Comment choisir son meilleur allié pour une attaque transférable ?

Thibault MAHO [†1] , Seyed-Mohsen MOOSAVI-DEZFOOLI[2] , Teddy FURON [*1]

[1]Univ. Rennes, Inria, CNRS, IRISA, Rennes, France

[2]Imperial College London, UK

Résumé – La transférabilité est une propriété des exemples adverses permettant à une perturbation de tromper plusieurs modèles, rendant la menace d'attaques plus réaliste. Ce papier propose une nouvelle évaluation de la transférabilité en réintégrant la distorsion. Ce nouvel outil montre les attaques transférables moins performantes qu'énoncé dans l'état de l'art surtout si l'attaquant choisit au hasard le modèle source. Nous proposons un outil de sélection, appelé `FiT`, qui vise à choisir le meilleur modèle source avec seulement quelques requêtes préliminaires à la cible.

Abstract – Transferability is a property of adversarial examples that allows a perturbation to deceive several models, making the threat of attacks more realistic. This paper proposes a new methodology for evaluating transferability by putting distortion in a central position. This new tool shows that transferable attacks may perform far worse than a black box attack if the attacker randomly picks the source model. To address this issue, we propose a new selection mechanism, called `FiT`, which aims at choosing the best source model with only a few preliminary queries to the target.

1 Introduction

La transférabilité est une propriété intrigante des exemples adverses qui, créés sur un modèle donné, peuvent également tromper d'autres modèles [7, 15]. La menace d'exemples adverses devient plus réaliste. Dans la pratique, le modèle ciblé est inconnu mais accessible en boîte noire. Les attaques en boîte blanche [11, 17] ne sont ainsi pas applicables. Des attaques en boîte noire existent, mais nécessitent des milliers de requêtes pour trouver un exemple adverse de faible distorsion [2, 9]. Les attaques transférables nécessitent pas ou peu de requêtes pour affiner un exemple adverse créé grâce à un modèle disponible suffisamment similaire à la cible.

La transférabilité est généralement mesurée par le taux de succès de l'attaque (ASR), soit la probabilité que l'exemple adverse conçu pour le modèle source trompe aussi le modèle cible. Nous soutenons que cette mesure est injuste. Dans le contexte des exemples adverses, il ne s'agit pas seulement de générer des données mal classées, mais de trouver la perturbation de distorsion minimale trompant un classifieur.

Pour illustrer, considérons deux modèles : le premier robuste dans le sens où les perturbations le trompant sont importantes, et un second faible. Si le modèle robuste est utilisé comme source, l'ASR de l'attaque transférable sera certainement élevé, mais cela ne signifie pas qu'il s'agit du bon choix. L'ASR est élevé car le modèle source robuste a besoin d'une grande perturbation pour être trompé, ce qui trompera n'importe quel modèle plus faible. La configuration inverse donnera un ASR faible. En résumé, l'ASR ne permet pas à lui seul de déterminer que la *direction de la perturbation* donné par la source est pertinente pour attaquer la cible.

La première contribution de ce papier est de réintégrer la distorsion. La section 3 évalue la transférabilité en comparant la distorsion d'une attaque transférable à celle d'attaques références : l'attaque en boîte blanche directement appliquée au modèle cible et l'attaque en boîte noire.

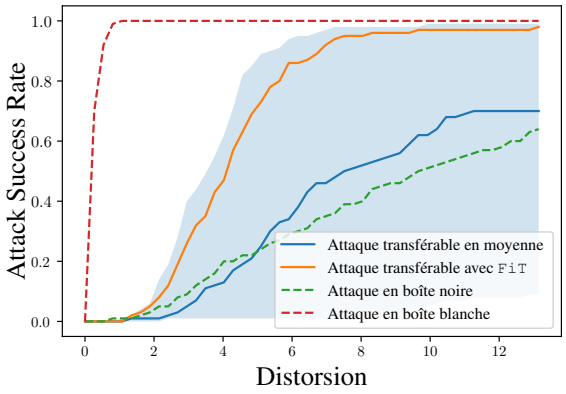

FIGURE 1 : ASR fonction de la distorsion d'attaques en boîte blanche, transférable et en boîte noire contre le modèle $\texttt{CoatLite}_{\texttt{small}}$ (Sec. 4.3.1). La zone bleue est la plage de résultat obtenu par une attaque transférable avec une source aléatoire. Une attaque transférable peut être pire qu'une attaque en boîte noire sans la bonne sélection de la source comme `FiT`.

