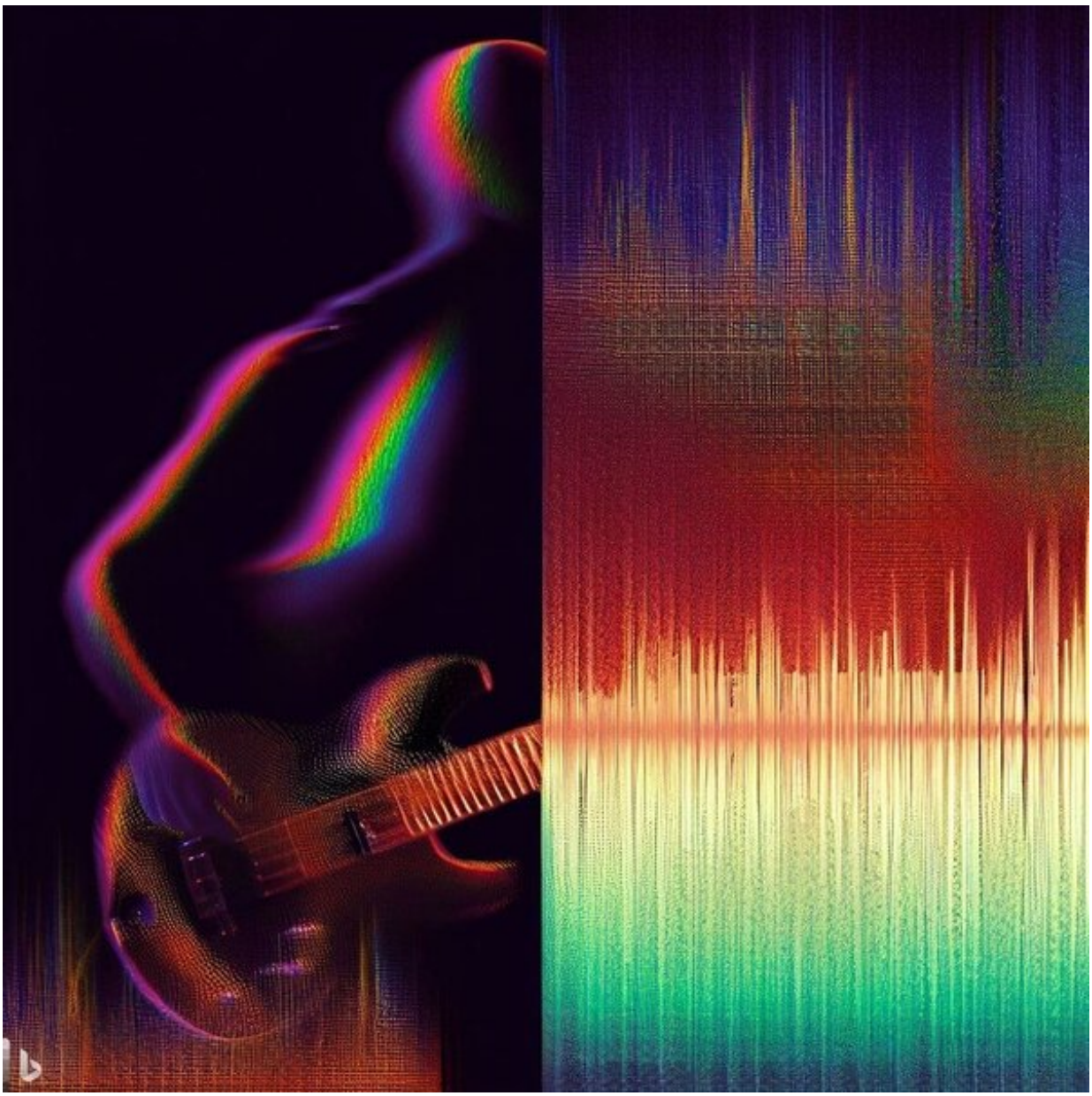
Multimodal technique for analyzing audio and visual data improves performance of machine-learning models

June 6 2023, by Lauren Hinkel



A new machine-learning technique allows for more efficient multi-modal learning. Credit: Lauren Hinkel/MIT-IBM Watson AI Lab, via Bing Create

Researchers from MIT, the MIT-IBM Watson AI Lab, IBM Research, and elsewhere have developed a new technique for analyzing unlabeled audio and visual data that could improve the performance of machine-learning models used in applications like speech recognition and object detection. The work, for the first time, combines two architectures of self-supervised learning, contrastive learning and masked data modeling, in an effort to scale machine-learning tasks like event classification in single- and multimodal data without the need for annotation, thereby replicating how humans understand and perceive our world.

"A larger portion of human knowledge is learned in a self-supervised way, because we don't always get supervision signals, and we want to enable the machine-learning model to have the same ability," says Yuan Gong, an MIT postdoc in the Computer Science and Artificial Intelligence Laboratory (CSAIL).

"So, another way to put it is that self-supervised learning often forms the foundation of an initial model, because it can learn on vast amounts of unlabeled data. And then you can use classical, supervised learning or reinforcement learning to fine tune the model to something particular if you want to," says Jim Glass, an MIT senior research scientist and member of the MIT-IBM Watson AI Lab.

