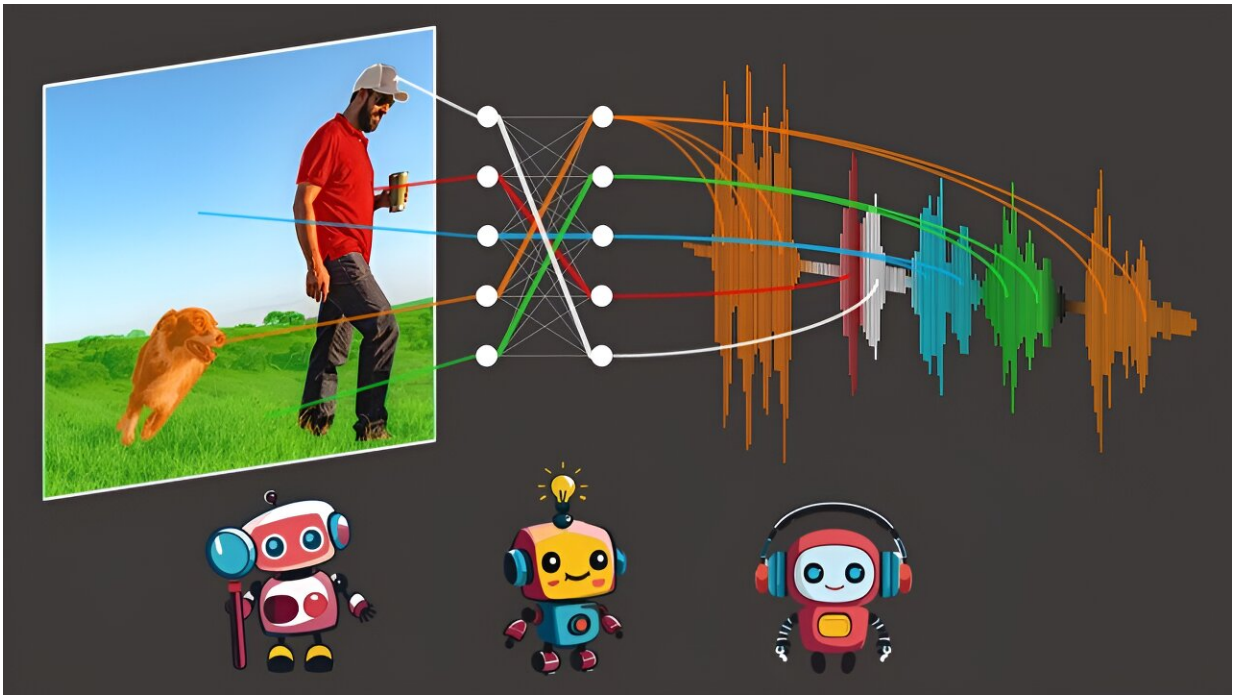
New algorithm discovers language just by watching videos

June 11 2024, by Rachel Gordon



The algorithm DenseAV learns the meaning of language solely by associating audio and video signals. Credit: Mark Hamilton

Mark Hamilton, an MIT Ph.D. student in electrical engineering and computer science and affiliate of MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL), wants to use machines to understand how animals communicate. To do that, he set out first to create a system that can learn human language "from scratch."

"Funny enough, the key moment of inspiration came from the movie 'March of the Penguins.' There's a scene where a penguin falls while crossing the ice, and lets out a little belabored groan while getting up. When you watch it, it's almost obvious that this groan is standing in for a four letter word. This was the moment where we thought, maybe we need to use audio and video to learn language." says Hamilton. "Is there a way we could let an algorithm watch TV all day and from this figure out what we're talking about?"

"Our model, DenseAV, aims to learn language by predicting what it's seeing from what it's hearing, and vice-versa. For example, if you hear the sound of someone saying 'bake the cake at 350' chances are you might be seeing a cake or an oven. To succeed at this audio-video matching game across millions of videos, the model has to learn what people are talking about," says Hamilton.

A paper describing the work appears on the *arXiv* preprint server.

Once they trained DenseAV on this matching game, Hamilton and his colleagues looked at which pixels the model looked for when it heard a sound. For example, when someone says "dog," the algorithm immediately starts looking for dogs in the video stream. By seeing which pixels are selected by the algorithm, one can discover what the algorithm thinks a word means.

Interestingly, a similar search process happens when DenseAV listens to a dog barking: It searches for a dog in the video stream.

