Emergence of attention in a neural model of visually grounded speech
William N Havard, Jean-Pierre Chevrot, Laurent Besacier

To cite this version:
William N Havard, Jean-Pierre Chevrot, Laurent Besacier. Emergence of attention in a neural model of visually grounded speech. Learning Language in Humans and in Machines 2018 conference, Jul 2018, Paris, France. hal-01970514

HAL Id: hal-01970514
https://hal.science/hal-01970514
Submitted on 5 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
1. Introduction

- Context provides children with all necessary information to build a coherent mental representation of the world. While acquiring their native language, children learn to map portions of this mental representation to whole or part of acoustic realisations they perceive from surrounding speech: this process is known as lexical acquisition.
- The goal of this research is to study how a visually grounded neural network segments a continuous flow of speech hoping it could shed light on how children perform the same process. To do so, we analyse how the final representation learnt by the attention mechanism changes over time and more specifically, we focus on the attention models of the network.
- An attention mechanism is a component that computes a vector representing the focus of the system on different parts of the speech input.

2. Architecture

3. Attention weights

Attention weights for GRU1 and GRU5 evolve differently over time suggesting they encode speech in a different manner.

Peaks are first randomly positioned and gradually shift to part of the speech signal that are useful in describing the image before reaching their final position.

4. Peak detection and statistics

- Peak detection for each level of attention
- Statistics on peak number and height and speech segments located underneath

5. Word level

<table>
<thead>
<tr>
<th>Ref. frequencies</th>
<th>1-gram</th>
<th>2-gram</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5 grams</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.5 grams</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.5 grams</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.5 grams</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

GRU1: word 1-gram with a skateboard
GRU5: word 2-gram with a toilet

Top 10 occurrences

- Word bigrams for GRU1 and GRU5 very different from reference frequencies
- GRU5 focuses on the end of the captions (<100>)

6. Phone level

<table>
<thead>
<tr>
<th>GRU1</th>
<th>1-gram</th>
<th>2-gram</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5 grams</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.5 grams</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.5 grams</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.5 grams</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

GRU5: peak mainly located above phones ending a word: cue for word segmentation?

7. Conclusion

- GRU5 highlights speech segments that match word bigrams and isolated word that correspond to the main concept of the image.
- Behaviour of the first attention mechanism (GRU1) is still unclear, and it is hard to tell so far which linguistic units are targeted by this attention mechanism. However, removing the attention layer on GRU1 yields worse results, suggesting it does bring useful information to the network.
- Further linguistics units could be considered, such as POS and syllables.
- Diachronic study: How does the distribution of the words under the peaks evolve over time?

References