Randomized Smoothing sous attaque: Théorie vs Pratique

Thibault Maho[1], Teddy Furon[1]*, Erwan Le Merrer[1]

[1]Univ. Rennes, Inria, CNRS
IRISA, Rennes, France
thibault.maho@inria.fr, teddy.furon@inria.fr, erwan.lemerrer@inria.fr

Résumé – Le Randomized Smoothing (RS) est une solution récente pour certifier la robustesse d'un classifieur contre les attaques adversaires. Ce papier s'interroge sur l'efficacité du randomized smoothing en tant que *défense* contre les attaques boîtes noires. Nous commençons par mettre en évidence formellement le décalage entre certification théorique et pratique. Nous réalisons ensuite des attaques pour observer qu'il existe un décalage important entre les paramètres pour obtenir une robustesse théorique certifiée élevée et ceux pour lutter efficacement contre les attaques boîtes noires.

Abstract – Randomized Smoothing (RS) is a recent solution to certify the robustness of a classifier against adversarial attacks. This paper investigates the effectiveness of randomized smoothing as a *defense*. We start by formally demonstrating the mismatch between theoretical and practical certification. We then perform attacks to observe that there is a significant mismatch between the parameters to obtain a high certified theoretical robustness and those to effectively fight against black-box attacks.

1 Introduction

L'adoption des réseaux de neurones a été cruciale pour la performance dans de multiples domaines, y compris ceux sensibles (voitures autonomes, systèmes de sécurité, reconnaissance faciale). Ce succès est entravé par leur vulnérabilité [1] face à des attaques dites adverses. Ceci est particulièrement critique dans les attaques boîtes noires basées sur la décision [2, 3, 4, 5, 6]. L'attaquant peut observer les résultats d'autant de requêtes qu'il souhaite pour créer une perturbation adverse de petit amplitude. Il se trouve que l'adversaire parvient toujours à mener des attaques puissantes.

Des défenses ont été proposées pour augmenter la robustesse de ces classifieurs face à ces attaques. Par exemple, l'adversarial training [7] implique un réentraînement avec des entrées adverses. C'est efficace mais coûteux.

Une nouvelle approche consiste à *certifier* localement la robustesse des classifieurs, avec en particulier le randomized smoothing [8, 9]. La certification est un paradigme général et agnostique par rapport au modèle. Il peut être appliqué sans ré-entrainement. Son avantage est de certifier théoriquement un niveau de robustesse aux attaques: Pour une entrée donnée, il n'existe pas de perturbation adverse (qui change la décision du classifieur) de norme (Euclidienne) plus petite que ce niveau certifié.

Le RS est une grande avancée pour la robustesse des classifieurs. Néanmoins, son application en tant que défense (et non plus comme une garantie théorique) a des angles morts: *i)* La robustesse exacte certifiée est impossible à calculer dû à la dimension de l'espace d'entrée des classifieurs actuels. Une méthode Monte Carlo est utilisée pour l'estimer. Il y a un manque de compréhension de l'interaction entre certification théorique et pratique où une quantité limitée d'échantillons est la clé de l'applicabilité. En outre, puisque cette défense est aléatoire par nature, la définition classique d'un exemple adverse [1] n'est plus applicable. *ii)* Le nombre d'échantillons requis est flou et varie selon les articles: entre 100 [10] et 100000 [9]. Aucun résultat à ce jour n'a montré l'influence de cette quantité sur l'efficacité des attaques. *iii)* Enfin, bien que cette défense soit en principe applicable sans ré-entraînement, il est néanmoins recommandé [9] pour limiter la baisse de précision induite. En effet, plus le rayon de bruit certifié est grand, plus le classifieur est robuste, au prix d'une importante chute de précision qui peut être limitée par un ré-entraînement sur des données bruitées. L'impact de la variance de bruit sur la précision finale ou l'efficacité des attaques est également floue.

