A Computational Acquisition Model for Multi-modal Word Learning from Scratch

Uri Berger[1*] **Gabriel Stanovsky**[1] **Omri Abend**[1] **Lea Frermann**[2]

[1]The Hebrew University Of Jerusalem
[2]University of Melbourne

{*uri.berger2, gabriel.stanovsky, omri.abend*}*@mail.huji.ac.il*
lea.frermann@unimelb.edu.au

Extended abstract

Human infants learn to interpret the external world remarkably well, and with little direct supervision or pre-training. As an example, consider first language acquisition. Infants with no prior knowledge of the meaning of words or the syntactic rules of the language learn to speak and understand natural languages from scratch. Some supervision might be provided from other modalities (e.g., vision), but at the beginning of the learning process, these modalities are untrained as well, and cannot provide supervision (e.g., an infant presented with an apple while hearing the word "apple" cannot identify that the referred object is indeed the apple; from her point of view, it might as well be the face or hand of the person holding the apple).

Clearly, after knowledge of language processing has been acquired, visual processing becomes much easier, and vice versa. However, at the beginning of the learning process, the infant is unable to process neither of the modalities. Thus, the infant needs to bootstrap both types of processing together.

Models of bootstrapping in early language acquisition have been studied extensively from a cognitive point of view, particularly in the context of bootstrapping of syntax rules with the semantic meaning of words (Desai, 2002; Alishahi and Stevenson, 2008; Frank et al., 2009; Maurits et al., 2009; Alishahi and Stevenson, 2010; Kwiatkowski et al., 2012; Alishahi and Chrupala, 2012; Abend et al., 2017; Nikolaus and Fourtassi, 2021). However, according to some psychological studies of early language acquisition, learning complex syntax rules might be an advanced stage of language acquisition, the first stage being the identification of nouns in a sentence, since nouns can be acquired efficiently in the absence of structural knowledge (Fisher et al., 1994).

These findings make the language-vision bootstrapping theory even more appealing; the visual modality presents entities that are, by definition, concrete. Words representing concrete entities are mostly nouns; In the concreteness dataset built by Brysbaert et al. (Brysbaert et al., 2013), in which human annotators rated the concreteness of words on a scale of 1 to 5, 85.6% of the words with an average concreteness rating above 4 are nouns. Therefore, in line with the above studies, a model that learns from visual-textual data will first acquire concrete words, which are mostly nouns. Moreover, empirical data show that infants' first words are words that represent concrete entities. According to Wordbank (Frank et al., 2016), a large database of infants' word usage, the most frequent words uttered by 16-month-old English-speaking infants are "mommy", "daddy", "ball", "dog", "baby", "book", "banana", "shoe", "bird", "duck" – all are words that represent concrete objects.

Recently, there has been an increasing interest in multi-modal bootstrapping models from a computational point of view, especially in the deep learning community. The main line of work is multi-modal semi-supervised models. In these studies, the input patterns are pairs of samples from different modalities (usually *image, text* pairs), with the sole supervision being the fact that these samples describe the same thing (e.g., a picture of a dog playing with a Frisbee and the caption "A photo of a dog playing with a Frisbee"). This framework loosely resembles the framework of infants language acquisition, where the child is exposed to directed speech that describes the scene which she is experiencing through different modalities (e.g., vision). Most studies (Vong and Lake, 2021; Joy et al., 2021; Radford et al., 2021; Nikolaus and Fourtassi, 2021) map both modalities to an unconstrained linear sub-

[*]Presenter, an early PhD student presenting preliminary results

space ($\mathbb{R}^n$) and the meaning of individual words is ignored, as the main downstream task is classification of unseen images (unlike human infants, who first learn basic words and use them as building blocks to learn more complex structures, see (Fisher et al., 1994)). In addition, these studies assume a pre-trained visual model, usually trained on the ImageNet dataset (Deng et al., 2009) – an unrealistic assumption in child language acquisition, since the supervision in the ImageNet dataset contains semantic categorization information to which the infant is not exposed.

