Automated Audio Captioning by Fine-tuning BART on AudioSet Tags

Félix Gontier, Romain Serizel and Christophe Cerisara
Université de Lorraine, CNRS, Inria, Loria

Task: Describe the contents of audio extracts with fluent English sentences

Current approaches:
• Encoder-decoder models with attention
  -> Language modeling is learned from scratch
• Textual conditioning provides vocabulary guidance in addition to audio embeddings
  -> Audio-based keyword extraction or caption retrieval systems trained on the captioning dataset

Proposition: Combine pre-trained audio tagging and large-scale language models

Proposed Approach

BART conditional language model [1]
• Standard transformer encoder-decoder architecture
• Byte-pair encoding tokenization: 50265 tokens in vocabulary
• Pre-training scheme: denoising heavily corrupted text
• Transfer learning to AAC by simply fine-tuning with multi-modal inputs

AudioCaps dataset [2]
• Training on 49000 clips, 10s each, single caption
• Validation/Evaluation on 485/955 clips, 5 captions
• Subset of AudioSet: in-domain audio for pre-trained tagging models

Training:
• Cross-entropy loss on BART token vocabulary
• Stable training observed until convergence, even when fully fine-tuning
• Results reported over 3 runs with different random seeds
• Sampling on YAMNet logits at training: data augmentation
• Most likely tags taken at inference

Evaluation:
• Machine translation metrics (n-gram matching): BLEU-1/2/3/4, METEOR, ROUGE-L, CIDEr
• SPICE: average of CIDEr and SPICE, main metric of DCASE task 6

Results

<table>
<thead>
<tr>
<th>VGGish</th>
<th>PANNs</th>
<th>YAMNet</th>
<th>CIDEr</th>
<th>SPICE</th>
<th>SPIDER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Koizumi et al. [3]</td>
<td>50.3</td>
<td>13.9</td>
<td>32.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eren et al. [4]</td>
<td>75.0</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BART + YAMNet + PANNs</td>
<td>75.3</td>
<td>17.6</td>
<td>46.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>91.3</td>
<td>21.6</td>
<td>56.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Model performance
• On par or better than the state of the art on AudioCaps
• Higher BLEU-1/2/3 than reference captions cross-validation (human score)

Complementary experiments
• Random initialization: The performance improvement from BART pre-training is limited with sufficient amounts of training data
• Freezing: BART decoder is already effective to model caption structure
• Capacity: A marginal decrease in performance is observed with 3 times fewer parameters
  -> Low diversity in caption structure and vocabulary?
• Task-specific fine-tuning: The initial training loss is lower with summarization checkpoints

References: