New research shows how child-like language learning is possible using AI tools

February 1 2024

Credit: NYU's Center for Data Science

AI systems, such as GPT-4, can now learn and use human language, but they learn from astronomical amounts of language input—much more than children receive when learning how to understand and speak a language. The best AI systems train on text with a word count in the trillions, whereas children receive just millions per year.

Due to this enormous data gap, researchers have been skeptical that recent AI advances can tell us much about human learning and development. An ideal test for demonstrating a connection would involve training an AI model, not on massive data from the web, but on only the input that a single child receives. What would the model be able to learn then?

A team of New York University researchers ran this exact experiment. They trained a multimodal AI system through the eyes and ears of a single child, using headcam video recordings from when the child was 6 months and through their second birthday. They examined if the AI model could learn words and concepts present in a child's everyday experience.

Their findings, reported in the journal *Science*, showed that the model, or neural network, could, in fact, learn a substantial number of words and concepts using limited slices of what the child experienced. That is, the video only captured about 1% of the child's waking hours, but that was sufficient for genuine language learning.

In this video, the researchers describe their work in greater detail:

"We show, for the first time, that a neural network trained on this developmentally realistic input from a single child can learn to link words to their visual counterparts," says Wai Keen Vong, a research scientist at NYU's Center for Data Science and the paper's first author.

"Our results demonstrate how recent algorithmic advances paired with one child's naturalistic experience has the potential to reshape our understanding of early language and concept acquisition."

"By using AI models to study the real language-learning problem faced by children, we can address classic debates about what ingredients

children need to learn words—whether they need language-specific biases, innate knowledge, or just associative learning to get going," adds Brenden Lake, an assistant professor in NYU's Center for Data Science and Department of Psychology and the paper's senior author. "It seems we can get more with just learning than commonly thought."

Vong, Lake, and their NYU colleagues, Wentao Wang and Emin Orhan, analyzed a child's learning process captured on first-person video—via a light, head-mounted camera—on a weekly basis beginning at 6 months and through 25 months, using more than 60 hours of footage.

Video frames captured from a child wearing a head-mounted camera. Credit: NYU's Center for Data Science

The footage contained approximately a quarter of a million word instances (i.e., the number of words communicated, many of them repeatedly) that are linked with video frames of what the child saw when those words were spoken and included a wide range of different activities across development, including mealtimes, reading books, and the child playing.

The NYU researchers then trained a multimodal neural network with two separate modules: One that takes in single video frames (the vision encoder) and another that takes in the transcribed child-directed speech (the language encoder).

These two encoders were combined and trained using an algorithm called "contrastive learning," which aims to learn useful input features and their cross-modal associations. For instance, when a parent says something in view of the child, it is likely that some of the words used are likely referring to something that the child can see, meaning comprehension is instilled by linking visual and linguistic cues.

"This provides the model a clue as to which words should be associated with which objects," explains Vong. "Combining these cues is what enables contrastive learning to gradually determine which words belong with which visuals and to capture the learning of a child's first words."

After training the model, the researchers tested it using the same kinds of evaluations used to measure word learning in infants—presenting the model with the target word and an array of four different image options and asking it to select the image that matches the target word.

Their results showed that the model was able to learn a substantial number of the words and concepts present in the child's everyday experience. Furthermore, for some of the words the model learned, it could generalize them to very different visual instances than those seen at training, reflecting an aspect of generalization also seen in children when they are tested in the lab.

"These findings suggest that this aspect of word learning is feasible from the kind of naturalistic data that children receive while using relatively generic learning mechanisms such as those found in neural networks," observes Lake.

More information: Wai Keen Vong et al, Grounded language acquisition through the eyes and ears of a single child, *Science* (2024). DOI: 10.1126/science.adi1374.
www.science.org/doi/10.1126/science.adi1374

Provided by New York University