How to choose your best allies for a transferable attack?

Thibault Maho, Seyed-Mohsen Moosavi-Dezfooli, Teddy Furon

How to choose your best allies for a transferable attack?

Thibault Maho*
Univ. Rennes, Inria, CNRS
IRISA, Rennes, France
thibault.maho@inria.fr

Seyed-Mohsen Moosavi-Dezfooli
Imperial College London, UK
seyed.moosavi@imperial.ac.uk

Teddy Furon †
Univ. Rennes, Inria, CNRS
IRISA, Rennes, France
teddy.furon@inria.fr

Abstract

The transferability of adversarial examples is a key issue in the security of deep neural networks. The possibility of an adversarial example crafted for a source model fooling another targeted model makes the threat of adversarial attacks more realistic. Measuring transferability is a crucial problem, but the Attack Success Rate alone does not provide a sound evaluation. 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. Our experimental results show that* `FiT` *is highly effective at selecting the best source model for multiple scenarios such as single-model attacks, ensemble-model attacks and multiple attacks.*

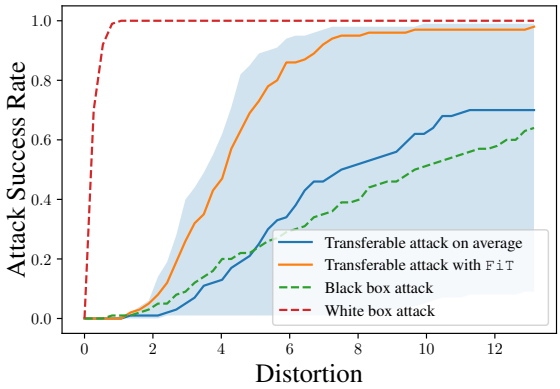


Figure 1: Evaluation of transferability by comparing the Attack Success Rate vs. distortion trade-off of a white box, transferable, and black box attacks against model $\texttt{CoatLite}_{\texttt{small}}$ (See Sect. 4.1 for details). The blue area is the range of trade-off operated by a transferable attack with random source models. A transferable attack may be worse than a black box attack without a good source selection (like `FiT`).

1. Introduction

Transferability is one of the most intriguing properties of adversarial examples. A white box attack crafting adversarial examples for an open-source model is likely to fool other models too [1, 14, 20, 25, 34]. This makes the threat of adversarial examples more realistic. In practice, the model targeted is usually unknown but accessible as a black box. This prevents directly applying any white box gradient-based attack [10, 16, 19, 35]. Black box attacks do exist but they require some thousands of queries to find an adversarial example of low distortion [4, 11, 17, 24]. Transferable attacks require no or few queries to fine-tune an adversarial example thanks to the help of a publicly available model similar enough to the target.

Transferability is usually measured by the Attack Success Rate (ASR), i.e., the probability that the adversarial example crafted for the source model also deludes the target model. We argue that this measure leads to an unfair evaluation of transferability. In the context of adversarial examples, it is not just a matter of discovering data that is not well classified, but rather identifying the perturbation that can fool a classifier with minimal distortion. This principle should also apply to transferable attacks.

For illustration purposes, let us consider two models, one is robust in the sense that the necessary amount of adversarial perturbation is large, whereas the other model is weak. If the attacker uses the robust model as the source to attack the weak target network, the ASR of the transferable attack will certainly be big. It does not mean that this is the right choice. The ASR is high because the robust source model needs large perturbation to be deluded, which will fool any weaker model. The ASR alone does not reflect the overshooting in distortion. The converse, using the weak to at-

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

tack the robust, would yield a low ASR. To summarize, the ASR alone fails to capture how relevant the *direction of the perturbation* given by the source is for attacking the target.

The first contribution of this paper is to put distortion back into the picture. Section 3 evaluates transferability by comparing the distortion of a transferable attack to the ones of two reference attacks: On one hand, the strongest attack, i.e., the white box attack directly applied on the target model; on the other hand, the weakest attack, i.e., the black box attack.

The second contribution shows the great variability of the performance of transferable attacks. Figure 1 summarizes this observation by plotting the ASR as a function of the distortion (the experimental protocol is explained in Sect. 4.1). Naturally, the black box attack needs much more distortion than the white box attack. For instance, the white box attack yields an ASR of 50% with a distortion of 0.19, whereas the black box attack needs a distortion of 9.7. The surprise is that if the attacker resorts to a transferable attack and picks a source model at random, there is almost a 50% chance that the attack performs even worse than the black box attack. Section 4 outlines a triad of factors: input, source model, and attack.