Cet article contribue à aborder ces trois questions en considérant le RS comme une défense pratique. Nous confrontons d'abord la théorie à la pratique pour la certification et la défense dans le contexte du RS. Nous évaluons ensuite la robustesse pratique de cette défense en ce qui concerne l'impact des tailles des échantillons du Monte Carlo et du paramètre de variance du bruit. Cette étude met en évidence l'efficacité du RS pour vaincre les meilleures attaques boîtes noires. Cependant, les meilleurs paramètres du RS comme défense se révèlent bien différents de ceux suggérés dans la littérature pour la certification théorique.

*Thanks to ANR and AID french agencies for funding Chaire SAIDA.

2 Travaux Associés

2.1 Attaque boite noire

Brendel *et al.* [11] a initié les attaques boîtes noires où seule la décision retournée par le modèle est disponible. `HSJA` [2] améliore l'efficacité en construisant des substituts du gradient. Le gradient est estimé en un point sur la frontière en bombardant le modèle avec une version bruité de ce point sensible. De meilleurs résultats sont obtenus en travaillant dans le domaine fréquentiel [3, 4]. `SurFree` [5] repose quant à lui sur une approximation géométrique de la frontière. Ces attaques créent de faibles perturbations adverses au sens ℓ_2. Il existe aussi des attaques pour la norme ℓ_∞ comme `RayS` [6]. Cet article exploite ces trois attaques pour tester la robustesse pratique du RS.

2.2 Randomized smoothing (RS)

Initié par Lecuyer *et al.* [8], RS est une méthode agnostique au modèle pour obtenir une robustesse locale certifiée. La prédiction est garantie inchangée dans un certain rayon autour d'une entrée donnée. Mesurer la robustesse d'un modèle avec des attaques adverses n'est donc plus nécessaire car la robustesse est formellement garantie. En pratique, RS est une simulation Monte Carlo nécessitant un grand nombre d'appels au modèle. Elle produit une borne inférieure de la robustesse et détériore la précision du modèle [9]. Tandis que [8, 10] ont amélioré la tolérance au bruit des classifieurs. Aucuns travaux n'a attaqué RS et comparé sa robustesse théorique à celle pratique.

3 RS: de la théorie à la pratique

3.1 Introduction au RS

Pour simplifier, on considère un classifieur binaire $f : \mathbb{R}^d \to \{0, 1\}$. RS définit un nouveau classifieur g_σ comme :

$$g_\sigma(\mathbf{x}) = \arg\max_{y\in\{0,1\}} \mathbb{P}[f(\mathbf{x}+\sigma\mathbf{N}) = y],\ \mathbf{N} \sim \mathcal{N}(0, I). \quad (1)$$

Le grand avantage de g_σ est que sa robustesse est certifiée. Supposons qu'un génie révèle la valeur des deux probabilités $\pi_0(\mathbf{x}) := \mathbb{P}[f(\mathbf{x}+\sigma\mathbf{N}) = 0]$ et $\pi_1(\mathbf{x}) := 1 - \pi_0(\mathbf{x})$, alors $\mathbf{x}$ est classé selon (1) avec une robustesse certifiée:

$$R(\mathbf{x}, \sigma) = \sigma\Phi^{-1}(\pi_{g_\sigma(\mathbf{x})}(\mathbf{x})). \quad (2)$$

où Φ est la fonction de répartition de la distribution normale standard $\mathcal{N}(0, I)$. Tous les points à une distance de $\mathbf{x}$ inférieure à $R(\mathbf{x}, \sigma)$ sont classés de la même manière, *i.e.* comme $g_\sigma(x)$. Malgré le terme 'randomized smoothing', g_σ est bien un classifieur déterministe. Sa frontière ∂g_σ est le lieu des points t.q. $\pi_0(\mathbf{x}) = 1/2$.

