https://doi.org/10.1038/d41586-024-00713-5

News & views

Machine learning

Can lessons from infants solve the problems of AI?

Linda B. Smith

Words and images experienced by an infant wearing sensors during their daily life have led to efficient machine learning, pointing to the power of multimodal training signals and to the potentially exploitable statistics of real-life experience.

Current advances in artificial intelligence (AI) seem to be transforming science into science fiction, as large-data machine-learning models approach and, in some ways, surpass human abilities. But such models are trained on vast amounts of data. Writing in *Science*, Vong *et al.*[1] have thrown down a challenge on behalf of humans by using 61 hours of one infant's real-life experiences to demonstrate the efficiency of a multimodal learning model.

The training data were captured by a head-mounted camera as multiple brief samples between the ages of 6 months and 25 months. This is the period of rapid vocabulary expansion at the start of language learning. Video clips from the head camera and transcripts of adults speaking to the infant were fed to the authors' model, which used a 'contrastive' approach for both visual and language learning.

Contrastive learning is a widely used method in machine learning[2]. It involves feeding pairs of training examples into an algorithm with a label indicating the similarity of the two examples. Evidence that two items are in the same category changes the model parameters to make those items more similar in the learnt representational space of all pairings. Evidence to the contrary changes the model parameters that define the space to push items apart.

Vong and colleagues' model integrated contrastive representational learning with associative learning — that is, the learnt links between the utterances and the images. It did so by using the learnt representations in one modality as the teaching signal (for increasing or decreasing the similarity of the pairs) for the other modality. In this way, learning was self-supervised because of the co-occurrence of text and images. This is a form of the bi-directional exchange of learning signals across different neural systems (such as the visual and language-processing parts of the brain) that was first introduced by Nobel laureate Gerald Edelman[3] and called re-entrance.

Re-entrant signals are a potentially powerful form of self-supervised learning because the teaching signals from one neural system change as a function of learning driven by the other system (Fig. 1). Multimodal models with re-entrance learn rapidly as a consequence of the simultaneous coordination of different representational components, as is demonstrated by Vong and co-workers' model.

The everyday experiences of infants are challenging for any learning algorithm to use successfully. Objects and co-occurring words are well known to produce noisy data, leading to many spurious pairings[4]. Moreover, the presence of object names in language heard by children is remarkably sparse. For example, the word 'basket' (which is comprehended by a child before the age of 25 months) occurs just 8 times in a 6-million-word corpus[5] of language that parents use when talking to children. Nonetheless, young children learn object names and immediately generalize those names to never-seen-before instances. The authors' model, which was trained with child-egocentric data, did the same. The re-entrance in the model provides a potential theoretical path for explaining infants' rapid learning and generalization.

Vong *et al.* characterize their contribution as a demonstration that object names and visual categories can be learnt from a small amount of sparse and noisy training data. Of course, infants have already provided this proof. The key question is what the authors' model tells us about how this can be achieved. It is likely that the success of the model reflects the computational power of the re-entrant signals,