This observation challenges the prevalent notion that adversarial examples transfer easily between models, and highlights the need to carefully choose the source model to attack a target. Under the assumption that the attacker has indeed several candidate models, our third contribution, named `FiT` in Sect. 5, provides an affordable measure for model selection, allowing the attacker to choose a good source model with only a few queries to the target.

2. Related work

2.1. White box, black box, and transferable attacks

The threat analysis related to adversarial examples usually considers two scenarios: the white box assumption where the attacker knows the internals of the target model, and the black box assumption where he does not but has limited access to the target.

White box attacks are now performing very well in terms of Attack Success Rate and distortion. We distinguish two kinds of attacks: 1) attacks constrained by a distortion budget like `PGD` [16] whose performance is measured by the ASR, and 2) attacks yielding almost surely an adversarial example like `DeepFool` [19] and `CW` [3] whose performance is measured by the distortion. The trend is to make them speed efficient as well with fewer computations of the gradient of the neural network, like `FMN` [23] and `BP` [35].

Our paper also considers so-called decision-based black box attacks like `GeoDA` [24], `SurFree` [17], and `RayS` [4], where the attacker only sees the class predicted by the model, but not the confidence score, neither the probit nor the logit. A decision-based black box attack typically needs thousands of queries to find an adversarial example with a relatively low distortion, yet, much higher than with a white box attack. This is the price the attacker has to pay when he does not know the target model.

The transferable attacks, which pertain to the black box scenario, disrupt this last statement. The assumption is that attacker knows another model, called source, trained for the same classification problem as the target. Yet, leading a white box attack on the source model with the hope that this adversarial example will also fool the target yields poor results because the adversarial perturbation is too specific to the source. Modern transferable attacks increase the ASR by avoiding the overfitting of examples on the source. Input transformations have been purposed by `DI` [32], `TI` [6] and `Admix` [27]. Papers [8, 13, 28] stabilized the gradient while [7,29,36,39] focus on the importance of intermediate features. The combination of several methods has also been investigated [9,40]. For instance, `TAIG` [9] uses input transformation, integrated gradients, and attention reduction.

Ensemble-model attacks make the assumption that the attacker leverages not a single source but several source models. Paper [28] averages the logits of several sources and leads a white box attack on this aggregation of models, whereas [33] reduces model variance. The main obstacle to these methods is the computational complexity. To counteract this drawback, some methods design a specific ensemble of sources like ghost networks [12] or pruned network [26]. This amounts to creating several sources without the need to compute the gradient of each model.

All this literature uses the ASR as the figure of merit for a given distortion while our paper draws the full operating characteristic of ASR vs. distortion. Moreover, the performance of the transferability is compared neither to the white box nor the decision-based black box attack.

2.2. Fingerprinting

Models are valuable assets because of their training costs, be it the expertise, the annotated data, or the computational power. Fingerprinting methods have been proposed to protect this intellectual property. They provide a similarity score between the two models by looking at similar decisions in submitted queries. They often use adversarial examples specifically designed to be unique for a model. The models are considered similar if they share the same decision for these inputs. `IPGuard` [2] creates adversarial examples close to the decision boundary to frame it. Universal adversarial perturbations are used to characterize the boundary in [21]. The paper [15] built adversarial examples with a transferable property called conferrable adversarial examples. `AFA` [37] crafts adversarial examples model-specific using dropout. `FBI` [18] is the only method using unmodified images. It estimates the mutual information to measure

the statistical dependence of two models' predictions.

Our paper investigates the idea that a model similar to the target may be a good source for transferable attacks.

3. Methodology

This paper introduces a new way to gauge transferability. The main motivation is that transferability should indicate whether a source s is useful for attacking a target t, independently from the inherent robustness of t.

3.1. Notations

Let $f : [0, 255]^N \to \mathcal{R}^K$ denote a classifier computing the predicted probabilities $f(x)$ of K classes for input x. For a given input x of class y, the adversarial example x_a is the result of an attack (be it white box, black box, or by transferability) such that

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

In the case of transferable attacks, the attacker uses a source model f_s to build a transferable adversarial example against the target model f_t. This paper considers a collection of m source models denoted by $\mathcal{F}_s = \{f_s^1, ..., f_s^m\}$, and a set of n inputs $\mathcal{X}$.

3.2. Measurement

Distortion. We measure the distortion of an adversarial perturbation using the Euclidean norm which is a common choice as we deal with natural images:

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

with $N = 3 \times H \times L$ pixels. This distortion can be seen as the average amplitude of the perturbation per pixel.

Operating characteristic. Given an attack (be it white box, black box, or by transferability) and a target model f_t, we call operating characteristic the following function:

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

It is thus the Attack Success Rate as a function of the distortion. Contrary to the literature which measures the ASR for a few distortion levels, we consider the full range of D values.

