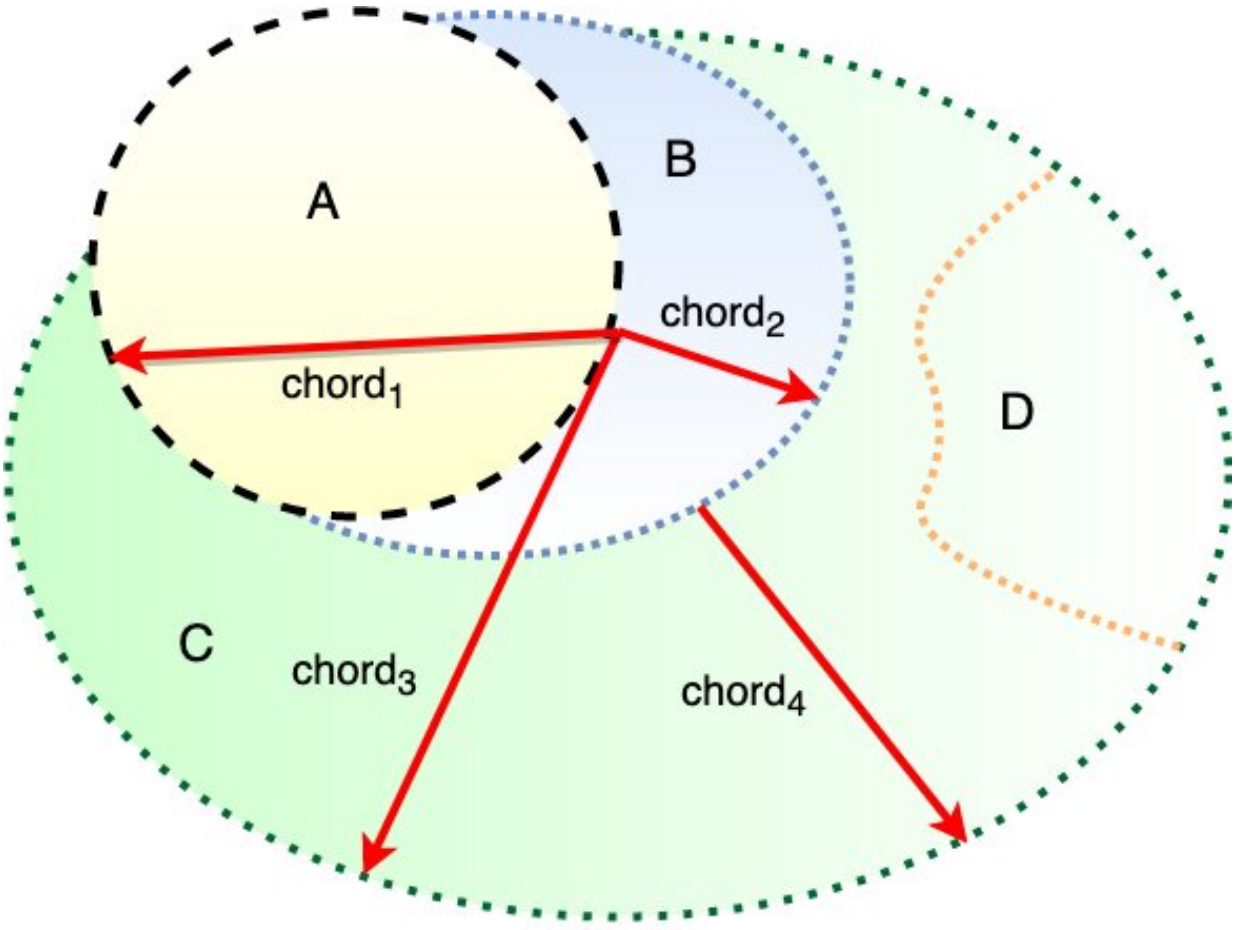
A method to protect audio classifiers against adversarial attacks

June 25 2020, by Ingrid Fadelli



(Chord) A typical visualization of adversarial, noisy and real subspaces with their associated chordal distances. The image denotes: A) legitimate manifold; B) noisy manifold; C) adversarial manifold; and D) ill-conditioned subspace. Credit: Esmaeilpour et al.

In recent years, machine learning algorithms have attained remarkable results in a variety of tasks, including the classification of both images and audio files. A class of algorithms that has proven to be particularly promising are deep neural networks (DNNs) that can automatically learn to solve specific problems by analyzing large quantities of data.

DNNs are data-driven techniques, which means that they need to be trained on large quantities of data to learn to classify new information most effectively. Their dependence on such training data makes this type of algorithm quite vulnerable. In fact, even If they are efficiently trained, DNNs can be easily tricked into classifying data incorrectly.