Ce nouvel outil met en évidence la grande variabilité des résultats des attaques transférables pour différents choix de source. La figure 1 résume cette observation en traçant l'ASR en fonction de la distorsion (protocole expérimental expliqué en section 4.3.1). Naturellement, l'attaque en boîte noire nécessite davantage de distorsion que l'attaque en boîte blanche. La surprise est que si l'attaquant recourt à une attaque transférable et choisit un modèle source au hasard, il y a environ 50% de chances que l'attaque donne des résultats encore plus mauvais que l'attaque en boîte noire.

Cette observation questionne l'idée répandue que les exemples adverses se transfèrent facilement d'un modèle à l'autre et souligne la nécessité de choisir avec soin le modèle source pour attaquer une cible. Dans l'hypothèse où l'attaquant dispose de plusieurs modèles candidats, notre deuxième

[†]Thanks to Rennes Métropole for its funding for international mobility.
[*]Thanks to ANR and AID french agencies for funding Chaire SAIDA.

contribution, nommée FiT, permet à l'attaquant de choisir un bon modèle source avec peu de requêtes à la cible.

2 Travaux associés

2.1 Attaque transférable

Lors d'une attaque transférable, l'attaquant dispose d'un accès limité à la cible placé dans une boîte noire mais dipose d'un autre modèle, appelé source, entraîné pour le même problème de classification que la cible. Les exemples adverses obtenus par une attaque en boîte blanche [11, 17] sont trop spécifiques à la source et sont peu transférables. Les attaques transférables améliorent l'ASR en évitant le surajustement des exemples sur la source. Les transformations d'entrée ont été proposées par DI [16] et TI [3]. Les articles [6, 13] ont stabilisé le gradient tandis que [4, 14] se concentrent sur l'importance des caractéristiques intermédiaires. Plusieurs de ces méthodes sont combinées par TAIG [5]. Des attaques ensemble ont également été proposées. Elles supposent que l'attaquant exploite l'information de plusieurs sources. L'article [13] fait la moyenne des logits de ces sources et mène une attaque en boîte blanche sur cette agrégation de modèles. Ces méthodes ont néanmoins été peu étudiées dû à leur complexité de calcul.

2.2 Empreintes de réseaux

Les modèles sont des actifs précieux et coûteux, dûs à l'expertise, aux données annotées ou à la puissance de calcul. Des méthodes d'empreintes ont été proposées pour protéger cette propriété intellectuelle. Elles fournissent un score de similarité entre deux modèles en examinant les décisions similaires sur des requêtes soumises. Les modèles sont considérés comme similaires s'ils partagent la même décision pour ces entrées. IPGuard [1] crée des exemples adverses proches de la frontière de décision pour l'encadrer. Des perturbations adverses universelles sont utilisées pour caractériser la frontière dans [12]. L'article [8] construit des exemples adverses avec une propriété transférable appelée exemples adverses conférables. FBI [10] est la seule méthode utilisant des images non modifiées pour mesurer la dépendance statistique des prédictions de deux modèles avec l'information mutuelle.

3 Méthodologie

3.1 Notations

Soit $f : [0, 255]^N \to [0, 1]^K$ un classifieur prédisant les probabilités $f(x)$ de K classes pour l'entrée x. Pour une entrée x de classe y, une attaque crée un exemple adverse x_a telle que :

$$\arg\max_{1\leq k\leq K} f_k(x_a) \neq \arg\max_{1\leq k\leq K} f_k(x) = y. \tag{1}$$

Lors d'une attaque transférable, un exemple adverse est construit sur un modèle source f_s et testé sur le modèle cible f_t. Cet article considère un ensemble de m modèles sources désignés par $\mathcal{F}_s = \{f_s^1, ..., f_s^m\}$, et un ensemble $\mathcal{X}$ de n entrées.

3.2 Mesure

Distorsion. La distorsion d'une perturbation est mesurée avec avec la norme Euclidienne, usuelle pour les images :

$$\mathsf{dist}(x_a, x) := \|x_a - x\|_2/\sqrt{N}, \tag{2}$$

avec $N = 3 \times H \times L$ pixels. Cette distorsion peut être considérée comme l'amplitude moyenne de la perturbation par pixel.

