EPFL-Inria
INTERNATIONAL LAB

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Inria
INVENTEURS DU MONDE NUMÉRIQUE

Workshop EPFL-Inria
January 30 and 31, 2019, Lausanne

Erwan Le Merrer, Inria

Title: « Tweaking neural models: watermarking and tamperproofing them »

Abstract:

This talk will relate some security problems to the decision boundaries obtained by deep learning classifiers. In particular, this talk will introduce the tampering detection and the watermarking of deep neural models, as well as a measure of input safety for those models.
Proposed techniques operate in a black-box setup, where solely queries and obtained labels are leveraged to gain some information on the remotely executed neural model.