3.3. Transferability

Given a white box and a black box attacks, The proposed methodology first computes the operating characteristic $P_t^{\mathsf{wb}}(D)$ of a state-of-the-art white box attack directly applied to the target model, and the operating characteristic $P_t^{\mathsf{bb}}(D)$ of a state-of-the-art black box attack. It then measures where the operating characteristic of the transferable attack from the source s to the target t lies in between the two characteristics as follows:

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

The proposed score is calculated as the ratio of the areas between the different operational characteristics, which are defined as Cumulative Distribution Functions (CDFs). The numerator is therefore close to the 1-Wasserstein distance [5] between the ASR function of the distortions of the transferable attack and that of the black box attack, where the absolute value is removed to obtain a signed score. If the transferable attack performs as well as the white box attack (resp. as bad as the black box attack), then $T_{s,t} = 1$ (resp. $T_{s,t} = 0$). Figure 1 shows that the numerator of (4) can indeed be negative as transferability can be even worse than the black box if the source s is not well chosen. Therefore, zero is not a lower bound for $T_{s,t}$. In the denominator, the 1-Wasserstein distance is obtained exactly between the white box and the black box because the white box consistently generates adversarial examples with lower distortion.

3.4. Practical implementation

Attacks. We only consider state-of-the-art attacks which almost surely deliver an adversarial example. As said in Sect. 2.1, this is usually the case for black box attacks, but we are limited to using white box attacks that are *not* subject to a distortion budget. As for transferability, the attack uses the publicly available model s and input x to craft an adversarial direction $u_{x,s}$ of Euclidean norm $\sqrt{N}$. The use of $\sqrt{N}$ for the norm is for the sake of simplicity in notation. We assume that there is an oracle giving the minimal distortion along this direction to fool the target t. In other words, $x_a = x + d\, u_{x,s}$, with

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

Note that $\mathsf{dist}(x_a, x) = d$. This definition favours transferability as its best. In practice, such an oracle does not exist but the attacker finds a good estimate of (5) thanks to a line search within a few queries to the target.

Transferability. We run a given attack over a collection $\mathcal{X}$ of n inputs correctly classified by the target model. We compute the distortions $d(j) = \mathsf{dist}(x_{a,j}, x_j)$ with $x_j \in \mathcal{X}$ and sort them so that $d(1) \leq d(2) \leq \ldots \leq d(n)$. We set $d(0) = 0$ and $d(n+1) = \infty$ to properly define the empirical operating characteristic by the following step function:

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

It is then easier to estimate the integrals appearing in (4), which are indeed areas under two curves, by a Lebesgue

sum rather than a Riemann sum. This gives:

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

where the distortions $(d_t^{\mathsf{bb}}(j))_j$ (resp. $(d_t^{\mathsf{wb}}(j))_j$) resulting from the black box (resp. white box) attack against model t are also sorted in increasing order.

4. Triad of transferability: data, model, attack

This section is an experimental investigation of the factors affecting transferability.

4.1. Experimental setup

The study assesses transferable attacks on a total of 48 models, with 47 of them sourced from the Timm library [30] and one very robust, namely $\texttt{ResNet50}_{\texttt{AdvTrain}}$, obtained from the GitHub repository [1]. Appendix A lists all the models. The experiments utilize 100 images from the validation set of ILSVRC'12, all of which are correctly classified by all models under consideration.

The study considers three transferable attacks - `DI` [32], `TAIG` [9], and `DWP` [26]- each using a different approach to improve transferability (see Sect. 2.1). These attacks are selected as the best in their categories in [38]. They all share an ϵ parameter to control the maximum perturbation added per pixel. The effect of the parameter ϵ indeed happens to be negligible in our protocol, as demonstrated in Appendix B. We choose $\epsilon = 8$.

To measure transferability (7), we need state-of-the-art black box and white box attacks. Certain methods may exhibit a preference towards one model over the other, necessitating the use of multiple attacks. Our study employs four white box attacks (`BP` [35], `DeepFool` [19], `I-FGSM` [10], and `PGD` [16]) and three black box attacks (`SurFree` [17], `RayS` [4], and `GeoDA` [24]). All black box attacks are run with 2,000 queries, which has been determined to be sufficient for achieving convergence. For distortion-constrained white box attack, the ϵ parameter is set to 4. We record the smallest perturbation distortion over the black box (white box) for each image and draw the operating characteristic (6) as appearing in green (resp. red) dashed line in Fig. 1.

As for the transferable attack, for a given target model, the attacker has access to a subset of all the other models whose architecture differs from the target. This amounts to an average of 45 models out of 48. For instance, in Fig. 1, $\texttt{CoatLite}_{\texttt{small}}$ being the target, we exclude all other `CoatLite` models from being a source. The light blue area delimits the operating characteristics of transferable attack `DI` [32] using as the source one of the 45 remaining models.

