Lexical emergence from context: exploring unsupervised learning approaches on large multimodal language corpora

William N. Havard

OUTLINE

1. **Context**

2. **Model & Data**

3. **Experiments**

   - **Behaviour of Attention**

     Havard, W. N., Chevrot, J.-P. & Besacier, L. (2019), **Models of Visually Grounded Speech Signal Pay Attention to Nouns: A bilingual Experiment on English and Japanese**, ICASSP2019

   - **Word Recognition, Competition, and Activation**

     Havard, W. N., Chevrot, J.-P. & Besacier, L. (2019), **Word Recognition, Competition, and Activation in a Model of Visually Grounded Speech**, CoNLL2019

   - **Introduction of Linguistic Information**

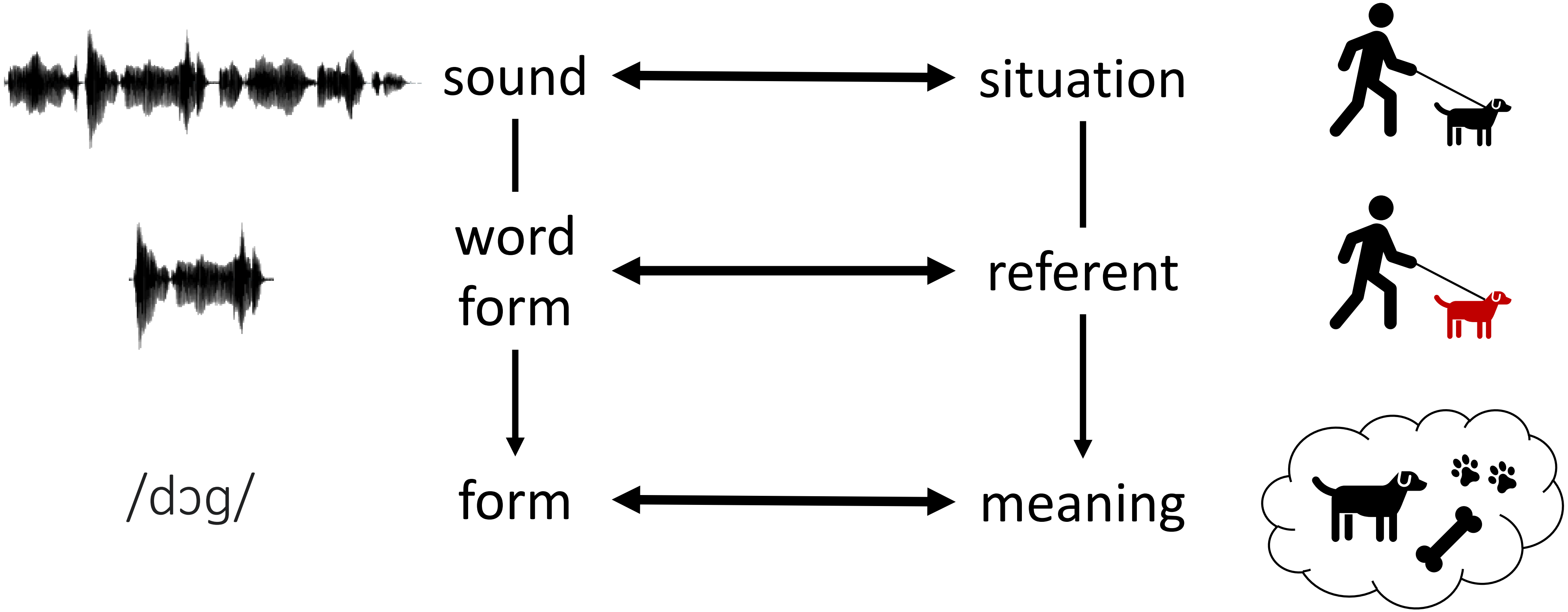
     Havard, W., , Chevrot, J.-P. & Besacier, L. (2020), **Catplayinginthesnow: Impact of Prior Segmentation on a Model of Visually Grounded Speech**, CoNLL2020

4. **Conclusion**

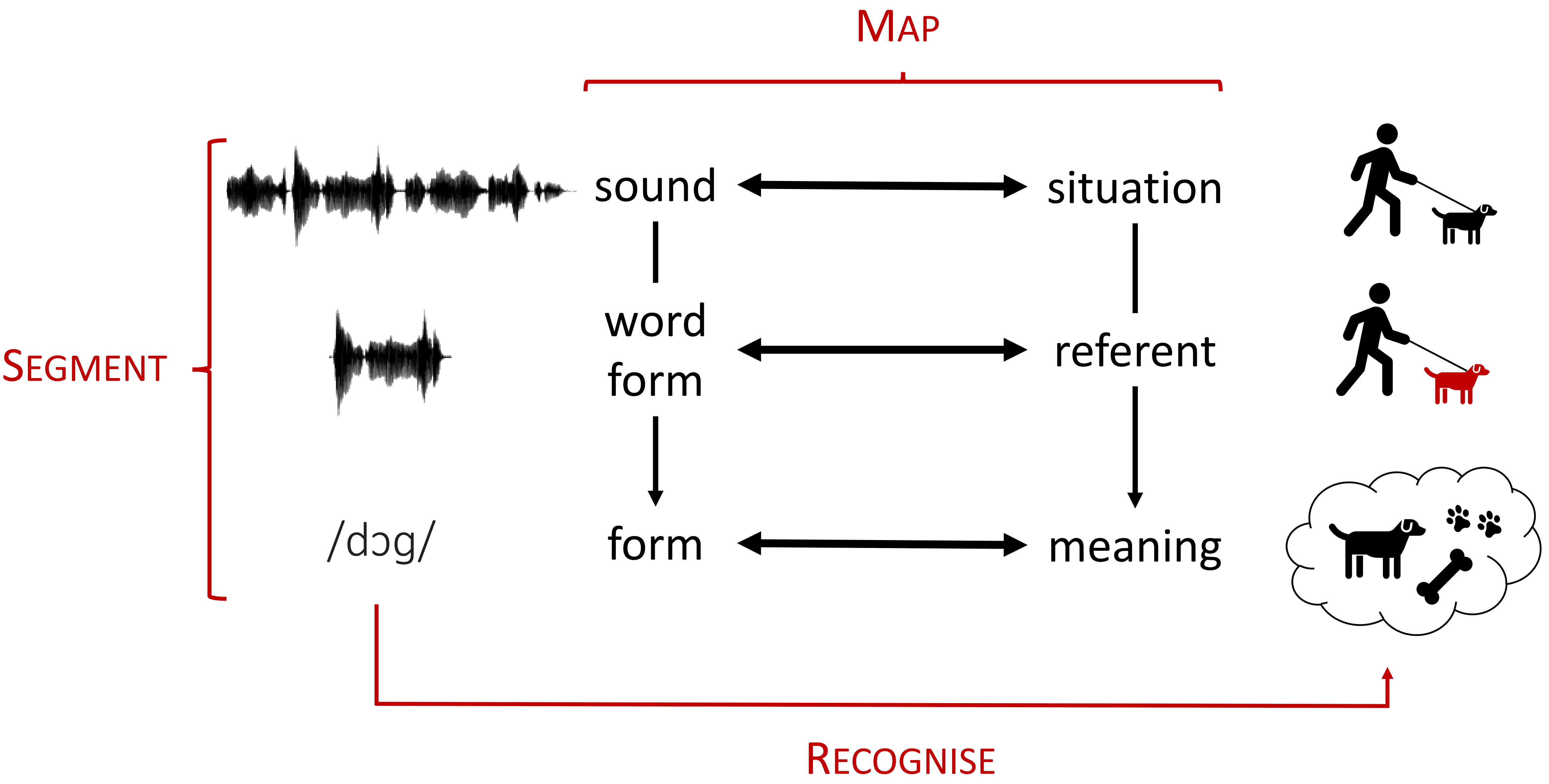
Context

Context — Language Acquisition

- "the child's input consists of **sound/situation pairs**, but his final output is a set of **form/meaning pairs**" [Landau & Gleitman, p.7, 1985]

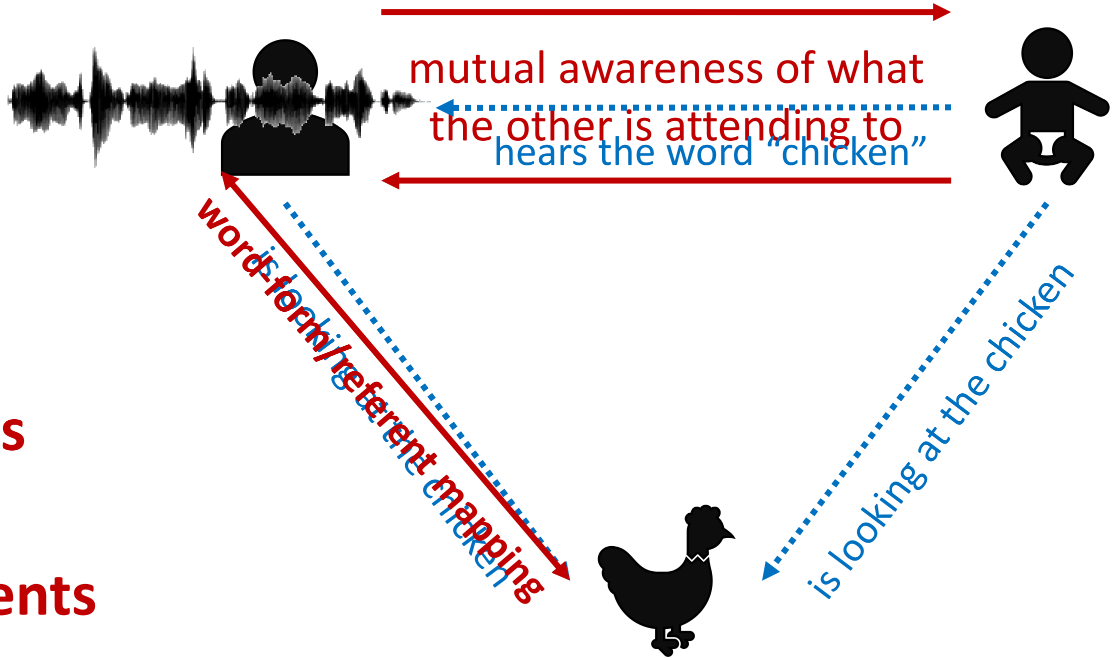


Context — Language Acquisition



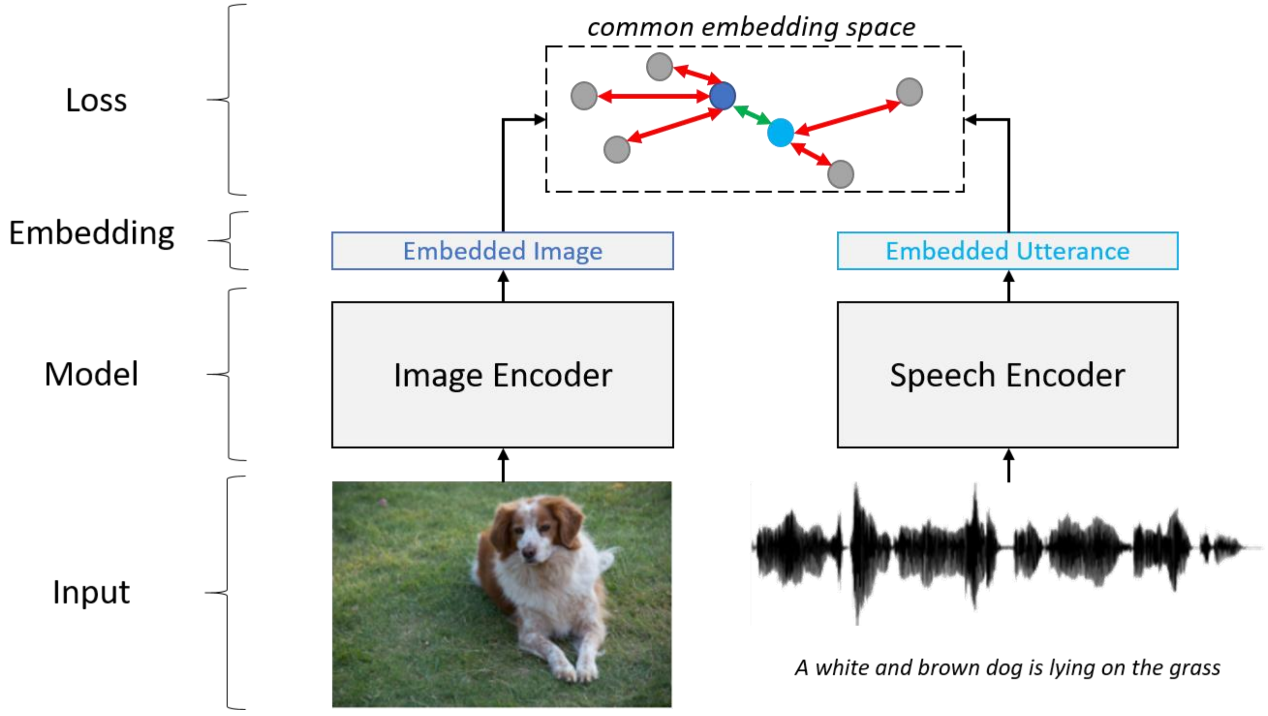
Context — Language Acquisition

- Contextual information
  - **Vision**, touch, smell, …
  - **Social interactions**



- Vision
  - essential to enter **joint attention frames**
  - used to **map word-forms to their referents**
  - **lack of visual input** slightly **hinders language acquisition** [Andersen et al., 1984; Dunlea, 1989]

VISUALLY GROUNDED SPEECH MODELS



- Trained on a **speech ⇄ image retrieval task**
- Either **CNN-based** [Gabriel et al., Harwath et al., Kamper et al.] or **RNN-based** [Chrupała et al., Merkx et al.]
- Project images and paired spoken captions in a **common representation space**

Visually Grounded Speech Models — Hypothesis

- **Hypothesis:** VGS models also have to **transition from sound/situation pairs** to **form/meaning pairs** to solve their task

- Same tasks as children
  - **Segmentation**
  - **Mapping**
  - **Recognition**

- Develop **linguistic abilities** as **a by-product** of their task

Previous works — CNN-Based Models

→ **fine-grained audio-visual mappings**
- map words to their visual referent

[Harwath et al., 2017]

→ **language agnostic**
- English, Hindi & Japanese

[Harwath et al., 2018]

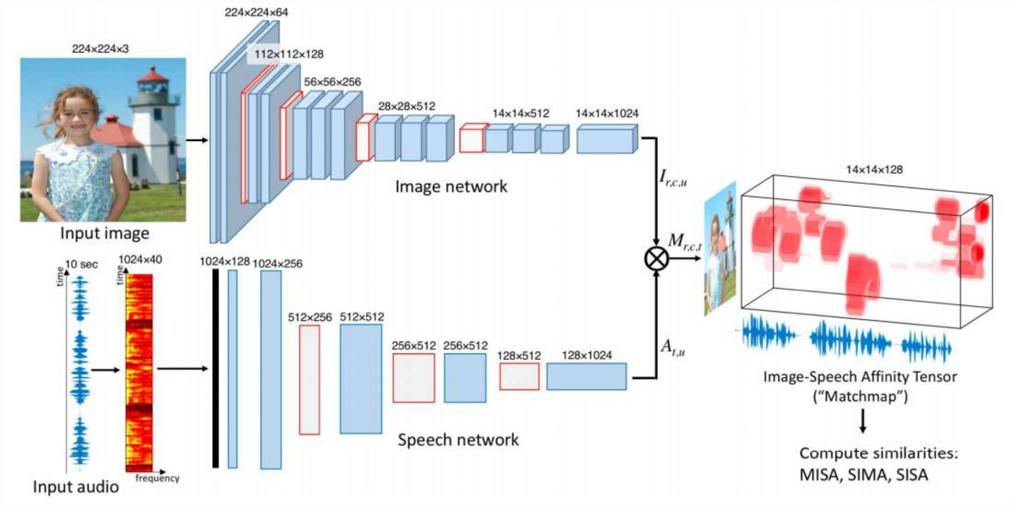


Fig. from Harwath et al., 2018

→ **form in lower layers, meaning in higher layers**
- lower layers: clusters according to speaker indentity
- upper layers: clusters according to meaning

[Drexler et al., 2017]

→ **implicit segmentation**
- sensitivity to phone boundaries

[Harwath et al., 2019]

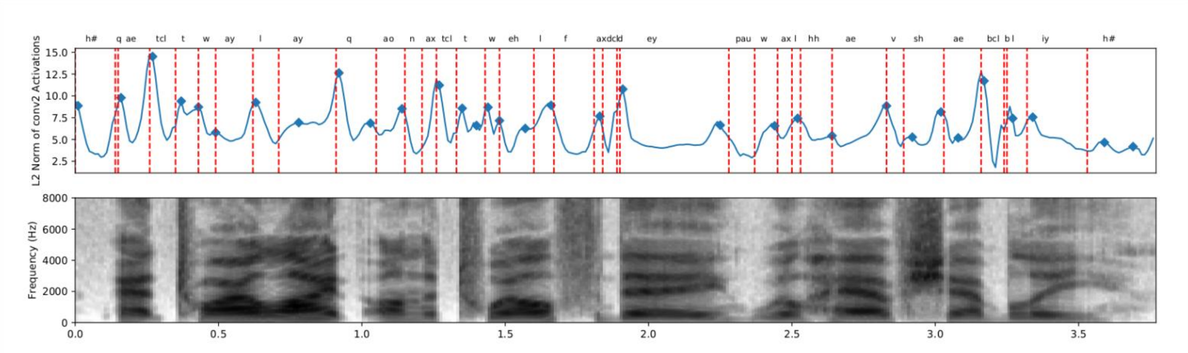


Fig. from Harwath et al., 2019

Previous works — RNN-based Models

→ models encode presence of individual words

- word presence/absence task
- not all layers are equally informative

[Chrupała et al., 2017; Merkx et al., 2019]

