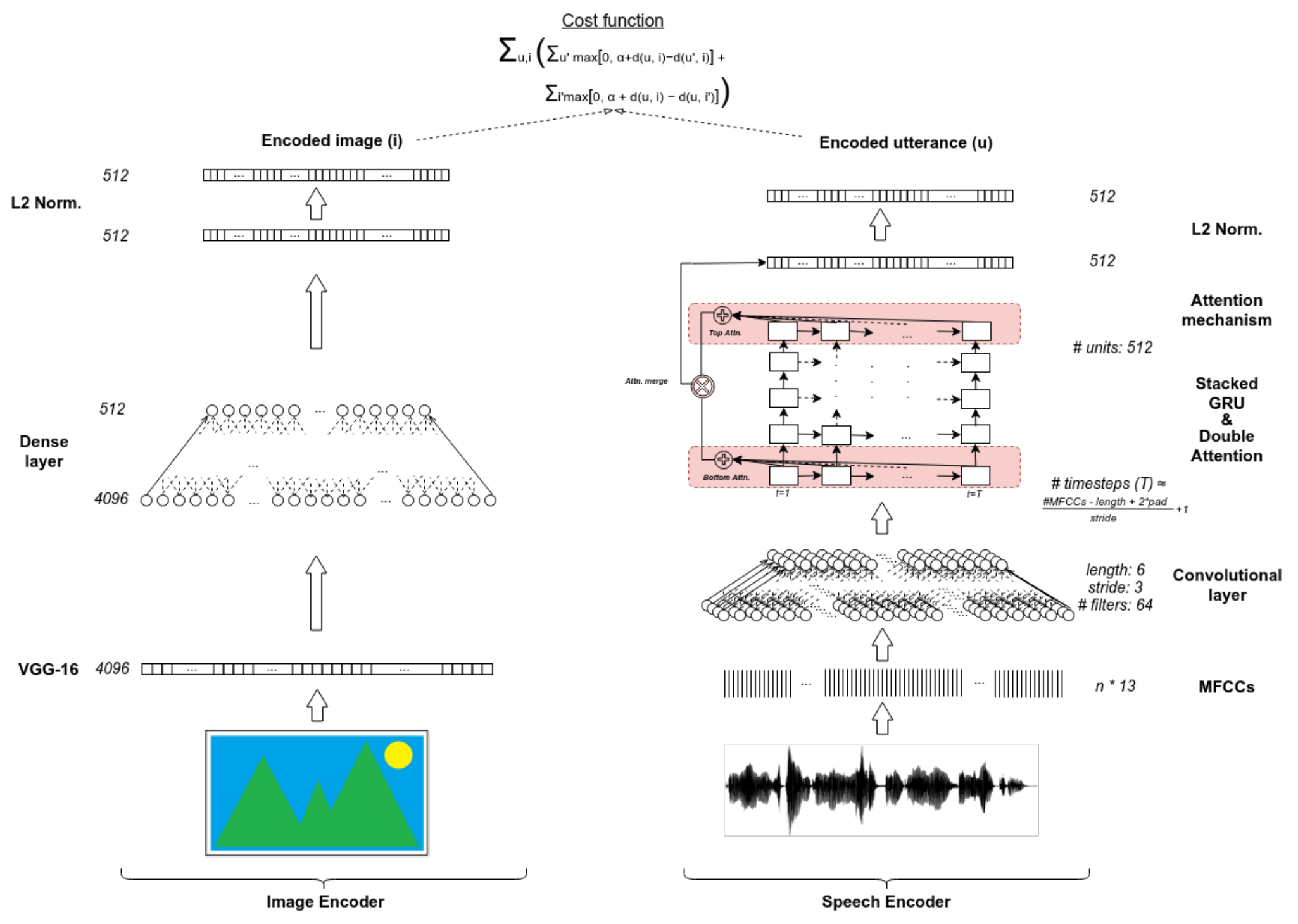
# Emergence of attention in a neural model of visually grounded speech