Courbe ROC. Pour une attaque et un modèle cible f_t donnés, la courbe ROC est définit comme :

$$P(D) := \mathbb{P}(\mathsf{dist}(x_a, x) < D). \tag{3}$$

Il s'agit de l'ASR en fonction de la distorsion. Contrairement à la littérature qui mesure l'ASR pour quelques niveaux de distorsion, nous considérons l'ensemble des valeurs de D.

3.3 Transférabilité

La méthodologie proposée calcule tout d'abord la courbe ROC $P_t^{\mathsf{wb}}(D)$ d'une attaque en boîte blanche directement appliquée au modèle cible, et celle d'une attaque en boîte noire $P_t^{\mathsf{bb}}(D)$. La courbe ROC mesurée pour l'attaque transférable de f_s vers f_t se situe entre les deux caractéristiques comme suit :

$$T_{s,t} := \frac{\int_0^\infty P_{s,t}(u) - P_t^{\mathsf{bb}}(u)du}{\int_0^\infty |P_t^{\mathsf{wb}}(u) - P_t^{\mathsf{bb}}(u)|du}. \tag{4}$$

Ce score est calculé comme le rapport des aires entre les différentes courbes ROC définissant des fonctions de répartition. Le dénominateur est ainsi la 1-Wasserstein entre les attaques en boîte blanche et boîte noire. Au numérateur, la valeur absolue est manquante, elle a été supprimée pour obtenir un score signé. La figure 1 montre que le numérateur de (4) peut en effet être négatif car la transférabilité peut être pire qu'une attaque en boîte noire et alors $T_{s,t} < 0$. Si l'attaque transférable donne des résultats aussi bons que l'attaque en boîte blanche (aussi mauvais qu'en boîte noire), alors $T_{s,t} = 1$ (resp. $T_{s,t} = 0$).

3.4 Implémentation pratique

Attaques. L'attaque transférable utilise le modèle disponible f_s et l'entrée x pour créer une direction adverse $u_{x,s}$ de norme Euclidienne $\sqrt{N}$ (cela simplifie les notations). Nous supposons qu'il existe un oracle donnant la distorsion minimale le long de cette direction pour tromper la cible t. En d'autres termes, $x_a = x + d\, u_{x,s}$, avec

$$d = \min\{\delta : \arg\max_k f_{t,k}(x + \delta\, u_{x,s}) \neq y\}. \tag{5}$$

Notons que $\mathsf{dist}(x_a, x) = d$. Cette définition privilégie la transférabilité. En pratique, un tel oracle n'existe pas mais l'attaquant trouve une bonne estimation de (5) grâce à une dichotomie avec quelques requêtes à la cible.

Transferabilité. Nous exécutons une attaque donnée sur un ensemble $\mathcal{X}$ de n entrées correctement classées par le modèle cible. Nous calculons les distorsions $d(j) = \mathsf{dist}(x_{a,j}, x_j)$ avec $x_j \in \mathcal{X}$ et les trions de sorte que $d(1) \leq d(2) \leq \ldots \leq d(n)$. Nous fixons $d(0) = 0$ et $d(n + 1) = \infty$ pour définir la fonction constante par morceau suivante :

$$\hat{P}(D) := j/n \quad \forall D \in [d(j), d(j + 1)]. \tag{6}$$

L'estimation des intégrales dans (4) devient facile par une somme de Lebesgue plutôt que par une somme de Riemann. On obtient ainsi :

$$\hat{T}_{s,t} := \frac{\sum_{j=1}^{n} d_{s,t}(j) - d_t^{\mathsf{bb}}(j)}{\sum_{j=1}^{n} d_t^{\mathsf{wb}}(j) - d_t^{\mathsf{bb}}(j)}, \tag{7}$$

où les distorsions $(d_t^{\mathsf{bb}}(j))_j$ (resp. $(d_t^{\mathsf{wb}}(j))_j$) résultant de l'attaque de la boîte noire (resp. de la boîte blanche) contre le modèle f_t sont également classées par ordre croissant.