Past studies have found that cyber attackers can easily trick DNNs by subtly modifying a real image or audio file and creating an artificial replica, known as an adversarial image/audio. The deep learning architecture would then incorrectly classify this adversarial data, allowing malicious users to access private information or disrupt the model's overall functioning. This method of fooling DNNs is known as adversarial attack.

Researchers at École de Technologie Supérieure (ÉTS) in Canada have recently developed a method to protect models for classifying environmental sounds from adversarial attacks. This method, presented at the 45th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), entails the use of a detector that can measure differences between legitimate and malicious sound representations, thus enhancing the reliability of audio classifiers.

"Generally, a classifier learns a decision boundary (a nonlinear function) among different classes to discriminate between them," Mohammad Esmaeilpour, one of the researchers who carried out the study, told TechXplore. "One can modify this decision boundary so that the sample goes over it by reducing sensitivity of the learned nonlinear function to

the correct class of the sample and increasing the chance of misclassification. This can be achieved by running optimization algorithms against victim DNNs, which is known as an adversarial attack."

In a simple example, DNNs can be trained to complete binary classification tasks, which involve classifying data into two categories, such as A and B. To carry out an adversarial attack, an attacker runs optimization algorithms on a DNN and generates samples that are visually similar to class A, but that the model will mistakenly and confidently classify as B.

Recent advances in computer science have enabled the development of increasingly advanced optimization algorithms, which greatly facilitates adversarial attacks. While several researchers have been trying to come up with techniques to protect classifiers against these attacks, none of these techniques has so far proved to be fully effective. In order to create an effective tool to protect classifiers against adversarial attacks, it is first necessary to better understand these attacks and their characteristics.

An original audio (top) vs an adversarial audio (bottom). The two audio signals look very similar, but the finely trained DNNs interprets them entirely differently. Credit: Esmaeilpour et al.

"Unfortunately, it is not really possible to demonstrate subspaces of adversarial examples in the Cartesian space (our natural living space) and compare it with the subspace of real samples, since they have too much overlap," Esmaeilpour explained. "Therefore, in our research, we ended up with unitary space of Schur decomposition for characterization of adversarial subspaces."

Esmaeilpour and his colleagues used a chordal distance metric to discriminate samples in nonadjacent subspaces and found that

adversarial audio representations diverged from both real and noisy audio samples in a number of ways. These differences ultimately allowed them to discriminate between adversarial and original audio files in the unitary Schur vector space.

Subsequently, the researchers devised a detector based on eigenvalues of samples represented in this vector space. This detector was found to outperform previously developed state-of-the-art techniques for detecting adversarial data in the vast majority of test cases.

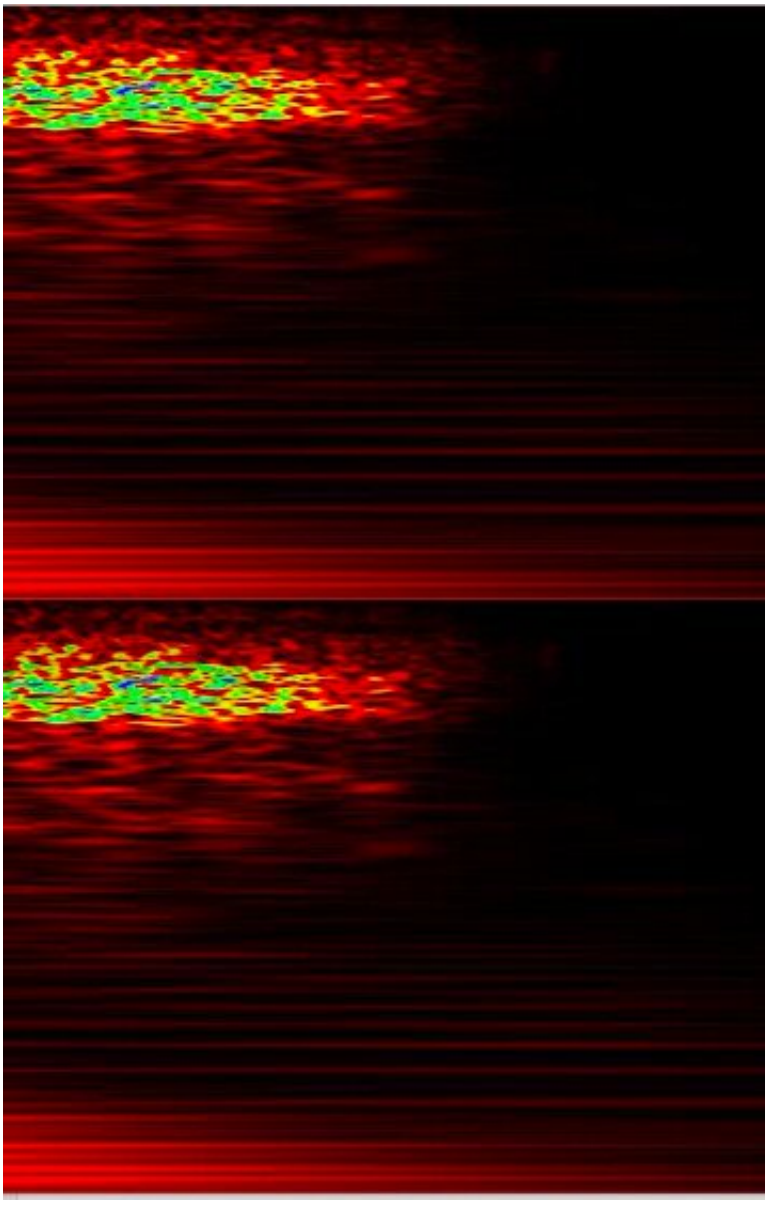
"We recently published a paper in the journal *IEEE Transactions on Information Forensics and Security*, where we used a decomposition approach similar to Schur," Esmaeilpour said. "We implemented singular values decomposition for spectrogram enhancement. While working on this approach, we noticed spectacular properties of unitary space. This aroused my personal interest to read more about these spaces, and eventually, I came up with the idea of exploring these spaces for adversarial example studies."

