DEEP LEARNING IN ADVERSARIAL CONTEXT

le 17 juin 2021

PhD defense of Hanwei ZHANG, first PROSFER thesis for ENS Rennes-East China Normal University!
Abstract:
This thesis is about the adversarial attacks and defenses in deep learning. We propose to improve the performance of adversarial attacks in the aspect of speed, magnitude of distortion, and invisibility. We contribute by defining invisibility with smoothness and integrating it into the optimization of producing adversarial examples. We succeed in creating smooth adversarial perturbations with less magnitude of distortion. To improve the efficiency of producing adversarial examples, we propose an optimization algorithm, i.e. Boundary Projection (BP) attack, based on the knowledge of the adversarial problem. BP attack searches against the gradient of the network to lead to misclassification when the current solution is not adversarial. It searches along the boundary to minimize the distortion when the current solution is adversarial. BP succeeds to generate adversarial examples with low distortion efficiently. Moreover, we also study the defenses. We apply patch replacement on both images and features. It removes the adversarial effects by replacing the input patches with the most similar patches of training data. Experiments show patch replacement is cheap and robust against adversarial attacks.
Mise à jour le 26 juin 2021

JURY MEMBERS:

Patrice Quinton, Prof. ENS Rennes
Cecilia Pasquini, Assistant prof., University of Trento
Michel Crucianu, Prof. CNAM
Aimin Zhou, Prof. ECNU
Changbo Wang, Prof. ECNU
Yuan Xie, Prof., ECNU