→ form in lower layers, meaning in higher layers

[Chrupała et al., 2017; Alishahi et al., 2017]

Attention: size 512
Recurrent 5: size 512
Recurrent 4: size 512
Recurrent 3: size 512
Recurrent 2: size 512
Recurrent 1: size 512
Convolutional: size 64, length 6, stride 3
Input MFCC: size 13

Fig. from Alishahi et al., 2017

Research Questions

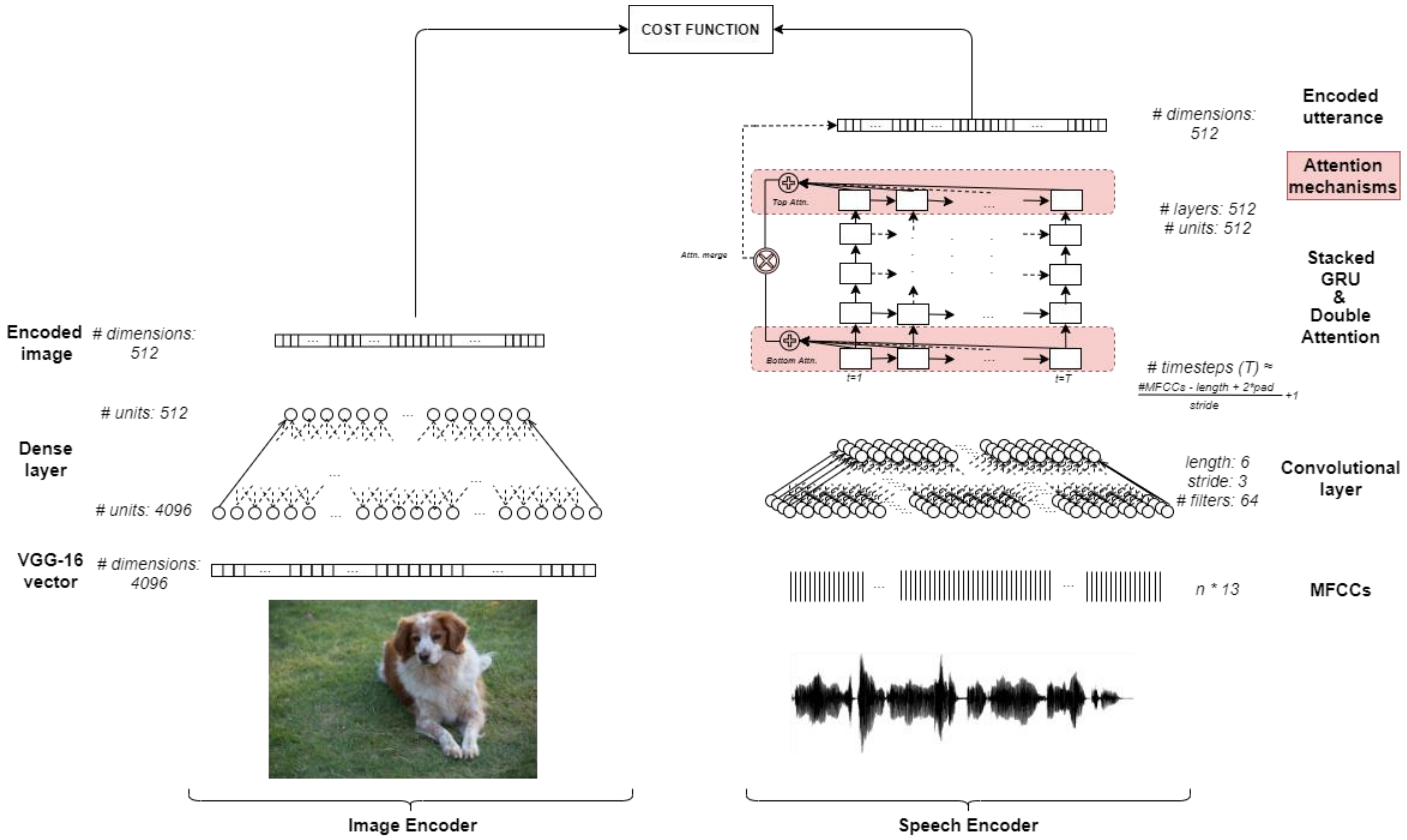
→ Do RNN-based VGS models learn to **detect specific words** in the speech signal?

→ **How** is the semantic representation of a **word activated**?

→ Is an **implicit segmentation** as efficient as an **explicit segmentation**?

Model & Data

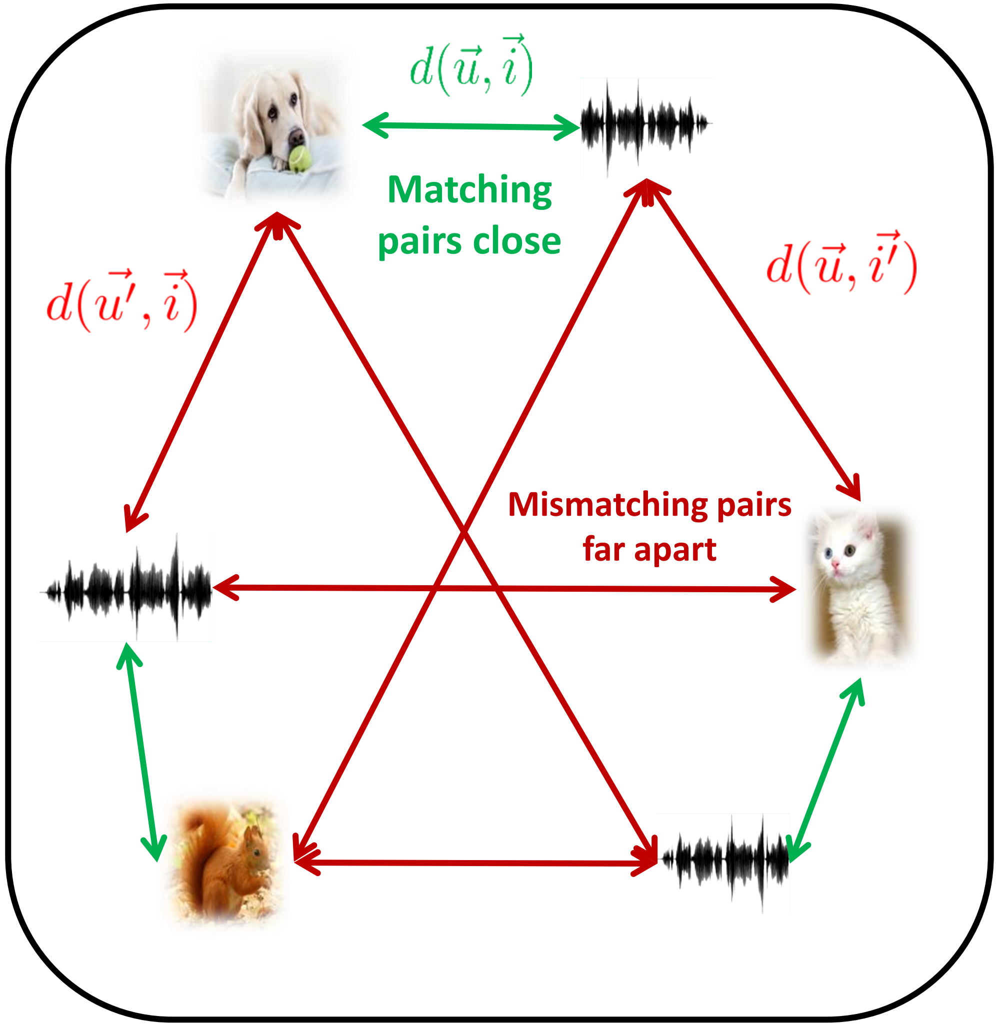
- Model by [Chrupała et al., 2017]



Triplet Loss

$$\mathcal{L}(u,i,\alpha) = \sum_{u,i}\left(\underbrace{\sum_{u'}\max[0,\alpha + d(\vec{u},\vec{i}) - d(\vec{u'},\vec{i})]}_{\substack{\text{maximise } d \\ \text{mismatching utterance} \\ \text{and anchor image}}} + \underbrace{\sum_{i'}\max[0,\alpha + d(\vec{u},\vec{i}) - d(\vec{u},\vec{i'})]}_{\substack{\text{maximise } d \\ \text{anchor utterance and} \\ \text{mismatching image}}}\right)$$

$\vec{u}$	utterance vector
$\vec{i}$	image vector
α	margin
$d(\cdot,\cdot)$	cosine distance
$d(\vec{u},\vec{i})$	distance between anchor utterance and anchor image
$d(\vec{u'},\vec{i})$	distance between **mismatching utterance** and anchor image
$d(\vec{u},\vec{i'})$	distance between anchor utterance and **mismatching image**



n dimensional space

Figure adapted from a presentation by Chrupała (2017)

Data Sets

- **Image Captionning Data Sets** and their **audio extensions**
  - **MSCOCO** [Lin et al., 2017] → **Synthetically Spoken COCO** [Chrupała et al., 2017]
  - **STAIR** [Yoshikawa et al., 2017] → **Synthetically Spoken STAIR** [Havard et al., 2019]
  - **FLICKR8k** [Hodosh et al., 2013] → **Flickr8k Audio Caption Corpus** [Harwath et al., 2015]

Same set of images
Synthetic speech
Natural speech

Photo by Martha de Jong-Lantink, CC BY-NC-ND 2.0
Fig. 492506 from MSCOCO

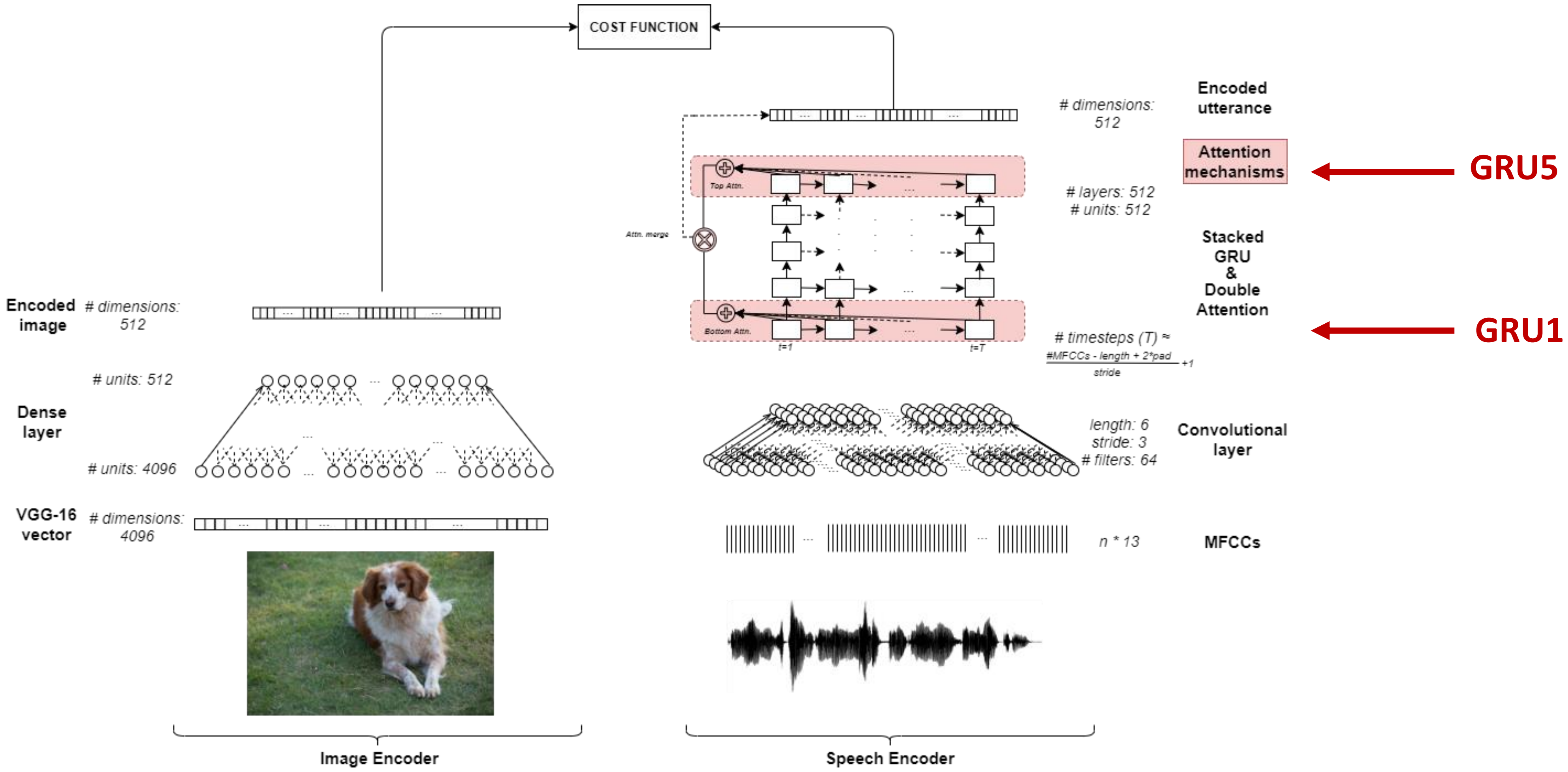
MSCOCO (English)
there are three giraffes together in the wild three giraffes are standing on the plains of africa *(sic)*. three giraffes walking in a field with trees in the background the three giraffes walk together in the safari. some animals that are around the grass together.

STAIR (Japanese)	STAIR (Translation)
キリンが3匹草原の中を歩いている	Three giraffes walking in the meadow
3頭のうち1頭のキリンは口を少し開けている	One of the three giraffes has a slightly open mouth
草地を三頭のキリンが歩いている	Three giraffes are walking in the grassland
キリンが3頭草原をあるいている	Three giraffes in the meadow
三頭のキリンが草原を歩いている	Three giraffes walking in the meadow

Behaviour of Attention

MODEL

- Model by [Chrupała et al., 2017]



Behaviour of Attention

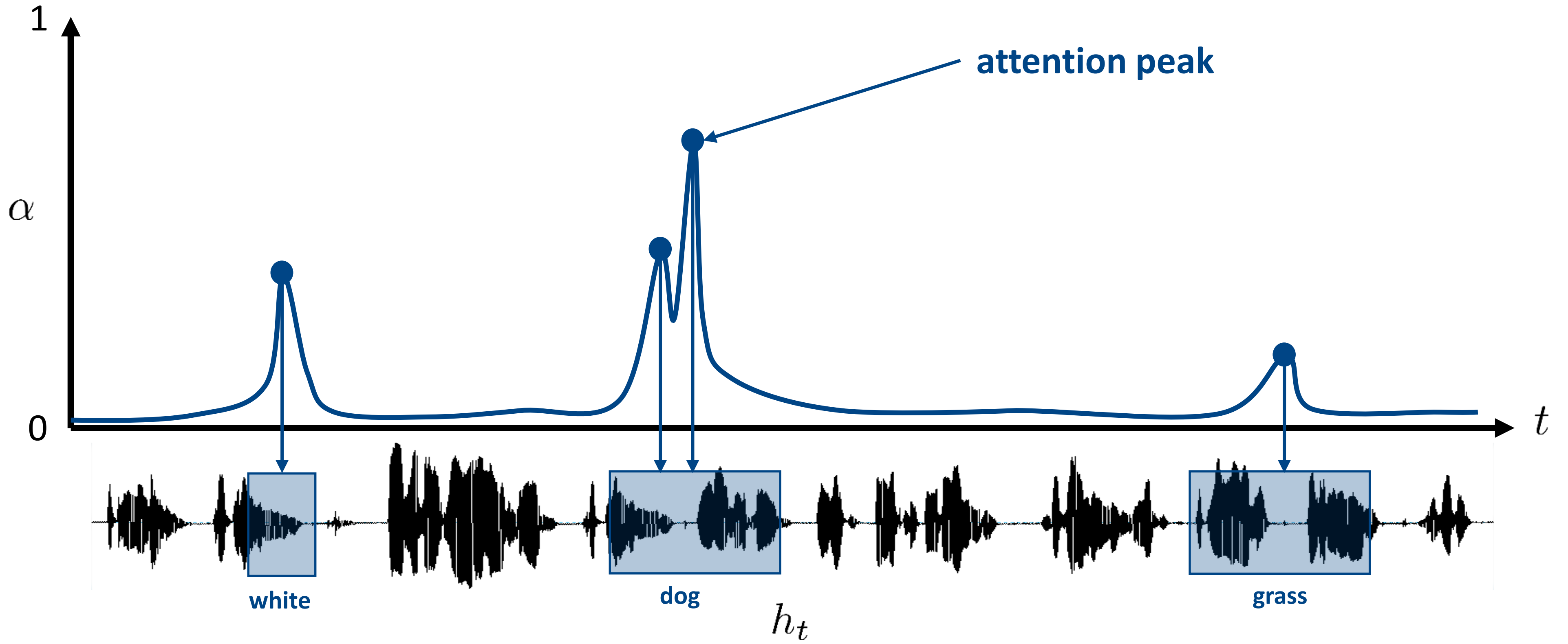
- Does the model rely on **specific parts** in the speech signal for its predictions?
  - Analysis of the **attention weights**
  - specific part-of-speech (nouns, adjectives, etc.) ?
  - specific words?

- How is this **different from chance**?