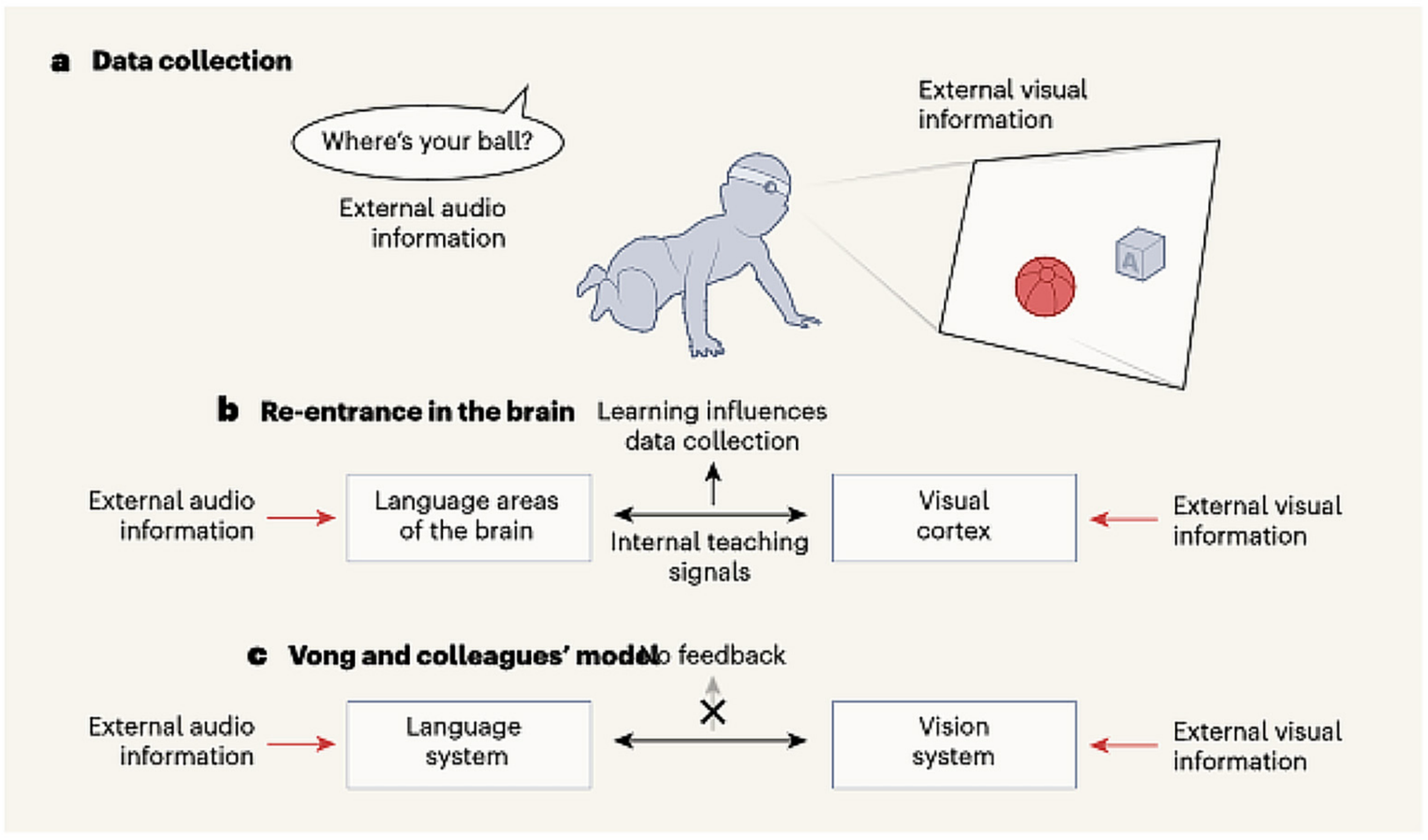

Figure 1 | Self-supervised learning. a, Vong *et al.*[1] used audio and video recordings from the daily life of an infant wearing a sensor to train an artificial-intelligence model to learn language. The co-occurrence of images and words made the training process 'self-supervised', meaning that it required only internal teaching signals. **b**, This is reminiscent of re-entrance[3], which is the continuous exchange of teaching signals between different neural systems, such as the visual system and the areas of the brain in which language develops. Re-entrance allows the brain to collect data that are relevant to its current learning state by actively selecting and creating sensory events. **c**, The authors' model does not include this feedback, but the data used by their model might still reflect some properties of infant-generated-data structures that benefited the model's ability to learn.

although that has not been shown. I think that the success of the model could also depend on the temporal and spatial statistics of the images captured by the head camera worn by the infant; this has also not been determined. Other studies have shown that training with infant-egocentric images outperforms training with other data sets, including adult-egocentric experiences[6,7].

A growing body of work in human developmental science uses wearable sensors to capture and quantify, at the scale of daily life, the statistics of infant and child experience[8]. This research is not conducted from the perspective of AI, but AI researchers might be wise to pay attention. The statistics of children's real-world experiences have some intriguing aspects[7,9] including an ordered curriculum of experience created by both motor development and the progression of learning itself. Moreover, infants instantiate Edelman's full model of re-entrance. At each moment in time, children select, elicit and create multimodal inputs from where they look, what they touch and what they do. This selection is dependent on the momentary activated representations in the multimodal neural systems, connecting in-the-moment samples of data to the current internal state of the learner[10].

A core problem for all learning systems is how, in a large and complex space of model parameters, the learner can find the optimal or near-optimal solution. Large-data models operate on the assumption that given enough data, a sufficiently powerful learner will find the optimal solution; in the limit in which the model has access to all the data, this might be true. However, in practice, there are many classes of learning problem for which there are not enough data, not enough time or not enough computational power.

The efficient learning that is evident in infants is probably a direct result of the multimodal statistics of natural experience and the infants' active participation in the collection of those data. Here is the conjecture: even a passive learner such as Vong and colleagues' model might benefit from infant-egocentric training data because the multimodal statistics constrain the search path. If that's the case, the many problems of data-greedy AI could be mitigated by determining and then exploiting the natural statistics of infant experience.

Linda B. Smith is in the Department of Psychological and Brain Sciences, Indiana University, Bloomington, Indiana 47405, USA.
e-mail: smith4@iu.edu

1. Vong, W. K., Wang, W., Orhan, A. E. & Lake, B. M. *Science* **383**, 504–511 (2024).
2. Hinton, G. E. *Neural Comput.* **14**, 1771–1800 (2002).
3. Edelman, G. M. *Neural Darwinism: The Theory of Neuronal Group Selection* (Basic, 1987).
4. Medina, T. N., Snedeker, J., Trueswell, J. C. & Gleitman, L. R. *Proc. Natl Acad. Sci. USA* **108**, 9014–9019 (2011).
5. MacWhinney, B. *The Childes Project*, Vol. 2 (Psychology, 2000).
6. Bambach, S., Crandall, D. J., Smith, L. B. & Yu, C. In *Proc. Advances in Neural Information Processing Systems* (eds Bengio, S. *et al.*) 1209–1218 (NeurIPS, 2018).
7. Sheybani, S., Hansaria, H., Wood, J. N., Smith, L. B. & Tiganj, Z. In *Proc. Advances in Neural Information Processing Systems* (eds Oh, A. *et al.*) 36 (NeurIPS, 2023).
8. de Barbaro, K. & Fausey, C. M. *Curr. Dir. Psychol. Sci.* **31**, 28–33 (2022).
9. Smith, L. B., Jayaraman, S., Clerkin, E. & Yu, C. *Trends Cogn. Sci.* **22**, 325–336 (2018).
10. Karmazyn-Raz, H. & Smith, L. B. *Phil. Trans. R. Soc. B* **378**, 20210358 (2023).

The author declares no competing interests.