All-in-One Transformer: Unifying Speech Recognition, Audio Tagging, and Event Detection

Niko Moritz, Gordon Wichern, Takaaki Hori, Jonathan Le Roux

Interspeech 2020
All-in-One Transformer: Unifying ASR, AT, and AED

• Motivation:
  – Automatic speech recognition (ASR), audio event detection (AED), and audio tagging (AT) are traditionally treated as separate problems with custom-made solutions.
  – In contrast, the human auditory system uses a single (binaural) pathway to process sound signals from different sources.

• Investigated Questions:
  – Can we develop a system that moves closer to the versatility of the human auditory system?
  – Can training on multiple heterogeneous tasks lead to a single system with performance similar to or better than systems developed independently for each task?
  – Can a single system successfully handle multiple tasks with widely varying characteristics, large length discrepancies, and w/ or w/o monotonicity?
Audio Tagging (AT)

cat meowing
speech
baby crying
dog barking
traffic sounds
home environment
Acoustic Event Detection (AED)

- cat meowing
- dog barking
- baby crying
- speech
- speech
- traffic sounds
Automatic Speech Recognition (ASR)

“It’s your turn to change diapers” “Okay, I’ll do that”
Baseline System Architectures of DCASE 2019 Task 1, 2, 4, and 5

<table>
<thead>
<tr>
<th>Audio sampling rates</th>
<th>Feature Extraction: Log-Mel Spectral Energies</th>
<th>Neural Network Models</th>
<th>Classifier Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>16 kHz, 22.05 kHz</td>
<td>40, 64, 96, or 128-dimensional; Different window and hop sizes;</td>
<td>CNNs, DNNs, RNNs, ...; MobileNet v1; VGGish;</td>
<td>Logistic regression; Max and average pooling; Attention-based pooling Clip- and frame-level classification;</td>
</tr>
<tr>
<td>44.1 kHz, or 48 kHz</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The Auditory Pathway

- Cochlear nucleus
- Super olivary nucleus
- Inferior colliculus
- Medial geniculate nucleus
- Auditory cortex

Decoder
Attention
Encoder

Cochlear nucleus → Medial geniculate nucleus → Inferior colliculus → Super olivary nucleus → Cochlear nucleus → Auditory cortex
Attention-based Encoder-Decoder

<asr> Hello World </s>

Encoder States

Encoder

Decoder

Acoustic Features

Audio waveform

Feature Extraction

© MERL
System Architecture: Attention-Based Encoder-Decoder

16 kHz sampling rate
80-dim. Log-Mel spectral energies + pitch features

Transformer

Feature Extraction
Encoder
Decoder

Audio waveform

<asr> hello <space> world </s>
<at> happy female_voice cat </s>
<aed> catS catC speechS catE speechC speechE </s>

Acoustic Event Detection

catS  catC  catE

speechS  speechC  speechE

eventS => event start
eventC => event continues
eventE => event end

© MERL
System Architecture: Hybrid CTC-Transformer

Connectionist Temporal Classification (CTC)

Acoustic Event Detection

event_s => event start
event_c => event continues
event_e => event end
Data Sets

- **Automatic Speech Recognition (ASR)**
  - Wall Street Journal (WSJ): “Read English newspapers”.
    Train: 81h; Dev.: 1.1h; Test: 0.7h

- **Acoustic Event Detection (AED)**
  - DCASE 2019 task 4 (DCASE19-4): “Sound event detection in domestic environments”.
    Train: 5.7h; Dev.: 2.9h; Test: 1.9h
• **Audio Tagging (AT):**
  - DCASE 2017 task 4 (**DCASE17-4**): “Large-scale weakly supervised sound event detection for smart cars”.
    Train: 140h; Dev.: 1.3h; Test: 3h
  - DCASE 2018 task 3 (**DCASE18-3**): “Bird audio detection”.
    Train: 99h; Dev.: n/a; Test: n/a
  - DCASE 2019 task 1 (**DCASE19-1**): “Audio scene classification”.
    Train: 25.5h; Dev.: 11.6h; Test: 9.8h
  - DCASE 2019 task 2 (**DCASE19-2**): “Audio tagging with noisy labels and minimal supervision”.
    Train: 90.8; Dev.: 3.1h; Test: 9.8h
  - DCASE 2019 task 4 (**DCASE19-4**): “Sound event detection in domestic environments”.
    Train: 9.8h; Dev.: 2.9h; Test: 1.9h
  - DCASE 2019 task 5 (**DCASE19-5**): “Urban sound tagging”.
    Train: 4.4h; Dev.: 1.2h; Test: 0.7h
  - The Ryerson Audio-Visual Database of Emotional Speech and Song (**RAVDESS**): Recognition of “emotion” + “vocal channel” + “gender”
    Train: 2.8h; Dev.: n/a; Test: n/a
All-in-One (AIO) Transformer