Behaviour of Attention

$$c = \sum_{t=1}^{T} \alpha_t h_t$$

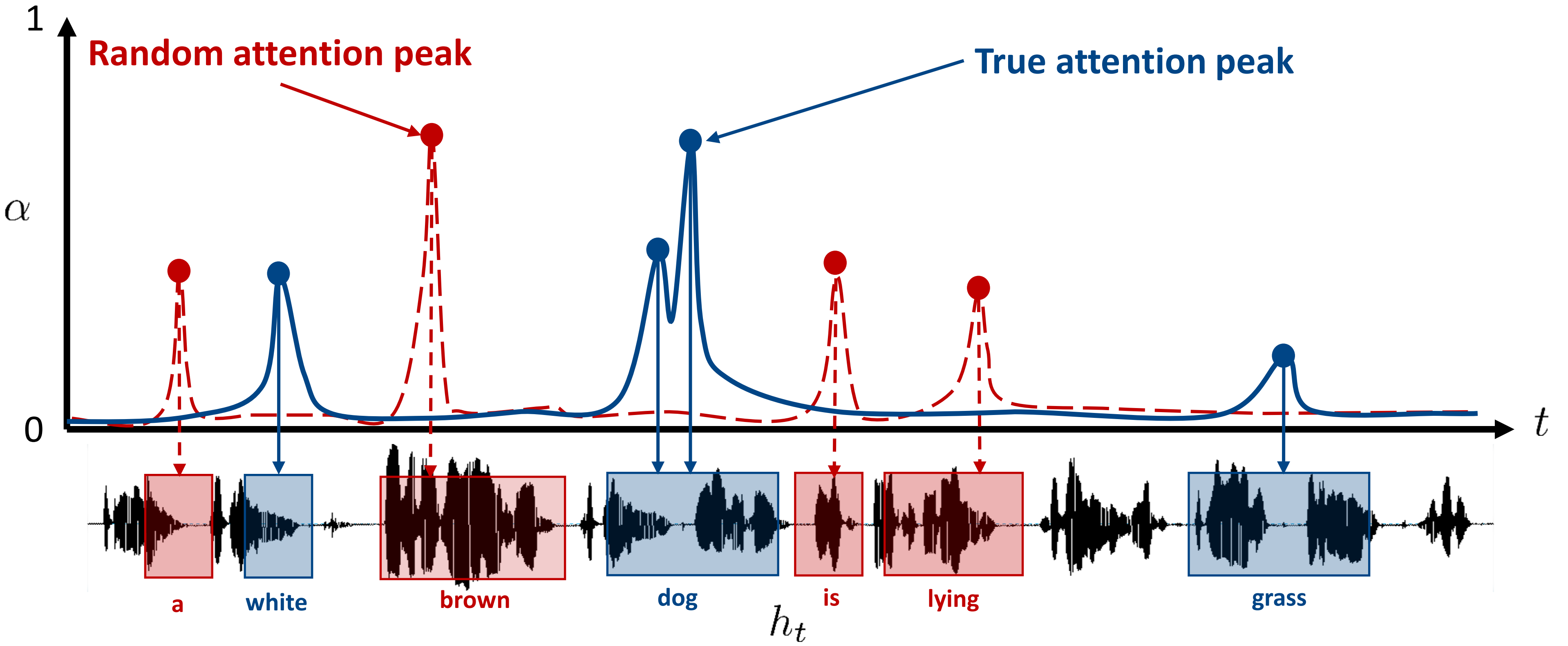
- Where α_t is the **attention weight** for the t^{th} vector h
- **Trainable component**: the network **learns** to assign a weight



Behaviour of Attention

$$c = \sum_{t=1}^{T} \alpha_t h_t$$

- Where α_t is the **attention weight** for the t^{th} vector h
- **True attention** v. **Random attention**

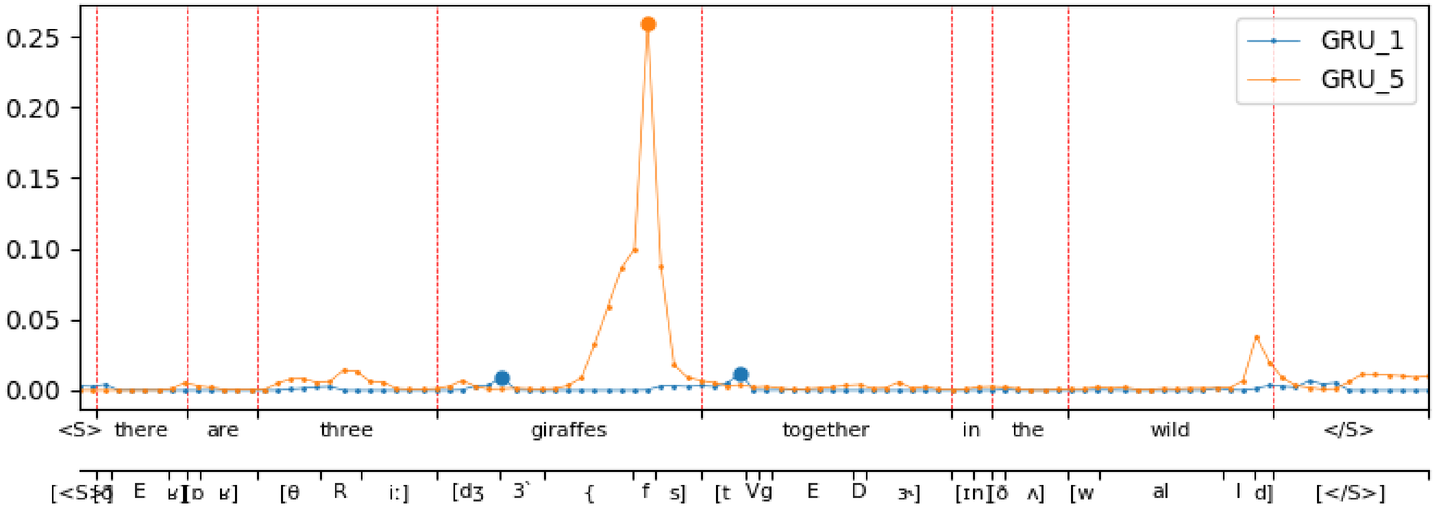


Main Task Results

- Recall@1
  - Evaluates model's ability to rank the target paired image as the **top 1 image**

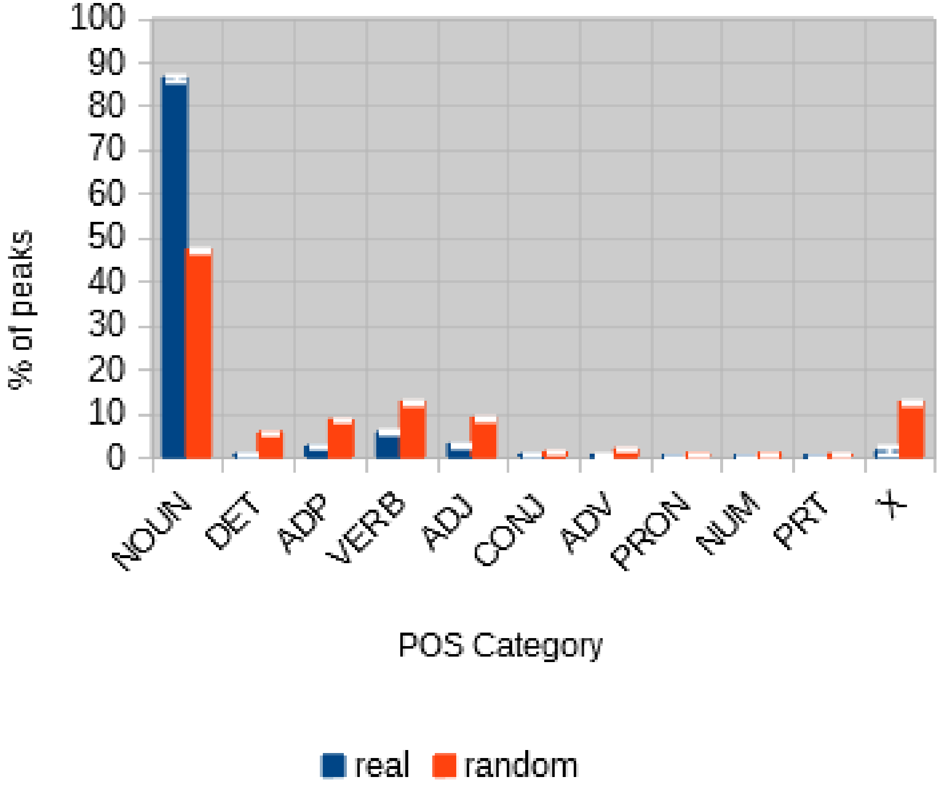
	Model	R@1	R@5	R@10	$\widetilde{r}$
GRU	English (ours)	5.52 ± 0.31	18.26 ± 0.82	28.64 ± 1.15	27.4 ± 1.51
GRU	Japanese (ours)	5.3 ± 0.16	17.88 ± 0.41	27.92 ± 0.36	29.2 ± 0.44
RHN	English (Chrupała, 2017)	11.1	31.0	44.4	13
	Random	0.02	0.1	0.2	2500.5

Behaviour of Attention — English



Behaviour of Attention — Results English

Peak Distribution



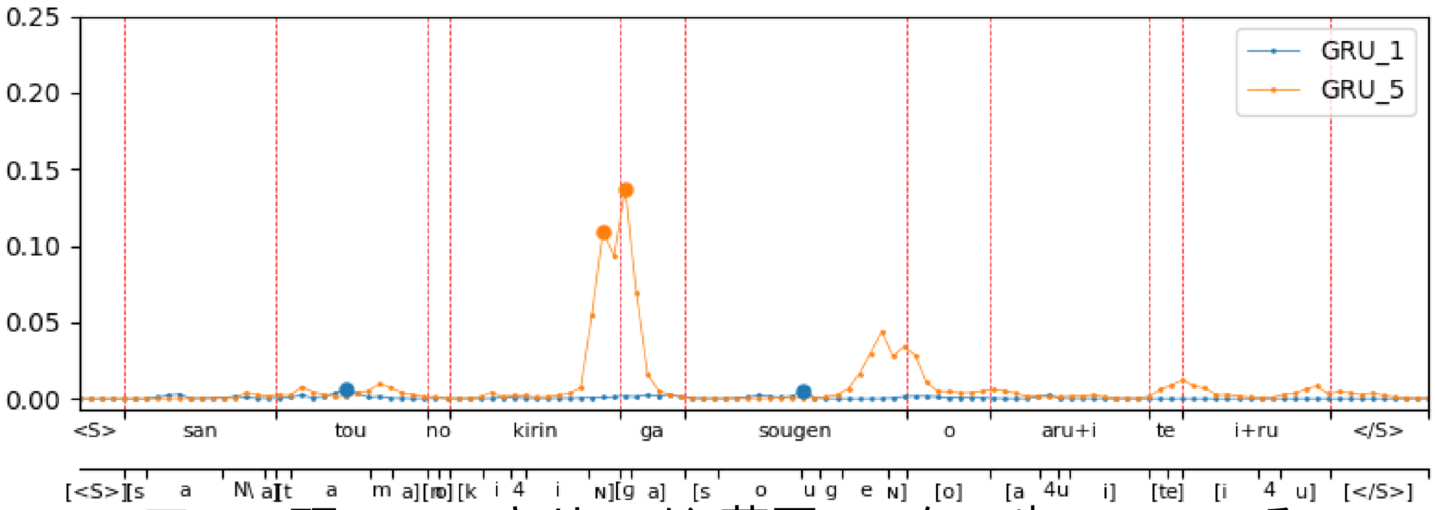
COCO

Top 10 Highlighted Words

Word Rank	TRUE Peaks		RANDOM Peaks	
	Word	% peak	Word	% Peak
1	train	2.04	</s>	9.15
2	tennis	1.73	a	3.47
3	toilet	1.53	<s>	2.47
4	baseball	1.50	on	1.96
5	skateboard	1.46	with	1.26
6	dog	1.45	in	1.22
7	cat	1.44	of	1.18
8	giraffe	1.39	man	1.08
9	pizza	1.35	and	1.04
10	kitchen	1.35	standing	1.02

$\neq$

Behaviour of Attention — Japanese



三　頭　の　キリン が 草原　を　歩い　て　いる

three　CLF　GEN giraffe SUJ meadow OBJ　exist　PROG

Behaviour of Attention — Japanese Particles

三	頭	の	キリン	が	草原	を	歩い	ている
san	tô	no	kirin	ga	sôgen	wo	arui	teiru
three	CLF	GEN	giraffe	SUJ	meadow	OBJ	be	PROG

subject
of the sentence

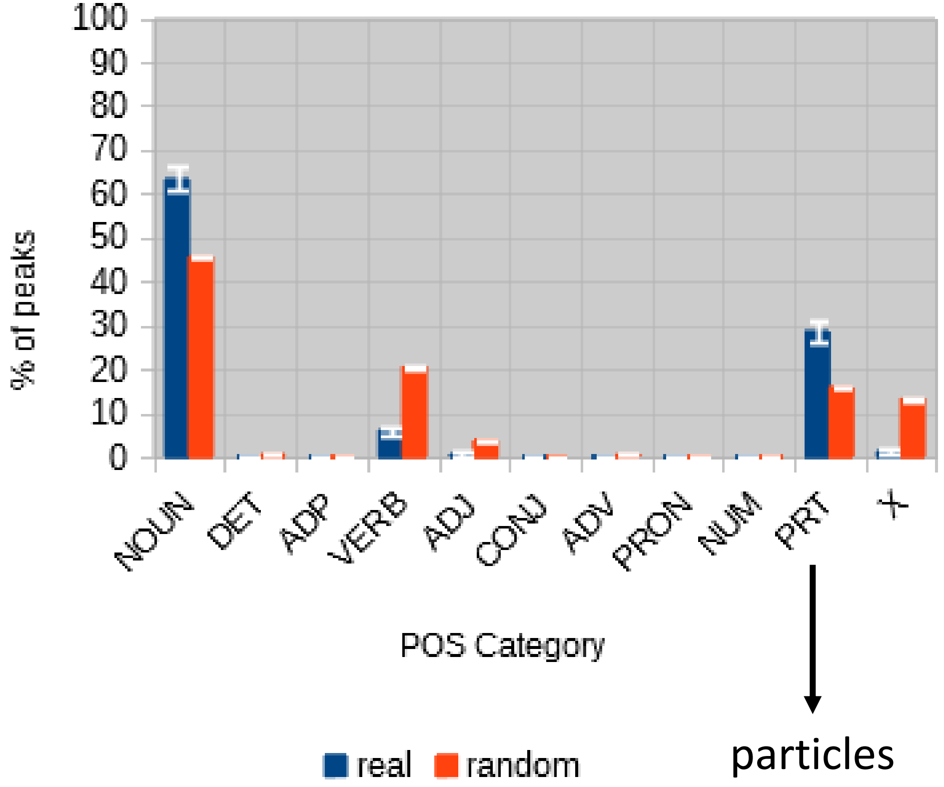
object
of the sentence

- Particles
  - **Grammatical words**: indicate the **function of the *preceding* word** in the sentence

- Main particles
  - **が (ga): subject**
  - **を (wo): object**
  - **は (ha): topic**
  - **の (no): genetive**

Behaviour of Attention — Results Japanese

Peak Distribution



particles

STAIR

Top 10 Highlighted Words

Word Rank	TRUE Peaks			RANDOM Peaks		
	Word	Translation	% peak	Word	Translation	% Peak
1	ga	SUBJ	13.18	</s>	*END OF SENTENCE*	9.97
2	no	GEN	5.35	i+ru	to be	5.65
3	o	OBJ	3.51	no	GEN	4.00
4	ni	LOC, ALL	3.34	ga	SUBJ	3.45
5	</s>	END OF SENTENCE	1.67	ni	LOC, ALL	2.55
6	kirin	giraffe	1.61	<s>	*START OF SENTENCE*	2.47
7	inu	dog	1.39	o	OBJ	2.05
8	uma	horse	1.27	dansei	man	1.79
9	baiku	bike	1.18	te	particle of reason, state	1.64
10	sukeetoboodo	skateboard	1.11	a+ru	to be	0.95

≠

Behaviour of Attention — Conclusion

- Models are **language agnostic**
  - Work equally well when trained on **English and Japanese** data
- Models use attention to focus on **specific words** and adopt...
  - a **language-general behaviour**: focus on **nouns**
  - a **language-specific behaviour**: focus on **particles**
- Reminds us of known psycholinguistic phenomena
  - **Noun bias** [Gentner, 1982]
  - **"ga" particle** [Haryu, 2016]

  "by 15 months of age, Japanese-learning infants become able to **use** the frequent particle ***ga*** **to segment adjacent nouns** from fluent speech"

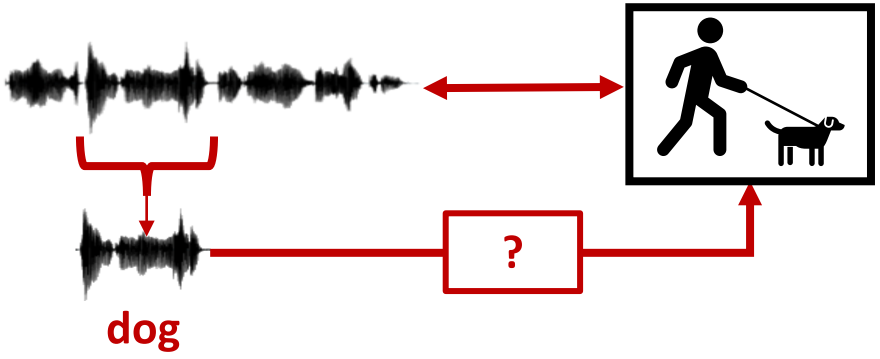
Word Recognition, Competition, and Activation

Word Recognition, Competition, and Activation

- Models are able to detect specific words …
  - Are they able to **map them to their visual referent?**
  - How do they **activate the semantic representation** of a word?

