Internship proposal:
Automata for Machine Learning Verification

Nathanaël Fijalkow

September 25, 2018

Encadrant: Nathanaël Fijalkow (CNRS and the Alan Turing Institute of data science and artificial intelligence)

Localisation: LaBRI (Bordeaux)

Sujet du stage: Ce stage propose d'étudier différents algorithmes pour la vérification de modèles issus de l'apprentissage, dont en particulier les réseaux de neurones. Plusieurs approches sont proposées, basées soit sur des abstractions d'automates ou sur la synthèse d'invariants. On adoptera un point de vue plus théorique ou plus pratique selon le déroulement du stage et l'intérêt du stagiaire.

Thèmes: Théorie des automates, Apprentissage, Systèmes Dynamiques, Théorie des jeux, Vérification, Interprétation Abstraite.

Neural networks and other machine learning models have proved to be very successful in a wide range of applications. However to be safely used in critical scenarios we need guarantees: how accurate and robust is your model? A very classical example shows a neural net classifying images being fooled by adding random noise, asserting that a panda becomes a gibbon with high confidence.

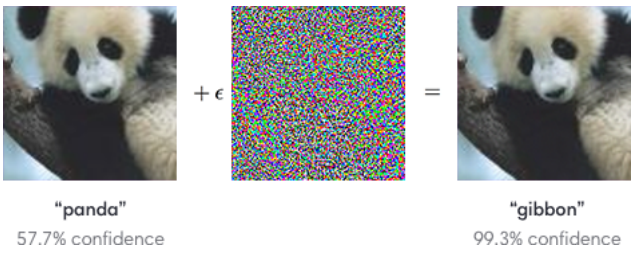

Figure 1: Taken from `openai.com`

Verification of machine learning models is a growing field, fostering ideas from automata theory, program verification, invariant synthesis, and dynamical systems.

Objectives of the internship

The goal of the internship is to study and construct algorithms for verifying neural networks with simple architectures. This in particular requires defining and comparing correctness and robustness notions. The outcome of the internship can be either theoretical or practical through implementation.

A PhD scholarship is available for starting in September 2019.