4 Comment choisir la meilleur source

La transférabilité dépend de trois facteurs : la source, l'attaque et l'image. Nous proposons une procédure qui combine la dépendance au modèle source et à l'image pour une attaque donnée.

$$\texttt{FiT}(s,t,x) := \texttt{ModSim}(s,t) \times \texttt{TransQ}(s,x). \tag{8}$$

la dépendance à l'image est mesurée par la qualité de l'exemple adverse, désignée par `TransQ`. La similarité entre la source et la cible, désignée par `ModSim`, mesure la dépendance au modèle. La transférabilité entre deux modèles de même architecture se révèle meilleure qu'entre deux architectures différentes [15]. Ces indicateurs doivent être faciles à calculer avec un nombre petit de requêtes à la cible.

Ce score ouvre la porte à une nouvelle stratégie pour l'attaquant qui sélectionne d'abord la meilleure source parmi les modèles disponibles

$$s^\star(t,x) = \arg\max_{\sigma \in \mathcal{F}_s} \texttt{FiT}(\sigma,t,x), \tag{9}$$

puis de créer la direction adverse $u_{x,s^\star(t,x)}$.

4.1 Critère **ModSim**(s,t)

La similarité entre modèles a déjà été étudiée dans le contexte d'empreintes de réseaux comme défense pour protéger la propriété intellectuelle (voir Sec. 2.2). Ces méthodes sont utilisées ici comme une attaque donnant des informations sur la cible. Nous considérons la méthode [10] qui nécessite uniquement la décision de la boîte noire. En interrogeant deux modèles f_s et f_t avec quelques images naturelles, elle calcule une distance $\texttt{Dist}(s,t) \in [0,1]$. Une similarité étant attendue, nous fixons $\texttt{ModSim}(s,t) = 1 - \texttt{Dist}(s,t)$. Cependant, la similarité est symétrique, c'est-à-dire $\texttt{ModSim}(s,t) = \texttt{ModSim}(t,s)$, alors que la transférabilité ne l'est pas [15]. Cela montre que ce critère seul n'est pas suffisant.

4.2 Critère **TransQ**(s,x)

Ce critère évalue la transférabilité générale d'un exemple adverse élaboré par une source. Nous utilisons l'hypothèse selon laquelle l'attaquant dispose d'un ensemble de modèles $\mathcal{F}_s$. La transférabilité est ainsi évaluée grâce aux autres modèles de cet ensemble, sans interroger la cible.

Pour une entrée donnée x, la source s fournit la direction adverse $u_{x,s}$ et nous calculons la distorsion $d_{s,\sigma}$ nécessaire pour tromper le classifieur $f_\sigma \in \mathcal{F}_s$ avec (5). Nous agrégeons ensuite ces distorsions en un score unique avec deux méthodes :

$$\texttt{TransQ}^{(1)}(s,x) := \left(\frac{1}{|\mathcal{F}_s|} \sum_{f_\sigma \in \mathcal{F}_s} d_{s,\sigma} \right)^{-1}. \tag{10}$$

Une bonne source pour l'entrée x donne lieu à des distorsions plus faibles, de sorte que $\texttt{TransQ}^{(1)}(s,x)$ est important.

$$\texttt{TransQ}^{(2)}(s,x) := \frac{\sum_{f_\sigma \in \mathcal{F}_s} d_{s,\sigma} - d_\sigma^{\mathsf{bb}}}{\sum_{f_\sigma \in \mathcal{F}_s} d_\sigma^{\mathsf{wb}} - d_\sigma^{\mathsf{bb}}}. \tag{11}$$

Cette mesure est similaire à (7), mais est calculée ici sur l'ensemble des modèles au lieu d'un ensemble d'entrées.

4.3 Evaluation

4.3.1 Configuration expérimental

L'étude porte sur deux attaques transférables - `DI` [16], `TAIG` [5]- utilisant deux approches différentes pour améliorer la transférabilité (voir Sect. 2.1). Elles partagent un paramètre ϵ permettant de contrôler la perturbation maximale ajoutée par pixel qui est fixé à $\epsilon = 8$. Ces attaques sont évaluées sur 48 modèles issus de la librairie Timm. Les expériences utilisent 100 images de l'ensemble de validation de l'ILSVRC'12 correctement classées par tous les modèles considérés.