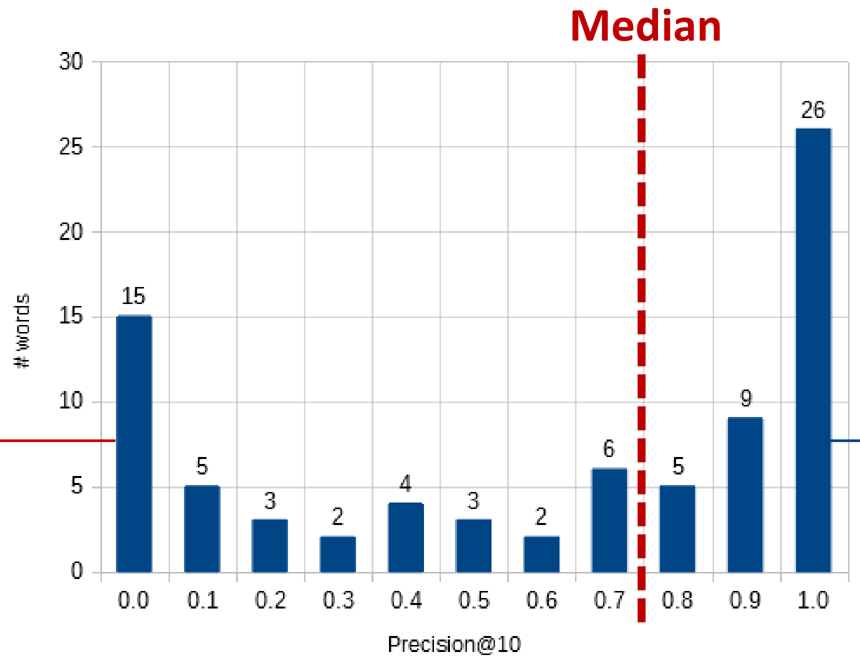
Isolated Word Recognition

- Input the network with **isolated words instead of full captions**
- If the network retrieves images with the target objet
  - **proof of (implicit) segmentation during training**
  - **correct word/object mapping**



- Experiment on **80 target words**
  - **P@10:** is there **at least one image that features the target object** among the 10 first?

bench
couch
tv
hot dog



zebra
truck
train
bicycle

WORD ACTIVATION

- **Gating Paradigm**

  "The gating paradigm involves the **repeated presentation of a spoken stimulus** (in this case, a word) **such that its duration from onset is increased with each successive presentation**."
  [Cotton & Grosjean, 1984]

→ **Measure** how **word activation** is carried out by the network

Word Activation

dʒɝæf

Left to Right

- Measure the importance of the **word's *onset***

dʒɝæf

- To **activate the semantic representation** of a word, what information is …
  - ***necessary?***
  - ***sufficient?***

Right to Left

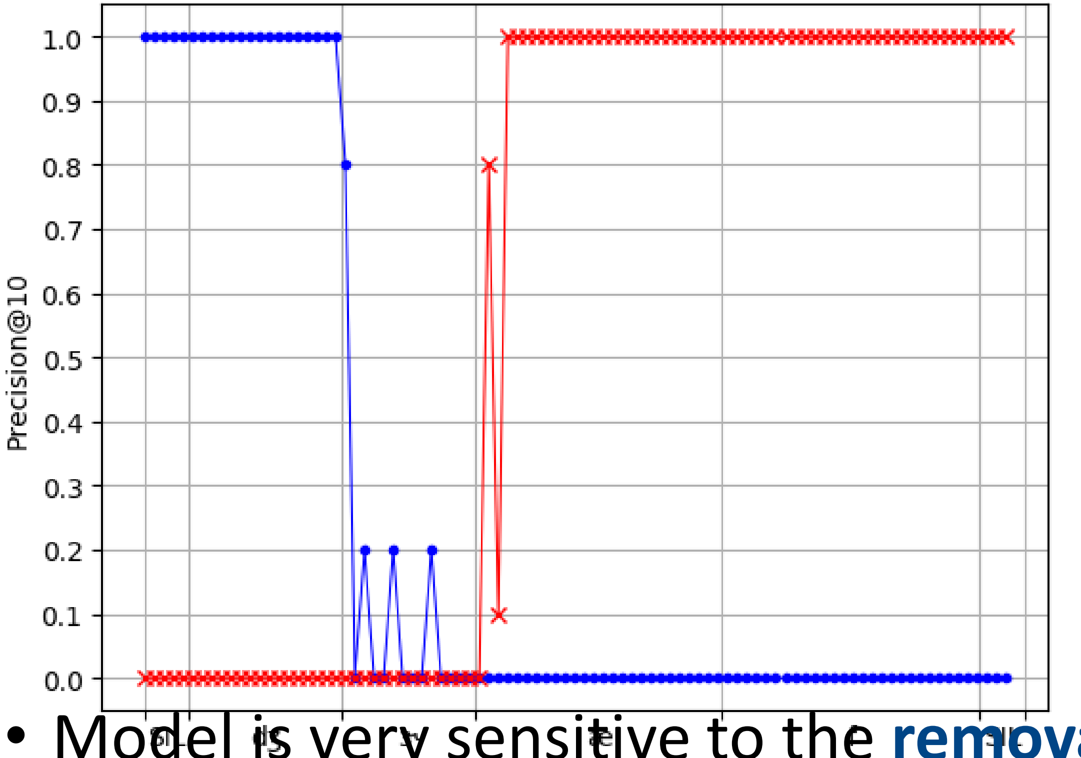
dʒɝæf

- Measure the importance of the **word's *offset***

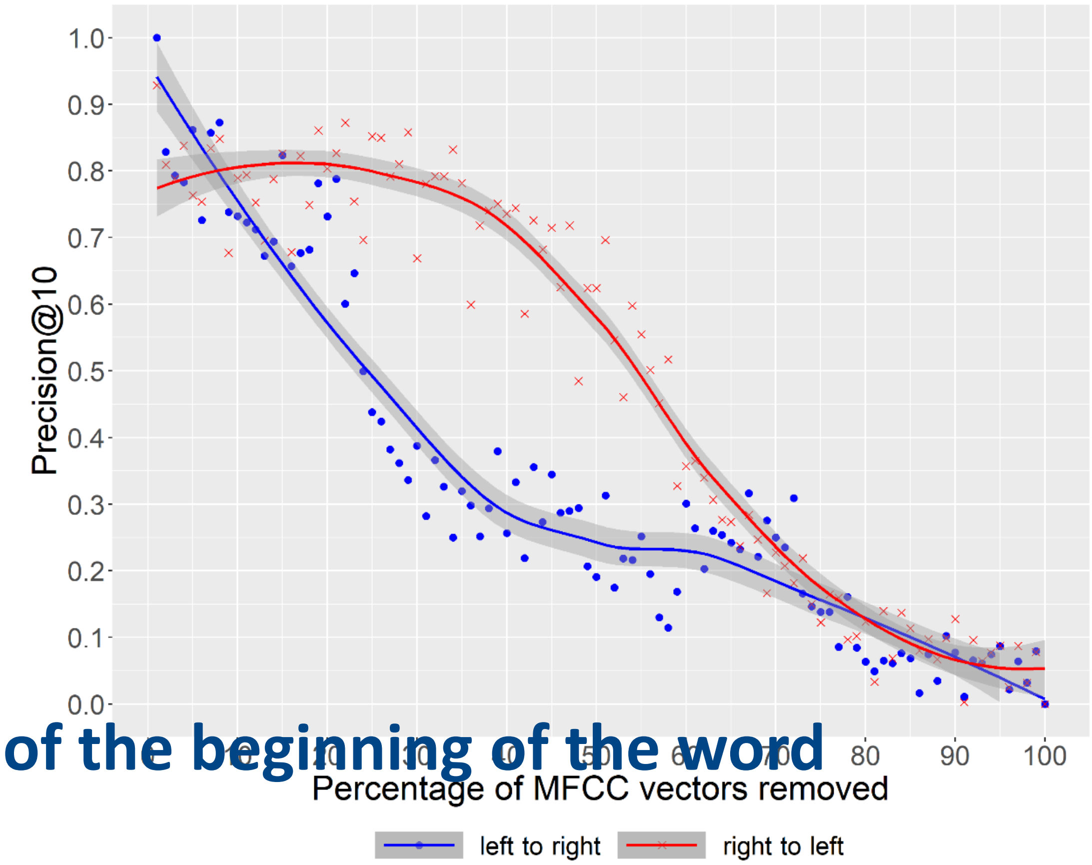
WORD ACTIVATION

dʒ ɜ˞ æ f

giraffe



Average over the 80 target words



- Model is very sensitive to the **removal of the beginning of the word**

dʒ ɜ˞ æ f

- Model is robust when **the end of the word is removed**

Word Activation: Conclusion

- Activation requires **access to the beginning of the word**
  - **COHORT**-like activation [Marslen-Wilson et al., 1978]

- Word recognition can occur from **a partial input (before its offset)**
  - Also happens in **human word recognition**

Introducing Prior Linguistic Information

Introducing Prior Linguistic Information

- Models trained on **full unsegmented captions**
  - Detect the relevant words
  - Map/recognised

→ **How efficient would the network be with segmented captions ?**

→ The network would "only" have to learn a better mapping

→ **Which segmentation would help the network best?**

Segment Boundaries

Example: *This is an article* [ðɪs # ɪz # ən # ɑɹtɪkəl]

- Phones

ð | ɪ | s | ɪ | z | ə | n | ɑ | ɹ | t | ɪ | k | ə | l

- Syllables-Connected (w/ resyllabification)

ðɪs| ɪz| ən| ɑɹ | tɪ | kəl

- Syllables-Word (w/o resyllabification)

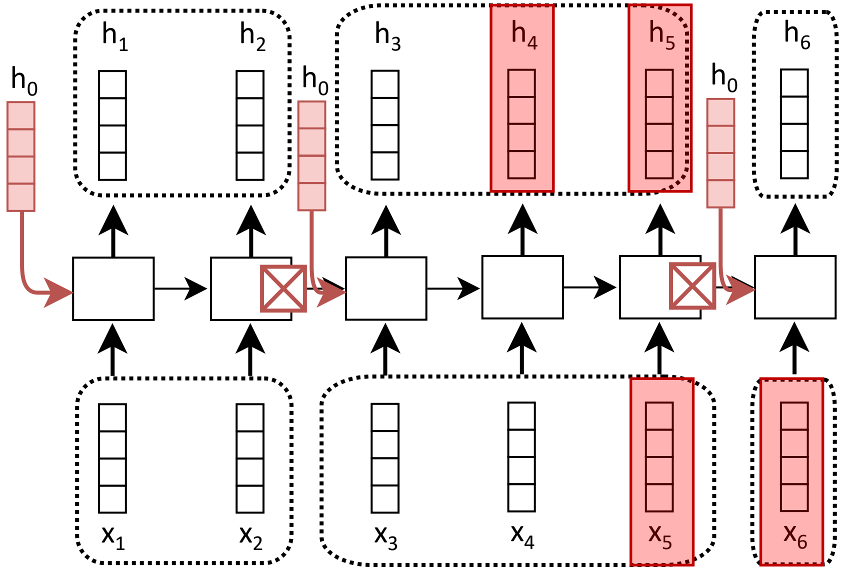
ðɪs | ɪz | ən | ɑɹ | tɪ | kəl

- Word

ðɪs | ɪz | ən | ɑɹtɪkəl

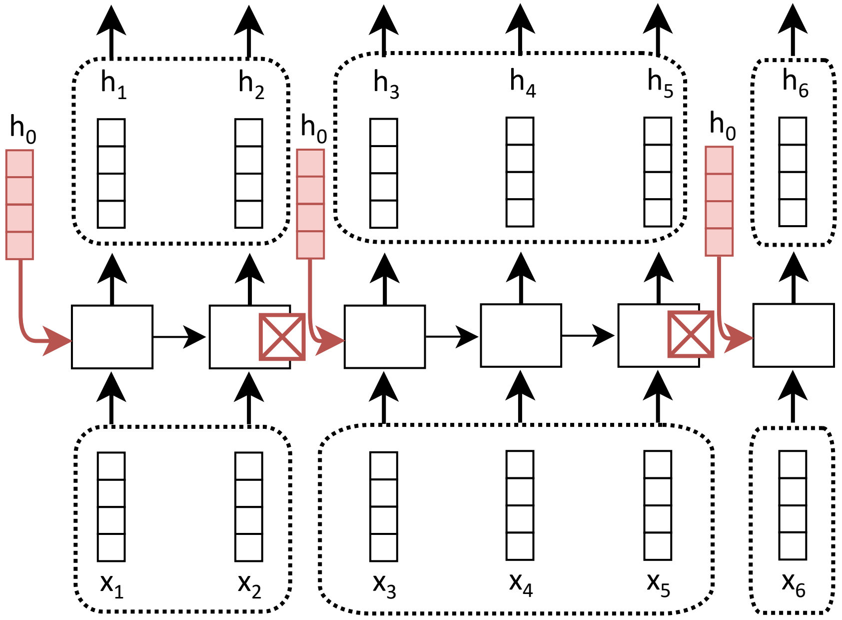
GRU Packager

- How to **introduce** segment boundary information?
  - Know where a segment begins/ends
  - Pass **h_0 instead of h_n**
  - Vectors belonging to the **same segment**: **temporally dependant**
  - Vectors belonging to a **different segment**: **temporally *in*dependant**

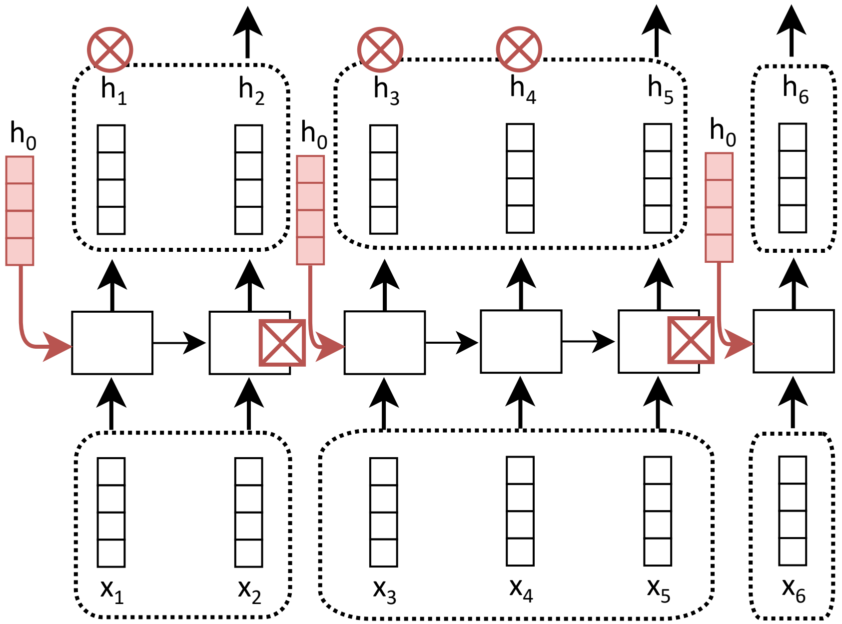


GRU Packager: All v. Keep

- ALL condition
  - **all the vectors** belonging to a segment are forwarded to the next layer
- KEEP condition
  - **only the last vector** belonging to a segment is forwarded to the next layer



ALL condition

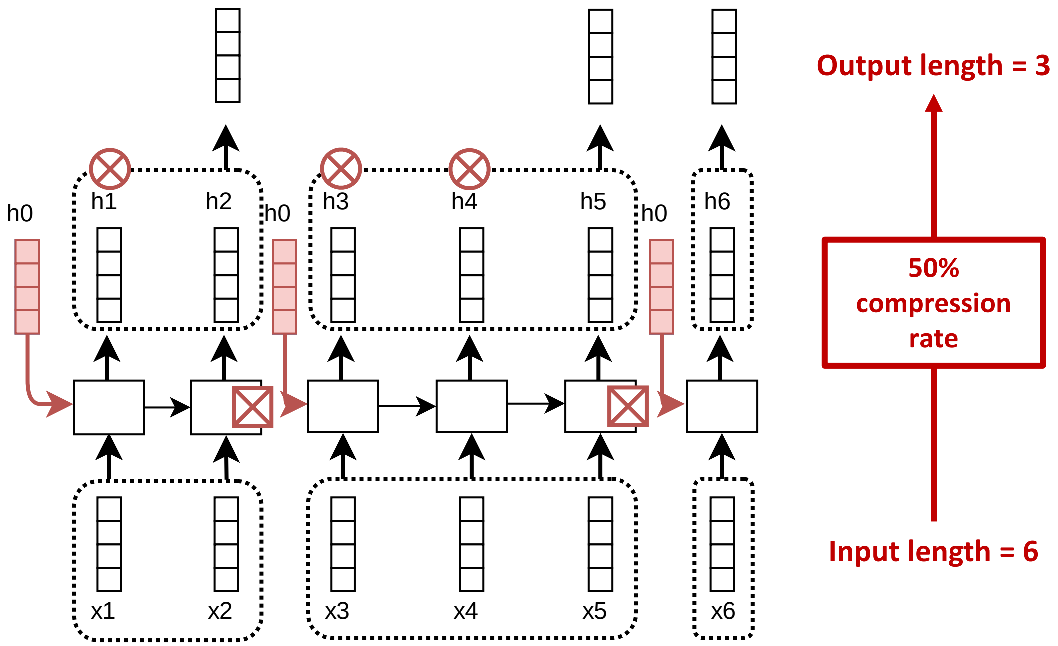


KEEP condition

GRU Packager: Keep

- KEEP condition : **compression**

- **Compression/sub-sampling rates**
  - phones 90.5%
  - syllables-connected 93.4%
  - syllables-word 94.3%
  - words 95%

