Titre: Vérification de propriétés de sûreté et de robustesse de réseaux de neurones

Mots-clés: Méthodes formelles, interpretation abstraite, zonotopes, intelligence artificielle

Lieu et équipe: LIX, Bâtiment Turing, campus de l'Ecole Polytechnique.
Au sein du laboratoire d'informatique de l'Ecole Polytechnique (LIX), le stagiaire intégrera l'équipe Cosynus, dont les recherches portent sur la sémantique et l'analyse statique des systèmes logiciels, distribués, hybrides et cyber-physiques.

Encadrement: Sylvie Putot & Eric Goubault – email: putot@lix.polytechnique.fr

Présentation générale: Les dernières années ont montré une large adoption des réseaux de neurones profonds dans les applications critiques pour la sécurité, typiquement dans les véhicules autonomes. Malgré leur succès, un défi fondamental reste à relever: veiller à ce que les systèmes d'apprentissage automatique, notamment les réseaux de neurones profonds, se comportent comme prévu. En effet, ces réseaux de neurones sont souvent vulnérables à des perturbations, qui peuvent entrainer des classifications incorrectes. Récemment, une approche [AI2 2018] a été proposée, permettant de prouver par interprétation abstraite la sûreté, c'est-à-dire la robustesse aux perturbations de réseaux de neurones de dimension réaliste. Cette approche se base sur une version du domaine abstrait des zonotopes que nous avons proposé avec Eric Goubault [FMSD 2016].

Objectifs du stage: Le stage vise à améliorer l'analyse de [AI2 2018] en explorant notamment des variations des domaines abstraits utilisés, en se basant sur notre très bonne connaissance de ces domaines [FMSD 2016]. Un objectif complémentaire et relié est d'étendre ces résultats très prometteurs à l'analyse de classes de programmes plus larges. En particulier, des aspects modélisation du système cyber-physique dans son entier apparaissent lorsque l'on souhaite s'intéresser au comportement des réseaux de neurones dans le cadre de perturbations spécifiques qui correspondent à des scénariis réalistes (par exemple la perturbation de l'image dans un algorithme de détection de piéton ne peut pas être raisonnablement modélisée par une perturbation trop quelconque). De façon plus générale, il s'agit de s'intéresser à la sureté des systèmes contenant des algos d'intelligence artificielle, différentes pistes peuvent être explores au-delà de celle proposée ci-dessus.

Possibilité de thèse.

Bibliographie:

[AI2 2018] AI2: Safety and Robustness Certification of Neural Networks with Abstract Interpretation, T. Gehr, M. Mirman, D. Drachsler-Cohen, P. Tsankov, S. Chaudhuri, M. Vechev, dans la conférence iEEE S&P 2018, 2018.

[FMSD 2016]. A zonotopic framework for functional abstraction, E. Goubault et S. Putot, dans le journal Formal Methods in System Design, 2016