Sujet de Thèse : Génération d'exemples adversariaux pour les réseaux de neurones inversibles

Janvier 2020

Cadre

— Encadrants de la thèse :
Yann Chevaleyre (Professeur), LAMSADE — UMR 7243
`yann.chevaleyre@dauphine.psl.eu`
Fabrice Rossi (Professeur), CEREMADE — UMR 7534
`fabrice.rossi@dauphine.psl.eu`
Benjamin Negrevergne (Maître de Conférences) LAMSADE — UMR 7243
`benjamin.negrevergne@dauphine.psl.eu`

— École doctorale : École doctorale de Dauphine - Université Paris Dauphine

— Laboratoire d'accueil : LAMSADE - Université Paris Dauphine. Équipe MILES.

Contexte sociétal

Un exemple *adversarial*[1] est une variation imperceptible d'un exemple naturel, construit dans le but de tromper un modèle de classification. L'existence d'exemples adversariaux dans les réseaux de neurones est un problème à la fois intriguant et paradoxal. En effet, on pourrait croire que les progrès récents des modèles de classification permettraient de venir naturellement à bout des exemples adversariaux, mais ça ne semble pas être le cas. Au contraire, depuis la découverte des exemples adversariaux dans les réseaux de neurones profonds [1], de nombreux travaux de recherche ont décrit de nouvelles techniques d'attaques, et proposé de nouvelles méthodes d'entraînement pour essayer de s'en défendre.

L'existence d'exemples adversariaux pose un problème sociétal grandissant à mesure que les réseaux de neurones sont déployés dans un nombre croissant d'applications. En effet, ils remettent en question la fiabilité des réseaux de neurones, et leur utilisation dans un contexte où la présence d'utilisateurs malveillants est à prévoir. Diffusés à la radio, à la télévision ou imprimés sur des panneaux publicitaires, les exemples adversariaux pourraient massivement dégrader les systèmes dont le bon fonctionnement s'appuie sur les réseaux de neurones (e.g. voitures autonomes).

Plus généralement, au delà de ces cas pathologiques, l'étude des exemples adversariaux permet également d'acquérir une meilleure compréhension des limites des réseaux de neurones actuels et de leurs procédures d'entraînement.

1. *Adversarial example* en anglais.

Contexte scientifique et objectifs de la thèse

L'objectif de cette thèse est d'étudier la génération d'exemples adversariaux dans le cadre des réseaux inversibles. Ces réseaux sont construits de telle sorte que la transformation habituellement appliquée à un exemple pour pouvoir être classé soit inversible. Dans le cadre d'une tâche de classification d'image, cela signifie qu'il est possible de passer de l'image à la représentation latente, (comme dans un réseau classique) ou de la représentation latente à l'image. Les réseaux inversibles ont de nombreuses applications, en premier lieu pour la génération de données [2], [3]. Ils peuvent également être utilisés pour économiser le stockage d'un grand nombre de résultats intermédiaires lors de l'entraînement du réseau, et permettre l'entraînement de très grands réseaux, sans être limité par les contraintes d'utilisation mémoire imposées par l'utilisation des GPUs [4]. Une augmentation importante de l'utilisation de ce types de réseaux est donc à prévoir dans les années à venir.

Malgré le développement de ce type de réseaux, il existe relativement peu de travaux qui étudient la robustesse des réseaux inversibles. La question est d'autant plus critique, que le caractère inversible de ces réseaux permet d'imaginer de nouvelles attaques, qui ne nécessitent pas d'avoir recours à des processus d'optimisation coûteux comme c'est le cas dans les réseaux classiques [5], ce qui les rend d'autant plus vulnérables. Dans un premier temps, nous chercherons à comprendre la structure des attaques dans ce type de réseaux pour développer des techniques d'attaques et de défenses spécialisées pour les réseaux inversibles. A plus long terme, nous souhaitons étudier les conséquences de la vulnérabilité des réseaux inversibles, pour les réseaux classiques, en s'appuyant sur les techniques de distillation [6]. S'il s'avère que les nouvelles attaques développées pour les réseaux inversibles peuvent être utilisées pour attaquer les réseaux classiques, la vulnérabilité des réseaux inversibles pourraient également affecter les réseaux classiques, bien plus largement utilisés aujourd'hui.