En pratique, le défenseur utilise une simulation Monte Carlo sur n échantillons aléatoires i.i.d. $\{\mathbf{n}_i\}_{i=1}^n$ distribués comme $\mathbf{N}$ donnant n décisions $\{y_i\}_{i=1}^n$. Ces n 'micro'-décisions sont agrégées pour obtenir la classe finale, *e.g.* par vote majoritaire. Ceci définit le classifieur $g_{\sigma,n}$, une implémentation pratique de la fonction idéale g_σ. La simulation Monte Carlo donne également un intervalle de confiance $\underline{\pi}_0(\mathbf{x}) < \pi_0(\mathbf{x}) < \overline{\pi}_0(\mathbf{x})$ (idem pour $\pi_1(\mathbf{x})$) pour un niveau de confiance donné. La robustesse est évaluée en utilisant (2) avec $\underline{\pi}_{g_{\sigma,n}(\mathbf{x})}$, ce qui donne $\underline{R}(\mathbf{x}, \sigma) < R(\mathbf{x}, \sigma)$. Maximiser la robustesse certifiée autour d'un point donné $\mathbf{x}$ avec un niveau de confiance élevé nécessite de grands n et σ [8, 9, 10].

3.2 Point de vue critique

L'argument principal du RS est le suivant : mener des attaques pour jauger la sécurité d'un classifieur n'est plus nécessaire car sa robustesse est certifiée. Ceci doit être pondéré : $\underline{R}(\mathbf{x}, \sigma)$ certifie la robustesse du classifieur g_σ qui n'existe pas en pratique. Le classifieur pratique $g_{\sigma,n}$ se comporte comme g_σ uniquement lorsque $n \to \infty$.

Plus important encore, $g_{\sigma,n}$ *n'est pas déterministe*. Pour $\mathbf{x} \in \partial g_\sigma$, $g_{\sigma,n}(\mathbf{x})$ agit comme une variable aléatoire d'un appel à l'autre. Cela remet en question le concept de frontières, et par extension celle des exemples adverses.

3.3 Pousser les frontières

Considérons un point $\mathbf{x}$ t.q. $f(\mathbf{x}) = 1$ et à une distance $\delta = \beta\sigma$ de la frontière ∂f du classifieur de base. Ce que l'on appelle SORM en ingénierie de la fiabilité statistique est l'approximation suivante :

$$\pi_0(\mathbf{x}) \approx \Phi(-\beta) \prod_{i=1}^{d-1} \frac{1}{\sqrt{1+\beta\kappa_i}}, \quad (3)$$

où les $\{\kappa_i\}$ sont les courbure principales signées de la surface ∂f au point le plus proche de $\mathbf{x}$. Si elle est plate, toutes les courbures sont égales à 0, et $\mathbf{x}$ se trouve sur la limite ∂g_σ du classifieur idéal RS si $\pi_0(\mathbf{x}) = 1/2$ impliquant $\delta = 0$. Si ∂f est convexe vers $\mathbf{x}$, les courbures sont toutes négatives. Le second terme devient plus grand et compense $\Phi(-\beta)$ de sorte que $\pi_0(\mathbf{x}) = 1/2$ pour un certain $\beta > 0$. Ceci montre que la frontière ∂g_σ est plus proche que ∂f lorsqu'elle se trouve dans une région convexe, et donc plus éloignée lorsqu'elle se trouve dans une région concave. Si les images originales se trouvent dans des régions concaves, alors le RS repousse la frontière et augmente ainsi la norme de la perturbation adverse. Dans la Fig. 1, une attaque en boîte blanche contre le modèle f trouve d'abord un $\mathbf{x}_a \in \partial f$ adverse. Nous voyons qu'en partant de $\mathbf{x}_o$ en suivant la direction $\mathbf{x}_a - \mathbf{x}_o$, nous traversons la frontière ∂g_σ (*i.e.* $\pi_0(\mathbf{x}) = 0.5$) après avoir dépassé $\mathbf{x}_a \in \partial f$ d'une distance de ≈ 4.0. Ceci est également illustré dans la Fig. 2 sur une coupe 2D de $\mathbb{R}^d$.

