Robot Language Acquisition Modelling via Cross-Situational Learning with Little Data

Xavier Hinaut

Robot Language Acquisition Modelling via Cross-Situational Learning with Little Data

Xavier Hinaut[1,2,3]

[1]Inria centre of Bordeaux University.
[2]LaBRI, Bordeaux University, Bordeaux INP, CNRS UMR 5800.
[3]Bordeaux University, CNRS, IMN, UMR 5293, Bordeaux, France
xavier.hinaut@inria.fr

Abstract

How do children bootstrap language through noisy supervision? Most prior works focused on tracking co-occurrences between individual words and referents. We model cross-situational learning (CSL) at sentence level with few (1000) training examples. We compare two recurrent neural network architectures often use as cognitive models: reservoir computing (RC) and LSTMs on three datasets including complex robotic commands. Surprisingly, reservoirs demonstrate robust generalization when increasing vocabulary size: the error grows slowly compared to an LSTM of fixed size. This suggests that that random projections used in RC helps to bootstrap generalization quickly. How robots acquire basics of language like in child-caregiver (Human-Human) interactions could give hints of how to link animal vocalisations with behaviour in ambiguous context. Cross-statistics between sequence of vocalisations and various contexts could probably be learnt in few trials by such Reservoir architecture.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Discussion

Comparison to other non-recurrent architectures It is likely that Transformers architecture [1] would require more data for training, thus the comparison at this tiny data scale (1000 examples) does not seem relevant. However, their attention mechanism is interesting, in particular to parse long sentences in some of the more challenging datasets that we tried [2]. In future work we will explore how such attention mechanisms can help reservoir computing to scale to much bigger datasets, enabling to have an architecture able to generalize from tiny to big datasets.

2. References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[2] S. R. Oota, F. Alexandre, and X. Hinaut, "Cross-situational learning towards robot grounding," *HAL preprint*, 2022.

[3] H. Jaeger, "The "echo state" approach to analysing and training recurrent neural networks-with an erratum note," *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, vol. 148, no. 34, p. 13, 2001.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] A. Juven and X. Hinaut, "Cross-situational learning with reservoir computing for language acquisition modelling," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[6] A. Variengien and X. Hinaut, "A journey in esn and lstm visualisations on a language task," *arXiv preprint arXiv:2012.01748*, 2020.

