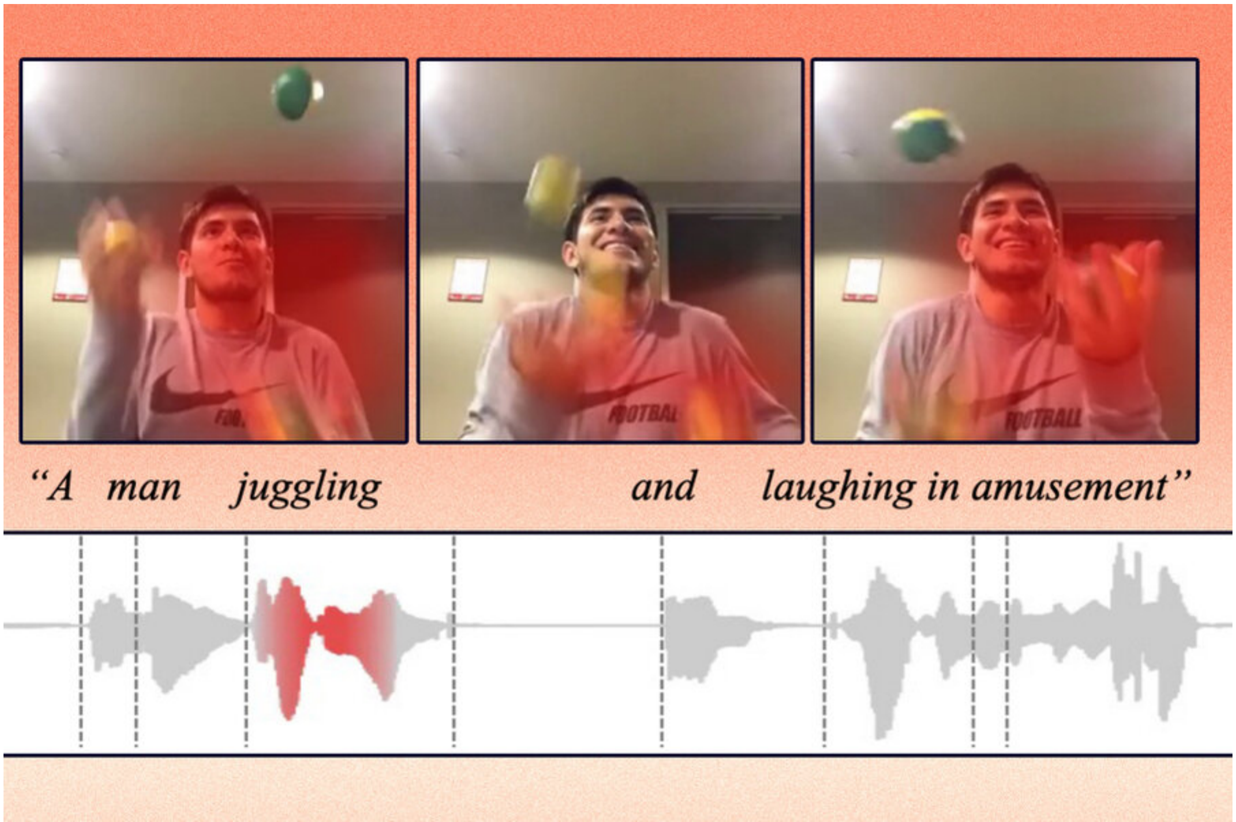
Machine-learning model can identify the action in a video clip and label it, without the help of humans

May 4 2022, by Adam Zewe



MIT researchers developed a machine learning technique that learns to represent data in a way that captures concepts which are shared between visual and audio modalities. Their model can identify where certain action is taking place in a video and label it. Credit: Massachusetts Institute of Technology

Humans observe the world through a combination of different modalities, like vision, hearing, and our understanding of language. Machines, on the other hand, interpret the world through data that algorithms can process.

So, when a machine "sees" a photo, it must encode that photo into data it can use to perform a task like image classification. This process becomes more complicated when inputs come in multiple formats, like videos, audio clips, and images.