[1] https://github.com/MadryLab/robustness

4.2. Model dependence

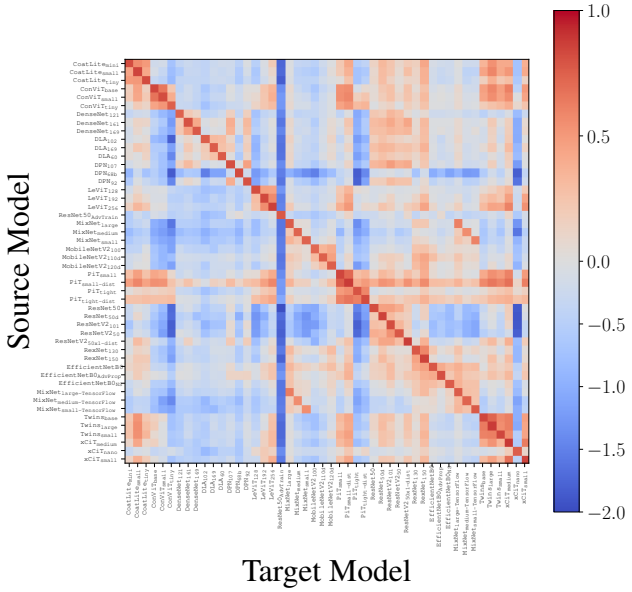


Figure 2: Transferability score $\hat{T}_{s,t}$ matrix of 48 sources and 48 targets listed in App. A with attack `DI` [32].

Large transferability variation. Figure 2 shows the matrix $\hat{T}_{s,t}$, where s and t are any of the 48^2 pairs, for attack `DI` [32]. Not all models possess the same transferability capabilities. At first look, the figure contains more blue than red cells which means that transferability takes negative value more often.

Some rare models, like $\texttt{PiT}_{\texttt{small-dist}}$, exhibit good transferability towards any target. On the contrary, `DPN68b` is always a bad source. On the other hand, $\texttt{ResNet50}_{\texttt{AdvTrain}}$ is a very difficult target (note how-

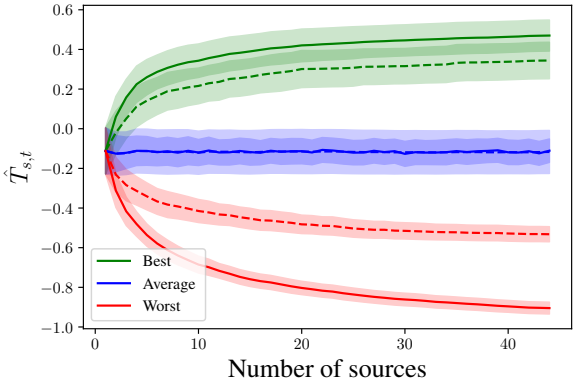


Figure 3: $\hat{T}_{s,t}$ function of the number of available sources and attack `DI` [32]. The best or worst model selected per image (solid line) or on average (dashed line).

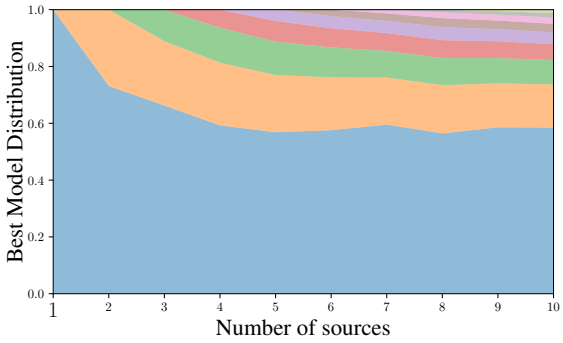


Figure 4: Distribution of best image-model selection function of the number of sources. Adversarial examples obtained with DI [32].

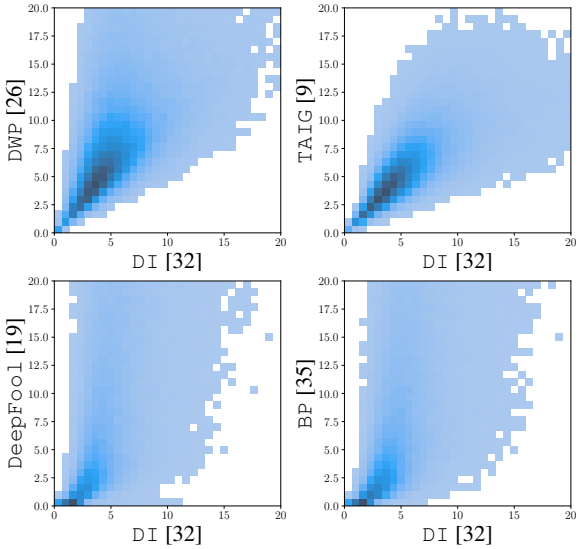


Figure 5: 2D Histogram of the minimum distortion for transferable and white box attacks.

ever that its accuracy is low), followed by $\texttt{PiT}_{\texttt{light}}$ and $\texttt{ConViT}_{\texttt{tiny}}$, whereas the family of ResNet are fairly easy target.