3.4 Confiance des exemples adverses

Nous proposons une nouvelle définition de l'attaque non ciblée : un exemple adverse de $\mathbf{x}_o$ de niveau $P_a \in [0, 1]$ est

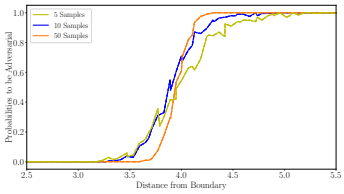
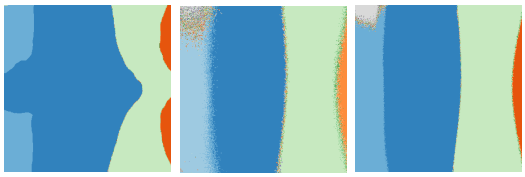


FIG. 1: Probabilité d'être adverse sur la direction $\mathbf{x}_a - \mathbf{x}_o$.



(a) No RS (b) $n = 10$ (c) $n = 200$

FIG. 2: Coupe 2D dans l'espace image de `ResNet50` avec et sans RS ($\sigma = 0.05$).Chaque point est une image, son label élu donne la couleur. Coupe obtenue avec une image central de l'ensemble de validation et 2 directions aléatoires.

un point $\mathbf{x}_a$ s.t.

$$\mathbb{P}[g_{\sigma,n}(\mathbf{x}_a) \neq g_\sigma(\mathbf{x}_o)] \geq P_a. \tag{4}$$

Si l'attaquant est satisfait d'un niveau $P_a = 1/2$, alors l'exemple adverse le plus proche se trouve sur la frontière de g_σ à une distance supérieure à $R(\mathbf{x}, \sigma) > \underline{R}(\mathbf{x}, \sigma)$. Nous pensons que les attaquants exigent une garantie plus forte $P_a > 1/2$, c'est pourquoi l'exemple adverse se trouve à une distance encore plus grande. L'équation (4) exige que $\sum y_i \sim \mathcal{B}(n, 1 - \pi_0(\mathbf{x}_a))$ prenne une valeur supérieure à $n/2$ (en raison du vote majoritaire) avec une probabilité supérieure à P_a. Ceci est valable pour :

$$\pi_0(\mathbf{x}_a) < 1 - I^{-1}_{P_a}(\tilde{n}, \tilde{n}) < 1/2, \tag{5}$$

avec $\tilde{n} = 1 + \lfloor n/2 \rfloor$ et $I_p^{-1}(a, b)$ est la fonction bêta incomplète inverse. Appliquer (2) sur $\mathbf{x}_o$ et $\mathbf{x}_a$ donne :

$$\begin{aligned} \|\mathbf{x}_o - \mathbf{x}_a\| &= \|\mathbf{x}_o - \mathbf{x}_b\| + \|\mathbf{x}_b - \mathbf{x}_a\| \\ &\geq R(\mathbf{x}_o, \sigma) + \sigma\Phi^{-1}(I^{-1}_{P_a}(\tilde{n}, \tilde{n})), \end{aligned} \tag{6}$$

où $\mathbf{x}_b \in [\mathbf{x}_o, \mathbf{x}_a] \cap \partial g_\sigma$. En conclusion, la robustesse $\underline{R}(\mathbf{x}_o, \sigma)$ certifiée par l'implémentation pratique de RS est encore moins rigoureuse en pratique.

3.5 Perturber les attaques boîtes noires

Les attaques boîtes noires reposent généralement sur deux hypothèses. Premièrement, à partir de 2 points de classes différentes, une dichotomie trouve un point $\mathbf{x}_b$ sur la frontière avec une précision contrôlée. Cependant, le classifieur pratique $g_{\sigma,n}$ est aléatoire. La Fig. 3 montre la distribution du résultat de la dichotomie. Plus n est petit, plus la dichotomie est perturbée.

