Representations of language in a model of visually grounded speech signal

Grzegorz Chrupała  Lieke Gelderloos  Afra Alishahi
Automatic Speech Recognition

A major commercial success story in Language Technology
Very heavy-handed supervision

I can see you
Grounded speech perception
Data

- Flickr8K Audio (Harwath & Glass 2015)
  - 8K images, five audio captions each
- MS COCO Synthetic Spoken Captions
  - 300K images, five synthetically spoken captions each
Project speech and image to joint space

- A bird walks on a beam
- Bears play in water
Image model
Speech model

- Input: **MFCC**
- Subsampling CNN
- Recurrent Highway Network (Zilly et al 2016)
- Attention
## Model settings

<table>
<thead>
<tr>
<th></th>
<th>Flickr8K Speech</th>
<th>COCO Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Attention</strong></td>
<td>128</td>
<td>512</td>
</tr>
<tr>
<td><strong>RHN depth</strong></td>
<td>2, 1024</td>
<td>2, 512</td>
</tr>
<tr>
<td><strong>RHN depth</strong></td>
<td>2, 1024</td>
<td>2, 512</td>
</tr>
<tr>
<td><strong>RHN depth</strong></td>
<td>2, 1024</td>
<td>2, 512</td>
</tr>
<tr>
<td><strong>Conv</strong></td>
<td>6x64, stride 2</td>
<td>6x64, stride 3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Flickr8K Text</th>
<th>COCO Text</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RHN depth</strong></td>
<td>1, 1024</td>
<td>1, 1024</td>
</tr>
<tr>
<td><strong>Embedding</strong></td>
<td>300</td>
<td>300</td>
</tr>
</tbody>
</table>
# Image Retrieval

<table>
<thead>
<tr>
<th></th>
<th>Model</th>
<th>R@10</th>
<th>(\tilde{r})</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Flickr8K</strong></td>
<td>Speech RHN(_{4,2})</td>
<td>0.253</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>Harwath &amp; Glass 2015</td>
<td>0.179</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Text RHN(_{1,1})</td>
<td>0.494</td>
<td>11</td>
</tr>
<tr>
<td><strong>MSCOCO</strong></td>
<td>Speech RHN(_{5,2})</td>
<td>0.444</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Text RHN(_{1,1})</td>
<td>0.565</td>
<td>8</td>
</tr>
</tbody>
</table>

Newer CNN architecture: Harwath et al 2016 (NIPS), Harwath and Glass 2017 (ACL)
Levels of representation

▪ What aspects of sentences are encoded?
▪ Which layers encode form, which encode meaning?
▪ Auxiliary tasks (Adi et al 2017)
Form-related aspects

Use activation vectors to decode

- Utterance length in words
- Presence of specific words
Number of words

- **Input**
  - Activations for utterance

- **Model**
  - Linear regression

![Graph showing sentence length prediction (R^2) vs. network layers for different data sets and model configurations.](image)
Word presence

- **Input**
  - Activations for utterance
  - MFCC for word
- **Model**
  - MLP

![Graph showing Word Prediction Accuracy vs Network layers for COCO and Flickr8k datasets. The graph has two curves, one for COCO in blue and one for Flickr8k in red. The y-axis represents Word Prediction Accuracy ranging from 45 to 80, and the x-axis represents Network layers from 0 to 6. The blue curve for COCO is generally higher than the red curve for Flickr8k.](image_url)
Semantic aspects
Representational Similarity

- Correlations between sets of pairwise similarities according to
  - Activations
  - Edit ops on written sentences
  - Human judgments

(SICK dataset)
Homonym disambiguation
Follow-up work

Afra Alishahi, Marie Barking and Grzegorz Chrupała. Encoding of phonology in a recurrent neural model of grounded speech

Friday, session #4 at CoNLL
Conclusion

Encodings of form and meaning emerge and evolve in hidden layers of stacked RHN listening to grounded speech

Code: github.com/gchrupala/visually-grounded-speech
Data: doi.org/10.5281/zenodo.400926
Error analysis

- Text usually better
- Speech better:
  - Long descriptions
  - Misspellings

Text: a yellow and white birtd is in flight
Speech: a yellow and white bird is in flight

Text
Speech
Length

![Graph showing the comparison between text and speech length]

Below is a list of words and phrases related to length:

- while
- in the
- immediate
- foreground
- juts
- a gnarled tree
- branch
- the majority
- of the view
- consists of a
- an expanse of short grass
- dotted with a few longer tufts of long grass
- scattered grazing sheep
Text model

- Convolution → word embedding
- No attention