Prior works have shown that models with architectures similar to the target transfer better [22, 31]. The red squares appearing in the matrix confirm this (models are ordered according to architecture family). For instance, MobileNetV2 and RexNet architectures exhibit transferability close to 0.7. However, this is not an absolute truth. For example, EfficientNet-B0 has better transferability against $\texttt{MixNet}_{\texttt{large}}$ than $\texttt{MixNet}_{\texttt{medium}}$, even though they have similar architectures. ConViT transfers exceptionally well to Twins although their architectures are very different. Similar observations hold for the other attacks but with lower transferability scores (see Appendix C).

From now on, we exclude models whose architecture is similar to the target.

Impact on the attack. Selecting the right source among several available models is critical for achieving high transferability. Figure 3 displays the transferability score (7) for the best and the worst choices (shown in dotted lines) as the number of source candidates increases. In this evaluation, a single model is selected as the source for building all adversarial examples. On average, the transferability is lower than zero, regardless of the attack method. This means that transferable attacks perform worst than black box attacks on average. If the attacker knows how to select the best source model, $\hat{T}_{s,t}$ quickly converges to a maximum which is positive but below 0.5. This means that transferability at its best performs closer to a black box attack rather than a white box attack. These remarks highly mitigate the threat of transferability and put emphasis on the crucial selection of the best source model. As far as we know, these facts are not reported in the adversarial examples literature.

4.3. Image dependence

The difficulty of transferring adversarial examples from a source to a target varies significantly depending on the input x. This is illustrated in Fig. 4, which shows the distribution of the best-performing model for each image based on the available sources. Even with a large number of sources available, there is typically one source that performs significantly better than the others, but only over 60% of the images on average. However, this superiority decreases rapidly as the number of available sources increases. It means that a better source exists for 40% of the images on average.

Supposing that the attacker knows the best source of each input, Figure 3 shows in solid line that the transferability converges to a value close to 0.5. The performance of the transferable attack lies in between the ones of the white box and black box attacks. This highlights the importance of selecting an appropriate source model for a given target and input.

4.4. Attack dependence

Transferable attack. Figure 5 compares transferable attacks with a 2D histogram of the distortion pair for two attacks. This is computed over all inputs in $\mathcal{X}$ and all pairs of source and target models. The results show that DWP [26] exhibits poor transferability compared to DI [32] and TAIG [9], which produce adversarial perturbations with similar distortion. Additionally, regardless of the attack complexity or method, the challenging images remain consistent. If one attack requires a high distortion for a given input / source / target, other attacks are likely to encounter

similar difficulties.

Traditional white box attack. We now compare the methods designed for transferability with a naive approach. A white box attack (ie. not specific to transferability) is executed on the source, and then the direction found serves as $u_{x,s}$ to place the adversarial examples on the target boundary with (5). The best $\hat{T}_{s,t}$ score is 0.27 for BP [35] compared to 0.52 for DI [32]. This confirms the superiority of the recent methods designed for transferability. However, even though DI [32] is on average better, Figure 5 shows that when the necessary distortion is small, traditional white box attacks like DeepFool [19] and BP [35] indeed beats DI [32]. On the other hand, DI [32] performs much better for inputs requiring more distortion.

5. How to choose the best source

Section 4 outlines that the choice of the source is of utmost importance thanks to the transferability measure defined in Sect. 3. This section investigates whether the attacker can guess which model is the best source.

We now propose a procedure which combines the dependences with respect to the source and target models (Sect. 4.2), and the input (Sect. 4.3). Model dependence is measured by evaluating the similarity between the source and target models, which is denoted as ModSim. On the other hand, the image dependence is measured by the quality of the adversarial example, denoted as TransQ. Both metrics are combined into the following score:

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

These indicators should be easy to compute. We especially pay attention to the number of queries to the target. This score opens the door to a new strategy for the attacker which first selects the best source among the available models

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

and then crafts the adversarial direction $u_{x,s^\star(t,x)}$.

5.1. Criterion $\texttt{ModSim}(s,t)$

Section 4.2 highlights the correlation between the transferability and the similarity between the source and target models. Gauging model similarity has been previously studied in the context of fingerprinting as a defense to protect intellectual property (see Sect. 2.2). This paper uses fingerprinting methods as an attack that leaks information about the target.