- Feature Extraction
- Encoder
- Decoder
- CTC

- <asr> hello world </s>
- <aed> cat_c cat_c speech_c cat_E speech_c speech_E </s>
- <at1> happy_at1 speech_at1 female_at1 </s>
- <at2> bus_at2 skateboard_at2 </s>
- <at3> bird_at3 </s>
- <at4> public_square_at4 </s>
- <at5> bus_at5 female_speech_at5 </s>
- <at6> cat_at6 speech_at6 </s>
- <at7> engine_presence_at7 </s>

- (WSJ)
- (DCASE 2019, task 4)
- (RAVDESS)
- (DCASE 2017, task 4)
- (DCASE 2018, task 3)
- (DCASE 2019, task 1)
- (DCASE 2019, task 2)
- (DCASE 2019, task 4)
- (DCASE 2019, task 5)
## Audio Tagging – Results

### Micro-averaged F1-scores [%]

<table>
<thead>
<tr>
<th>System</th>
<th>Training data</th>
<th>DCASE19</th>
<th>DCASE18</th>
<th>DCASE17</th>
<th>RAVDESS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Task 1</td>
<td>Task 2</td>
<td>Task 4</td>
<td>Task 5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Task 1</td>
<td>Task 2</td>
<td>Task 4</td>
<td>Task 5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Task 3</td>
<td>Task 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>System</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline Systems</td>
<td>single</td>
<td>AT</td>
<td>AED</td>
<td>ASR</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>dev</td>
<td>dev</td>
<td>test</td>
<td>dev</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dev</td>
<td>test</td>
<td>dev</td>
<td>test</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dev</td>
<td>test</td>
<td>dev</td>
<td>test</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dev</td>
<td>test</td>
<td>dev</td>
<td>test</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n/a</td>
<td>n/a</td>
<td>19.0</td>
<td>29.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>62.5</td>
<td>39.8</td>
<td>38.8</td>
<td>71.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>66.8</td>
<td>73.0</td>
<td>68.9</td>
<td></td>
</tr>
</tbody>
</table>

**Note:** Single baseline systems are marked with an 'X'.
## Automatic Speech Recognition – Results

<table>
<thead>
<tr>
<th>System</th>
<th>Training data</th>
<th>Word Error Rates [%]</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>WSJ</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>System</strong></td>
<td><strong>AT</strong></td>
<td><strong>AED</strong></td>
</tr>
<tr>
<td>CTC-Transformer</td>
<td>✓</td>
<td>✓</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>CTC-Transformer</td>
<td>✓</td>
<td>✔</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>AIO Transformer</td>
<td>multi</td>
<td>✓</td>
<td>✔</td>
<td></td>
</tr>
</tbody>
</table>

Multi-condition training using DEMAND and NOISEX data sets.
Noisy test conditions using the DCASE data sets.

* Multi-condition training
<table>
<thead>
<tr>
<th>System</th>
<th>Training data</th>
<th></th>
<th></th>
<th></th>
<th>Event-based</th>
<th>Segment-based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AT</td>
<td>AED</td>
<td>ASR</td>
<td>dev</td>
<td>test</td>
</tr>
<tr>
<td>Baseline system</td>
<td>X</td>
<td>✓</td>
<td>X</td>
<td></td>
<td>29.0</td>
<td>24.0</td>
</tr>
<tr>
<td>CTC-Transformer</td>
<td>X</td>
<td>✓</td>
<td>X</td>
<td></td>
<td>16.0</td>
<td>10.6</td>
</tr>
</tbody>
</table>

Event-based F1-scores: 200 ms collar for on- and offsets
Segment-based F1-scores: 1 sec. long segments
* Multi-condition training
• ASR, AED, and AT tasks can be unified under a single system architecture, where model parameters are shared across all tasks.

• Multi-task learning has shown to improve results for individual tasks.

• The AIO Transformer model has achieve competitive or better results compared to all tested DCASE challenge baseline systems, as well as to an ASR baseline system of similar architecture.

• The proposed system can be used to perform the total transcription of an acoustic scene, i.e., a single system can be used to transcribe speech as well as other acoustic events.