Déroulement de la thèse

La thèse pour laquelle nous demandons un financement au sein de l'école doctorale de Dauphine vise à se placer dans un contexte de réseaux de neurones inversibles. La plupart des papiers traitant d'attaques adversariales se placent dans un cadre de structures non inversibles bien que l'inversibilité permette un lien entre l'information contenue à tous les niveaux. Nous étudierons le lien entre ces réseaux et les attaques. Plus précisément les objectifs de la thèse seront les suivants.

Génération d'attaques par inversion. Dans les réseaux classiques la plupart des attaques optimisent une perturbation sur l'image initiale de sorte à maximiser la fonction de perte. Avec un réseau inversible, l'attaquant peut directement spécifier la perte souhaitée dans l'espace latent, et inverser la représentation latente pour obtenir un exemple adversarial. Nos travaux préliminaires sur ce sujet montrent le bien fondé de l'approche, et un première implémentation suggère que cette approche peut être utilisée pour générer très efficacement des exemples particulièrement nuisibles. La difficulté devient alors de retrouver une image proche de l'image originale, et pour y parvenir, plusieurs pistes doivent être explorées. Une première piste consiste à s'appuyer sur sur nos travaux de recherche récents qui ont permis d'entraîner des réseaux disposant d'une constante de Lipschitz faible. Une seconde piste serait d'utiliser des réseaux dont le Jacobien est facilement calculable afin de l'évaluer simplement l'effet que peut avoir une perturbation dans l'espace des images.

Génération d'attaques rapides et efficaces pour les réseaux non inversibles. Dans un deuxième temps, nous souhaiterons comprendre dans quelle mesure la vulnérabilité des réseaux inversibles peut impacter celle des réseaux classiques. A priori, les attaques par inversion ne pourront pas être directement mises en œuvre dans les réseaux classiques, mais grâce aux techniques de distillation [6], il est possible d'entraîner un réseau inversible à mimer le comportement du réseau original. (Pour y parvenir, on entraîne le nouveau réseau avec une fonction de perte qui caractérise la différence entre ses prédictions, et celles du réseau initial). La technique est avant tout utilisée pour entraîner des réseaux plus compacts, mais pourra être utilisée ici pour entraîner un réseau inversible au comportement similaire à celui du réseau classique. Grâce à ce nouveau réseau, il devient alors possible d'attaquer le réseau inversible, et de transférer l'attaque sur le réseau classique.

Caractériser la robustesse des réseaux de neurones et fournir de nouvelles garanties de robustesses. Des résultats récents visent à caractériser la robustesses des réseaux de neurones classiques en analysant leur constante de Lipschitz [7] ou l'information mutuelle entre les vecteurs dans l'espace de départ et l'espace latent. Or ces même grandeurs interviennent également dans les réseaux inversibles [8]. Dans cette troisième partie, nous nous appuierons donc sur les résultats produits dans les deux premières parties pour identifier les grandeurs directement responsables de la robustesse des réseaux de neurones. À long terme, ces travaux nous permettrons de développer des nouvelles procédures d'entraînement, capables d'offrir des garanties fortes sur la robustesse des modèles qu'elles permettent d'obtenir.

Références

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2013.

[2] J. Behrmann, W. Grathwohl, R. T. Q. Chen, D. Duvenaud, and J.-H. Jacobsen, "Invertible residual networks," 2018.

[3] R. T. Q. Chen, J. Behrmann, D. Duvenaud, and J.-H. Jacobsen, "Residual flows for invertible generative modeling," 2019.

[4] N. Kitaev, Łukasz Kaiser, and A. Levskaya, "Reformer : The efficient transformer," 2020.

[5] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," 2016.

[6] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015.

[7] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier, "Parseval networks : Improving robustness to adversarial examples," 2017.

[8] J.-H. Jacobsen, J. Behrmann, R. Zemel, and M. Bethge, "Excessive invariance causes adversarial vulnerability," 2018.