Contribution ID: **7** Type: **Spotlight talk**

Geometric approach for Identification of Adversarial Examples

Monday 6 December 2021 11:15 (15 minutes)

Deep learning methods have found profound success in recent years in solving complex tasks such as in the field of computer vision, speech recognition, and security applications. The robustness of these deep learning models has been found to be vulnerable to adversarial examples. These are perturbed samples, which are imperceptible to the human eye, that lead the model to erroneous output decisions. In this study, we adapt and introduce two geometric metrics, namely density, and coverage, and evaluate their use in detecting adversarial samples in batches. We empirically study the metrics using MNIST and real-world biomedical datasets, from MedMNIST, subjected to two different adversarial attacks. Our experiments show promising results.

Physical Presentation

I would not feel comfortable to present in front of an audience and prefer a video (call) presentation.

Primary authors: VENKATESH, Danush Kumar; STEINBACH, Peter (FWCC)

Presenter: VENKATESH, Danush Kumar

Session Classification: Parallel Session