We consider the fingerprinting method [18] because it works in a decision-based setup since the target is a black box in our application. Querying two models s and t with few natural images, it computes a distance $\texttt{Dist}(s,t) \in [0,1]$ by comparing their outputs. Since we look for a similarity, we set $\texttt{ModSim}(s,t) = 1 - \texttt{Dist}(s,t)$. However, this provides a symmetrical similarity, ie. $\texttt{ModSim}(s,t) = \texttt{ModSim}(t,s)$, while transferability is not (see Fig. 2). This shows that this criterion alone is not sufficient.

5.2. Criterion $\texttt{TransQ}(s,x)$

This criterion evaluates the general transferability of a given adversarial example crafted by a source. Our idea is to leverage the assumption that the attacker has a set of models $\mathcal{F}_s$. Consequently, we can evaluate the transferability thanks to the other models of this set, without querying the target.

For a given input x, the source s provides the adversarial direction $u_{x,s}$ and we compute the distortion $d_{s,\sigma}$ necessary to delude classifier $\sigma \in \mathcal{F}_s$ with (5). We then aggregate these distortions into a single score with two flavours:

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

A good source for input x gives birth to lower distortions, so that $\texttt{TransQ}^{(1)}(s,x)$ is large.

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

This measure is similar to (7) except that it is computed over the set of models instead of a set of inputs.

5.3. Results

The experimental setup is the same as in Sect. 4.1. We select the fingerprinting method FBI [18] with 200 benign natural images to compute the criterion $\texttt{ModSim}(s,t)$. It implies that the attacker first makes 200 queries to the target in a preliminary step before forging any adversarial example. Appendix D.1 shows that more images improve the fingerprinting accuracy hence the model selection, but the results converge after a number of 200 images.

This section is structured as follows: we use FiT to select the best model for single-model attacks, then we employ it to identify the best subset of models for ensemble-model attacks and different combinations of attacks.

5.3.1 Single-model attacks

Criterion $\texttt{ModSim}(s,t)$. Table 1 indicates that architectural similarity is a reliable measure of transferability between two models. It can drive the selection of a good source giving birth to a transferable attack outperforming the black box attack since $\hat{T}_{s,t}$ is larger than 0 (except for DWP [26]). Yet, the results remain low compared to the best results obtained. As discussed in Sec 4.2, similarity may not suffice because it implies the selection of one unique for a given target. Better results are achieved when adapting the source to the input.

Table 1: Transferability $\hat{T}_{s,t}$ for DI [32], TAIG [9] and DWP [26] for single and ensemble-model attacks.

Category	Selection Method		DI [32]		TAIG [9]		DWP [26]	
Single-model attack	Best		0.52	±0.12	0.46	±0.12	0.34	±0.13
	Random		−0.16	±0.26	−0.12	±0.21	−0.72	±0.32
	ModSim	FBI [18]	0.18	±0.17	0.12	±0.18	−0.39	±0.19
	TransQ	ASR	−0.21	±0.39	−0.24	±0.30	−1.14	±0.95
		TransQ$^{(1)}$(10)	0.38	±0.15	0.24	±0.16	0.10	±0.20
		TransQ$^{(2)}$(11)	0.37	±0.18	0.23	±0.18	0.08	±0.22
	FiT	TransQ$^{(1)}$(10)	**0.40**	**±0.13**	**0.27**	**±0.17**	**0.12**	**±0.20**
		TransQ$^{(2)}$(11)	0.39	±0.16	0.25	±0.18	0.10	±0.22
Ensemble-model attack with three sources	Best		0.72	±0.08	0.62	±0.08	0.46	±0.1
	Random		0.43	±0.12	0.40	±0.11	0.02	±0.17
	ModSim	FBI [18]	0.59	±0.10	0.49	±0.10	0.05	±0.11
	TransQ	ASR	0.53	±0.21	0.45	±0.27	0.06	±0.27
		TransQ$^{(1)}$(10)	0.62	±0.11	0.54	±0.13	0.33	±0.13
		TransQ$^{(2)}$(11)	0.61	±0.11	0.54	±0.12	0.33	±0.17
	FiT	TransQ$^{(1)}$(10)	**0.64**	**±0.11**	0.55	±0.14	0.35	±0.15
		TransQ$^{(2)}$(11)	**0.64**	**±0.10**	**0.57**	**±0.14**	**0.36**	**±0.15**

