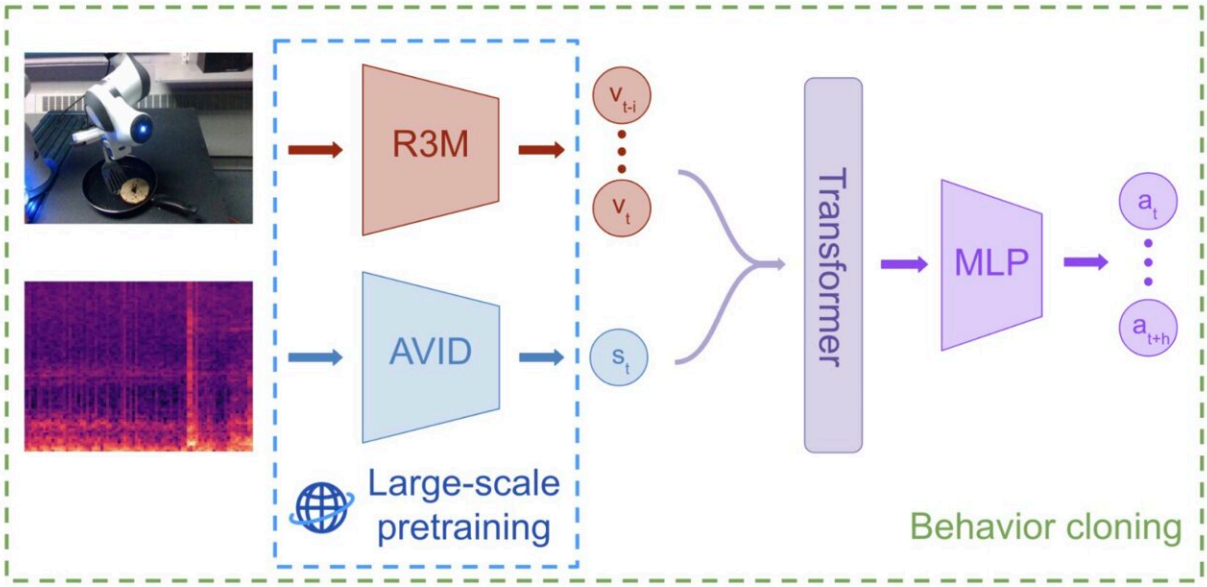
Using contact microphones as tactile sensors for robot manipulation

May 30 2024, by Ingrid Fadelli



Two-stage model training. AVID and R3M pretraining leverages the large scale of internet video data (blue dashed box). We initialize the vision and audio encoders with the resulting pre-trained representations and then train the entire policy end-to-end with behavior cloning from a small number of in-domain demonstrations. The policy takes image and spectrogram inputs (left) and outputs a sequence of actions in delta end effector space (right). Credit: Mejia et al.

To complete real-world tasks in home environments, offices and public

spaces, robots should be able to effectively grasp and manipulate a wide range of objects. In recent years, developers have created various machine learning–based models designed to enable skilled object manipulation in robots.

While some of these models achieved good results, to perform well they typically need to be pre-trained on large amounts of data. The datasets used to train these models are primarily comprised of visual data, such as annotated images and video footage captured using cameras, yet some approaches also analyze other sensory inputs, such as tactile information.

Researchers at Carnegie Mellon University and Olin College of Engineering recently explored the possibility of using contact microphones instead of conventional tactile sensors, thus enabling the use of audio data to train machine learning models for robot manipulation. Their paper, posted to the preprint server *arXiv*, could open new opportunities for the large-scale multi-sensory pre-training of these models.

"Although pre-training on a large amount of data is beneficial for robot learning, current paradigms only perform large-scale pretraining for visual representations, whereas representations for other modalities are trained from scratch," Jared Mejia, Victoria Dean and their colleagues wrote in the paper.

"In contrast to the abundance of visual data, it is unclear what relevant internet-scale data may be used for pretraining other modalities such as tactile sensing. Such pretraining becomes increasingly crucial in the low-data regimes common in robotics applications. We address this gap using contact microphones as an alternative tactile sensor."

As part of their recent study, Mejia, Dean and their collaborators pre-trained a self-supervised machine learning approach on audio-visual

representations from the Audioset dataset, which contains more than 2 million 10-second video clips of sounds and music clips collected from the internet. The model they pre-trained relies on audio-visual instance discrimination (AVID), a technique that can learn to distinguish between different types of audio-visual data.

The researchers assessed their approach in a series of tests, where a robot was tasked with completing real-world manipulation tasks relying on a maximum of 60 demonstrations for each task. Their findings were highly promising, as their model outperformed policies for robot manipulation that only rely on visual data, particularly in instances where objects and locations were markedly different from those included in the training data.

"Our key insight is that contact microphones capture inherently audio-based information, allowing us to leverage large-scale audio-visual pretraining to obtain representations that boost the performance of robotic manipulation," Mejia, Dean and their colleagues wrote. "To the best of our knowledge, our method is the first approach leveraging largescale multisensory pre-training for robotic manipulation."

In the future, the study by Mejia, Dean and their colleagues could open a new avenue for the realization of skilled robot manipulation utilizing pre-trained multimodal machine learning models. Their proposed approach could soon be improved further and tested on a broader range of real-world manipulation tasks.

"Future work may investigate which properties of pre-training datasets are most conducive to learning audio-visual representations for manipulation policies," Mejia, Dean and their colleagues wrote. "Further, a promising direction would be to equip end-effectors with visuo-tactile sensors and contact microphones with pre-trained audio representations to determine how to leverage both for equipping robotic

agents with a richer understanding of their environment."

More information: Jared Mejia et al, Hearing Touch: Audio-Visual Pretraining for Contact-Rich Manipulation, *arXiv* (2024). DOI: 10.48550/arxiv.2405.08576