Le score de transférabilité (7) requiert des attaques en boîte noire et boîte blanche. Certaines méthodes peuvent favoriser un modèle, nécessitant l'utilisation de plusieurs attaques. Notre étude utilise deux attaques en boîte blanche (`BP` [17], `DeepFool` [11]) et deux en boîte noire (`SurFree` [9], `RayS` [2]), toutes faisant partie de l'état de l'art. Les attaques en boîte noire sont exécutées avec 2 000 appels, ce qui est jugé suffisant pour leur convergence. Nous enregistrons la plus petite perturbation boîte noire (boîte blanche) pour chaque image et calculons la courbe (6) telle qu'elle apparaît dans la ligne pointillée verte (resp. rouge) de la figure 1.

En ce qui concerne l'attaque transférable, pour un modèle cible donné, l'attaquant a accès à un sous-ensemble de tous les autres modèles dont l'architecture diffère de celle du modèle cible. Cela représente en moyenne 45 modèles sur 48.

Nous sélectionnons la méthode `FBI` [10] avec 200 images naturelles bénignes pour calculer le critère $\texttt{ModSim}(s,t)$. Cela implique que l'attaquant effectue 200 appels à la cible dans une étape préliminaire avant de forger un exemple adverse.

4.3.2 Résultat

Critère $\texttt{ModSim}(s,t)$. Le tableau 1 indique que la similarité architecturale est une mesure fiable de la transférabilité entre deux modèles. Elle peut conduire à la sélection d'une bonne source donnant naissance à une attaque transférable plus performante que l'attaque en boîte noire puisque $\hat{T}_{s,t}$ est supérieure à 0. Cependant, les résultats restent faibles par rapport aux meilleurs résultats obtenable. La similarité peut ne pas suffire car elle implique la sélection d'un modèle unique pour une cible donnée. La figure 2 montre que de meilleurs résultats sont obtenus en adaptant la source à l'entrée.

Critère $\texttt{TransQ}(s,x)$. L'amélioration de la transférabilité sans interroger le modèle cible dans une étape préliminaire est possible grâce à $\texttt{TransQ}(s,x)$. Les exemples adverses

TABLE 1 : Mesure de transférabilité pour DI [16] et TAIG [5].

Méthode de sélection		DI [16]	TAIG [5]
Meilleur source par image		0.52	0.46
Source aléatoire par image		-0.16	-0.12
ModSim	FBI [10]	0.18	0.12
TransQ	ASR	-0.21	-0.24
	TransQ$^{(1)}$(10)	0.38	0.24
	TransQ$^{(2)}$(11)	0.37	0.23
FiT	TransQ$^{(1)}$(10)	**0.40**	**0.27**
	TransQ$^{(2)}$(11)	0.39	0.25

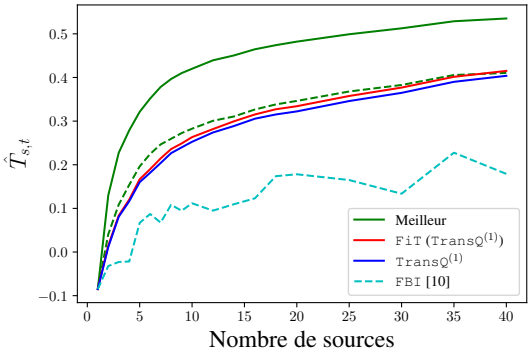

FIGURE 2 : $\hat{T}_{s,t}$ en fonction du nombre de sources disponibles pour plusieurs méthodes de sélection. En pointillé, sélection d'un modèle unique pour toutes les images ; en plein sélection d'un modèle par image. L'attaque est DI [16].

qui présentent une bonne transférabilité sur plusieurs modèles sont plus susceptibles d'également tromper le modèle cible inconnu. La figure 2 montre qu'une amélioration significative de la transférabilité est obtenue même avec seulement quelques modèles disponibles. Le tableau 1 confirme cette observation pour TAIG [5]. Cette stratégie est en effet meilleure que la sélection basée sur la similarité des modèles.

Score FiT (s,t,x). La combinaison des deux critères (8) conduit à une légère amélioration de la transférabilité par rapport à TransQ(s, x) seul. La figure 2 confirme que cela vaut pour une large gamme de nombres de sources disponibles. Pour les attaques à modèle unique, TransQ$^{(1)}$donne des résultats légèrement meilleurs que TransQ$^{(2)}$.

