An Encoder-Decoder Based Audio Captioning System with Transfer and Reinforcement Learning

Xinhao Mei1, Qiushi Huang1,4, Xubo Liu1, Gengyun Chen2, Jingqian Wu3, Yusong Wu3, Jinzheng Zhao1, Shengchen Li3, Tom Ko4, H Lilian Tang1, Xi Shao2, Mark D. Plumbley5, Wenwu Wang6

1University of Surrey, Guildford, United Kingdom, 2Nanjing University of Posts and Telecommunications, Nanjing, China, 3Xi’an Jiaotong-Liverpool University, Suzhou, China, 4Southern University of Science and Technology, Shenzhen, China

Introduction:

- Automated audio captioning aims to use natural language to describe the content of audio data.
- Existing audio captioning systems almost follow an encoder-decoder architecture, where the decoder predicts words based on audio features extracted by the encoder.
- In this work, we propose to use transfer learning and reinforcement learning to improve an audio captioning system.
- The resulting system was ranked 3rd in DCASE 2021 Task 6, and it was the best system without using ensemble technique.
- Reinforcement learning may impact adversely on the quality of the generated captions.

Model and methods:

- CNN-Transformer model is used as our baseline.
- Transfer learning is used to solve the data scarcity problem.
- Transferring from upstream task (PANNs)
- Transferring from a large in-domain dataset (AudioCaps)
- Reinforcement learning is used to address the ‘exposure bias’ problem and directly optimize the evaluation metric CIDEr.

Issues:

- The official dataset Clotho is limited, which just contains 5929 audio clips.
- Maximum likelihood training introduces ‘exposure bias’.
- Training objective mismatches with the evaluation metrics.

Results:

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU1</th>
<th>BLEU2</th>
<th>BLEU3</th>
<th>BLEU4</th>
<th>ROUGE_L</th>
<th>METERO</th>
<th>CIDEr</th>
<th>SPICE</th>
<th>SPIDER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.525</td>
<td>0.344</td>
<td>0.237</td>
<td>0.163</td>
<td>0.359</td>
<td>0.154</td>
<td>0.352</td>
<td>0.100</td>
<td>0.226</td>
</tr>
<tr>
<td>B+PANNs</td>
<td>0.564</td>
<td>0.375</td>
<td>0.255</td>
<td>0.171</td>
<td>0.383</td>
<td>0.172</td>
<td>0.421</td>
<td>0.120</td>
<td>0.270</td>
</tr>
<tr>
<td>B+PANNs+AC</td>
<td>0.561</td>
<td>0.374</td>
<td>0.257</td>
<td>0.174</td>
<td>0.379</td>
<td>0.171</td>
<td>0.426</td>
<td>0.124</td>
<td>0.275</td>
</tr>
<tr>
<td>B+PANNs+RL</td>
<td>0.639</td>
<td>0.415</td>
<td>0.276</td>
<td>0.174</td>
<td>0.401</td>
<td>0.186</td>
<td>0.452</td>
<td>0.131</td>
<td>0.292</td>
</tr>
<tr>
<td>B+PANNs+AC+RL</td>
<td>0.634</td>
<td>0.423</td>
<td>0.288</td>
<td>0.185</td>
<td>0.410</td>
<td>0.187</td>
<td>0.476</td>
<td>0.134</td>
<td>0.305</td>
</tr>
</tbody>
</table>

Conclusion:

- Transfer learning and reinforcement learning both significantly improve the score of the evaluation metrics.
- Reinforcement learning may impact adversely on the quality of the generated captions.
- Conventional evaluation metrics may not correlate well with human judgements.