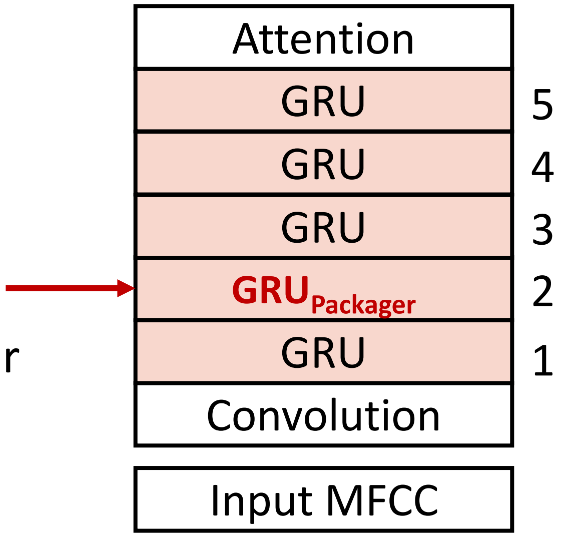


KEEP condition

EXPERIMENTAL CONDITIONS

- Use **random boundaries** as a control condition
  - sample as many random boundaries as true boundaries

- Experiments:
  - Vary the **type of boundaries** used (phones, syllables, and words)
  - Use either **TRUE or RANDOM** boundaries
  - Vary the **position** of the $\mathrm{GRU}_{\mathrm{Packager}}$ layer (from layer 1 to 5)
    - Instead of having 5 vanilla GRU layers, change one of them for a $\mathrm{GRU}_{\mathrm{Packager}}$ layer
  - Vary the type of $\mathrm{GRU}_{\mathrm{Packager}}$: **ALL or KEEP**



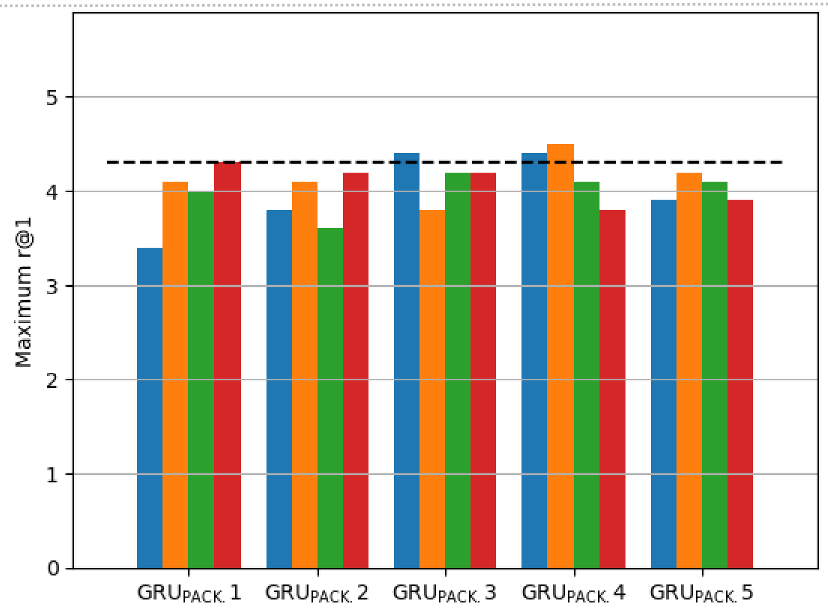
→ **80 models in total**

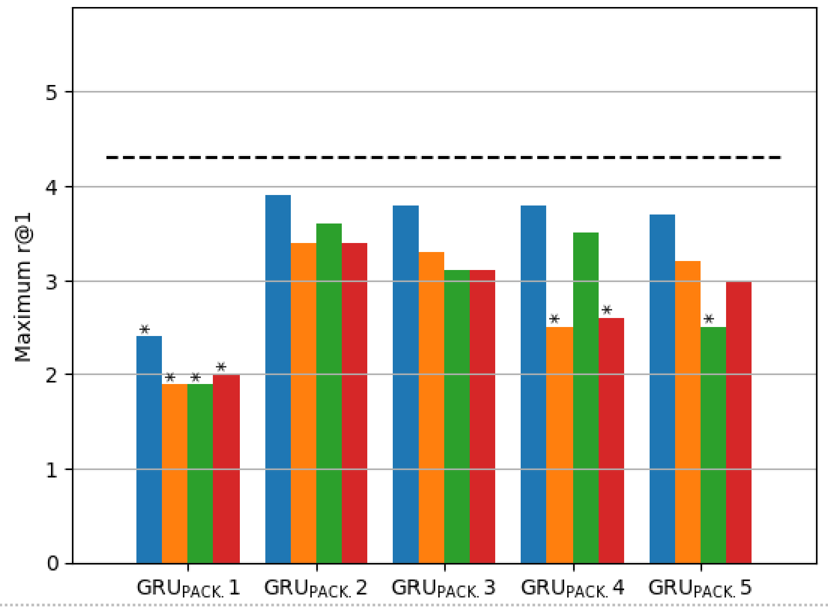
+ BASELINE model (no boundary used, 5 vanilla GRU layers)

Results

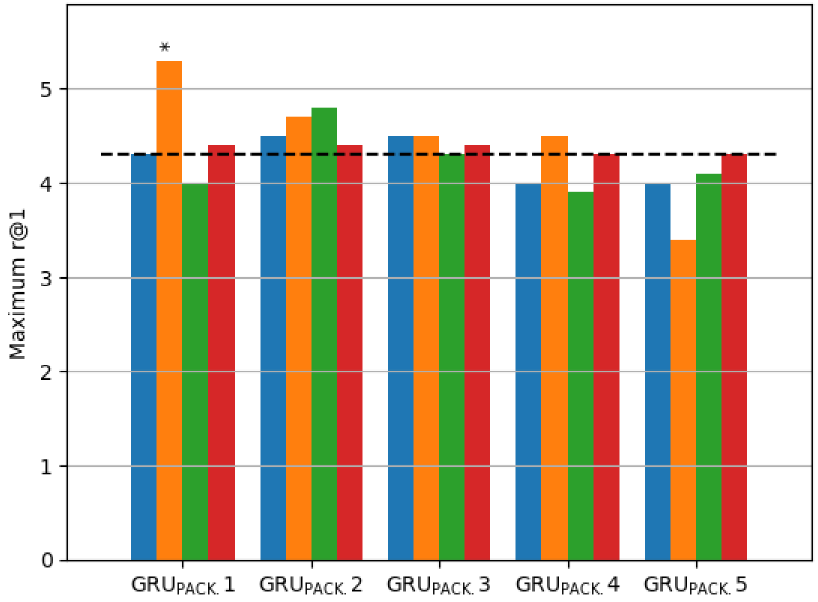
ALL Condition | KEEP Condition

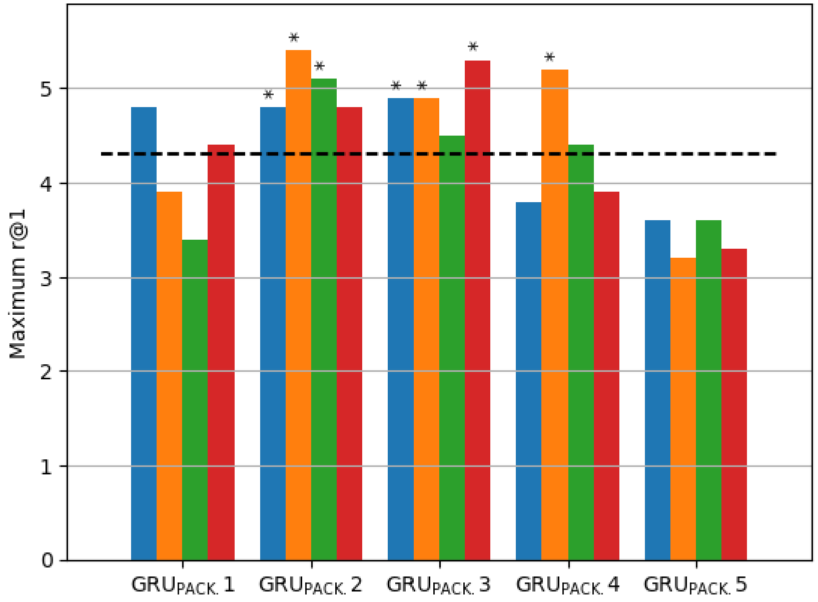
Random Boundaries





True Boundaries





Phones

Words

Syllables Connected

Syllables Word

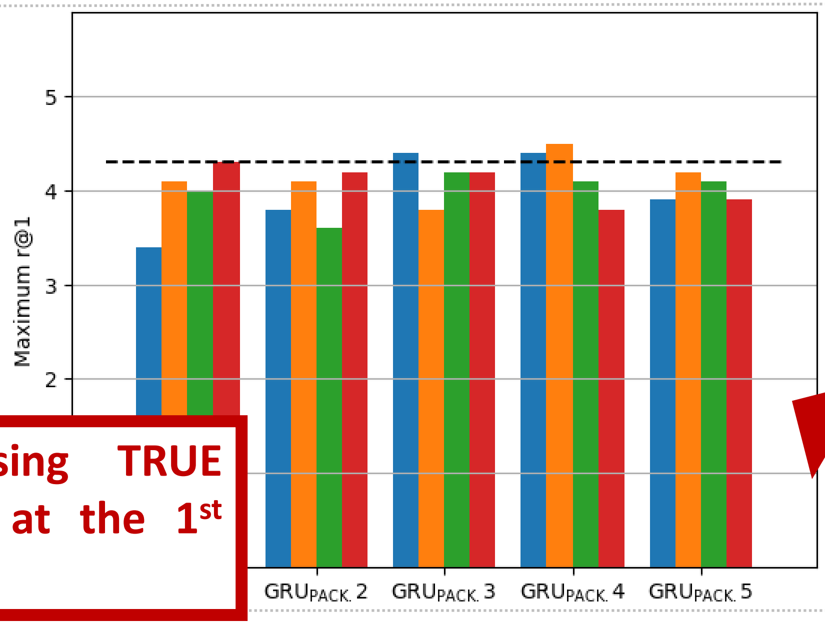
Baseline Result

* significativity (Z-Test, p<0.01)

Results

ALL Condition | **KEEP Condition**

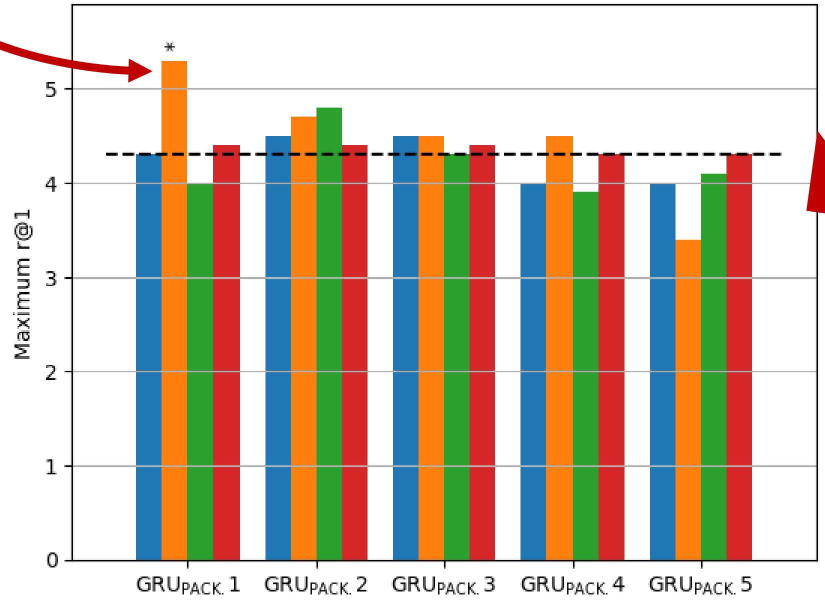
Random Boundaries

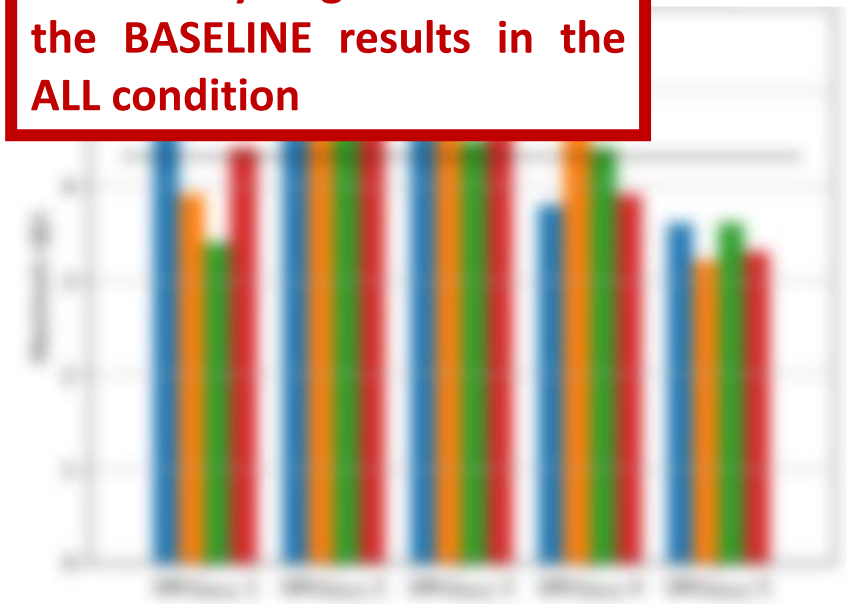


Except when using TRUE word boundaries at the 1st layer

Difference between RANDOM and TRUE overall not statistically significant from the BASELINE results in the ALL condition

True Boundaries





Phones

Words

Syllables Connected

Syllables Word

Baseline Result

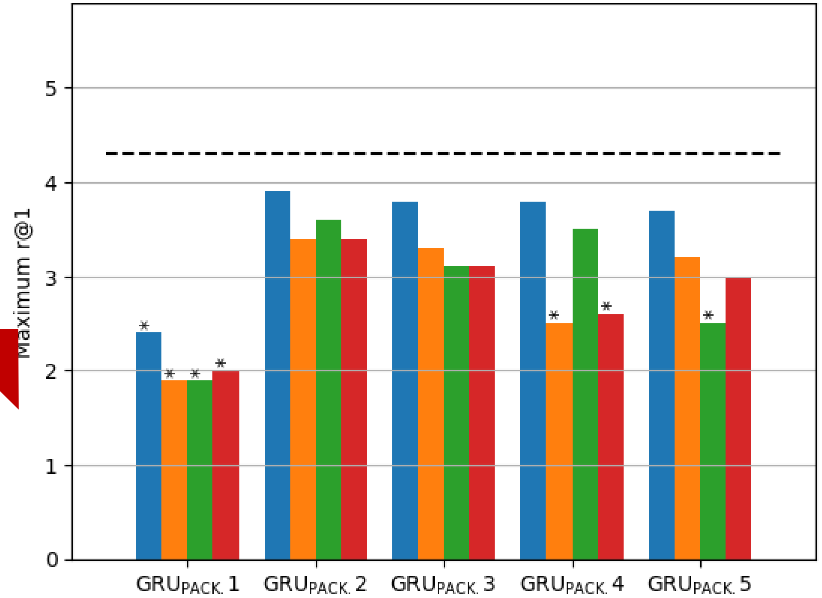
* significativity (Z-Test, p<0.01)

RESULTS

ALL Condition

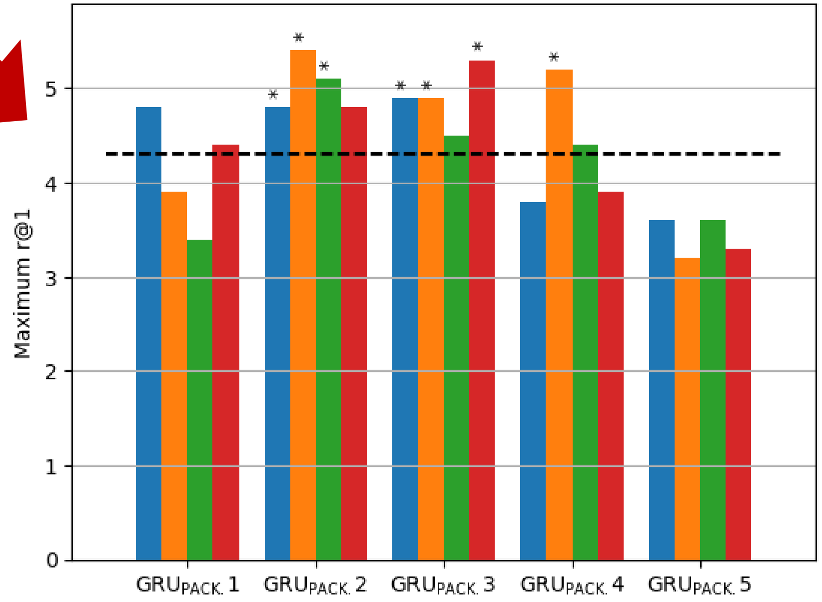
KEEP Condition

Random Boundaries



Large difference between RANDOM and TRUE in the KEEP condition

True Boundaries



Phones

Words

Syllables Connected

Syllables Word

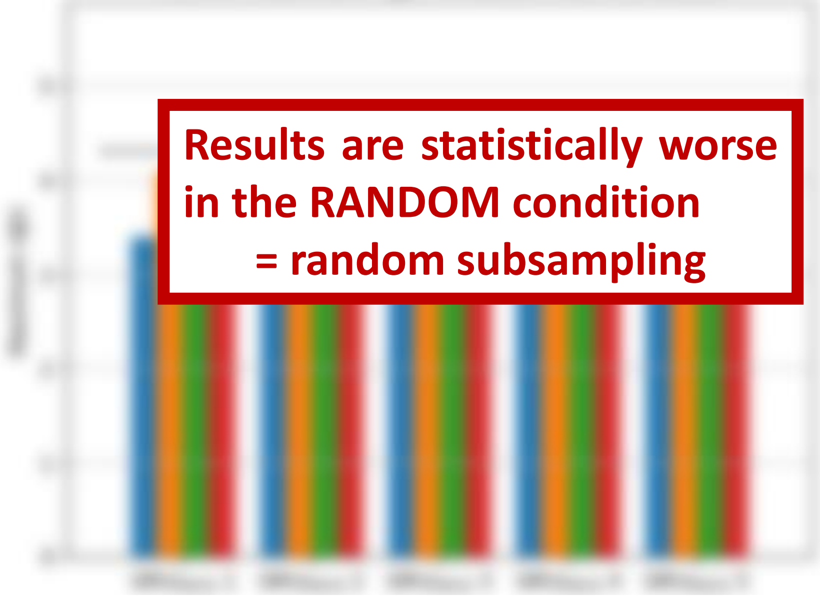
Baseline Result

* significativity (Z-Test, p<0.01)