5 Conclusion

La transférabilité est une caractéristique essentielle des exemples adverses. Elle permet à une perturbation unique de tromper plusieurs modèles. Cependant, la distorsion nécessaire pour tromper un modèle est négligée avec l'ASR. Cet article présente une nouvelle approche pour évaluer la transférabilité en la comparant à la distorsion des attaques en boîte blanche et boîte noire. Nous montrons que les attaques transférables peuvent être moins performantes que les attaques en boîte noire sans une sélection appropriée du modèle source, ce qui souligne la nécessité de choisir le meilleur modèle source pour cibler un modèle spécifique. La solution proposée, appelée FiT, permet à l'attaquant de choisir l'un des meilleurs modèles sources avec un minimum de requêtes à la cible.

Références

[1] Xiaoyu CAO, Jinyuan JIA et Neil Zhenqiang GONG : Ipguard : Protecting intellectual property of deep neural networks via fingerprinting the classification boundary. *In ACM ASIACCS*, 2021.

[2] Jinghui CHEN et Quanquan GU : Rays : A ray searching method for hard-label adversarial attack. *In ACM SIGKDD*, 2020.

[3] Yinpeng DONG, Tianyu PANG, Hang SU et Jun ZHU : Evading defenses to transferable adversarial examples by translation-invariant attacks. *In CVPR*, 2019.

[4] Qian HUANG, Isay KATSMAN, Horace HE, Zeqi GU, Serge BELONGIE et Ser-Nam LIM : Enhancing adversarial example transferability with an intermediate level attack. *In ICCV*, 2019.

[5] Yi HUANG et Adams Wai-Kin KONG : Transferable adversarial attack based on integrated gradients. *In ICLR*, 2022.

[6] Jiadong LIN, Chuanbiao SONG, Kun HE, Liwei WANG et John E HOPCROFT : Nesterov accelerated gradient and scale invariance for adversarial attacks. *In ICLR*, 2020.

[7] Yanpei LIU, Xinyun CHEN, Chang LIU et Dawn SONG : Delving into transferable adversarial examples and black-box attacks. *In ICLR*, 2017.

[8] Nils LUKAS, Yuxuan ZHANG et Florian KERSCHBAUM : Deep neural network fingerprinting by conferrable adversarial examples. *In ICLR*, 2021.

[9] Thibault MAHO, Teddy FURON et Erwan LE MERRER : Surfree : a fast surrogate-free black-box attack. *In CVPR*, 2021.

[10] Thibault MAHO, Teddy FURON et Erwan Le MERRER : Fbi : Fingerprinting models with benign inputs. *arXiv preprint arXiv :2208.03169*, 2022.

[11] Seyed-Mohsen MOOSAVI-DEZFOOLI, Alhussein FAWZI et Pascal FROSSARD : Deepfool : a simple and accurate method to fool deep neural networks. *In CVPR*, 2016.

[12] Zirui PENG, Shaofeng LI, Guoxing CHEN, Cheng ZHANG, Haojin ZHU et Minhui XUE : Fingerprinting deep neural networks globally via universal adversarial perturbations, 2022.

[13] Xiaosen WANG, Jiadong LIN, Han HU, Jingdong WANG et Kun HE : Boosting adversarial transferability through enhanced momentum. *In BMVC*, 2021.

[14] Zhibo WANG, Hengchang GUO, Zhifei ZHANG, Wenxin LIU, Zhan QIN et Kui REN : Feature importance-aware transferable adversarial attacks. *In ICCV*, 2021.

[15] Lei WU et Zhanxing ZHU : Towards understanding and improving the transferability of adversarial examples in deep neural networks. *In ACML*. PMLR, 2020.

[16] Cihang XIE, Zhishuai ZHANG, Yuyin ZHOU, Song BAI, Jianyu WANG, Zhou REN et Alan L YUILLE : Improving transferability of adversarial examples with input diversity. *In CVPR*, 2019.

[17] Hanwei ZHANG, Yannis AVRITHIS, Teddy FURON et Laurent AMSALEG : Walking on the edge : Fast, low-distortion adversarial examples. *IEEE T-IFS*, 2020.