In this study, we present a language-vision bootstrapping algorithm, that learns to identify concrete words and localize visual objects, without pre-training and without any supervision except the matching of image-caption pairs. Unlike previous cognitive studies (Maurits et al., 2009; Abend et al., 2017; Nikolaus and Fourtassi, 2021), we use naturalistic data with a large vocabulary (over 10,000 words). Moreover, our model is more realistic than recent computational models: first, we constrain the shared, latent space to a binary space ($\{0,1\}^n$), which forces clustering of image/text into conceptual classes. Forcing the model to cluster inputs is a desirable feature, since clustering objects into semantic classes allows infants to generalize well. For example, after learning the semantic class *dog*, when a new dog is encountered, it will be clustered into this semantic class, allowing the infant to label this unseen animal as a dog. Second, we map each word (rather than the entire sentence) to the shared space and only later aggregate the mappings into a single representation of the entire sentence, allowing the model to share the single-word learning mechanism of human infants. Third, we use a visual model that was not trained on any labelled data, thereby excluding assumptions that are unlikely to be realized in infant language acquisition. The meaning and concreteness of words is learned in the process, enabling the most elementary syntactic task: identifying concrete words that will be marked as "nouns" by the child.

Experiments on the MSCOCO dataset (Lin et al., 2014) show that our model learns to identify concrete words even though it does not use concreteness supervision or any pre-trained models. Furthermore, the use of CAM (class activation mapping (Zhou et al., 2016)) allows us to analyze the visual knowledge acquired by our model and use it to localize objects (see figure 1). Note that not only no localization supervision is given, but also no supervision is given regarding to which types, or even how many objects, are in the image.

Figure 1: Examples of our model's object localization capabilities. Using the CAM technique from (Zhou et al., 2016), pixels with high class activation values are highlighted, creating the object localization heatmap

In future work we hope to expand our approach, using this basic task as a building block and target the acquisition of more complex syntax and semantics.

References

Omri Abend, Tom Kwiatkowski, Nathaniel J. Smith, Sharon Goldwater, and Mark Steedman. 2017. Bootstrapping language acquisition. *Cognition*, 164:116–143.

Afra Alishahi and Grzegorz Chrupala. 2012. Concurrent acquisition of word meaning and lexical categories. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 643–654, Jeju Island, Korea. Association for Computational Linguistics.

Afra Alishahi and Suzanne Stevenson. 2008. A computational model of early argument structure acquisition. *Cognitive Science*, 32(5):789–834.

Afra Alishahi and Suzanne Stevenson. 2010. A computational model of learning semantic roles from child-directed language. *Language and Cognitive Processes*, 25(1):50–93.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2013. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46(3):904–911.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*.

Rutvik Desai. 2002. Bootstrapping in miniature language acquisition. *Cognitive Systems Research*, 3(1):15–23.

Cynthia Fisher, D.geoffrey Hall, Susan Rakowitz, and Lila Gleitman. 1994. When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth. *Lingua*, 92:333–375.

Michael C. Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A. Marchman. 2016. Wordbank: an open repository for developmental vocabulary data. *Journal of Child Language*, 44(3):677–694.

Michael C. Frank, Noah D. Goodman, and Joshua B. Tenenbaum. 2009. Using speakers referential intentions to model early cross-situational word learning. *Psychological Science*, 20(5):578–585.

Tom Joy, Yuge Shi, Philip H. S. Torr, Tom Rainforth, Sebastian M. Schmon, and N. Siddharth. 2021. Learning multimodal VAEs through mutual supervision.

Tom Kwiatkowski, Sharon Goldwater, Luke Zettlemoyer, and Mark Steedman. 2012. A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 234–244, Avignon, France. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. *Computer Vision – ECCV 2014 Lecture Notes in Computer Science*, page 740–755.

Luke Maurits, Amy F. Perfors, and Daniel J. Navarro. 2009. Joint acquisition of word order and word reference. In *Proceedings of the 31st annual conference of the Cognitive Science Society*, page 1728–1733.

Mitja Nikolaus and Abdellah Fourtassi. 2021. Evaluating the acquisition of semantic knowledge from cross-situational learning in artificial neural networks.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Wai Keen Vong and Brenden M. Lake. 2021. Cross-situational word learning with multimodal neural networks.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.