Seeking a way of preventing audio models for AI machine learning from being fooled

January 6 2022

Jon Vadillo, in his office at the University of The Basque Country. Credit: Nagore Iraola, UPV/EHU

Artificial intelligence (AI) is increasingly based on machine learning models, trained using large datasets. Likewise, human-computer interaction is increasingly dependent on speech communication, mainly

due to the remarkable performance of machine learning models in speech recognition tasks.

However, these models can be fooled by "adversarial" examples; in other words, inputs intentionally perturbed to produce a wrong prediction without the changes being noticed by humans. "Suppose we have a model that classifies audio (e.g., voice command recognition) and we want to deceive it; in other words, generate a perturbation that maliciously prevents the model from working properly. If a signal is heard properly, a person is able to notice whether a signal says 'yes,' for example. When we add an adversarial perturbation we will still hear 'yes,' but the model will start to hear 'no,' or 'turn right' instead of left or any other command we don't want to execute," explained Jon Vadillo, researcher in the UPV/EHU's Departament of Computer Science and Artificial Intelligence.

This could have "very serious implications at the level of applying these technologies to real-world or highly sensitive problems," added Vadillo. It remains unclear why this happens. Why would a model that behaves so intelligently suddenly stop working properly when it receives even slightly altered signals?

Deceiving the model by using an undetectable perturbation

"It is important to know whether a model or a program has vulnerabilities," added the researcher from the Faculty of Informatics. "Firstly, we investigate these vulnerabilities, to check that they exist, and because that is the first step in eventually fixing them." While much research has focused on the development of new techniques for generating adversarial perturbations, less attention has been paid to the aspects that determine whether these perturbations can be perceived by

humans and what these aspects are like. This issue is important, as the adversarial perturbation strategies proposed only pose a threat if the perturbations cannot be detected by humans.

This study has investigated the extent to which the distortion metrics proposed in the literature for audio adversarial examples can reliably measure the human perception of perturbations. In an experiment in which 36 people evaluated adversarial examples or audio perturbations according to various factors, the researchers showed that "the metrics that are being used by convention in the literature are not completely robust or reliable. In other words, they do not adequately represent the auditory perception of humans; they may tell you that a perturbation cannot be detected, but then when we evaluate it with humans, it turns out to be detectable. So we want to issue a warning that due to the lack of reliability of these metrics, the study of these audio attacks is not being conducted very well," said the researcher.

In addition, the researchers have proposed a more robust evaluation method that is the outcome of the "analysis of certain properties or factors in the audio that are relevant when assessing detectability, for example, the parts of the audio in which a perturbation is most detectable." Even so, "this problem remains open because it is very difficult to come up with a mathematical metric that is capable of modeling auditory perception. Depending on the type of audio signal, different metrics will probably be required or different factors will need to be considered. Achieving general audio metrics that are representative is a complex task," concluded Vadillo.

More information: Jon Vadillo et al, On the human evaluation of universal audio adversarial perturbations, *Computers & Security* (2021). DOI: 10.1016/j.cose.2021.102495

Provided by University of the Basque Country