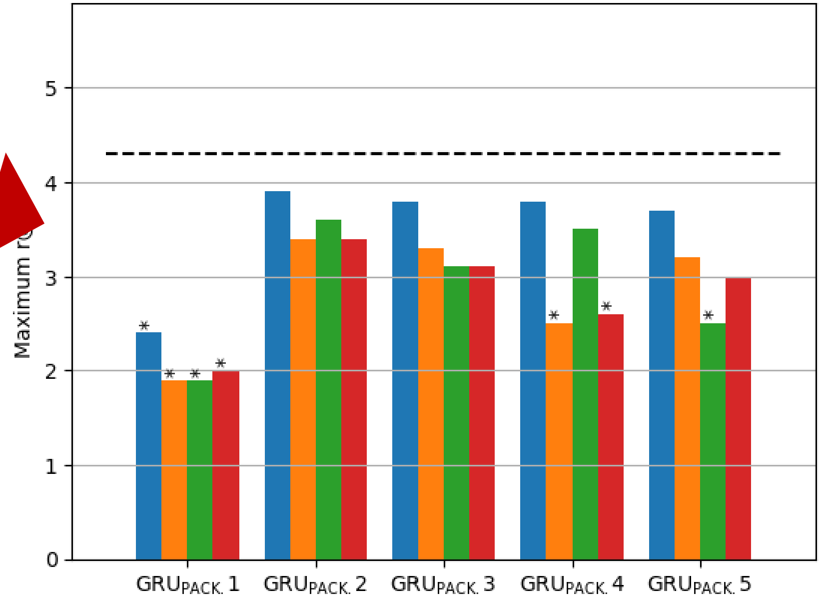
Results

ALL Condition | **KEEP Condition**

Random Boundaries

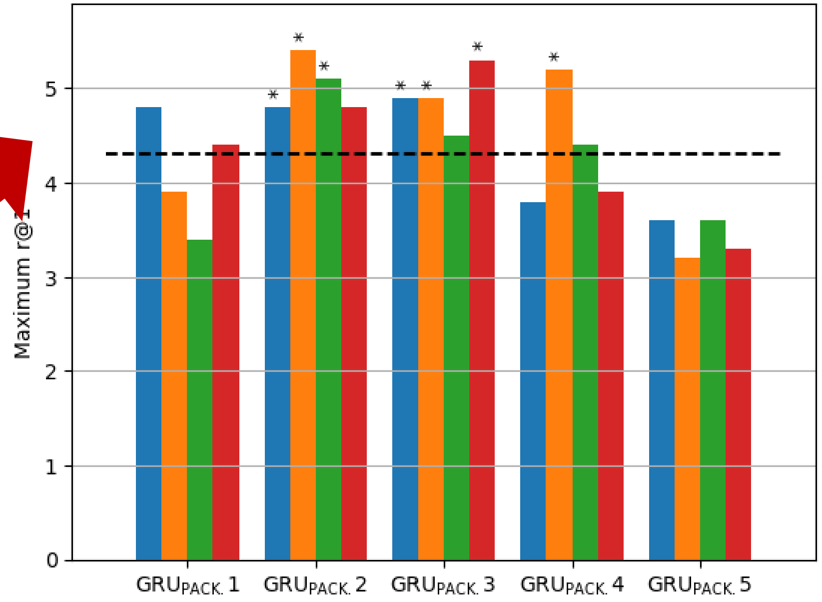


Results are statistically worse in the RANDOM condition = random subsampling



True Boundaries

Results are statistically better in the TRUE condition



Phones

Words

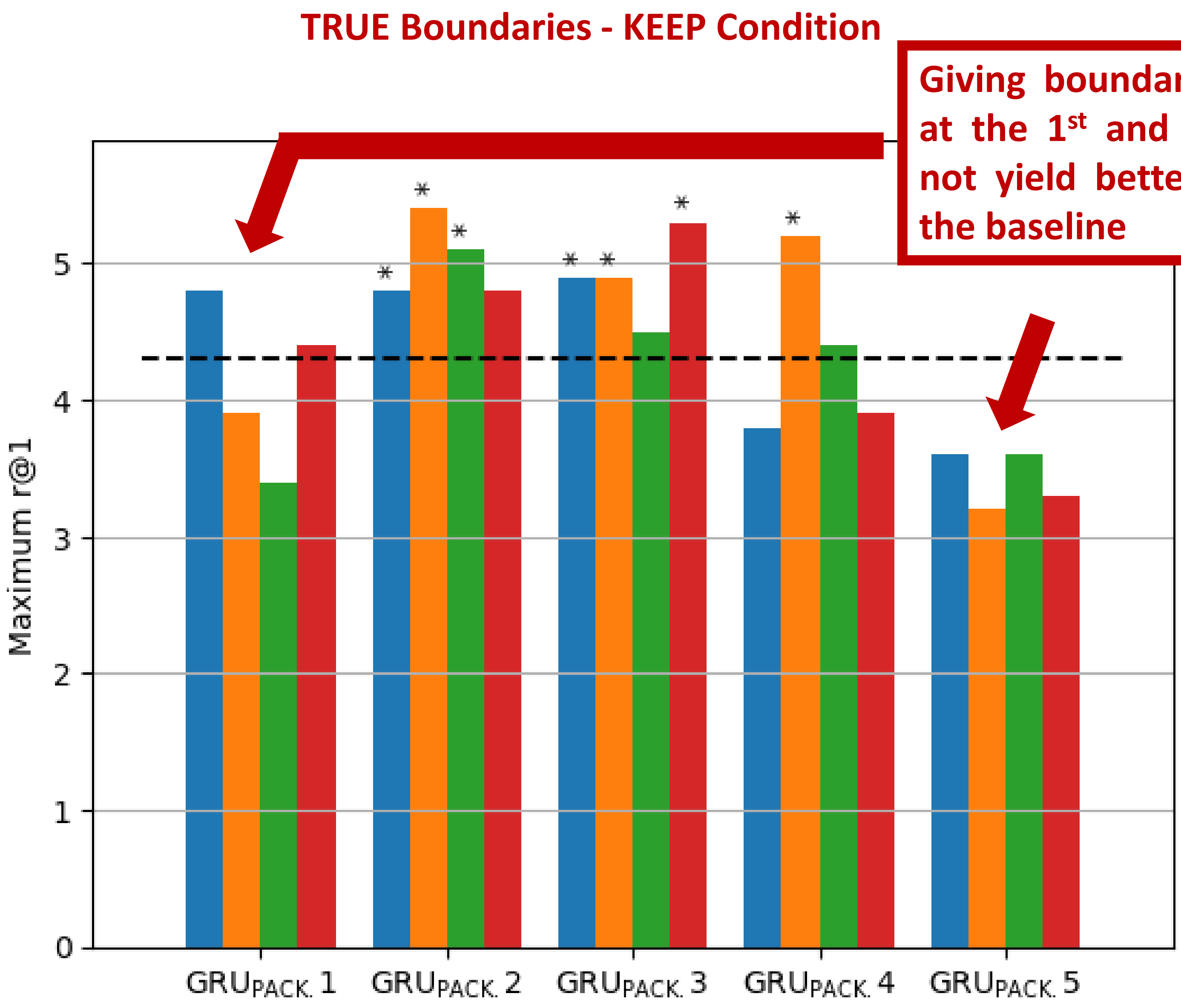
Syllables Connected

Syllables Word

Baseline Result

* = significative (Z-Test, p<0.01)

RESULTS



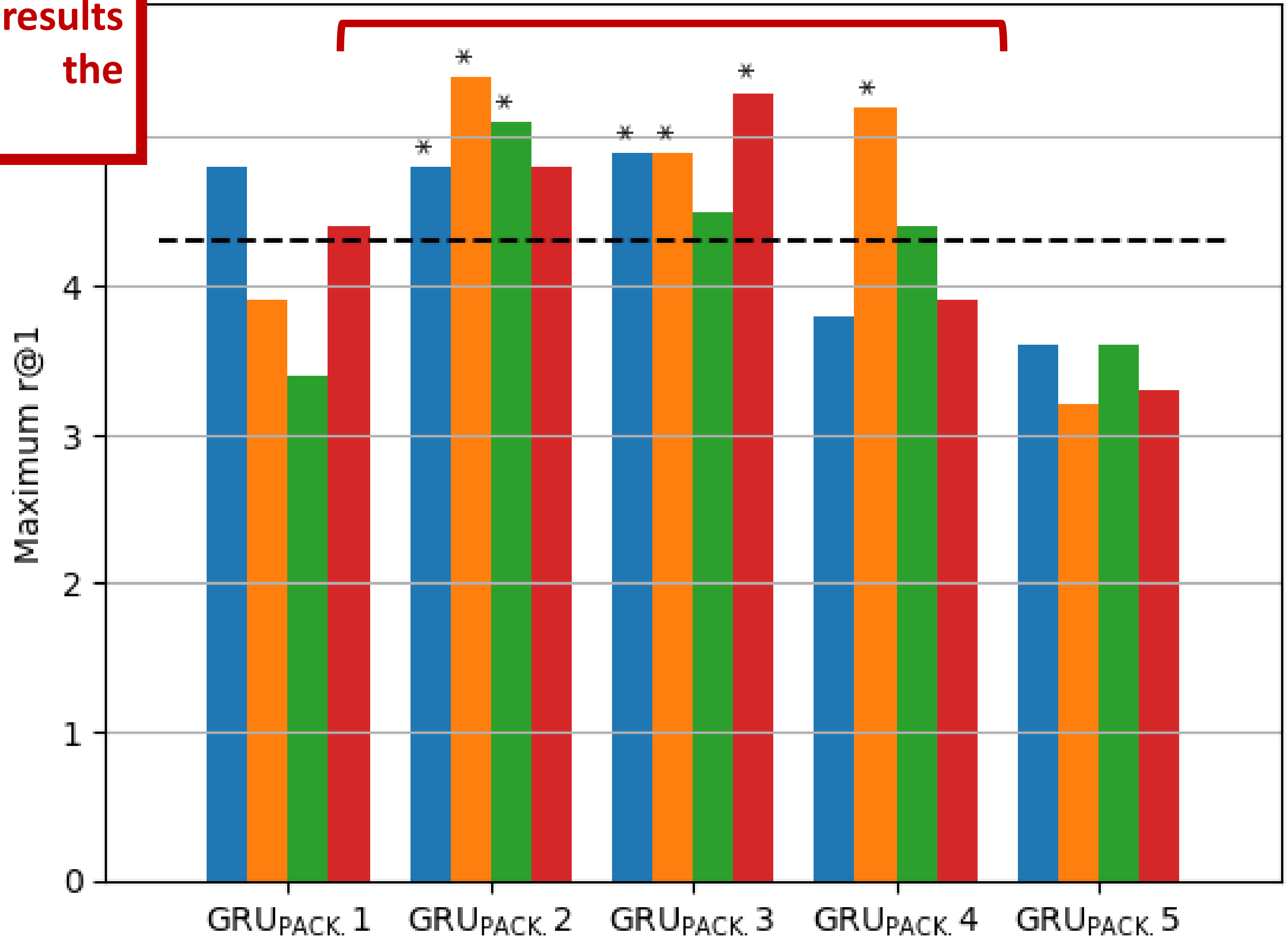
Giving boundary information at the 1st and 5th layer does not yield better results than the baseline

* significativity
(Z-Test, p<0.01)

RESULTS

TRUE Boundaries - KEEP Condition

Boundaries only yield significantly better results when used by the intermediate layers



Phones

Words

Syllables Connected

Syllables Word

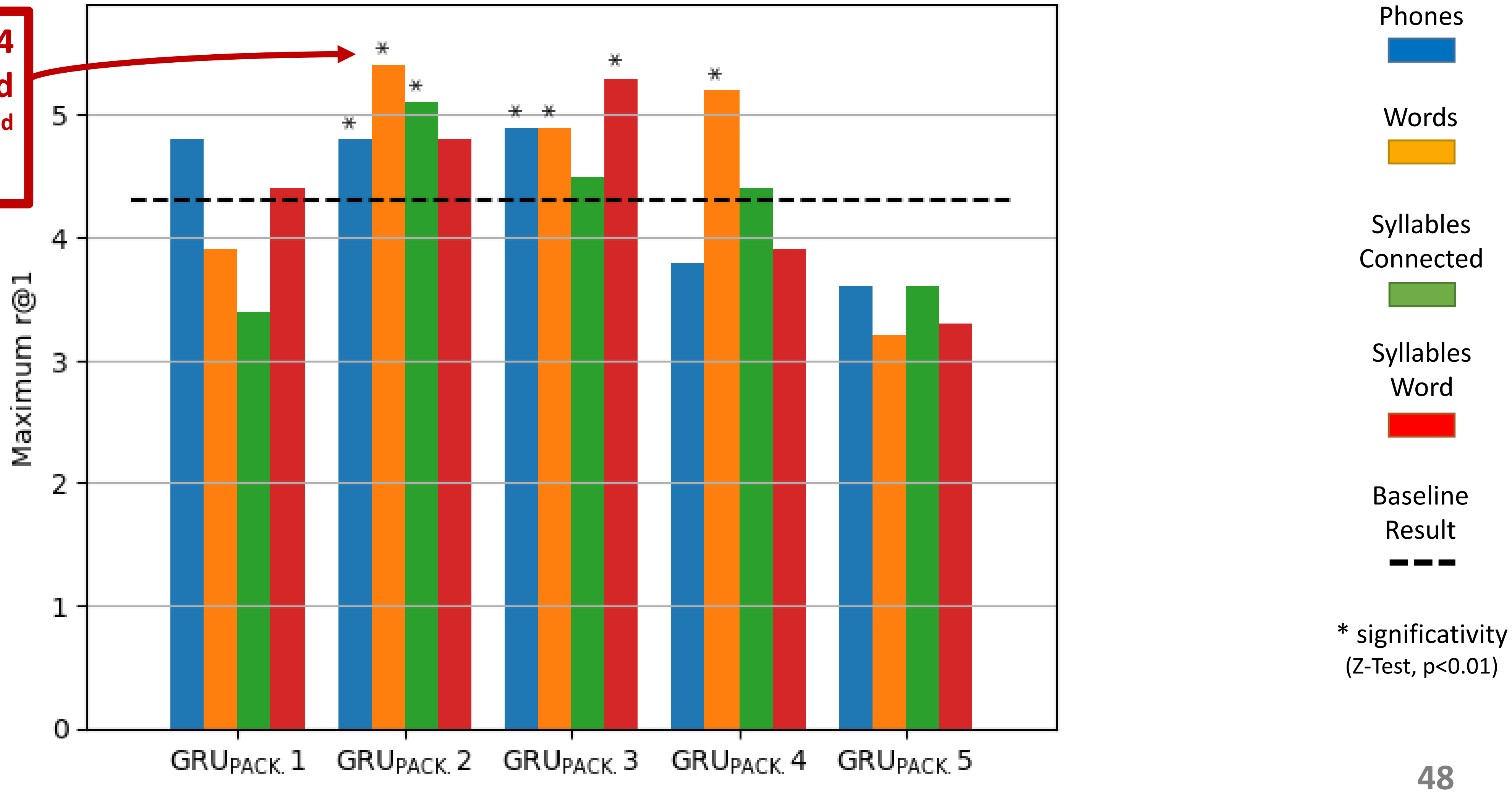
Baseline Result

* significativity (Z-Test, p<0.01)

Results

Units that preserve word boundaries (syllables-word and words) obtain the best results across all three middle layers.

Best R@1 = 5.4 when using word boundaries at 2nd layer



CONCLUSION

- Is segmenting speech into sub-units beneficial?
  - **Yes! + 1.1pp over the baseline**
  - **Large units** that preserve **word boundaries** yield the best results
- **Introducing hierarchy yields even better results!**
  - **+3.9pp** over the baseline architecture when using **2 $\text{GRU}_{\text{Packager}}$**
  - **+5.3pp** over the baseline architecture when using **3 $\text{GRU}_{\text{Packager}}$**
- Strong **difference** between **ALL** and **KEEP**
  - KEEP enforces the network to learn **better representations**

CONCLUSION

Main Contributions of this Thesis

- **Synthetically Spoken STAIR** data set

- Analysis of **Attention in an RNN-based VGS models**
  - **Focus on nouns**
  - **Focus on particles**
  - **Quickly acquired behaviour**

- Analysis of **individual word knowledge and word/referent mapping**
  - Taking inspiration from **methodologies** stemming from the **psycholinguistics literature**
  - May occur from a **partial input**
  - Sensitive **to the presence/absence of words' onsets**

- Effect of the Incorporation of **Prior Linguistic Information**
  - **Models fare better when speech is explicitly segmented**
  - **Even if the input if strongly compressed/subsampled (KEEP condition)**

Regarding Language Acquisition

- "children learn the meanings of words through theory of mind. If this is right, then **a direct connectionist implementation of word learning, in which sounds are associated with percepts, is unfeasible**. (And this does preclude all connectionist theories of word learning that I'm aware of.)".