William N. Havard[1][2], Jean-Pierre Chevrot[1], Laurent Besacier[2]
[1]LIDILEM, *Univ. Grenoble Alpes, 38000 Grenoble, France*
[2]*LIG, Univ. Grenoble Alpes, CNRS, Grenoble INP, 38000 Grenoble, France*

## 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 analyses **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



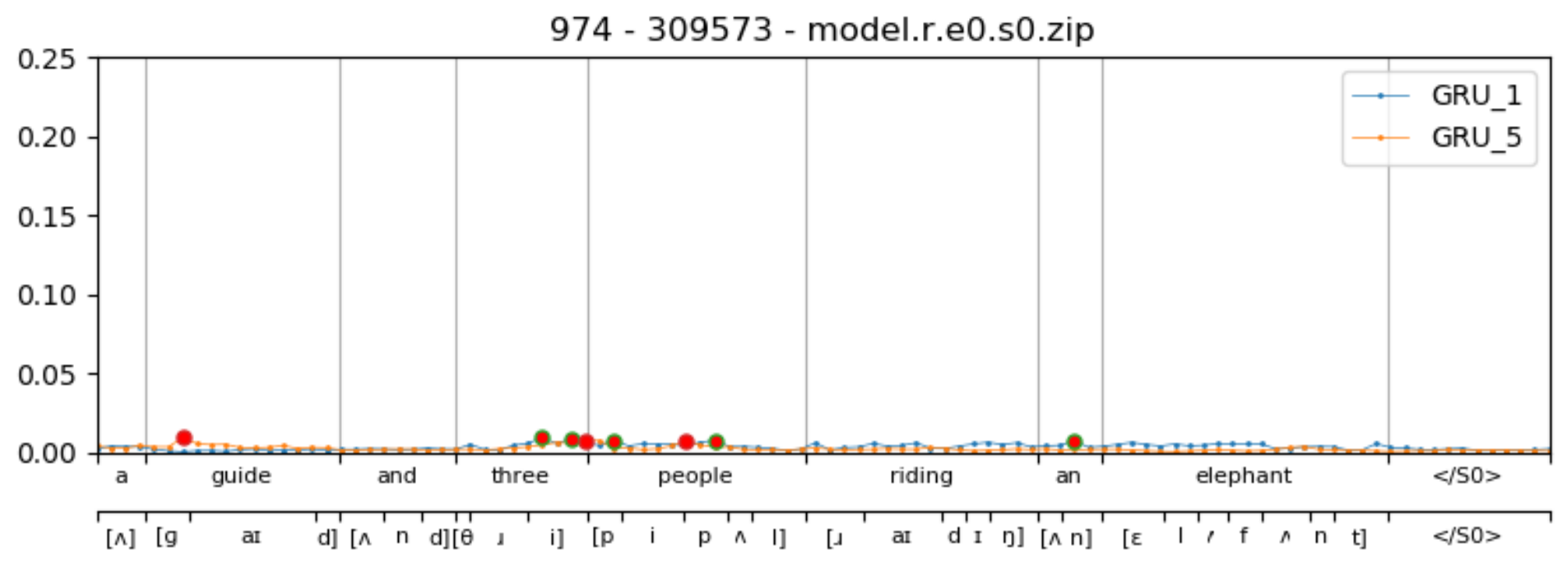
| | r@1 | r@5 | r@10 | Avg. rank |
|---|---|---|---|---|
| Base model | 0.108 | 0.305 | 0.443 | 13 |
| Our model | 0.059 | 0.193 | 0.301 | 25 |

- Based on Chrupała *et al.* (2017)
- Two attentions mechanisms (inspired by Matthys *et al.*, 2005)
- Model regularly saved during training