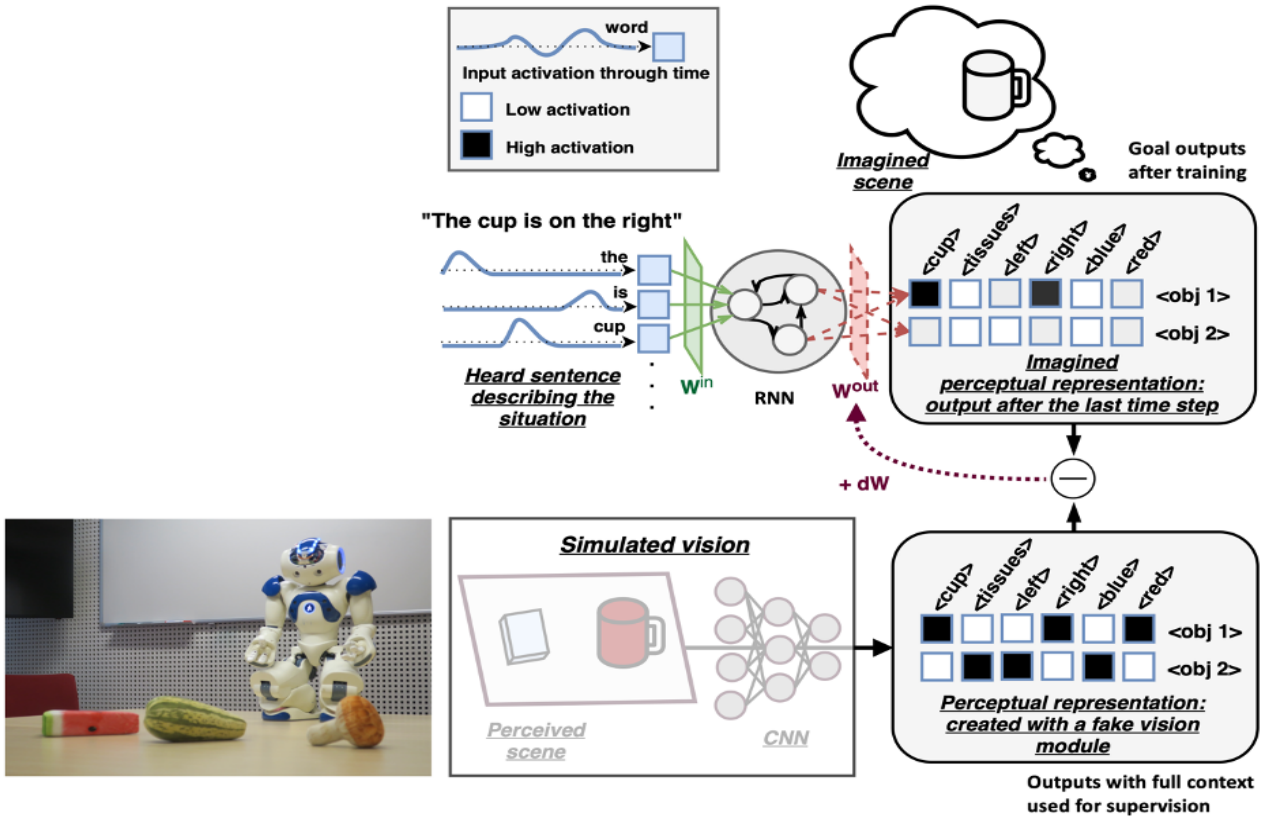

Figure 1: *The Cross-Situational Learning (CSL) learning procedure for a Recurrent Neural Network (RNN) architecture. We compare two RNNS: Reservoir Computing (RC) [3] and Long Short-Term Memory network (LSTM) [4]. The model has to reconstruct an imagined scene from the sentence given word by word. The simulated vision creates a perceptual representation corresponding to the full description of objects in the scene. This representation is used as target outputs for the reservoir, even if the sentence only partially describes the objects in the scene, or if it describes only one object. This particular set-up creates cross-situational learning conditions similar to the ones children are facing. The set-up, input and target outputs were the same for the LSTM experiments. (Image adapted from [5].*

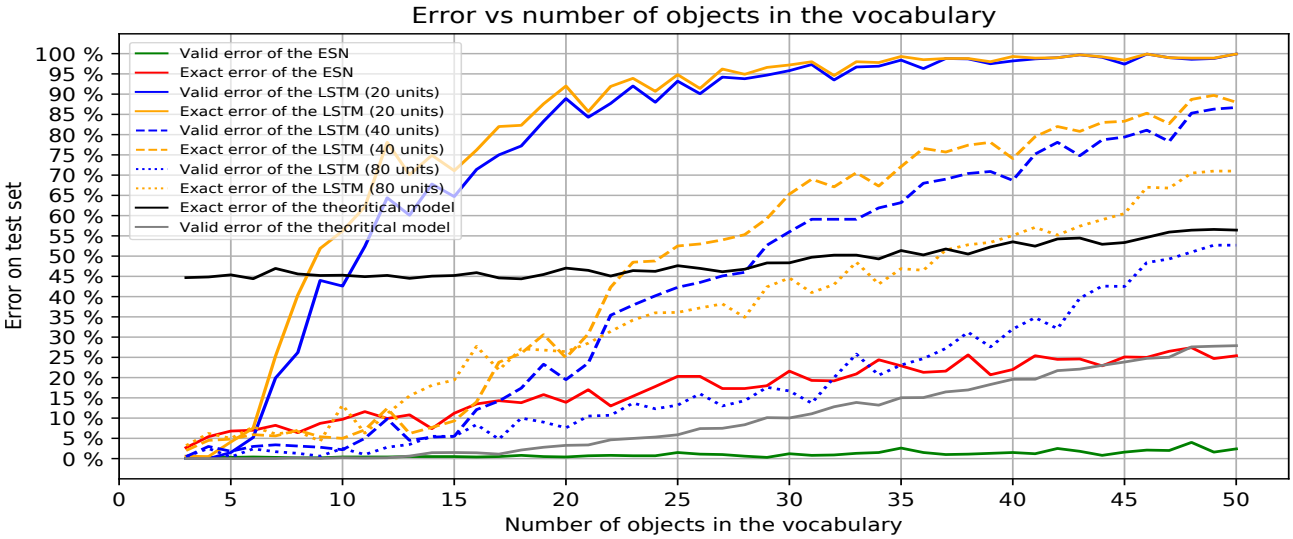

Figure 2: *Comparison of the performance of 5 models (1 ESN + 3 LSTMs + theoretical) for different number of objects in the dataset. Echo State Network (ESN) is a particular instance of the Reservoir Computing paradigm. The small LSTM (20 units), optimized to perform well on a dataset with 4 objects, is not able to keep good performance with a higher number of objects. The medium LSTM (40 units) trained for longer with dropout is able to outperform the ESN until 15 objects. The bigger LSTM (80 units) limits the rise of the error compared to the other LSTM. However, it comes with poorer performances even for a small number of objects. The ESN is able to keep an error below the theoretical model and all the LSTMs despite the fact that its hyper-parameters were optimized for the 4-object dataset. Image from [6].*