Emotion recognition has a privacy problem—here's how to fix it

February 21 2020, by Steve Crang

Credit: CC0 Public Domain

With devices listening everywhere you go, privacy concerns are endemic to advancing technology. Especially sensitive are different techniques powered by audio from your smartphones and speakers, putting consumers in a constant cost-benefit analysis between privacy and utility.

Take, for instance, a mobile app or virtual assistant that can learn to adapt to a users' mood and recognize emotions in real time. This sort of adaptation can create more naturally flowing conversations, and more useful, human-like understanding from voice assistants. But where does the user draw the line if the audio powering these insights was stored full of identifiers about their gender and demographic information?

A new paper by CSE Ph.D. student Mimansa Jaiswal and Prof. Emily Mower Provost proposes a method to remove this barrier and enable more secure technologies built on machine learning (ML). Through the use of adversarial ML, they've demonstrated the ability to "unlearn" these sensitive identifiers from audio before it's stored, and instead use stripped-down representations of the speaker to train emotion recognition models.

Emotion recognition, sentiment analysis, and other techniques for automatically identifying different complex features of speech are powered by ML models trained on huge stores of labeled data. In order to reliably pick out patterns in a user's speech, the model has to have significant training experience with similar speech that helps it identify certain common features.

These systems that deal with the day-to-day lives of typical smartphone users will then have to be trained on a wide range of ordinary human speech—essentially, recordings of conversations.

"The hope of this paper is to show that these machine learning algorithms end up encoding quite a lot of information about a person's gender or demographic information," says Jaiswal. This demographic information is stored on company servers that power a particular mobile app or voice assistant—leaving the user open to identification by the company or, worse, any malicious eavesdroppers.

"The implications of sensitive information leakage is profound," the authors write. "Research has shown that discrimination occurs across variables of age, race, and gender in hiring, policing, and credit ratings."

This identifying audio data, stored in its raw form, could even override opt-out options selected by the user elsewhere in the app. To handle this, services moved to storing representations obtained after pre-processing on the cloud, to avoid information leakage.

Previous work on encoding audio data with privacy in mind tried adding random noise to the dataset. While the technique worked if the listener had no knowledge of what sort of noise was used, the instant the attacker was able to access the network generating the anonymity the method fell apart.

Instead, Jaiswal and Mower Provost use adversarial ML techniques to reduce the encoding of demographic and private features from the raw audio before it's ever stored. What remains is an abstracted data representation of the original recording. The authors use these representations to partially obfuscate the actual content of the conversation, eliminating the risks to privacy that come with wholesale data storage.

The challenge was, then, to ensure that this new format of privacy-protected data could still be used to train ML models effectively on their main task. What the researchers found was that as the strength of the adversarial component increases, the privacy metric mostly increases—and the performance on the primary task is unchanged, or is only minorly affected.

"We find that the performance is either maintained, or there is a slight decrease in performance for some setups," the authors write. In multiple cases they even identified a significant increase in performance,

implying that making the model blind to gender increases its robustness by not learning associations between gender and emotion labels.

Jaiswal hopes to use these findings to make machine learning research safer and more secure for users in the real world.

"ML models are mostly black box models," she says, "meaning you don't usually know what exactly they encode, what information they have, or whether that information can be used in a good or malicious way. The next step is to understand the difference in information being encoded between two models where the only difference is that one has been trained to protect privacy."

"We want to improve how humans perceive and interact with these models."

This research was published in the paper "Privacy Enhanced Multimodal Neural Representations for Emotion Recognition," published at the 2020 Association for the Advancement of Artificial Intelligence (AAAI) Conference.

More information: Privacy Enhanced Multimodal Neural Representations for Emotion Recognition, arXiv:1910.13212 [cs.LG] arxiv.org/abs/1910.13212

Provided by University of Michigan