מכון ויצמן למדע

THE WEIZMANN INSTITUTE OF SCIENCE

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Vision and Robotics Seminar

on Thursday, Apr 22, 2021
at 12:15

Schmidt Hall and via Zoom:
https://weizmann.zoom.us/j/97445179862?pwd=VGVTdzNVU2VnTU1USmdtemNiUEtNUT09
** Note that attendance is limited to 50 people under the 'green and purple badges' **
Covid-19 instructions for Schmidt hall: 1. Only vaccinated/recovered people can enter. 2. At least one empty chair is required between each two participants. 3. Sitting is not allowed in the first two rows. 4. All participants should wear a mask (also during the lecture).

Adi Shamir
Weizmann Inst. of Science

A New Theory of Adversarial Examples in Machine Learning

Abstract:

The extreme fragility of deep neural networks when presented with tiny perturbations in their inputs was independently discovered by several research groups in 2013. Due to their mysterious properties and major security implications, these adversarial examples had been studied extensively over the last eight years, but in spite of enormous effort they remained a baffling phenomenon with no clear explanation. In particular, it was not clear why a tiny distance away from almost any cat image there are images which are recognized with a very high level of confidence as cars, planes, frogs, horses, or any other desired class, why the adversarial modification which turns a cat into a car does not look like a car at all, and why a network which was adversarially trained with randomly permuted labels (so that it never saw any image which looks like a cat being called a cat) still recognizes most cat images as cats. The goal of this talk is to introduce a new theory of adversarial examples, which we call the Dimpled Manifold Model. It can easily explain in a simple and intuitive way why they exist and why they have all the bizarre properties mentioned above. In addition, it sheds new light on broader issues in machine learning such as what happens to deep neural networks during regular and during adversarial training. Experimental support for this theory, obtained jointly with Oriel Ben Shmuel and Odelia Melamed, will be presented and discussed in the last part of the talk.