Grounded Sequence to Sequence Transduction (Multi-Modal Speech Recognition)

Florian Metze
July 17, 2018

Imagine How-To Videos

- Lots of potential for multi-modal processing & fusion
- For Speech-to-Text and beyond
Audio-Visual ASR vs Multi-modal ASR

- Traditional audio-visual ASR based on speakers’ lip/mouth movement
  - **Sub-phonetic synchronicity** required, fusion a problem
- Lip/mouth information not always available in how-to videos
  - Humans are usually present, but often they “do things”
  - Instead: fuse information at the **semantic level** (**words, …**)

  e.g. AVASR “Grid” Corpus
    
  “How-To” Video

Multi-Modal MT – Example

- **SRC**: Three children in **football uniforms** of two different teams are playing **football** on a **football field**, while another player and an adult stand in the background.
- **TXT**: Drei Kinder in **Fußballtrikots** zweier verschiedener Mannschaften spielen **Fußball** auf einem **Fußballplatz** während ein weiterer Spieler und eine Erwachsene im Hintergrund stehen.
- **IMG**: Drei Kinder in **Footballtrikots** zweier verschiedener Mannschaften spielen **Football** auf einem **Footballplatz** während ein weiterer Spieler und ein Erwachsener im Hintergrund stehen.

Courtesy of Lucia Specia
Two (+) Types of Features

- **Object Features**
  - monitor, mouse, keyboard, ...
  - 1000 classes [Deng et al., 2009]
  - Could also do Actions, ...

- **Place Features (Scenes)**
  - train (office, baseball field, airport apron, …)
  - 205 classes [Zhou et al., 2014]

How-to Video Corpus

- "How-to" dataset of instructional videos
  - Harvested from the web *(2000h+ available)*
  - "Utterance" (from caption) is 8s...10s
  - On average 18 words
  - ~55,000 videos
  - 300h+ have been translated into Portuguese
  - 4h dev & eval set; ~20k+ vocabulary size
  - Extract one quasi-static visual "context" vector per utterance
  - Pick frame randomly (for now)
  - Object/ place detection, or action recognition
The Goal

- Have a corpus of 2000h of how-to videos
- Fully transcribed in English
- Partially translated into Portuguese (and Turkish)
- With short descriptions of videos
- Learn shared audio-visual (or text-visual) representations to help us understand video
- Recognize, translate, and summarize videos
- Use sequence-to-sequence models (S2S) as unified architecture
Preliminary Experiments: ASR Adaptation

- All is standard error back-propagation
- Independent of the structure & features, context
  - SAT technique can be naturally applied to CNNs, RNNs
  - Also tried: speaker microphone distance, speaker features (age, gender, race; 61-dimensional) [Miao et al., 2016]

Comparison of Approaches

<table>
<thead>
<tr>
<th>Model</th>
<th>Features</th>
<th>WER(%)</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNN (Baseline)</td>
<td>------</td>
<td>23.4</td>
<td></td>
</tr>
<tr>
<td>Adaptive Training</td>
<td>161-dim visual features</td>
<td>22.3</td>
<td>4.7%</td>
</tr>
<tr>
<td>Adaptive Training</td>
<td>100-dim speaker i-vectors</td>
<td>22.0</td>
<td>6.0%</td>
</tr>
<tr>
<td>Adaptive Training</td>
<td>261-dim fused features</td>
<td>21.5</td>
<td>8.1%</td>
</tr>
</tbody>
</table>

[Gupta et al., 2017]

- AV adaptation does not beat i-Vector adaptation, but is in ballpark, somewhat complementary
Language Modeling

- Context aware language models easy with RNNs
- [Zweig et al., 2012; …]
- Append context vector to word embeddings
- NMT of image captions [Specia et al., 2016]

---

Bi-LSTM LM (5-fold CV)

- 30-best lists from 23.4% WER DNN baseline
- Re-score and re-rank with LSTM-LM
- 22.6% WER (15.6% Oracle WER)
- Small but consistent improvements
Result Analysis – “indoor” vs “outdoor”

- Using object and place features only
  - AM+LM adapt.: 23.4% → 21.5% WER on 4h dev set (90h training)
- LM adaptation improves results across the board
  - 126/ 156 videos improve
- AM improves “noisy” videos
  - 55/ 156 videos improve (most are “outdoor”, according to their category)

Video as side-information in S2S ASR?

\[ n \in \{\text{all utterances}\} \]

\[ X_{t,n}, c_n, T_{i,n}, S_{j,n} \]

**INPUT**

**OUTPUT**

“Hey, now, I got those nice ripe cherries and I will decorate the cake with it”

“Decorate the cake with cherries”

Summary

Translation

Retrieval

Q\&A
Adaptive Seq-2-Seq with Attention

6+ ways of incorporating "visual context":

- Encoder feature shifts and appending features (AM)
- Input layer, pyramid output
- At decoder (LM)
  - With attention mechanism
  - Co-Attention (2 sequences)
- At softmax layer (1G LM)

S2S Training Results (90h How-To)

- Appending 100d adaptation vector to 120d IMEL feature
- Best TER observed for later epochs, where perplexity increases
- Nice improvement in TER (17.5% ➔ 16.8%)
- Also works for CTC models, but somewhat inconsistent
Audio-Visual ASR Results

- It is possible to adapt (condition) a E2E ASR Model to static context, like a domain
  - CTC and S2S models both work
  - The error rates improve, integration with an adapted language model gives further gains
- More experimentation is needed, but models seem to learn semantic properties of the (correlated) video
  - Multi-task (CTC+S2S) training?
  - Determine best units: chars, BPE, words, …
  - Shared representations have been learned?

Can you fly this thing?

Not yet. 
[...]
Let’s go!
Multimedia Summarization

- Which how-to videos to watch, and why?

S2S Summarization
Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Bleu-1</th>
<th>Bleu-2</th>
<th>Bleu-3</th>
<th>Bleu-4</th>
<th>Meteor, penalty=0</th>
<th>Rouge-L</th>
<th>Avg. words replaced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (original)</td>
<td>52.282</td>
<td>41.929</td>
<td>35.652</td>
<td>31.214</td>
<td>0.52</td>
<td>0.506</td>
<td>-</td>
</tr>
<tr>
<td>Without catch-</td>
<td>33.811</td>
<td>22.731</td>
<td>16.699</td>
<td>12.862</td>
<td>0.36</td>
<td>0.370</td>
<td>6.70</td>
</tr>
<tr>
<td>phrases</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rule-based</td>
<td>22.152</td>
<td>10.059</td>
<td>5.527</td>
<td>3.345</td>
<td>0.21</td>
<td>0.164</td>
<td>-</td>
</tr>
<tr>
<td>Without catch-</td>
<td>19.483</td>
<td>8.656</td>
<td>4.800</td>
<td>2.904</td>
<td>0.19</td>
<td>0.155</td>
<td>1.25</td>
</tr>
<tr>
<td>phrases</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Ongoing Experiments

- Multi-Document Summarization
  - Take **triplets** of videos (anchor/ same/ different)
  - Use a sequence-to-sequence model to generate **two** "descriptions" for **three** videos together
    - "similar" (portions of) videos or
    - "different" videos
  - Experiment with different architectures ongoing
- Triplet loss to encourage sharing and learning
- Multi-modal features
Where To?

- Conversational Search: UIs without Screens
- Robotics – see what Humans see
- Explainable AI

Questions?

https://www.clsp.jhu.edu/workshops/18-workshop

fmetze@cs.cmu.edu
Bibliography ASR


Bibliography (Video) Summarization