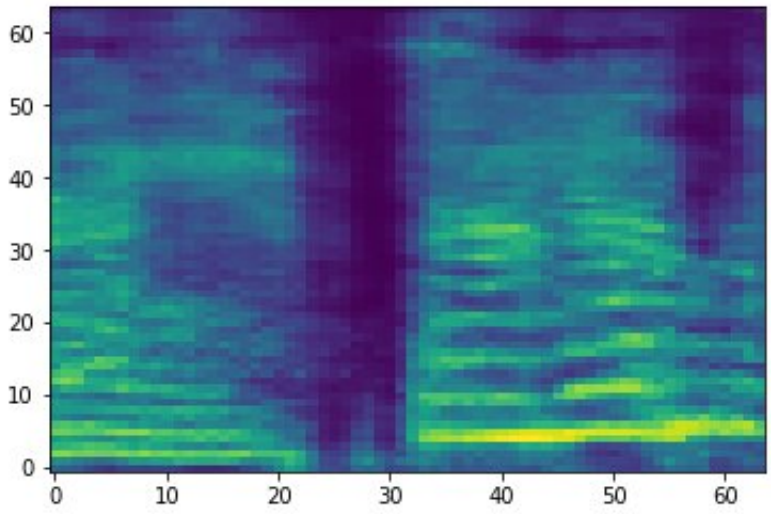
An approach for securing audio classification against adversarial attacks

May 7 2019, by Ingrid Fadelli



Spectrogram of a random audio signal. Credit: Esmailpour, Cardinal & Lemeiras Koerich.

Adversarial audio attacks are small perturbations that are not perceivable by humans and are intentionally added to audio signals to impair the performance of machine learning (ML) models. These attacks raise serious concerns about the security of ML models, as they can cause them to make mistakes and ultimately generate wrong predictions.

Researchers at École de Technologie Supérieure, part of the University of Quebec in Canada have recently developed a new approach that could help to secure audio classification tools against adversarial attacks. In

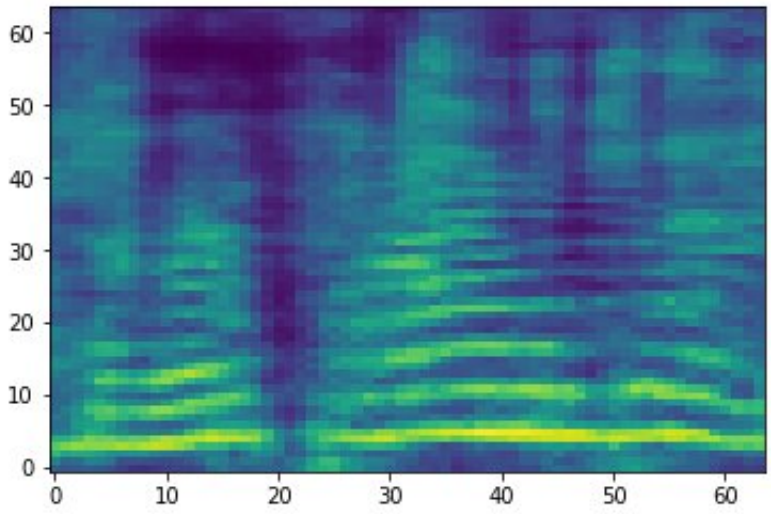
their paper, pre-published on arXiv, they review some of the strongest existing adversarial attacks and their impact on the performance of common ML models, then propose an approach that could counteract these attacks.

"At the moment, there are many strong and fast (at runtime) classifiers in terms of accuracy, namely deep learning classifiers (e.g. convolutional neural networks), which can even outperform human level of media (e.g. speech, image, video, animation, text, etc.) recognition and regression," Mohammad Esmaeilpour, one of the researchers who carried out the study, told TechXplore. "The Achilles heel of these advanced algorithms is their vulnerability to inputs that contain carefully crafted perturbations, known as adversarial attacks."

Adversarial attacks work by producing samples that closely resemble legitimate training samples, but that actually lead a ML model or models to generate wrong labels with high confidence levels. In ML research, if there is enough data to train a classifier, the key challenge is no longer improving its recognition accuracy, but ensuring its resilience against adversarial attacks.

