Neural audio captioning
- Generating text describing non-speech audio -

Abstract

Recently, detection and classification of various sounds has attracted many researchers' attention. We propose an audio captioning system that can describe various non-speech audio signals in the form of natural language. Most existing audio captioning systems have mainly focused on "what the individual sound is," or classifying sounds to find object labels or types. In contrast, the proposed system generates (1) an onomatopoeia, i.e. a verbal simulation of non-speech sounds, and (2) a sentence describing sounds, given an audio signal as an input. This allows the description to include more information, such as how the sound sounds and how the tone or volume changes over time. Our approach also enables directly measuring the distance between a sentence and an audio sample. The potential applications include sound effect search systems that can accept detailed sentence queries, audio captioning systems for videos, and AI systems that can hear and represent sounds as humans do.

Detailed audio captioning

- Given an audio input, the system describes not only what the sound source is, but also how it is sounding and how it is changing over time, as a natural language sentence.

Method

1. Onomatopoeic mode: output a word simulating the sound
2. Sentence mode: output a sentence describing the sound

Point: No unique correct answer for captioning → Proposal: Can control the degree of detail by conditioning the decoder by "Specificity" input

Specificity: Sum of the amount of information contained in the output text.

Experimental results

The description matches the sound:

Approx. 3/4

Appropriateness of the output sentence (without specificity conditioning)

With the specificity control, the sentence is:

Approx. 2/3

Effectiveness of the specificity conditioning

Examples of description for a base drum sound

(English translation from Japanese)

<table>
<thead>
<tr>
<th>C</th>
<th>Generated sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>A low sound rings for a moment</td>
</tr>
<tr>
<td>20</td>
<td>A low sound sounds for a moment</td>
</tr>
<tr>
<td>50</td>
<td>A low, striking sound sounds as if something is dashed on a mat</td>
</tr>
<tr>
<td>80</td>
<td>A very low-pitched drum is played only once</td>
</tr>
<tr>
<td>110</td>
<td>A faint, low-pitched sound sounds as if something is hit dully, and it soon disappears</td>
</tr>
</tbody>
</table>

References


Contact

Kunio Kashino Email: cs-liaison-ml at hco.ntt.co.jp
Media Information Laboratory