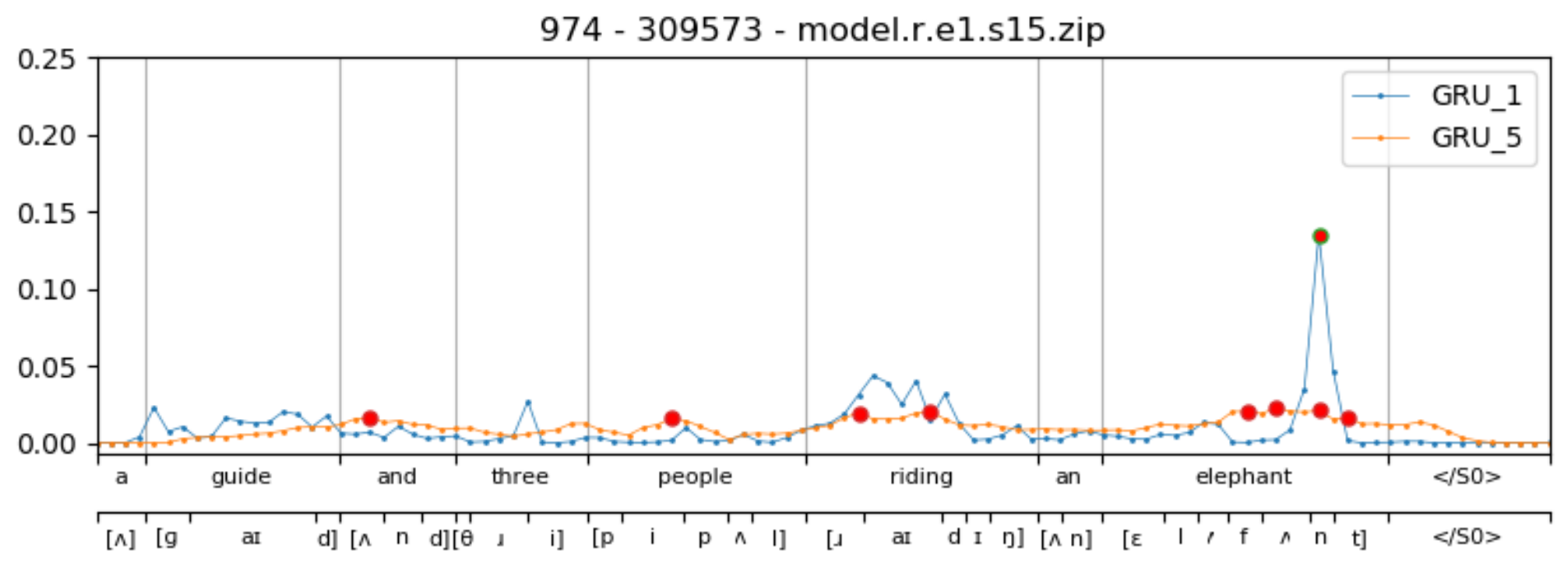
## 3. Attention weights

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

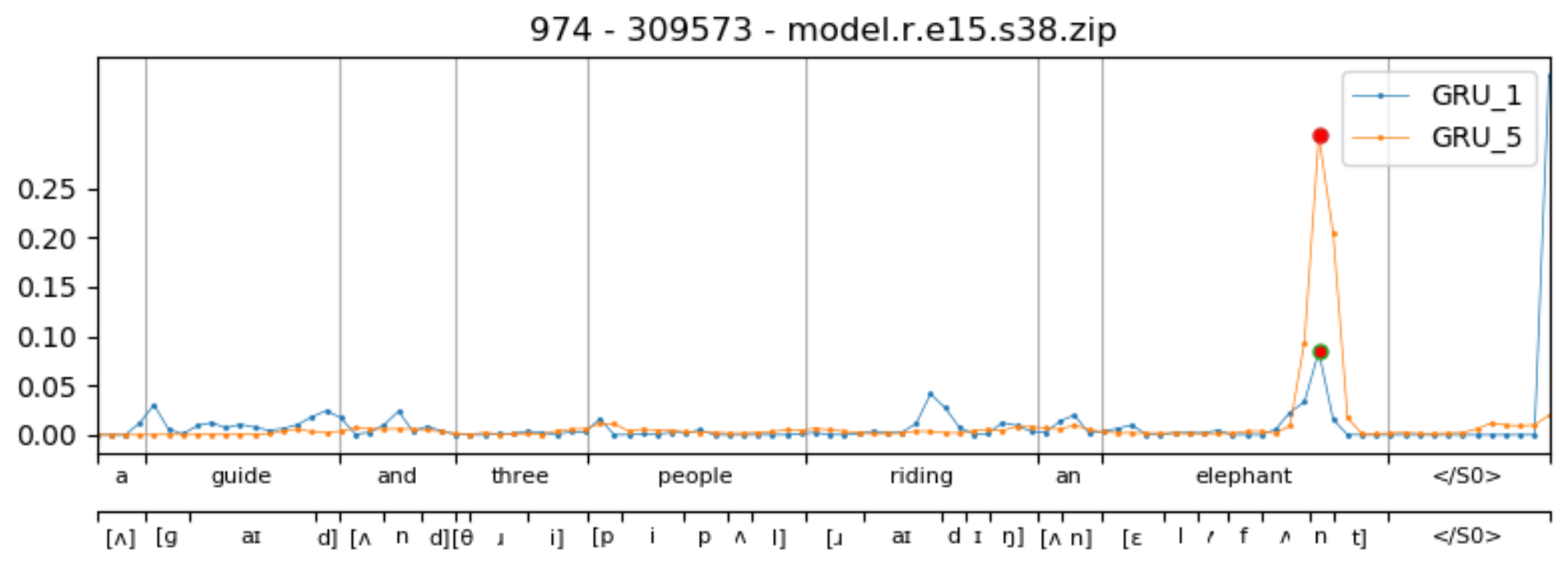
Peaks are first randomly positioned and gradually shift to part of the speech signal that are useful in describing the image before reaching their final position.