  [Bloom, 2002]

- Apparently **it is feasible… for a neural network**

- A purely associative learning mechanism *could* **bootstrap lexical acquisition** in children

Future Works

- Incorporate a **segmenting mechanism** into the network
  [Kreutzer, 2019; Shain, 2017; Shain 2020]
  - Similar patterns as child language acquisition?
  - What units are segmented?
- Work on **child language acquisition** data sets
  - SEEDlingS data set [Bergelson et al., 2017] or data set by [Tsutsui, 2020]

PERSONAL BIBLIOGRAPHY

Havard, W. N., Besacier, L. & Chevrot, J.-P. (2020), **Catplayinginthesnow: Impact of Prior Segmentation on a Model of Visually Grounded Speech**, in Proceedings of the 24th Conference on Computational Natural Language Learning (CoNLL), Association for Computational Linguistics, pp. 291--301. URL: https://www.aclweb.org/anthology/2020.conll-1.22

*Zanon Boito M., *Havard W. N., Garnerin M., Le Ferrand E & Besacier L. (2020), **MaSS: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the Bible**. In Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC2020), Marseille, France. European Language Resources Association (ELRA), pp. 6486--6493. URL: https://www.aclweb.org/anthology/2020.lrec-1.799

Havard, W. N., Chevrot, J.-P. & Besacier, L. (2019), **Models of Visually Grounded Speech Signal Pay Attention to Nouns: A bilingual Experiment on English and Japanese**, in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8618--8622. URL: https://doi.org/10.1109/icassp.2019.8683069

Havard, W. N., Chevrot, J.-P. & Besacier, L. (2019), **Word Recognition, Competition, and Activation in a Model of Visually Grounded Speech**, in Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Association for Computational Linguistics, Hong Kong, China, pp. 339--348. URL: https://www.aclweb.org/anthology/K19-1032

He, X., Tran, Q., Havard, W. N., Besacier, L., Zukerman, I. & Haffari, G. (2018) **Exploring textual and speech information in dialogue act classification with speaker domain adaptation**. In `Proceedings of the Australasian Language Technology Association Workshop 2018', Dunedin, New Zealand, pp. 61--65. URL: https://www.aclweb.org/anthology/U18-1007/

Havard, W. N., Besacier, L. & Rosec, O. (2017), SPEECH-COCO: 600k Visually Grounded Spoken Captions Aligned to MSCOCO data Set, in `Proceedings of the GLU 2017 International Workshop on Grounding Language Understanding', pp. 42--46. URL: http://dx.doi.org/10.21437/GLU.2017-9

Havard, W. N., Chevrot, J.-P., & Besacier L. (2018), `Emergence of attention in a neural model of visually grounded speech'. Learning Language in Humans and in Machines 2018 conference, ENS Paris, Poster (Peer Reviewed Abstract).URL: https://hal.archives-ouvertes.fr/hal-01970514/document

Thank You!

REFERENCES

Alishahi, A., Barking, M. & Chrupa la, G. (2017), Encoding of phonology in a recurrent neural model of grounded speech, in `Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)', Association for Computational Linguistics, Vancouver, Canada, pp. 368-378.

Andersen, E. S., Dunlea, A. & Kekelis, L. S. (1984), `Blind children's language: resolving some differences', Journal of Child Language 11(3), 645-664. Dunlea, A. (1989), Vision and the emergence of meaning: blind and sighted children's early language, Cambridge University Press, Cambridge [England] ; New York

Bergelson, E. & Aslin, R. N. (2017), `Nature and origins of the lexicon in 6-mo-olds', Proceedings of the National Academy of Sciences 114(49), 12916-12921.

Chrupala, G., Gelderloos, L. & Alishahi, A. (2017a), Representations of language in a model of visually grounded speech signal, in `Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)', Association for Computational Linguistics, pp. 613-622.

Drexler, J. & Glass, J. (2017), Analysis of Audio-Visual Features for Unsupervised Speech Recognition, in `Proc. GLU 2017 International Workshop on Grounding Language Understanding', pp. 57-61.

Gabriel, S., Maarten, V. & Dupoux, E. (2014), Learning Words from Images and Speech, in `NIPS Workshop on Learning Semantics'.

Gentner, D. (1982), `Why nouns are learned before verbs: Linguistic relativity versus natural partitioning', Language 2, 301-334

Harwath, D. & Glass, J. R. (2017), Learning Word-Like Units from Joint Audio-Visual Analysis, in R. Barzilay & M. Kan, eds, `Proceedings of the 55th Annual Meeting of the Association for Coputational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers', Association for Computational Linguistics, pp. 506-517.

Harwath, D. & Glass, J. R. (2019), Towards Visually Grounded Sub-word Speech Unit Discovery, in `IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019', Ieee, pp. 3017-3021.

Harwath, D., Chuang, G. & Glass, J. (2018), Vision as an Interlingua: Learning Multilingual Semantic Embeddings of Untranscribed Speech, in `2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 4969-4973.

Haryu, E. & Kajikawa, S. (2016), `Use of bound morphemes (noun particles) in word segmentation by Japanese-learning infants', Journal of Memory and Language 88(C), 18-27.

Havron, N., Raviv, L. & Arnon, I. (2018), `Literate and preliterate children show different learning patterns in an artificial language learning task', Journal of Cultural Cognitive Science 2(1-2), 21-33.

Kreutzer, J. & Sokolov, A. (2018), `Learning to Segment Inputs for NMT Favors Character-Level Processing', Proceedings of the International Workshop on SpokenLanguage Translation October 29-30, 2018 Bruges, Belgium 1, 166-172.

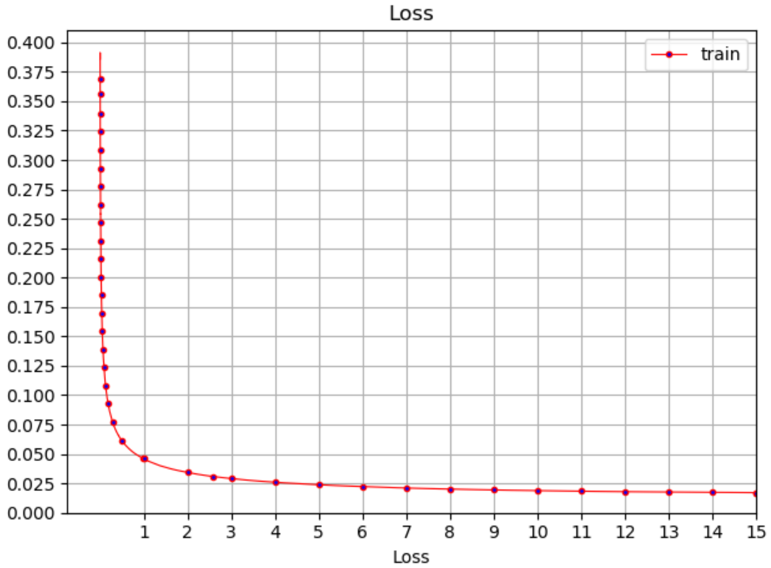
Landau, B. & Gleitman, L. (1985), Language and Experience: Evidence from the Blind Child, Cognitive Science Series, Harvard University Press

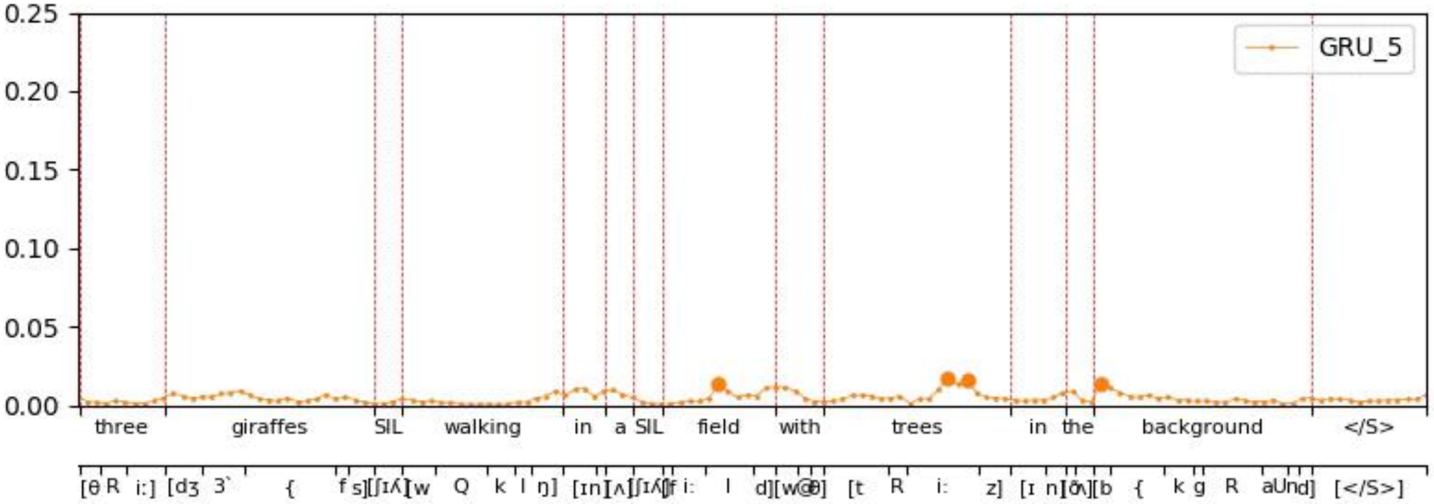
Merkx, D., Frank, S. L. & Ernestus, M. (2019), Language Learning Using Speech to Image Retrieval, in `Interspeech 2019, 20st Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019', pp. 1841-1845.

Shain, C. & Elsner, M. (2020), Acquiring language from speech by learning to remember and predict, in `Proceedings of the 24th Conference on Computational Natural Language Learning', Association for Computational Linguistics, Online, pp. 195-214.

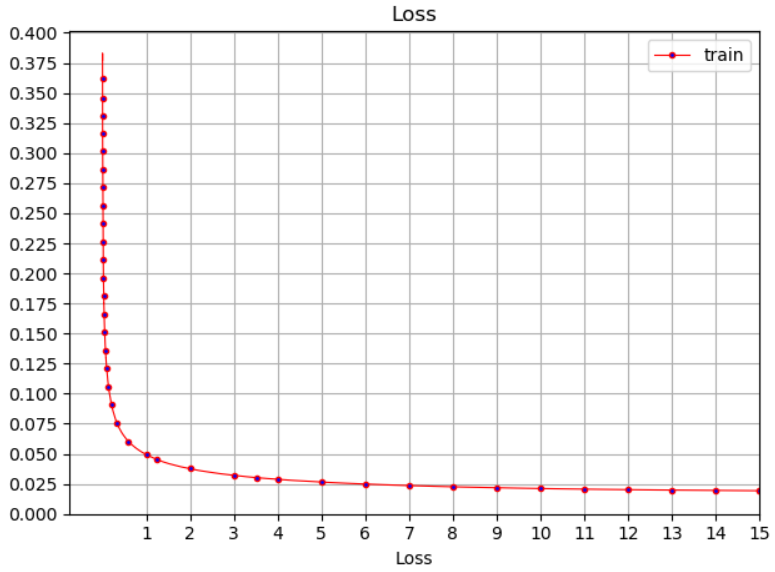
Tsutsui, S., Chandrasekaran, A., Reza, M. A., Crandall, D. J. & Yu, C. (2020), A computational Model of Early Word Learning from the Infant's Point of View, in `Proceedings of the 42th Annual Meeting of the Cognitive Science Society – Developing a Mind: Learning in Humans, Animals, and Machines, CogSci 2020, virtual, July 29 - August 1, 2020', cognitivesciencesociety.org.

