Cross-situational Word Learning from Naturalistic Headcam Data

Wai Keen Vong, Emin Orhan and Brenden Lake (New York University)

One of the challenges of word learning is the problem of reference. When a child hears a word like “ball”, how are they able to figure out which referents in the world this word refers to? One proposed learning mechanism for resolving this is through cross-situational learning: rather than learning from an ambiguous single instance, children can aggregate information across multiple ambiguous co-occurrences of the word “ball” to correctly determine the underlying referent. While the topic of cross-situational word learning has received significant attention, both empirically (Yu & Smith, 2007), and in the development of various computational models (Frank, Goodman & Tenenbaum, 2009; Stevens et al., 2017), many well-known models require the visual referents to be preprocessed as discrete entities so it is unclear how these models could scale to explain cross-situational word learning from naturalistic data.

Recently, researchers have begun to combine convolutional neural networks with egocentric headcam data, and have shown that such models can learn useful visual representations (Bambach et al., 2018; Orhan, Gupta & Lake, 2020; Tsutsui et al., 2020). One limitation of these approaches is that they are trained using supervised feedback on category labels. In contrast, cross-situational learning provides no direct supervision akin to supervised feedback during the learning process, but only a form of weak supervision based on which words co-occur with which referents in the scene.

Inspired by these challenges, we recently developed a computational model to perform cross-situational word learning using the SAYCam dataset, a large-scale longitudinal egocentric headcam dataset (Sullivan et al., 2020). We trained our model using data from a single child, by creating a dataset of roughly 35000 image-utterance pairs extracted from roughly 60 hours of raw video with transcribed utterance data. Our model architecture is a multimodal neural network model, which embeds images using a pre-trained convolutional neural network trained only from the raw visual data from the same child (Orhan et al., 2020), and separately embeds utterances using a language encoder consisting of either a single embedding layer or an LSTM. The model is trained via a contrastive loss, learning to pair images with their corresponding utterances in the embedding space, which allows the model to learn word-referent mappings.

The model was evaluated on a separate dataset of frames extracted from 22 common visual categories in SAYCam, by presenting the model with a target word and four images (one target, and three foils), and asking the model to select the correct target referent. Our results show that the model is above chance in selecting the correct referent for more than half of these categories. We also show that qualitatively, the model can also localize the referents in a given scene (as shown below). To our knowledge, this is the first model that successfully captures cross-situational word learning using longitudinal egocentric data from a single child, and demonstrates that such learning is possible from raw inputs using recent advances in deep learning.

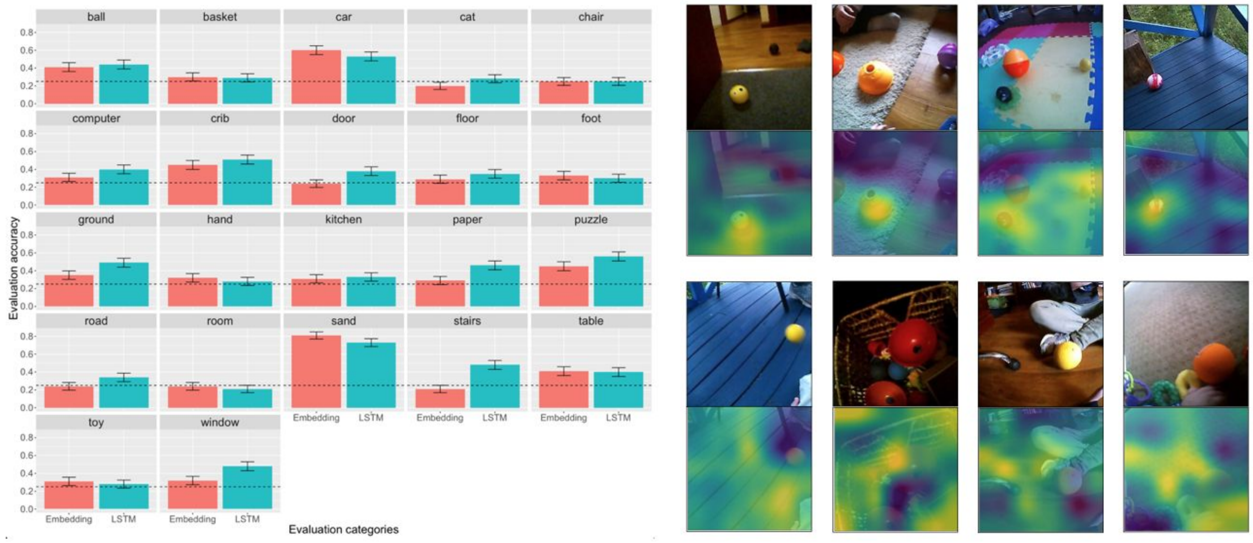

Figure 1. The left figure shows the evaluation performance of the two models for the 22 visual categories in SAYCam, with error bars as standard error and the dotted line representing chance. The right figure shows examples of localization of referents using attention maps extracted from our model, with the top row showing successes for the word “ball”, and the bottom showing some failures.

References

- Bambach, S., Crandall, D., Smith, L., & Yu, C. (2018). Toddler-inspired visual object learning. *Advances in Neural Information Processing Systems*, 31, 1201-1210.
- Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20(5), 578-585.
- Orhan, E., Gupta, V., & Lake, B. M. (2020). Self-supervised learning through the eyes of a child. *Advances in Neural Information Processing Systems*, 33.
- Stevens, J. S., Gleitman, L. R., Trueswell, J. C., & Yang, C. (2017). The pursuit of word meanings. *Cognitive Science*, 41, 638-676.
- Sullivan, J., Mei, M., Perfors, A., Wojcik, E. H., & Frank, M. C. (2020). SAYCam: A large, longitudinal audiovisual dataset recorded from the infant’s perspective.
- Tsutsui, S., Chandrasekaran, A., Reza, M. A., Crandall, D., & Yu, C. (2020). A Computational Model of Early Word Learning from the Infant's Point of View. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society.*
- Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18(5), 414-420.