The technique, called the contrastive audio-visual masked autoencoder (CAV-MAE), is a type of neural network that can learn to extract and map meaningful latent representations into high-dimensional space from

acoustic and visual data by training on large YouTube datasets of audio and video 10-second clips. The researchers say the technique is more effective than previous approaches because it explicitly models the relationships between audio and visual data in a way that other methods do not.

Joining Gong and Glass on the study are graduate students Andrew Rouditchenko and Alexander H. Liu of MIT, David Harwath Ph.D. '18 of the University of Texas at Austin, and MIT-IBM Watson AI Lab members Leonid Karlinsky and Hilde Kuehne. Kuehne is also affiliated with Goethe University Frankfurt. The method was recently presented at the International Conference on Learning Representations.

A joint and coordinated approach

The CAV-MAE works by "learning by prediction" and "learning by comparison," says Gong. The masked data modeling, or the prediction method, takes a video along with its coordinated audio waveform, converts the audio to a spectrogram, and masks 75% of both. The unmasked data is tokenized, then fed into separate audio and visual encoders before entering a joint encoder/decoder, where the model is asked to recover the missing data. The difference (reconstruction loss) between the resulting reconstructed prediction and the original audio-visual combination is then used to train the model for better performance.

An example of this would be covering part of a video of a piano and part of a spectrogram of piano music, and then asking the model to try to determine the masked inputs. Unfortunately, this method may not capture the association between the video and audio pair, whereas contrastive learning leverages this, but may discard some modality-unique information, like the background in a video.

Contrastive learning aims to map representations that are similar close to each other. For example, the model will attempt to place different video and audio data of different parrots close to each other and further away from pairs of video and audio of guitars playing. In a similar fashion to masked autoencoding, audio-visual pairs are passed into separate modality encoders; however, the audio and visual components are kept separately within the joint encoder before the model performs pooling and contrastive loss. In this way, contrastive learning tries to identify the parts of each audio or video that are most relevant to the other.

For example, if a video shows someone speaking and the corresponding audio clip contains speech, the autoencoder will learn to associate the mouth movements of the speaker with the words being spoken. It will then adjust the model's parameters so that those inputs are represented close to each other. Ultimately, the CAV-MAE method combines both techniques with multiple forward data streams with masking as a first step, modality-specific encoders, and layer normalization so that the representation strengths are similar.

"We [then] wanted to compare the proposed CAV-MAE with a model trained only with a masked autoencoder and a model trained only with contrastive learning, because we want to show that by combining masked autoencoder and contrastive learning, we can get some performance improvement," says Gong, "and the results support our hypothesis that there's obvious improvement."

The researchers tested CAV-MAE—as well as their method without contrastive loss or a masked autoencoder—against other state-of-the-art methods on audio-visual retrieval and audio-visual event classification tasks using standard AudioSet (20K and 2M) and VGGSound datasets—labeled, realistic short clips, which could include multiple sounds. Audio-visual retrieval means that the model sees either the audio or visual component of a query pair and searches for the missing one;

event classification includes identifying actions or sounds within data, like a person singing or a car driving.

Overall, they found that contrastive learning and masked data modeling are complementary methods. CAV-MAE was able to outperform previous techniques (with fully self-supervised pre-training) by about 2% for event classification performance verses models with comparable computation and, more impressively, kept pace with or outperformed models with industry-level computational resources. The team's model ranked similarly to models trained with only the contrastive loss. And surprisingly, the team says, the incorporation of multi-modal data into CAV-MAE pre-training greatly improves the fine-tuning of single-modality representation via supervised learning (with some labeled data) and performance on audio-only event classification tasks.

This demonstrates that, like humans, multi-modal information provides an additional "soft label" boost even for audio or visual only tasks; for instance, it helps the model to understand if it's looking for an electric or acoustic guitar—a richer supervision signal.

"I think people like the elegance of this model for combining information in the different audio and visual streams. It has the contrastive and the reconstruction loss, and compared to models that have been evaluated with similar data, it clearly does very well across a range of these tasks," says Glass.

Building on this, "one special thing is, our model can do both classification and the retrieval, which is not common," Gong adds. "Before this work, these methods are used separately, but after this work, I see that most of the audio-visual learning frameworks use contracting loss and the masked autoencoder together, implicitly or explicitly."

Bringing self-supervised audio-visual learning into our world

The researchers see their contribution of the contrastive audio-visual masked autoencoder (CAV-MAE) as an important milestone and a step forward for applications, which are increasingly moving from single modality to multi-modality and which require or leverage audio-visual fusion. They hypothesize that one day it could be used for action recognition in realms like sports, education, entertainment, motor vehicles, and public safety. It could also, one day, extend to other modalities.

At this time, the fact that, "this only applies to audio-visual data may be a limitation, but we are targeting multi-modal learning, which is trend of machine learning," says Gong. "As humans, we have multi-modalities—we have smell, touch—many more things that just audio-visual. So, when we try to build AI, we try to mimic humans somehow, not necessarily from the biological perspective, and this method could [potentially be] generalized to other unexplored modalities."

As machine-learning models continue to play an increasingly important role in our lives, techniques like this one will become increasingly valuable.

More information: Yuan Gong et al, Contrastive audio-visual masked autoencoder. openreview.net/pdf?id=QPtMRyk5rb

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