) \]
Prompt-based Music Generation

Prompt Conditioned Music Generation

Given a prompt (Input)

Generate the continuation (Output)

Music Transformer

I'm Dancing in the dark

You between my arms

Barefoot on the grass

Listening to our favourite song

When you said you looked a mess

I whispered
Theme-based Music Generation

- Problem for prompt-based method

**Self-attention**

Extract information based on attention scores

\[ b^1 = \sum_{i} \alpha'_{1,i} v^i \]

The NLL loss can be minimized without considering the “themes”

Figure from Prof. Hung-yi Lee

\[ v^1 = W^v a^1 \quad v^2 = W^v a^2 \quad v^3 = W^v a^3 \quad v^4 = W^v a^4 \]
Theme-based Music Generation

**Theme (Input)**

Dancing in the dark

**Output (shifted left)**

**Entire Song (Output)**

- Encoder $L$ Blocks
- Decoder $L$ Blocks
- Cross Attention
- Positional Encoding
- Segment Embedding
- Token Embedding
- Theme Region Mask $m^e$
- Theme Tokens $x^e$
- Full Composition Tokens $x^{d:1:N-1}$
Theme-based Music Generation

- Propose Theme Transformer
- Gating Mechanism

\[ h^l_t = \begin{cases} m_t h_t^{l,(cross)} + (1 - m_t) h_t^{l,(self)}, & l > L/2 \\ m_t h_t^{l,(cross)} + h_t^{l,(self)}, & l \leq L/2 \end{cases} \]

- Theme Positional Encoding

\[ p^{\text{self}}_i = i, \quad p^{\text{cross}}_i = i - \max_{(m_k^d=0) \land (0 \leq k < i)} k \]

- 2 Memory Networks
  - Theme
  - Non-Theme
Outline

• Overview
• Technical Background
• Results
• Conclusion & Contribution
Results

• Evaluation Metrics
  • Pitch Class Consistency
    • Overlapping Area of chroma histograms of two bars
  • Melody Inconsistency
    • The min distance of all the segments compared to the first one
      \[ D(S_1, S_*) \quad D(S_i, S_j) = \|\text{Emb}(S_i) - \text{Emb}(S_j)\|_2 \]
  • Grooving consistency: coherence in rhythm
Results

• Proposed Evaluation Metrics
  • Theme Inconsistency
    • the inconsistency between theme regions
      \[
      \frac{2}{N(N-1)} \sum_{i,j}^{\infty} D_{ij}(\Gamma_i, \Gamma_j)
      \]
  • Theme Uncontrollability
    • the differences between theme regions and the given condition
      \[
      \frac{1}{N} \sum_{i=1}^{N} D(c_{1:\tau}, \Gamma_i)
      \]
  • Theme Gap
    • Gaps between Theme Regions
Results
Results

• Objective Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Pitch class consistency↑</th>
<th>Melody inconsistency↓</th>
<th>Grooving consistency↑</th>
<th>Theme inconsistency↓</th>
<th>Theme uncontrollability↓</th>
<th>Theme gap (in # bars)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (prompt-based) [13], [17]</td>
<td>.59±.07</td>
<td>.33±.38</td>
<td>.84±.09</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Seq2seq Transformer [13]</td>
<td>.61±.04</td>
<td>.46±.28</td>
<td>.90±.06</td>
<td>1.01±.05</td>
<td>1.10±.14</td>
<td>6.02±1.91</td>
</tr>
<tr>
<td>Theme Transformer (proposed)</td>
<td>.61±.06</td>
<td>.13±.24</td>
<td>.92±.07</td>
<td>0.27±.26</td>
<td>0.24±.20</td>
<td>9.48±3.59</td>
</tr>
<tr>
<td>Original pieces</td>
<td>.65±.05</td>
<td>.09±.18</td>
<td>.74±.10</td>
<td>0.05±.05</td>
<td>0.04±.04</td>
<td>12.24±11.32</td>
</tr>
</tbody>
</table>

• Subjective Evaluation (Total 50 participants)

<table>
<thead>
<tr>
<th>User group 1 (33 subjects)</th>
<th>C control</th>
<th>R speat</th>
<th>T iming</th>
<th>V ariation</th>
<th>S tructure</th>
<th>Q uality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (prompt-based) [13], [17]</td>
<td>3.01±1.08</td>
<td>2.55±1.18</td>
<td>2.73±1.06</td>
<td>2.65±1.06</td>
<td>3.06±0.94</td>
<td>3.19±0.98</td>
</tr>
<tr>
<td>Seq2seq [13]</td>
<td>2.52±1.10</td>
<td>2.12±1.08</td>
<td>2.27±1.08</td>
<td>2.41±1.18</td>
<td>3.10±0.99</td>
<td>3.23±0.92</td>
</tr>
<tr>
<td>Theme Transformer (proposed)</td>
<td>3.63±1.10</td>
<td>3.55±1.22</td>
<td>3.27±1.03</td>
<td>3.03±1.11</td>
<td>3.33±0.99</td>
<td>3.38±0.97</td>
</tr>
</tbody>
</table>