The generalized Schur decomposition, also known as QZ decomposition, is a mathematical method that transforms a given matrix into three subsequent pseudo-normal matrices (i.e., eigenvectors and eigenvalues) with perpendicular spans. This method can serve as a baseline to reconstruct any matrix using eigenvectors promoted by eigenvalue coefficients.

In this context, eigenvalues hold structural components of a given sample and can represent them based on a number of dimensions. Ultimately, this can help to get rid of the subspace overlap, highlighting differences between different items.

The technique devised by Esmaeilpour and his colleagues uses Schur decomposition to discern between original and adversarial audio files.

The detector processes test samples, extracts their Schur eigenvalues and then verifies whether they are original or adversarial in real time using a pre-trained regression model.



Spectrogram of an original audio (top) vs spectrogram of an adversarial audio (bottom). The two spectrograms look very similar, but the finely trained DNNs interprets them entirely differently. Credit: Esmaeilpour et al.

This regression model is fast at runtime and can also be used as a proactive module for any classifier. It is particularly well suited for the analysis of spectrograms associated with short audio signals. Spectrograms are 2-D representations of audio and speech signals that illustrate their frequency information.

"The main contribution of our recent paper is the study of the adversarial subspace and characterization of adversarial examples in a non-Cartesian space where the majority of the introduced detectors do not work," Esmaeilpour said. "We hypothesized that difficulties in generalizing common adversarial detectors to other datasets or tasks are due to measuring sample similarities/distributions in non-orthonormal Cartesian space."

In a series of initial evaluations, the researchers found that their method can finely discriminate between any adversarial and legitimate audio samples in a vector space. Interestingly, it can also be encoded into almost any classifier, and could thus potentially prevent a number of DNN-based techniques from being fooled by adversarial attacks.

"Without lack of generalizability, since our proposed detector is primarily developed for spectrogram (short-time Fourier transformation, Mel-frequency cepstral coefficients, discrete wavelet transformation, etc.), audio and speech processing systems could use this metric for improving the robustness of their DNNs against both targeted/non-targeted, white/black-box adversarial attacks," Esmaeilpour said.

In the future, the reported technique could reduce the vulnerability of existing or newly developed classifiers to adversarial attacks, which could have implications for several applications. For instance, the detector could increase the reliability of biometric identification tools based on DNNs.

"Adversarial detection is an open problem, and the path toward developing a robust and multipurpose classifier is still long," Esmaeilpour said "In my next studies, I would like to improve our proposed detector using an enhanced version of the chordal distance encoding. Moreover, I am really keen on exploring other vector spaces to even better characterize and visualize adversarial manifolds."

More information: Mohammad Esmaeilpour et al. Detection of Adversarial Attacks and Characterization of Adversarial Subspace, *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020). DOI: 10.1109/ICASSP40776.2020.9052913

Mohammad Esmaeilpour et al. A Robust Approach for Securing Audio Classification Against Adversarial Attacks, *IEEE Transactions on Information Forensics and Security* (2019). DOI: 10.1109/TIFS.2019.2956591