Title: Modeling Situated Language Learning in Early Childhood via Hypernetworks

Authors:

Byoung-Tak Zhang[1, 2], Eun-Seok Lee[1], and Min-Oh Heo[2]

(1) Graduate Program in Cognitive Sciences, Seoul National University, Seoul 151-742, Korea

(2) School of Computer Science and Engineering, Seoul National University, Seoul 151-742, Korea

Text:

Human language is grounded, i.e. embodied and situated in the environment (Barsalou, 2008; Zwaan & Kaschak, 2008), and grounded language models rely on multimodal sensory data, such as gaze and gestures (Knoeferle & Crocker, 2006; Spivey, 2007; Yu et al. 2008). From the computational point of view, language grounding requires a flexible modeling tool that can deal with high-dimensional data. Here we explore a machine learning-based computational modeling method for situated language learning. The crux of the method is the hypernetwork model (Zhang, 2008), a probabilistic graphical model that learns multimodal associative memory incrementally from multisensory data.

We report on two sets of situated-language experiments simulating language acquisition in early childhood. We collected 10 series of cartoon videos for children of ages 3 to 7 and use them as a surrogate for large-scale multimodal experiment data. The first experiment uses dialogue sentences only, while the second experiment uses the visual scenes aligned to the sentences. The hypernetwork language model is learned serially on the videos of increasing order of age. We analyzed the concept maps (Griffiths et al., 2007) constructed by the probabilistic hypernetwork model to investigate the evolution of the concepts (or "visually-grounded" concepts in the second set of experiments). Using the generative property of the hypernetworks we generated the sentences and "mental" images from the learned model and analyzed them to study the evolution of grounded linguistic concepts as learning experience proceeds.

We found that the complexity of the sentences grows and more complex concepts emerge from simple ones as the learner observes diversified situations, as expected in language acquisition of infants and children (Bowerman & Levinson, 2001). Grammatical sentences were generated even though learning used neither lexical nor part-of-speech information. The *post hoc* analysis of the structure of the generated sentences showed the emergence of grammar rule-like patterns in the language model. Similar evolution of the multimodal visuo-linguistic concept maps was observed in the visually-grounded language learning experiments, though with a higher level of uncertainty due to the high noise ratio in the image data.

Acknowledgments:

This work was supported by the National Research Foundation (NRF) grants (2011-0016483-Videome, 2010-0018950-BrainNet), the IT R&D Program of KEIT (10035348-mLife), and the BK21-IT Program.

References:

Barsalou, L. W. (2008). Grounded cognition, *Ann. Rev. Psych.*, 59, 617-645.

Griffiths, T., Steyvers, M., & Tenenbaum, J., (2007) Topics in semantic representation, *Psych.*

Rev., 114(2), 211–244.

Knoeferle, P., & Crocker, M. W. (2006). The coordinated interplay of scene, utterance, and world knowledge: Evidence from eye-tracking, *Cog. Sci.*, 30, 481-529.

Spivey, J. M. (2007). *The Continuity of Mind*, Oxford Univ. Press.

Yu, C, Smith, L. B., & Pereira, A. F. (2008). Grounding word learning in multimodal sensorimotor interaction, *CogSci-2008*, pp. 1017-1022.

Zhang, B.-T. (2008). Hypernetworks: a molecular evolutionary architecture for cognitive learning and memory. *IEEE Comp. Intell. Mag.*, 3(3), 49-63.

Zwann, R. A. & Kaschak, M. P. (2008). Language in the brain, body, and world. Chap. 19, *Cambridge Handbook of Situated Cognition*.

Keywords: language learning in children, visually-grounded language, video corpus, probabilistic graphical models, emergence of grammar, concept maps.

Wordcount: 492