| User group 2 (17 subjects) |      |         |         |            |            |           |
|----------------------------|      |         |         |            |            |           |
| Baseline (prompt-based) [13], [17] | 2.90±1.09 | 2.39±0.97 | 2.76±1.26 | 3.22±1.24 | 2.78±1.09 | 2.78±1.00 |
| Theme Transformer (proposed) | 3.49±1.11 | 3.39±1.12 | 3.27±1.25 | 3.25±1.06 | 3.16±1.00 | 3.16±1.00 |
| Original pieces            | 3.61±1.17 | 3.37±1.14 | 3.53±1.11 | 3.29±1.11 | 3.39±0.97 | 3.41±1.11 |
## Results

- **Ablation Studies on Temperature and Sampling**

<table>
<thead>
<tr>
<th>$\epsilon$</th>
<th>$t$</th>
<th>Pitch class consistency↑</th>
<th>Melody inconsistency↓</th>
<th>Grooving consistency↑</th>
<th>Theme inconsistency↓</th>
<th>Theme uncontrollability↓</th>
<th>Theme gap (in # bars)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Theme Transformer</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.13</td>
<td>1.2</td>
<td>.61±.06</td>
<td>.13±.24</td>
<td>.92±.07</td>
<td>.07±0.26</td>
<td>0.24±0.20</td>
<td>9.48±3.59</td>
</tr>
<tr>
<td>0.25</td>
<td>1.2</td>
<td>.63±.05</td>
<td>.23±.20</td>
<td>.91±.08</td>
<td>0.42±0.23</td>
<td>0.66±0.42</td>
<td>8.41±3.05</td>
</tr>
<tr>
<td>0.13</td>
<td>1.8</td>
<td>.62±.07</td>
<td>.19±.25</td>
<td>.92±.06</td>
<td>0.40±0.28</td>
<td>0.38±0.26</td>
<td>9.43±3.56</td>
</tr>
<tr>
<td><strong>Original pieces</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.13</td>
<td></td>
<td>.65±.05</td>
<td>.09±.18</td>
<td>.74±.10</td>
<td>.05±0.05</td>
<td>0.04±0.04</td>
<td>12.24±11.32</td>
</tr>
<tr>
<td>0.25</td>
<td></td>
<td>.65±.05</td>
<td>.09±.18</td>
<td>.74±.10</td>
<td>.31±0.27</td>
<td>0.57±0.45</td>
<td>9.91±9.29</td>
</tr>
</tbody>
</table>

- **Ablation Studies on Model Architecture**

<table>
<thead>
<tr>
<th>sequence length $N$</th>
<th>#self-att layers $L$</th>
<th>SE</th>
<th>separate PEs</th>
<th>Melody inconsistency↓</th>
<th>Theme inconsistency↓</th>
<th>Theme uncontrollability↓</th>
<th>Theme gap (in # bars)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Theme Transformer</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>512</td>
<td>6</td>
<td></td>
<td></td>
<td>.13±.24</td>
<td>0.27±.26</td>
<td>0.24±0.20</td>
<td>9.48±3.59</td>
</tr>
<tr>
<td>1,024</td>
<td>6</td>
<td></td>
<td></td>
<td>.07±.15</td>
<td>0.27±.21</td>
<td>0.26±0.19</td>
<td>13.70±8.34</td>
</tr>
<tr>
<td>512</td>
<td>6</td>
<td></td>
<td></td>
<td>.19±.23</td>
<td>0.55±.27</td>
<td>1.07±.26</td>
<td>7.70±3.68</td>
</tr>
<tr>
<td><strong>Baseline [13], [17]</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>512</td>
<td>6</td>
<td></td>
<td></td>
<td>.33±.38</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>512</td>
<td>12</td>
<td></td>
<td></td>
<td>.33±.38</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><strong>Original pieces</strong></td>
<td></td>
<td></td>
<td></td>
<td>.09±.18</td>
<td>.05±.05</td>
<td>0.04±0.04</td>
<td>12.24±11.32</td>
</tr>
</tbody>
</table>
Outline

• Overview
• Technical Background
• Results
• Contribution
Contributions

• Proposed an Unsupervised Method for Theme Retrieval
• The first work to introduce Theme-based Symbolic Music Generation
• Design Theme-based Evaluation Metrics
• Our method outperform previous music generation works
Thanks for listening

Ian Shih
Email: yjshih23@gmail.com
Website: atosystem.github.io