"The main challenge here is, how can a machine align those different modalities? As humans, this is easy for us. We see a car and then hear the sound of a car driving by, and we know these are the same thing. But for machine learning, it is not that straightforward," says Alexander Liu, a graduate student in the Computer Science and Artificial Intelligence Laboratory (CSAIL) and first author of a paper tackling this problem.

Liu and his collaborators developed an artificial intelligence technique that learns to represent data in a way that captures concepts which are shared between visual and audio modalities. For instance, their method can learn that the action of a baby crying in a video is related to the spoken word "crying" in an audio clip.

Using this knowledge, their machine-learning model can identify where a certain action is taking place in a video and label it.

It performs better than other machine-learning methods at cross-modal retrieval tasks, which involve finding a piece of data, like a video, that matches a user's query given in another form, like spoken language. Their model also makes it easier for users to see why the machine thinks the video it retrieved matches their query.

This technique could someday be utilized to help robots learn about concepts in the world through perception, more like the way humans do.

Joining Liu on the paper are CSAIL postdoc SouYoung Jin; grad students Cheng-I Jeff Lai and Andrew Rouditchenko; Aude Oliva, senior research scientist in CSAIL and MIT director of the MIT-IBM Watson AI Lab; and senior author James Glass, senior research scientist and head of the Spoken Language Systems Group in CSAIL. The research will be presented at the Annual Meeting of the Association for Computational Linguistics.

Learning representations

The researchers focus their work on representation learning, which is a form of machine learning that seeks to transform input data to make it easier to perform a task like classification or prediction.

The representation learning model takes raw data, such as videos and their corresponding text captions, and encodes them by extracting features, or observations about objects and actions in the video. Then it maps those data points in a grid, known as an embedding space. The model clusters similar data together as single points in the grid. Each of these data points, or vectors, is represented by an individual word.

For instance, a video clip of a person juggling might be mapped to a vector labeled "juggling."

The researchers constrain the model so it can only use 1,000 words to label vectors. The model can decide which actions or concepts it wants to encode into a single vector, but it can only use 1,000 vectors. The model chooses the words it thinks best represent the data.

Rather than encoding data from different modalities onto separate grids, their method employs a shared embedding space where two modalities can be encoded together. This enables the model to learn the relationship between representations from two modalities, like video that shows a person juggling and an audio recording of someone saying "juggling."

To help the system process data from multiple modalities, they designed an algorithm that guides the machine to encode similar concepts into the same vector.

"If there is a video about pigs, the model might assign the word 'pig' to one of the 1,000 vectors. Then if the model hears someone saying the word 'pig' in an audio clip, it should still use the same vector to encode that," Liu explains.

A better retriever

They tested the model on cross-modal retrieval tasks using three datasets: a video-text dataset with video clips and text captions, a video-audio dataset with video clips and spoken audio captions, and an image-audio dataset with images and spoken audio captions.

For example, in the video-audio dataset, the model chose 1,000 words to represent the actions in the videos. Then, when the researchers fed it audio queries, the model tried to find the clip that best matched those spoken words.

"Just like a Google search, you type in some text and the machine tries to tell you the most relevant things you are searching for. Only we do this in the vector space," Liu says.

Not only was their technique more likely to find better matches than the models they compared it to, it is also easier to understand.

Because the model could only use 1,000 total words to label vectors, a user can more see easily which words the machine used to conclude that the video and spoken words are similar. This could make the model easier to apply in real-world situations where it is vital that users understand how it makes decisions, Liu says.

The model still has some limitations they hope to address in future work. For one, their research focused on data from two modalities at a time, but in the real world humans encounter many data modalities simultaneously, Liu says.

"And we know 1,000 words works on this kind of dataset, but we don't know if it can be generalized to a real-world problem," he adds.

Plus, the images and videos in their datasets contained simple objects or straightforward actions; real-world data are much messier. They also want to determine how well their method scales up when there is a wider diversity of inputs.

More information: Alexander H. Liu et al, Cross-Modal Discrete Representation Learning. arXiv:2106.05438v1 [cs.CV], arxiv.org/abs/2106.05438

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

Citation: Machine-learning model can identify the action in a video clip and label it, without the help of humans (2022, May 4) retrieved 25 April 2024 from https://techxplore.com/news/2022-05-machine-learning-action-video-humans.html