Sujet de stage de Master 2

Robustesse des réseaux de neurones : approche par l'optimisation en variables mixtes

1 Contexte scientifique

Les approches récentes en Intelligence Artificielle (IA), et en particulier celles basées sur l'entraînement de réseaux de neurones, ont de plus en plus d'impact que ce soit dans des domaines applicatifs ou théoriques. L'immense quantité de données disponible permet maintenant d'entraîner des modèles particulièrement performants et présentant des capacités de généralisation pour une grande variété de tâches : agents conversationnels [5, 12], reconnaissance d'objets [11], classification [6, 2, 7] ou génération de données [9, 10]. Cependant, le manque de garanties quant à la robustesse des décisions prises par ces modèles reste particulièrement problématique dans les domaines d'applications nécessitant un haut niveau de confiance [8, 4]. C'est le cas en particulier pour le développement de véhicules autonomes, le traitement de données médicales ou de surveillance.

Si les réseaux de neurones sont à même d'approcher des fonctions particulièrement complexes, leurs décisions restent intrinsèquement inexplicables. Cette inexplicabilité rend impossible la construction de garanties théoriques sur leurs performances. Par ailleurs, la plupart des modèles entraînés pour des tâches de classification ou de détection restent extrêmement confiants dans leurs prédictions même lorsqu'ils se trompent. Un réseaux de neurones pourra, par exemple, donner des probabilités élevées de reconnaître une classe particulière dans une image constituée uniquement de bruit [14, 15]. Cette sur-confiance peut également être exploitée pour construire artificiellement des exemples trompeurs pour une IA. Ceci peut être simplement réalisé en ajoutant une perturbation à l'entrée fournie à un réseau de neurones pour garantir la prédiction d'une classe erronée, comme illustré en Figure 1. Ce type de perturbations porte le nom d'*attaque adversaire*. Les attaques adversaires et en particulier la résistance à ce type d'attaques sont un enjeu majeur pour le déploiement d'IA sur le terrain. En effet, si les paramètres d'une IA utilisée pour une application sensible sont connus, il devient possible par attaques adverses d'usurper une identité, de tromper un système de pilotage, ou encore d'établir de faux documents.

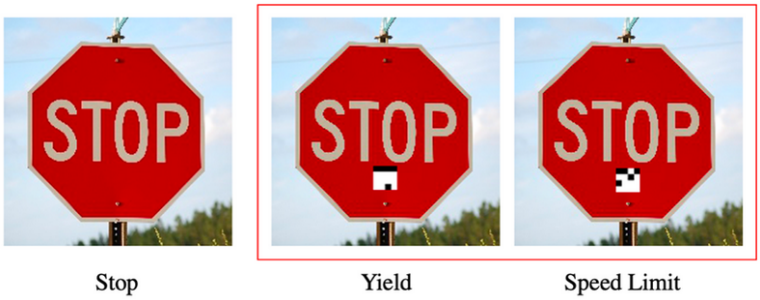

FIGURE 1 – Un exemple d'attaque adverse pour la reconnaissance d'image obtenue par le simple ajout d'un autocollant sur un panneau de circulation [13].

La robustesse des IA à ce type d'attaques a donné lieu à une littérature foisonnante ces cinq dernières années [3, 1, 13, 16, 14]. Pour rendre un réseau de neurones robuste, le principe de base est de l'entraîner à résoudre sa tâche tout en assurant que ses scores de détection restent stables pour des données adverses. Ce problème nécessite donc de minimiser l'erreur associée à la tâche devant être résolue tout en étant robuste aux perturbations d'impact maximal. Ceci peut se modéliser naturellement par une formulation minmax et se résout encore aujourd'hui très difficilement. La méthode la plus utilisée est basée sur une méthode de descente de gradient stochastique où, à chaque itération, la valeur du problème interne de maximisation est évaluée. En pratique, des approches locales basées sur le calcul d'une relaxation ou d'une solution réalisable sont utilisées pour la résolution du problème interne. Récemment, des approches exactes utilisant des problèmes d'optimisation en variables mixtes ont également été introduites dans ce but [1, 7]. Pour de petits jeux de données, ces méthodes sont efficaces, mais un challenge est de les faire passer à l'échelle. De plus, les modèles actuellement proposés ne résolvent pas le problème de maximisation interne dans le cas multiclasse, mais permettent uniquement de déterminer l'existence d'un exemple adverse pour une unique fixée.

2 Objectifs du stage

Notre objectif dans ce stage est d'explorer l'utilisation d'approches récentes en optimisation pour améliorer la résolution du problème de maximisation interne. Nous nous focaliserons sur l'élaboration d'algorithmes de résolution du problème minmax, avec comme domaine d'application la reconnaissance d'images. Ainsi, le stage se découpera en deux étapes :

1. Améliorer la résolution exacte du problème de maximisation interne par le biais d'une meilleure borne que celle de la formulation existante ou en utilisant une nouvelle formulation basée sur l'optimisation non-linéaire.

2. Mettre en oeuvre l'algorithme de résolution du problème de minimisation externe avec, à chaque itération, l'utilisation des résultats de la première étape.

Le travail à effectuer sera à la fois théorique et expérimental. La partie théorique portera sur l'étude de nouvelles modélisations et le travail expérimental consistera à implanter ces modélisations.

Connaissances requises : Cours de programmation mathématique, Connaissance d'un langage quelconque de programmation

Encadrants :
Zacharie Alès, MCF, zacharie.ales@ensta-paris.fr
Amélie Lambert, MCF HDR, amelie.lambert@cnam.fr
Clément Rambour, MCF, clement.rambour@cnam.fr
Lieu : CEDRIC-Cnam (Paris)
Durée : 6 mois
Poursuite en thèse possible : oui

Références

[1] Ross ANDERSON et al. « Strong mixed-integer programming formulations for trained neural networks ». In : *Mathematical Programming* 183.1 (2020), p. 3-39.

[2] Liang-Chieh CHEN et al. « Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs ». In : *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2017), p. 834-848.

[3] Jeremy COHEN, Elan ROSENFELD et Zico KOLTER. « Certified adversarial robustness via randomized smoothing ». In : *International Conference on Machine Learning*. PMLR. 2019, p. 1310-1320.

[4] Charles CORBIÈRE et al. « Addressing failure prediction by learning model confidence ». In : *Advances in Neural Information Processing Systems* 32 (2019).

[5] Jacob DEVLIN et al. « Bert: Pre-training of deep bidirectional transformers for language understanding ». In : *arXiv preprint arXiv:1810.04805* (2018).

[6] Alexey DOSOVITSKIY et al. « An image is worth 16x16 words: Transformers for image recognition at scale ». In : *arXiv preprint arXiv:2010.11929* (2020).

[7] Matteo FISCHETTI et Jason JO. « Deep neural networks and mixed integer linear optimization ». In : *Constraints* 23.3 (2018), p. 296-309.

[8] Dan HENDRYCKS, Mantas MAZEIKA et Thomas DIETTERICH. « Deep anomaly detection with outlier exposure ». In : *arXiv preprint arXiv:1812.04606* (2018).

[9] Jonathan HO, Ajay JAIN et Pieter ABBEEL. « Denoising diffusion probabilistic models ». In : *Advances in Neural Information Processing Systems* 33 (2020), p. 6840-6851.

[10] Tero KARRAS, Samuli LAINE et Timo AILA. « A style-based generator architecture for generative adversarial networks ». In : *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, p. 4401-4410.

[11] Shaoqing REN et al. « Faster r-cnn: Towards real-time object detection with region proposal networks ». In : *Advances in neural information processing systems* 28 (2015).

[12] Hao TAN et Mohit BANSAL. « Lxmert: Learning cross-modality encoder representations from transformers ». In : *arXiv preprint arXiv:1908.07490* (2019).

[13] Ruixiang TANG et al. « An embarrassingly simple approach for trojan attack in deep neural networks ». In : *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, p. 218-228.

[14] Vincent TJENG, Kai XIAO et Russ TEDRAKE. « Evaluating robustness of neural networks with mixed integer programming ». In : *arXiv preprint arXiv:1711.07356* (2017).

[15] Lily WENG et al. « Towards fast computation of certified robustness for relu networks ». In : *International Conference on Machine Learning*. PMLR. 2018, p. 5276-5285.

[16] Han XU et al. « Adversarial attacks and defenses in images, graphs and text: A review ». In : *International Journal of Automation and Computing* 17.2 (2020), p. 151-178.