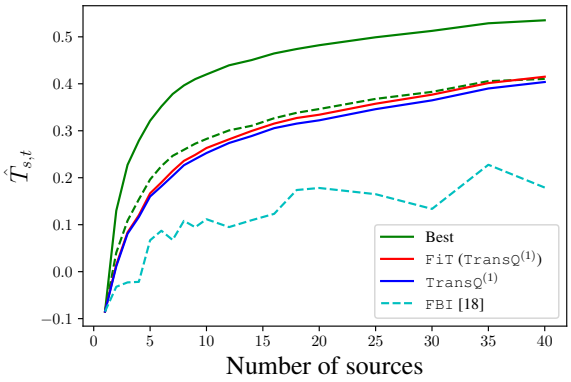


Figure 6: $\hat{T}_{s,t}$ as a function of the number of available sources for several selection methods. Dotted lines refer to the selection of a unique model for all images and solid lines refer to a model selected per image. Attack is DI [32].

Criterion TransQ(s, x). Improving transferability without querying the target model in a preliminary step is possible thanks to TransQ(s, x). Adversarial examples that exhibit good transferability on multiple models are more likely to also deceive the unknown targeted model. Figure 6 shows that a significant improvement in transferability is achieved even with only a few available models. Table 1 confirms this observation for the two other attack methods. This strategy is indeed better than the selection based on model similarity.

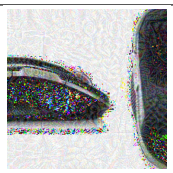
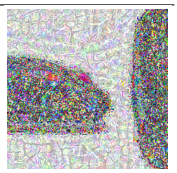
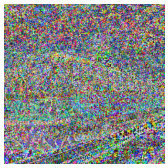
Score FiT (s,t,x). Combining both criteria together as in (8) leads to a slight improvement in transferability compared to TransQ(s, x) alone. Fig. 6 confirms this holds over a wide range of numbers of available sources. For single-model attacks, TransQ$^{(1)}$gives slightly better results than TransQ$^{(2)}$.

Visual results Figure 2 visually demonstrates the impact of different model selection methods on the quality of adversarial examples. Even when FiT is not accurate, the resulting adversarial examples are still close to the best ones obtained from the source models, and the perturbation remains imperceptible. However, random selection generates noisy perturbations, and the worst-case scenario destroys the image entirely.

5.3.2 Ensemble-model attacks

Importance of model selection in ensemble-model attack. Ensemble attacks remain an under-studied area due to the significant computational resources required for evaluating attacks. Consequently, these attacks have only been evaluated with a limited number of sources. Appendix D.2 shows that increasing the number of sources is not necessarily beneficial. The FiT score provides a scalable solution: we select a small subset of sources based on their FiT scores (top-3) from the bigger set of available sources. In a way, it is better to put quality above quantity. For example, when running an ensemble-model attack against $\text{xCiT}_{\text{nano}}$, selecting only the three best models using the FiT measure can lead to a significant improvement in transferability compared to using a larger set of models. In our experiments, an ensemble-model of 20 random models achieved a $\hat{T}_{s,t}$ of 0.47, while an ensemble-model of only three models se-

Table 2: Visual impact of the source selection with DI [32] attacking $\texttt{ConViT}_{\texttt{base}}$ (first row) and $\texttt{DPN}_{92}$ (second row).

	Original	Best	FiT ($\texttt{TransQ}^{(1)}$)	Random	Worst
					
label	mouse	oil filter	purse	purse	jigsaw puzzle
source		$\texttt{PiT}_{\texttt{small-dist}}$	$\texttt{ConViT}_{\texttt{base}}$	$\texttt{MobileNetV2}_{\texttt{110d}}$	$\texttt{DenseNet}_{121}$
distortion		5.62	8.13	35.9	87.8
					
label	tram	elec. locomotive	elec. locomotive	racing car	jigsaw puzzle
source		$\texttt{ResNetV2}_{\texttt{50x1-dist}}$	$\texttt{PiT}_{\texttt{small-dist}}$	$\texttt{DLA}_{60}$	$\texttt{MixNet}_{\texttt{medium}}$
distortion		6.94	7.50	33.3	128.7

lected with FiT was able to achieve a $\hat{T}_{s,t}$ of 0.56 against the same target.

Performance of ensemble-model attacks. Table. 1 shows that ensemble-model transferability surpasses that of single-model attacks. While the average results over a random selection of 3 sources increase, they remain closer to the black box results than the white box ones (transferability lower than 0.5). Choosing the top-3 sources returned by our scores for leading the ensemble-model attack yields better performance, comparable to the best possible results obtainable with ensemble-model attacks. Notably, for the DI [32] and TAIG [9] method, it approaches the performance of white box results (transferability greater than 0.5).

Table 3: Transferability $\hat{T}_{s,t}$ for white box and transferable attacks (single-model). The FiT score uses $\texttt{TransQ}^{(1)}$.