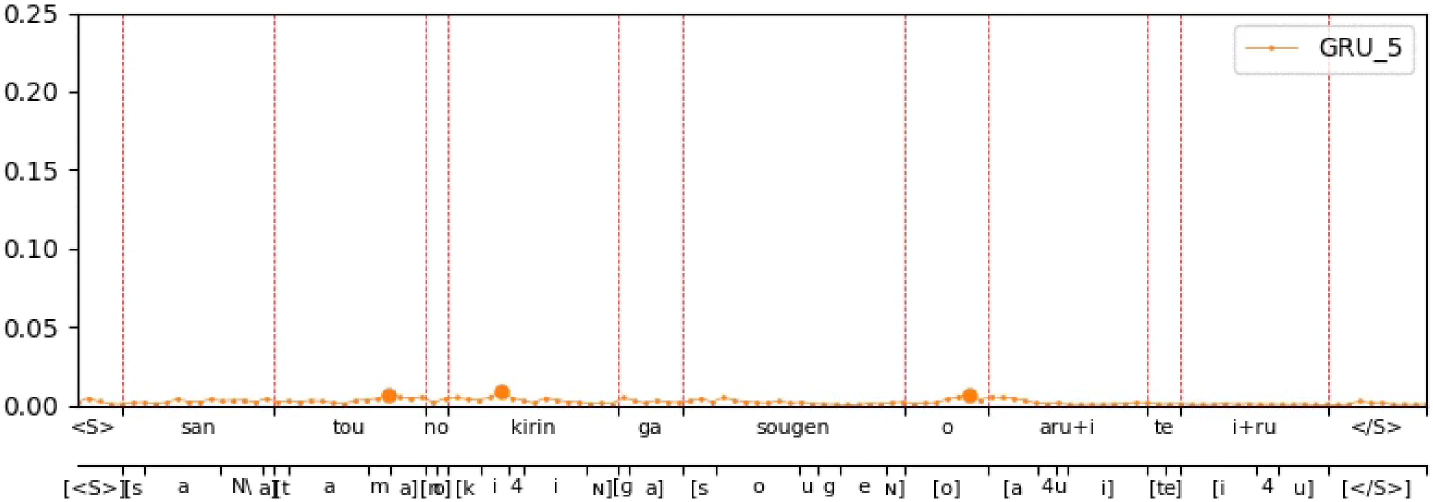
Behaviour of Attention — English





Behaviour of Attention — Japanese

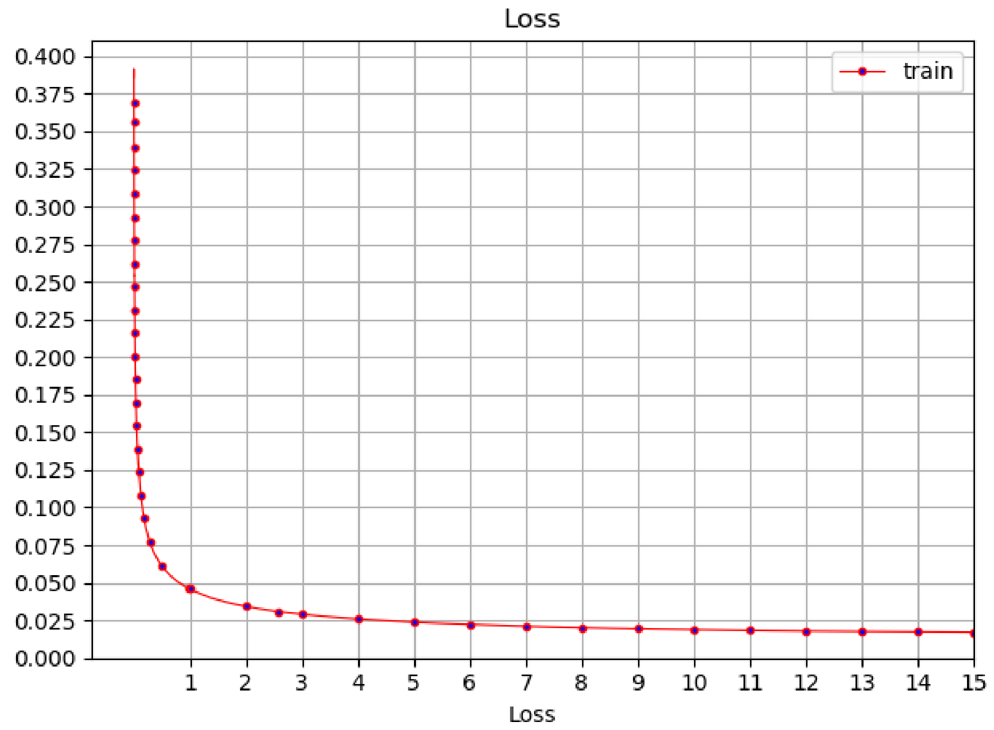


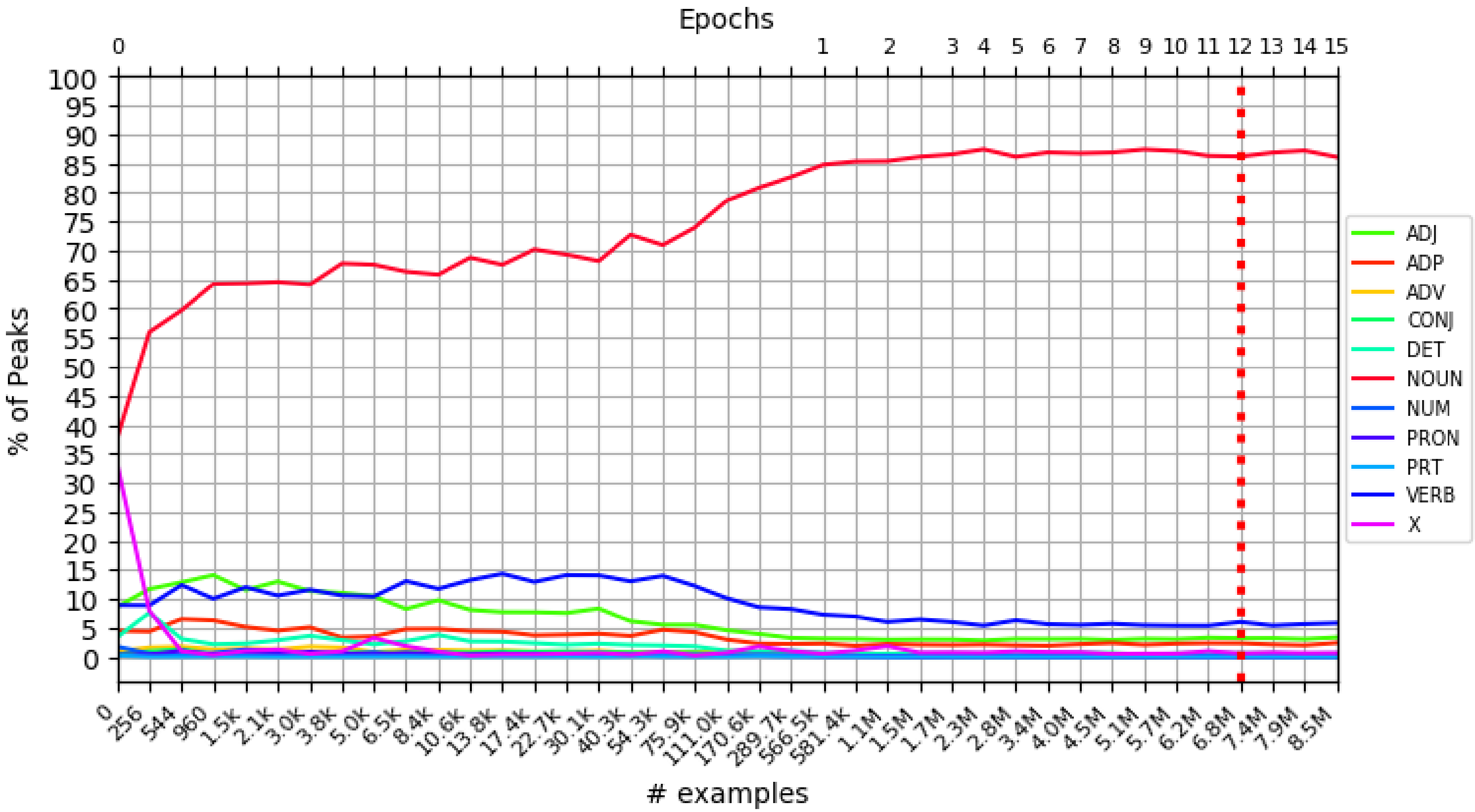


三 頭 の キリン が 草原 を 歩い て いる

three CLF GEN giraffe SUJ meadow OBJ exist PROG

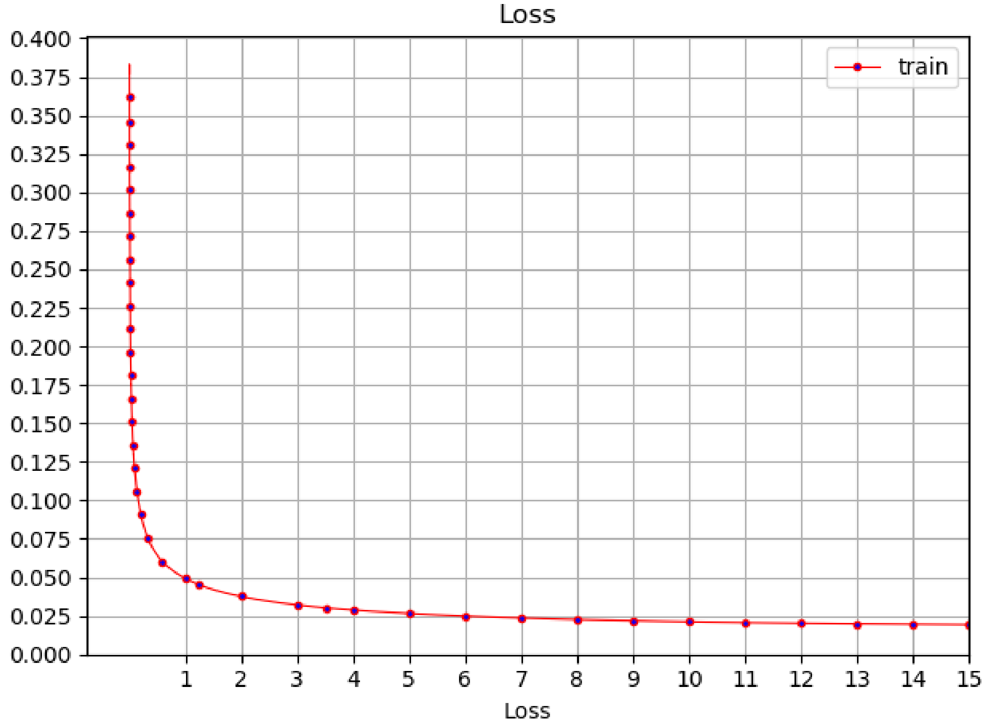
Behaviour of Attention Over Time

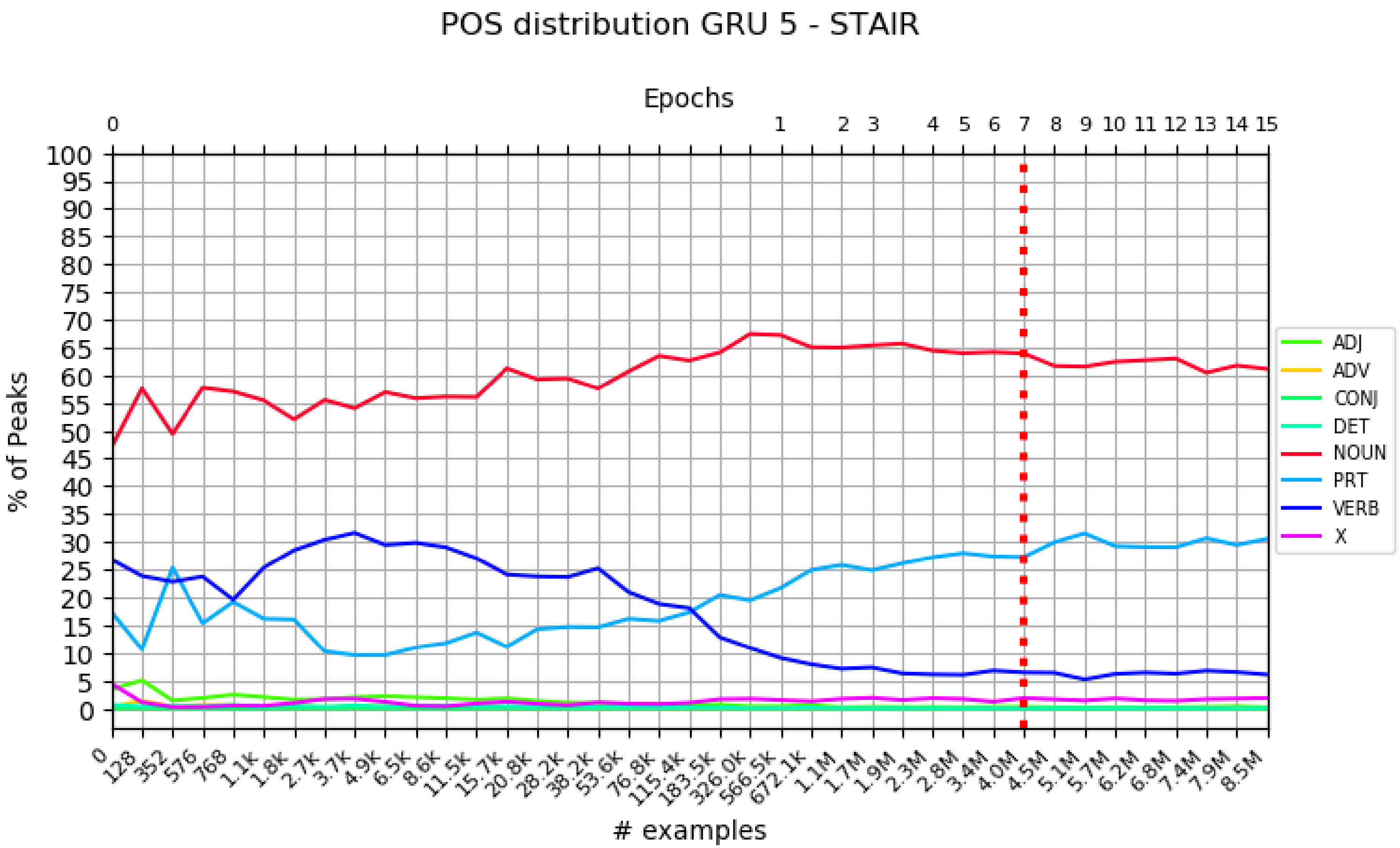




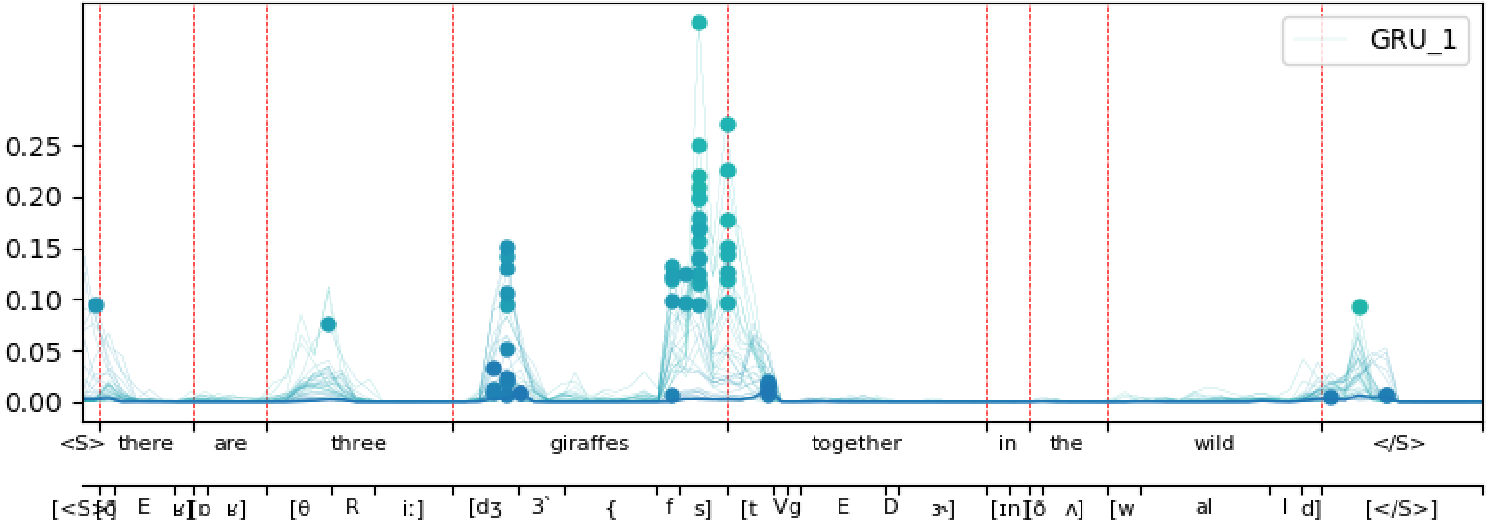
- Models quickly learn to focus on **important nouns (English)**
- Visible with **only 256 examples**

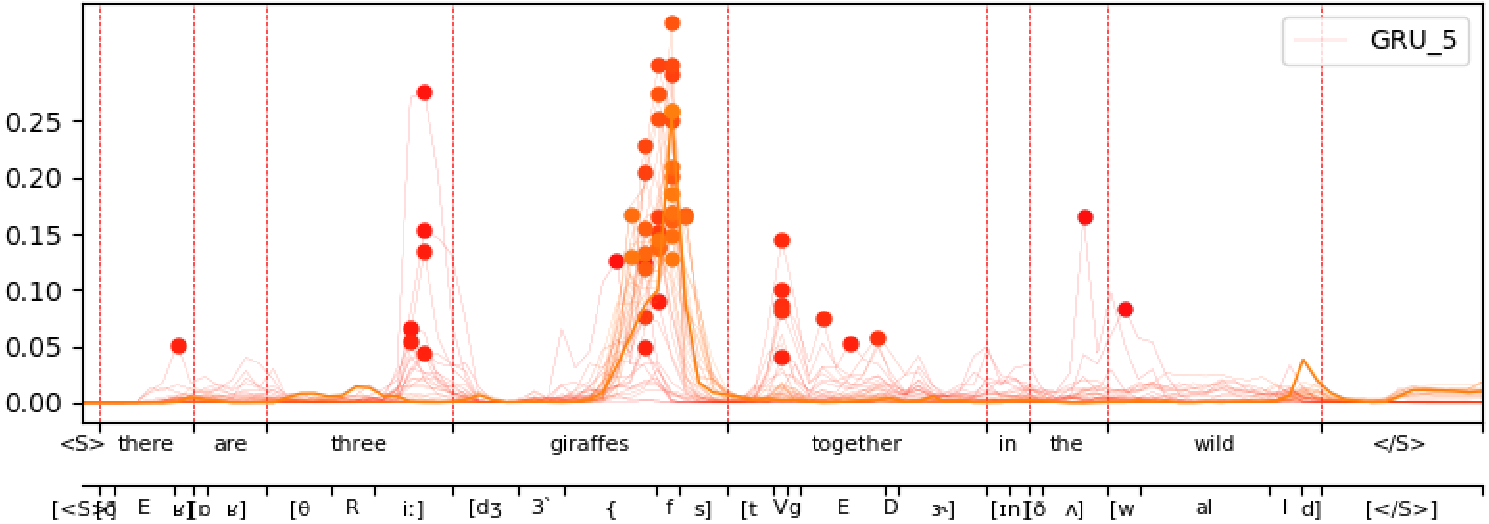
BEHAVIOUR OF ATTENTION OVER TIME

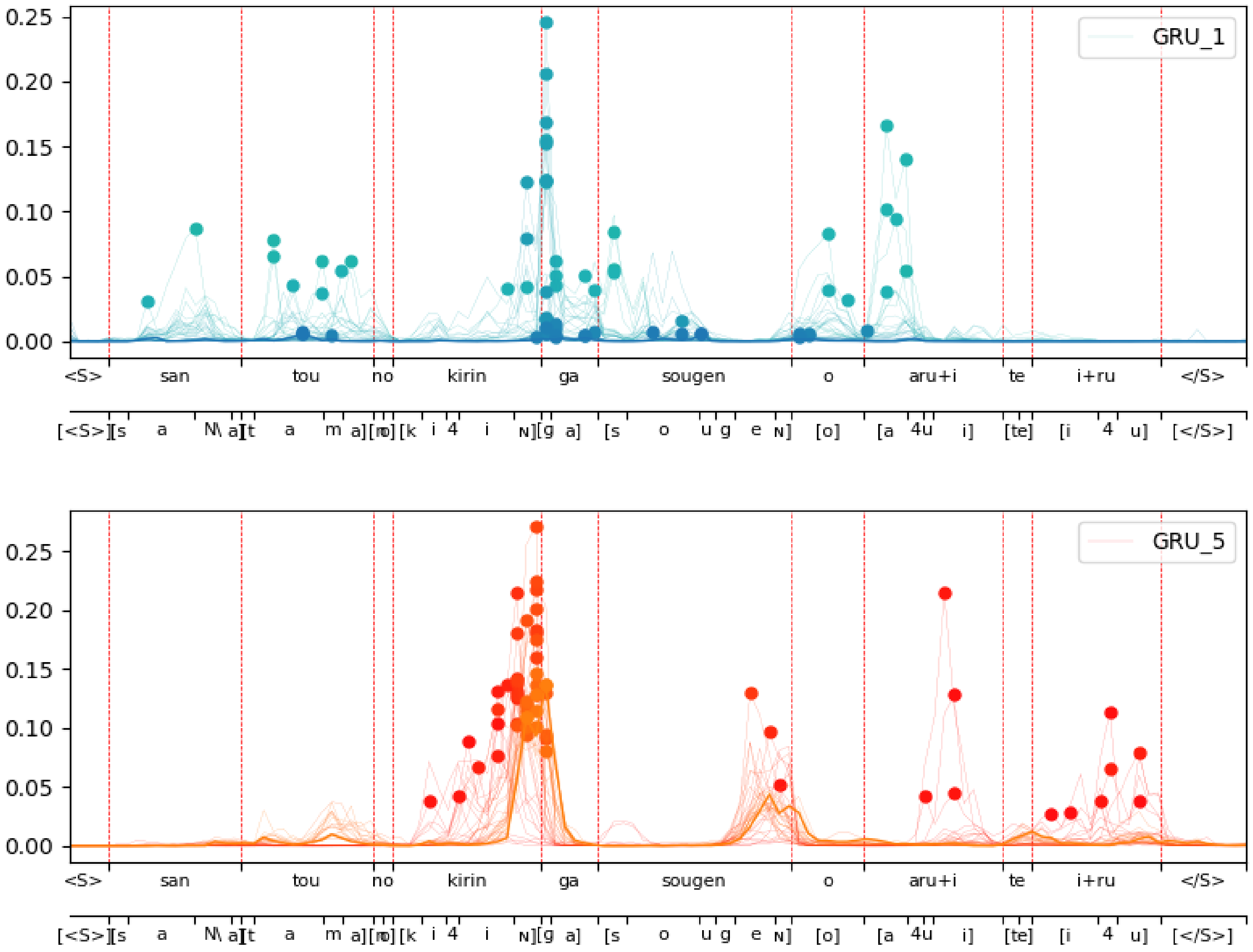




- Behaviour less clear-cut
- Focus on **particles is gradual**


GRU_1
0.25
0.20
0.15
0.10
0.05
0.00
<S> there are three giraffes together in the wild </S>


GRU_5
0.25
0.20
0.15
0.10
0.05
0.00
<S> there are three giraffes together in the wild </S>


GRU_1
0.25
0.20
0.15
0.10
0.05
0.00
<S> san tou no kirin ga sougen o aru+i te i+ru </S>
GRU_5
0.25
0.20
0.15
0.10
0.05
0.00
<S> san tou no kirin ga sougen o aru+i te i+ru </S>