0th epoch (random weights)



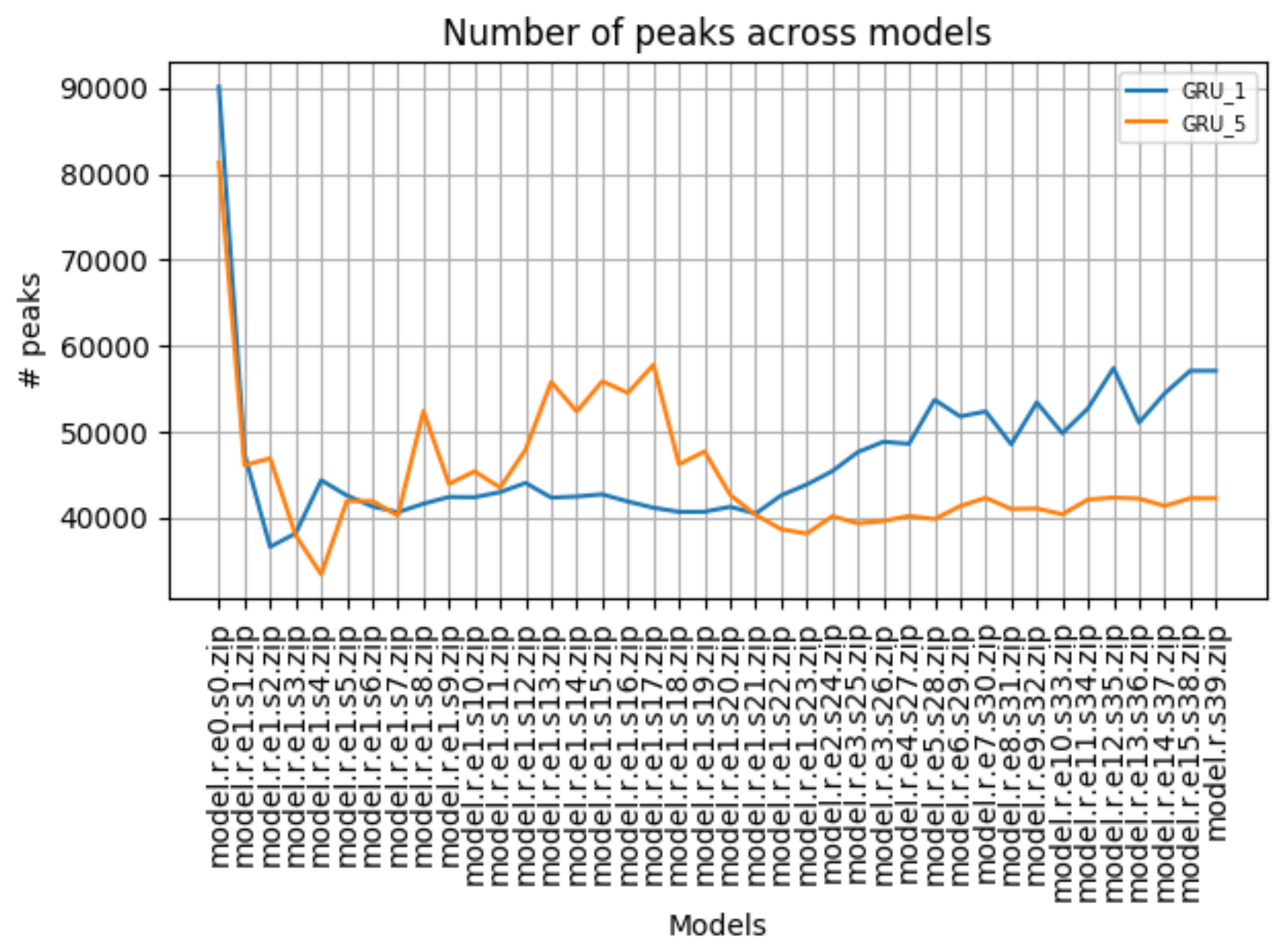
1th epoch - 26k training instances

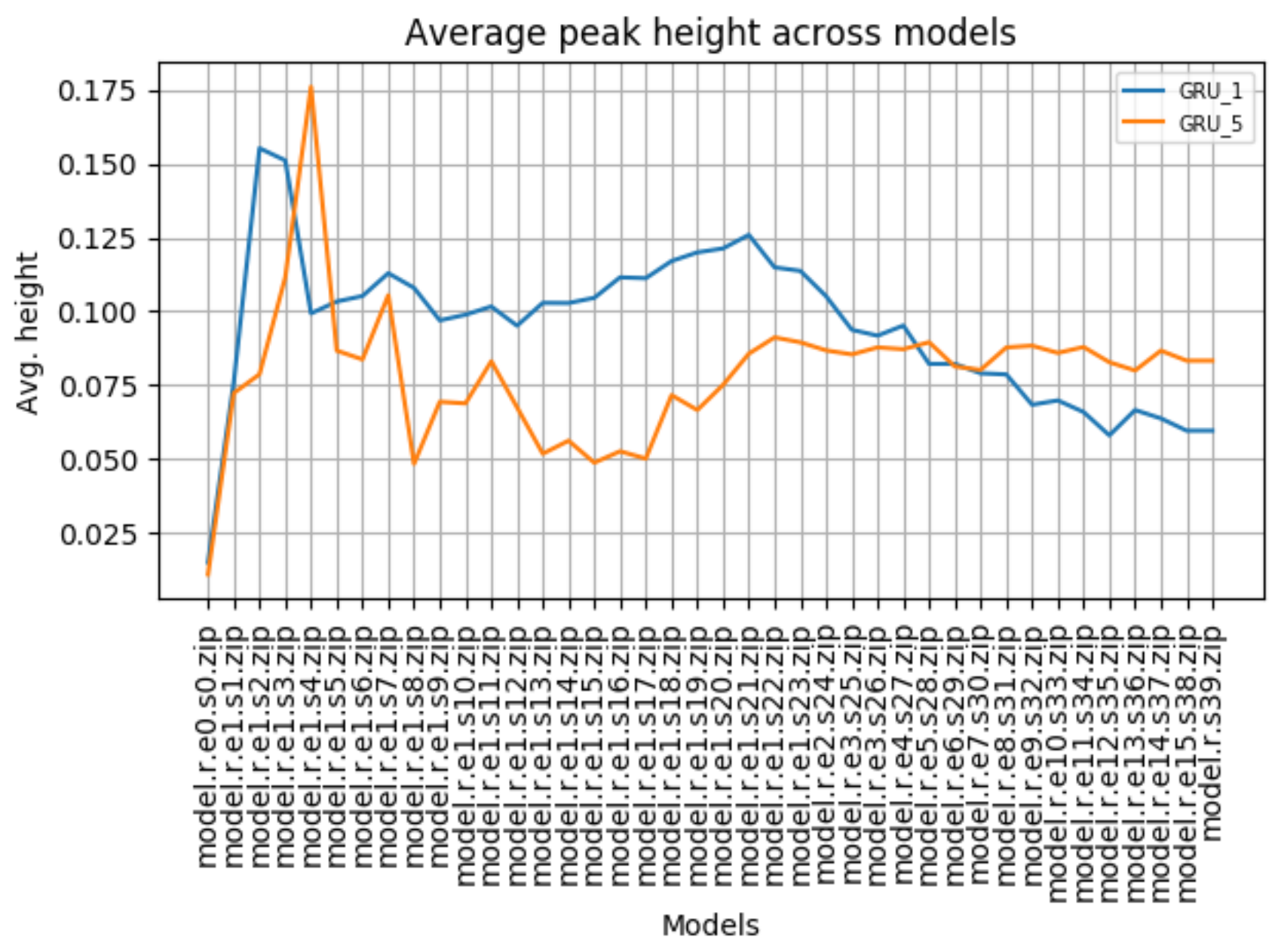


15th epoch - 8.4M training instances

## 4. Peak detection and statistics

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





- **Attention unevenly distributed over time** between GRU1 and GRU5:
  - Network gradually reduces the number of peaks for GRU5 while making them higher. **The network focuses on very one specific part of the speech signal**: the **main concept** of the image
  - On the contrary, the number of peaks for GRU1 increases while also making them smaller. **GRU1 attention is distributed over the whole speech signal and is thus less specific**

## 5. Word level

| Ref. frequencies | | GRU1 | | GRU5 | |
|---|---|---|---|---|---|