"This piqued our interest. We wanted to see if the algorithm knew the difference between the word 'dog' and a dog's bark," says Hamilton. The team explored this by giving the DenseAV a "two-sided brain." Interestingly, they found one side of DenseAV's brain naturally focused on language, like the word "dog," and the other side focused on sounds

like barking. This showed that DenseAV not only learned the meaning of words and the locations of sounds, but also learned to distinguish between these types of cross-modal connections, all without human intervention or any knowledge of written language.

One branch of applications is learning from the massive amount of video published to the internet each day.

"We want systems that can learn from massive amounts of video content, such as instructional videos," says Hamilton. "Another exciting application is understanding new languages, like dolphin or whale communication, which don't have a written form of communication. Our hope is that DenseAV can help us understand these languages that have evaded human translation efforts since the beginning. Finally, we hope that this method can be used to discover patterns between other pairs of signals, like the seismic sounds the earth makes and its geology."

A formidable challenge lay ahead of the team: Learning language without any text input. Their objective was to rediscover the meaning of language from a blank slate, avoiding using pre-trained language models. This approach is inspired by how children learn by observing and listening to their environment to understand language.

To achieve this feat, DenseAV uses two main components to process audio and visual data separately. This separation made it impossible for the algorithm to cheat, by letting the visual side look at the audio and vice versa. It forced the algorithm to recognize objects and created detailed and meaningful features for both audio and visual signals. DenseAV learns by comparing pairs of audio and visual signals to find which signals match and which signals do not. This method, called contrastive learning, doesn't require labeled examples, and allows DenseAV to figure out the important predictive patterns of language itself.

One major difference between DenseAV and previous algorithms is that prior works focused on a single notion of similarity between sound and images. An entire audio clip like someone saying "the dog sat on the grass" was matched to an entire image of a dog. This didn't allow previous methods to discover fine-grained details, like the connection between the word "grass" and the grass underneath the dog.

The team's algorithm searches for and aggregates all the possible matches between an audio clip and an image's pixels. This not only improved performance, but allowed the team to precisely localize sounds in a way that previous algorithms could not.

"Conventional methods use a single class token, but our approach compares every pixel and every second of sound. This fine-grained method lets DenseAV make more detailed connections for better localization," says Hamilton.

The researchers trained DenseAV on AudioSet, which includes 2 million YouTube videos. They also created new datasets to test how well the model can link sounds and images. In these tests, DenseAV outperformed other top models in tasks like identifying objects from their names and sounds, proving its effectiveness.

"Previous datasets only supported coarse evaluations, so we created a dataset using semantic segmentation datasets. This helps with pixel-perfect annotations for precise evaluation of our model's performance. We can prompt the algorithm with specific sounds or images and get those detailed localizations," says Hamilton.

Due to the massive amount of data involved, the project took about a year to complete. The team says that transitioning to a large transformer architecture presented challenges, as these models can easily overlook fine-grained details. Encouraging the model to focus on these details was

a significant hurdle.

Looking ahead, the team aims to create systems that can learn from massive amounts of video- or audio-only data. This is crucial for new domains where there's lots of either mode, but not together. They also aim to scale this up using larger backbones and possibly integrate knowledge from language models to improve performance.

"Recognizing and segmenting visual objects in images, as well as environmental sounds and spoken words in audio recordings, are each difficult problems in their own right. Historically researchers have relied upon expensive, human-provided annotations in order to train machine learning models to accomplish these tasks," says David Harwath, assistant professor in computer science at the University of Texas at Austin who was not involved in the work.

"DenseAV makes significant progress towards developing methods that can learn to solve these tasks simultaneously by simply observing the world through sight and sound—based on the insight that the things we see and interact with often make sound, and we also use spoken language to talk about them. This model also makes no assumptions about the specific language that is being spoken, and could therefore in principle learn from data in any language. It would be exciting to see what DenseAV could learn by scaling it up to thousands or millions of hours of video data across a multitude of languages."

Additional authors are Andrew Zisserman, professor of computer vision engineering at the University of Oxford; John R. Hershey, Google AI Perception researcher; and William T. Freeman, MIT electrical engineering and computer science professor and CSAIL principal investigator.

More information: Mark Hamilton et al, Separating the "Chirp" from the "Chat": Self-supervised Visual Grounding of Sound and Language, *arXiv* (2024). arxiv.org/abs/2406.05629

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology