Titre du projet : Robustesse des réseaux de neurones par l'optimisation en variables mixtes

Mots Clés : Deep Neural Networks, Adversarial examples, Mixed-Integer Optimization, Outer Approximation.

Encadrement de la thèse :

- Amélie Lambert - Maitresse de Conférences HDR au Cnam-Cédric
- Clément Rambour - Maître de Conférences au Cnam-Cédric
- Zacharie Ales - Professeur associé HDR à l'ENSTA Paris-Cédric

Introduction

L'apprentissage profond (*Deep Learning*) s'est montré très efficace durant la dernière décennie pour résoudre de nombreux problèmes. Dans le domaine de la vision par ordinateur, il a par exemple permis de réaliser des percées majeures en matière de segmentation d'images [CZP+18, TRT+23], de reconnaissance d'objets [Gir15] ou encore de détection de visages [par15]. De même, en traitement automatique du langage, l'apprentissage profond est largement utilisé pour la classification [YML19] ou l'analyse de texte [ANMC21]. Depuis peu, des Intelligences Artificielles (IA) obtenues grâce à l'entraînement de larges réseaux de neurones réussissent à produire des résultats particulièrement réalistes en matière de création de contenu, générant des images [RBL+22] ou des textes [SDHL22, ZKW+] difficilement différenciables de créations humaines. Ces IA commencent ainsi à être déployées par différents acteurs industriels pour automatiser certaines tâches d'analyse ou de généréation de contenu telles que la création ou l'édition d'images et de vidéos, l'intégration d'agents conversationnels, le contrôle de robots ou drones, le pilotage de voiture autonomes, l'inspection de contenu web… Cette intégration à marche forcée n'est cependant pas accompagnée d'autant de précautions en matières de sécurité ou d'éthique que le requerrait l'impact de ces technologies. En particulier, la place grandissante de l'IA au sein de notre tissu social et économique rend de plus en plus nécessaire la construction de systèmes fiables ne pouvant être détournés à des fins malveillantes.

Contexte national et international

Malgré d'excellentes performances, les réseaux de neurones manquent cruellement d'explicabilité. Ainsi, il est impossible de construire des garanties quant à leurs résultats. Cet aspect boite noire freine leur utilisation en particulier dans des domaines sensibles tels que la santé ou la sécurité dans lesquels une prise de décision doit pouvoir être justifiée. Outre leur complexité, ces modèles peuvent aisément être dupés par des données bien choisies compromettant d'autant plus leurs potentielles applications. Ces attaques dites adverses consistent à appliquer d'imperceptibles changements dans un contenu (image ou texte) pour en modifier le sens que lui attribue le modèle. En pratique, un attaquant ayant la connaissance de paramètres du modèle peut facilement tromper un classifieur sans qu'un utilisateur ne puisse s'en rendre compte. Ces attaques présentent des implications importantes dans de nombreux domaines, tels que la sécurité, la confidentialité et la fiabilité. Des exemples concrets peuvent être l'usurpation d'identité par attaque sur un système de reconnaissance faciale ou la mise en défaut de système de conduite autonome. Pour cette raison, il est essentiel que les réseaux de neurones soient conçus et entraînés pour être robustes contre ces types d'attaques. Diverses techniques ont été mises au point pour assurer la robustesse d'un réseau de neurones aux attaques adverses. Celles-ci se décomposent entre approches post-hoc cherchant à détecter des anomalies par rapport aux données d'entraînement [Ehl17, FJ18, TXT17] et des approches favorisant la robustesse des réseaux par une stratégie d'entraînement spécifique. Les premières ont l'avantage de pouvoir être appliquées à n'importe quel modèle mais présentent une efficacité limitée pour la prévention d'attaques. Les secondes sont basées sur plusieurs approches : augmenter la base d'apprentissage par des exemples adversaires [Xia22], ou bien introduire des régularisations visant à stabiliser la sortie du modèle autour des données d'apprentissage [CRK19], ou enfin construise un réseau robuste aux attaques adverses ayant le plus d'impact [WK18]. Cette dernière méthode assure d'avoir des réseaux parfaitement robustes mais ne peut être envisagée que pour des modèles de taille restreinte. Quelques techniques ont cependant été récemment proposées pour diminuer la taille du

problème [TXT17, Xia22]. Cette thèse s'inscrit dans ce cadre et a pour but de proposer des approches efficaces fournissant des modèles certifiés robustes pour permettre un passage à l'échelle.

Objectif du projet doctoral

L'objectif de cette thèse est de proposer des stratégies efficaces d'entraînement de réseaux de neurones robustes qui exploitent l'optimisation en variables mixtes. Pour construire un réseau de neurones robuste, le principe est de l'entraîner à résoudre sa tâche tout en assurant que ses prédictions restent stables pour des exemples adverses. Ce problème consiste donc à minimiser l'erreur associée à la prédiction, tout en étant robuste aux attaques adverses d'impact maximal, et se modélise par une formulation minmax. La méthode la plus utilisée pour résoudre le problème de minimisation est une méthode de descente de gradient 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 basées sur l'optimisation linéaire en variables mixtes (PL) ont été introduites pour construire des exemples adverses [AHM+20, FJ18, TXT17, Xia22]. Ces approches qui permettent en théorie de certifier la robustesse d'un réseau sont pénalisées par la faiblesse de leur borne et ne permettent pas le passage à l'échelle. Des modèles quadratiques et non convexes (QP) mais en variables continues ont ensuite été introduits [Zha20]. Ceux ci permettent d'obtenir des relaxations plus serrées en résolvant leurs relaxations Semi-Définies Positives (SDP) [Zha20, Lan, CZ23] qui sont connues pour fournir d'excellentes bornes [Ans09] mais qui sont très coûteuses en terme de temps de calcul. Ainsi, résoudre successivement plusieurs SDP pour rafiner la solution du QP est impraticable même pour des petits réseaux. Les approches les plus efficaces de résolution globale de modèles d'optimisation quadratique sont basées sur des reformulations quadratiques convexes qui sont calculées en résolvant une seule fois un SDP [EL19, Lam23]. L'objectif de cette thèse est de construire un algorithme permettant de certifier les réseaux de neurones de tailles significatives. Pour cela, nous projettons tout d'abord de construire un algorithme de résolution efficace du problème interne basé sur des approches de reformulations quadratiques convexes dédiées au modèle QP. Nous intégrerons ensuite cet algorithme dans l'apprentissage robuste du réseau afin de résoudre efficacement le problème minmax et de pouvoir ensuite le généraliser à des architectures de réseaux modernes.

Échéancier

- Étape 1 : Revue de la littérature (4 mois)
  - Étude des méthodes de référence pour la robustesse de réseaux peu profonds ;
  - Comparaisons entre méthodes résolvant exactement le problème de maximisation interne [FJ18] à d'autres approches impliquant une relaxation convexe [WK18, Zha20, Lan, CZ23] ou une sélection des échantillons d'entraînement [ZZG+20] en vu d'un passage à l'échelle ;
  - Mise en place de protocoles d'entraînement et de test de modèles robustes et implantations des modèles pour évaluer leurs performances.
- Étape 2 : Conception d'un algorithme d'apprentissage robuste (20 mois)
  - Extension des modèles de résolution du problème de maximisation interne de la littérature et étude théorique de leurs propriétés avec comme objectif préliminaire de concevoir une approche basée sur des reformulation quadratiques convexes des modèles quadratiques [Zha20, Lan, CZ23] au cas de réseaux de neurones pré-entraînés ;
  - Adaptation des approches classiques pour la résolution du problème de minimisation externe et étude des propriétés de convergence de l'algorithme de résolution du problème minmax en se concentrant initialement sur l'échantillonnage de données ambiguës fourni par la résolution du problème interne ;
  - Évaluation des méthodes proposées sur des applications dans le domaine de la reconnaissance d'images.
- Étape 3 : Généralisation à des architectures de réseaux modernes (8 mois)

  - Étendre nos résultats à d'autres familles de données (textes, graphes) ;
  - Évaluation des méthodes proposées sur des applications réelles.

- Étape 4 : Rédaction de la thèse (4 mois).

Références

[AHM+20] Ross Anderson, Joey Huchette, Will Ma, Christian Tjandraatmadja, and Juan Pablo Vielma. Strong mixed-integer programming formulations for trained neural networks. *Mathematical Programming*, 183(1-2) :3–39, 2020.

[ANMC21] Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. Transformer models for text-based emotion detection : a review of bert-based approaches. *Artificial Intelligence Review*, pages 1–41, 2021.

[Ans09] K. M. Anstreicher. Semidefinite programming versus the reformulation-linearization technique for non-convex quadratically constrained quadratic programming. *Journal of Global Optimization*, 43(2) :471–484, 2009.

[CRK19] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019.

[CZ23] Hong-Ming Chiu and Richard Y. Zhang. Tight certification of adversarially trained neural networks via non-convex low-rank semidefinite relaxations. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 5631–5660. PMLR, 23–29 Jul 2023.

[CZP+18] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[Ehl17] Ruediger Ehlers. Formal verification of piece-wise linear feed-forward neural networks. In *Automated Technology for Verification and Analysis : 15th International Symposium, ATVA 2017, Pune, India, October 3–6, 2017, Proceedings 15*, pages 269–286. Springer, 2017.

[EL19] Sourour Elloumi and Amélie Lambert. Global solution of non-convex quadratically constrained quadratic programs. *Optimization Methods and Software*, 34(1) :98–114, 2019.

[FJ18] Matteo Fischetti and Jason Jo. Deep neural networks and mixed integer linear optimization. *Constraints*, 23(3) :296–309, 2018.

[Gir15] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[Lam23] Amélie Lambert. Using general triangle inequalities within quadratic convex reformulation method. *Optimization Methods and Software*, 38(3) :626–653, 2023.

[Lan] Tight neural network verification via semidefinite relaxations and linear reformulations. 36.

[par15] Deep face recognition. 2015.

[RBL+22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[SDHL22] Sami Sarsa, Paul Denny, Arto Hellas, and Juho Leinonen. Automatic generation of programming exercises and code explanations using large language models. In *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 1*, pages 27–43, 2022.

[TRT+23] Loic Themyr, Clément Rambour, Nicolas Thome, Toby Collins, and Alexandre Hostettler. Full contextual attention for multi-resolution transformers in semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3224–3233, 2023.

[TXT17] Vincent Tjeng, Kai Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. *arXiv preprint arXiv :1711.07356*, 2017.

[WK18] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International conference on machine learning*, pages 5286–5295. PMLR, 2018.

[Xia22] Kai Yuanqing Xiao. *Probing, Improving, and Verifying Machine Learning Model Robustness*. PhD thesis, Massachusetts Institute of Technology, 2022.

[YML19] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7370–7377, 2019.

[Zha20] Richard Zhang. On the tightness of semidefinite relaxations for certifying robustness to adversarial examples. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3808–3820. Curran Associates, Inc., 2020.

[ZKW+] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore : Evaluating text generation with bert. In *International Conference on Learning Representations*.

[ZZG+20] Haizhong Zheng, Ziqi Zhang, Juncheng Gu, Honglak Lee, and Atul Prakash. Efficient adversarial training with transferable adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1181–1190, 2020.