| 1-gram | 2-gram | 1-gram | 2-gram | 1-gram | 2-gram |
| a | <S0> a | a | a cat | with | a skateboard |
| <S0> | on a | of | a man | toilet | a toilet |
| </S0> | in a | cat | a dog | </S0> | a giraffe |
| on | of a | dog | a tennis | <unk> | living room |
| of | with a | is | a giraffe | skateboard | a baseball |
| the | a man | white | on a | in | stop sign |
| in | in the | tennis | <S0> a | is | baseball player |
| with | on the | in | a white | sign | a kitchen |
| and | next to | man | a table | and | fire hydrant |
| is | <S0> two | with | a plate | giraffe | a laptop |

Top 10 occurrences

- Word bigrams for GRU1 and GRU5 very different from reference frequencies
- GRU5 focuses on the end of the captions (</S0>)

## 6. Phone level

| GRU1 | | GRU5 | |
|---|---|---|---|
| 1-gram | 2-gram | 1-gram | 2-gram |
| æ | ɪ ŋ] | n] | ɔ ɹ |
| t] | [k æ | t] | eɪ n] |
| eɪ | [ʌ v] | ɹ | ɪ ŋ] |
| n] | ɔ ɹ | ʌ | ʌ n] |
| ŋ] | aɪ t] | k] | u m] |
| t | [d ɔ | d] | aɪ n] |
| ɛ | u m] | s] | t ɝ] |
| ɪ | [dʒ ɝ | ɝ] | ʌ l] |
| n | l eɪ | z] | k ʌ |
| a ɪ | l ɑ | sp | ʌ s] |

Top 10 occurrences

- GRU5 peaks 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

[1] Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. Representations of language in a model of visually grounded speech signal. *arXiv:1702.01991 [cs]*, February 2017. arXiv: 1702.01991.

[2] Sven L. Mattys, Laurence White, and James F. Melhorn. Integration of Multiple Speech Segmentation Cues: A Hierarchical Framework. *Journal of Experimental Psychology: General*, 134(4):477–500, 2005.

This work was supported by grants from NeuroCoG IDEX UGA in the framework of the "Investissements d'avenir" program (ANR-15-IDEX-02)