Deuxièmement, la frontière est supposée lisse afin d'estimer le vecteur normal de l'hyperplan tangent localement autour de $\mathbf{x}_b$ sur la frontière. Cela se fait généralement en

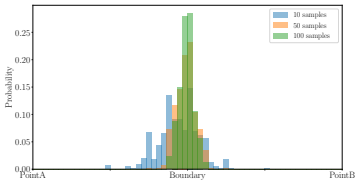


FIG. 3: Distribution de la sortie d'une dichotomie avec RS

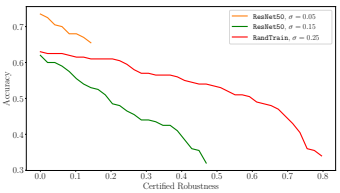


FIG. 4: Certification avec `ResNet50` et `RandTrain`

bombardant le classifieur de versions bruitées de $\mathbf{x}_b$ et en observant ses sorties. Cependant, RS fractalise le voisinage proche des frontières, comme on le voit sur la Fig. 2. Cela ne gâche pas l'estimation. Nous remarquons que les estimations du vecteur normal pour $\mathbf{x}_b \in \partial g_\sigma$ avec et sans RS sont très bien corrélées. La dichotomie peut donner un point $\mathbf{x}_b$ qui n'est pas exactement sur la frontière et cela biaise l'estimation. Par exemple, nous remarquons que `HSJA` [2] se trompe parfois car toutes les versions bruitées de $\mathbf{x}_b$ donnent la même sortie.

4 Attaque boite noire *vs.* RS

4.1 Configuration expérimentale

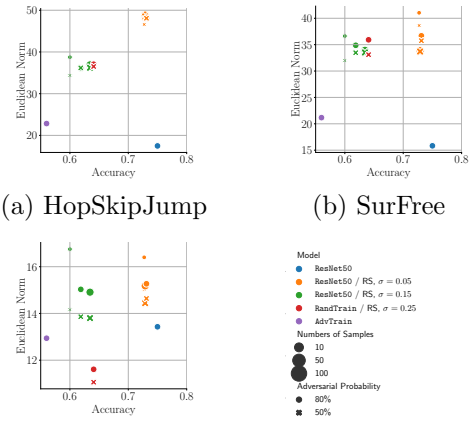
Nous attaquons les modèles de classification avec 200 images aléatoires provenant de l'ensemble de validation de ImageNet ILSVRC2012 de taille $d = 3 \times 224 \times 224$.

Classifieurs. Le classifieur de base est `ResNet50`. Le RS est effectué avec deux écarts types de bruit: $\sigma = 0.05$ donne une baisse acceptable de précision de 3%, alors que $\sigma = 0.15$ donne une plus grande valeur de robustesse certifiée avec une perte de précision de 12% (voir Fig. 4).

L'article [9] propose de réentraîner le modèle avec des données bruitées afin d'utiliser un σ plus grand sans sacrifier trop de précision. Ce modèle est appelé `RandTrain`. Avec $\sigma = 0.25$, la perte de précision est aussi de l'ordre de 12% mais il offre une plus grande robustesse certifiée (voir Fig. 4). Nous comparons le RS au `ResNet50` avec de l'adversarial training [7] désigné par `AdvTrain`.

Attaque boite noire. La section 2 mentionne trois attaques de l'état de l'art. `RayS` [6], `SurFree` [5] et `HSJA` [2] ont de bons résultats en moins de 1000 appels au classifieur. Pour atteindre leur plein potentiel, 2000 sont utilisés .

Protocole. La distorsion est mesurée comme la norme ℓ_2



(a) HopSkipJump (b) SurFree

(c) RayS

FIG. 5: Distorsion ℓ_2 des exemples adverses *vs.* précision

de la perturbation adverse dans le domaine $[0,1]^d$. Pour évaluer qu'un point $\mathbf{x}_a$ est conforme à (4), l'attaquant doit interroger $\ell = O(1/P_a)$ fois le classifieur $g_{\sigma,n}$. Nous accélérons la simulation en considérant que (4) tient si $\lceil nP_a \rceil$ micro-décisions sont incorrectes.