"Adversarial attacks are active threats for all data-driven algorithms, even those trained on small datasets," Esmaeilpour said. "This sparked our interest to study the threat of adversarial attacks for audio and speech recognition applications, since all smartphones are now equipped with a virtual speech assistant such as Siri, Google assistant and Cortana."

In their study, Esmaeilpour and his colleagues carried out experiments involving environmental audio datasets, rather than speech datasets. Nonetheless, in the future their approach could also potentially be extended to speech recognition, which would help to secure voice assistants against adversarial attacks.



Crafted adversarial spectrogram associated with the audio signal in the first image. While the two images are similar, they have different labels, suggesting that an attack is taking place. Credit: Esmailpour, Cardinal & Lemeiras Koerich.

"Our main objective in this paper was to study the threat of adversarial attacks for both conventional and deep learning audio classifiers and ideally propose a more reliable algorithm in terms of resiliency against some common attacks as a baseline towards real robust audio classification," Esmaeilpour explained. "We wanted to make a fair balance for classifiers in recognition accuracy, computational complexity, and robustness against adversarial attacks."

Generally, classifiers that are more robust against adversarial attacks attain lower recognition accuracy, and vice versa. In their study, the researchers focused on adversarial retraining, one of the most valid existing defense techniques that do not obfuscate gradient information. Despite its benefits, this particular defense strategy is costly (as strong attacks are costly, adversarial retraining using these attacks will be more costly) and can negatively affect a classifier's recognition performance.

"The ideal case for us would be to propose a gradient obfuscation-free and adversarial retraining-free audio classifier which inherently learns 'robust features'," Esmaeilpour said. "Our classification scenario includes several steps, mainly spectrogram (2D representation for audio signals) enhancement, dimensionality reduction using an algebraic decomposition technique, and smoothing by utilizing a convolutional de-noising autoencoder, where the last two steps (stacked together) have shown positive impacts on removing small unknown potential adversarial perturbations."

After reviewing some of the strongest adversarial attacks out there and their effects on the performance of ML models, the researchers extracted features from the spectrograms processed by the models, organized them into a codebook and trained a support vector machine (SVM) algorithm on this codebook. In their training pipeline, they did not implement any proactive or reactive adversarial attack detection techniques or defense algorithms.

"Our main goal was to 'learn robust feature vectors' without any pre- or post-processing overhead for detecting potential adversarial samples," Esmaeilpour explained. "Our results show that our proposed classifier outperforms state-of-the-art deep learning and conventional algorithms against five types of strong adversarial attacks for some practical environmental audio datasets."

Esmaeilpour and his colleagues statistically proved the vulnerability of both conventional classifiers (i.e. classifiers that learn from feature space) and deep learning algorithms (i.e. algorithms that learn from raw data) against adversarial attacks. According to the researchers, there is currently no reliable data-driven algorithm for audio classification that is also robust against adversarial attacks. Among existing models, deep-learning-based approaches appear to be the least secure against these attacks, even if they typically attain the highest recognition accuracy.

"The classification scenario we proposed in our paper uses a SVM with polynomial kernel as a final classifier," Esmaeilpour said. "However, applying a convolutional de-noising autoencoder on top of singular value decomposition followed by an unsupervised clustering of extracted speeded-up robust feature vectors could help to learn more structural components and probably robust features, which could allow us to attain a reasonable balance between recognition accuracy (comparable to state-of-the-art performance) and robustness against five common strong adversarial attacks."

While the results gathered by the researchers are very promising, they might vary according to the dataset used or a classifier's specific application, hence they are not yet generalizable. In the future, their study could inform the development of other classifiers that are better equipped against adversarial attacks, without presenting substantial losses in performance (i.e. recognition accuracy).

"Learning robust features is an open problem and we still do not have a clear idea of how to properly address it; it is being studied by our research team and some results will be released soon," Esmaeilpour said. "Meanwhile, we are working on a new, strong and fast adversarial attack technique aimed at utilizing this attack to adversarially train the learning model (which improves its robustness) and also save the recognition performance of the model before training it."

More information: A robust approach for securing audio classification against adversarial attacks. arXiv:1904.10990v1 [cs.LG]. arxiv.org/abs/1904.10990