4.2 Résultats

Certification vs Pratique. L'écart entre la théorie et la pratique est flagrant lorsque l'on considère la Fig. 4 et la Fig. 5 : La Fig. 5 fait état de distorsions au moins 30 fois plus importantes que la robustesse certifiée dans la Fig. 4.

Exemples adverses, nouvelle définition. Les attaques ne sont pas impactées par le niveau P_a (4). Le fait d'être 80% adverse oblige à s'éloigner un peu (Fig. 1) surtout pour les petits n. La robustesse est légèrement meilleure.

Un bruit faible est suffisant. un grand σ ne rend pas le réseau plus robuste, mais nuit à sa précision. Cela est vrai même si le réseau a appris à gérer le bruit : `RandTrain` a la même robustesse et précision que `ResNet50` avec RS $\sigma = 0.15$. La situation est encore pire contre `RayS` [6]: `RandTrain` est nettement moins robuste que `ResNet50` sans RS.

Un faible nombre d'échantillons est suffisant. Un grand nombre d'échantillons est la clé pour obtenir une robustesse certifiée potentiellement grande. La Fig. 5 montre une autre réalité. La robustesse contre les 3 attaques est meilleure avec moins d'échantillons. Cela confirme les explications de la Sect. 3.5. Moins d'échantillons rend la prédiction à la frontière plus aléatoire, ce qui compromet davantage les attaques boîtes noires.

La dichotomie est le seul outil commun aux 3 attaques. Un point exactement sur la frontière est crucial pour `HSJA` [2] car il estime le gradient. `SurFree` [5] ne fait pas cela mais se repose sur l'hypothèse d'une frontière lisse. Quant à `RayS` [6], sa dichotomie améliore la distorsion mais n'est pas cruciale pour sa convergence. Cela explique pourquoi `RayS` [6] est moins impacté que `SurFree` [5] et `HSJA` [2].

5 Conclusion

La certification avec RS est une grande avancée pour la robustesse des classifieurs. Pourtant, elle n'a pas été considérée comme défense pratique. Ce papier révèle sa réelle robustesse face aux attaques boîtes noires de l'état de l'art. Nous avons *i)* illustré de manière formelle le fossé entre certification théorique et défense pratique, et redéfini les exemples adverses face à une défense aléatoire. Nous avons constaté que les recommandations pour la certification sont souvent antagonistes à celles pratiques: *ii)* une faible quantité d'échantillons est suffisante pour perturber les frontières; c'est essentiel pour gêner les attaques boîte noire, et *iii)* une variance de bruit élevée ne renforce pas le classifieur mais fait chuter sa précision.

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *ICLR*, 2014.

[2] J. Chen, M. I. Jordan, and M. J. Wainwright, "HopSkipJumpAttack: A query-efficient decision-based attack," in *IEEE S&P*, 2020.

[3] A. Rahmati, S.-M. Moosavi-Dezfooli, P. Frossard, and H. Dai, "Geoda: a geometric framework for black-box adversarial attacks," in *CVPR*, 2020.

[4] H. Li, X. Xu, X. Zhang, S. Yang, and B. Li, "Qeba: Query-efficient boundary-based blackbox attack," in *CVPR*, 2020.

[5] T. Maho, T. Furon, and E. Le Merrer, "Surfree: a fast surrogate-free black-box attack," in *CVPR*, 2021.

[6] J. Chen and Q. Gu, "Rays: A ray searching method for hard-label adversarial attack," in *SIGKDD*, 2020.

[7] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.

[8] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, "Certified robustness to adversarial examples with differential privacy," in *S&P*, 2019.

[9] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *ICML*, 2019.

[10] G. Yang A. Kapoor H. Salman, M. Sun and J. Z. Kolter, "Denoised smoothing: A provable defense for pretrained classifiers," in *NIPS*, 2020.

[11] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," in *ICLR*, 2018.