Attack		Best		FiT	
W.B.	BP [35]	0.27	±0.18	−0.06	±0.31
	DeepFool [19]	0.03	±0.20	−0.34	±0.34
	PGD [16]	0.29	±0.20	−0.01	±0.36
	I-FGSM [10]	0.29	±0.20	−0.03	±0.34
Trans.	DI [32]	0.52	±0.12	0.40	±0.13
	TAIG [9]	0.46	±0.12	0.27	±0.17
	DWP [26]	0.34	±0.13	0.12	±0.20
All Attacks		0.65	±0.09	**0.48**	**±0.14**

5.3.3 White box vs. transferable attacks

Our last result is that the selection of a single source with the FiT score does not make traditional white box attacks transferable. Table 3 shows that BP [35], DeepFool [19], PGD [16], and I-FGSM [10] yield negative transferability values. More precisely, BP [35], I-FGSM [10], and PGD [16] perform better with FiT than the transferable attacks without any selection mechanism (ie. on average with a random source). However, section 4.4 highlights that the traditional white box attacks may be competitive for some inputs demanding low adversarial perturbation distortion. It is valuable to add them to the options of the attacker and let the FiT score decide the preferable option (source and attack method). This provides our best transferability score for a single model attack, close to 0.5.

6. Conclusion

Transferability is a crucial feature of adversarial examples as it allows a single perturbation to deceive multiple models. However, solely relying on Attack Success Rate (ASR) to measure transferability overlooks the degree of distortion needed to fool a model. This paper introduces a novel approach to assess transferability by comparing it to the distortion of two reference attacks: white box and black box attacks. We show that transferable attacks can perform worse than black box attacks without an appropriate selection of the source model, highlighting the need to choose the best source model to target a specific model.

The proposed solution, named FiT, allows the attacker to choose one of the best source models with minimal

queries to the target. Our experiments demonstrate that the proposed solution performs well in multiple attack scenarios.

This study has highlighted the differences in transferability between images for the same source model and their particularity for this specific network. Further research could focus on addressing this issue and investigating its underlying causes, with the hope of designing an even better selection mechanism able to spot the best source model.

References

[1] Mauro Barni, Kassem Kallas, Ehsan Nowroozi, and Benedetta Tondi. On the transferability of adversarial examples against CNN-based image forensics. In *ICASSP*. IEEE, 2019.

[2] Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Ipguard: Protecting intellectual property of deep neural networks via fingerprinting the classification boundary. In *ACM Asia Conference on Computer and Communications Security*, 2021.

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Symposium on Security and Privacy*. IEEE, 2017.

[4] Jinghui Chen and Quanquan Gu. Rays: A ray searching method for hard-label adversarial attack. In *ACM SIGKDD*, 2020.

[5] Marco De Angelis and Ander Gray. Why the 1-wasserstein distance is the area between the two marginal cdfs, 2021.

[6] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *CVPR*, 2019.

[7] Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *ICCV*, 2019.

[8] Tianjin Huang, Vlado Menkovski, Yulong Pei, Yuhao Wang, and Mykola Pechenizkiy. Direction-aggregated attack for transferable adversarial examples. *JETC*, 2022.

[9] Yi Huang and Adams Wai-Kin Kong. Transferable adversarial attack based on integrated gradients. In *ICLR*, 2022.

[10] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018.

[11] Huichen Li, Xiaojun Xu, Xiaolu Zhang, Shuang Yang, and Bo Li. Qeba: Query-efficient boundary-based blackbox attack. In *CVPR*, 2020.

[12] Yingwei Li, Song Bai, Yuyin Zhou, Cihang Xie, Zhishuai Zhang, and Alan Yuille. Learning transferable adversarial examples via ghost networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[13] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *ICLR*, 2020.

[14] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *ICLR*, 2017.

[15] Nils Lukas, Yuxuan Zhang, and Florian Kerschbaum. Deep neural network fingerprinting by conferrable adversarial examples. In *ICLR*, 2021.

[16] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2017.

[17] Thibault Maho, Teddy Furon, and Erwan Le Merrer. Surfree: a fast surrogate-free black-box attack. In *CVPR*, 2021.

[18] Thibault Maho, Teddy Furon, and Erwan Le Merrer. Fbi: Fingerprinting models with benign inputs. *arXiv preprint arXiv:2208.03169*, 2022.

[19] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, 2016.

[20] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Fahad Khan, and Fatih Porikli. On improving adversarial transferability of vision transformers. In *ICLR*, 2022.

[21] Zirui Peng, Shaofeng Li, Guoxing Chen, Cheng Zhang, Haojin Zhu, and Minhui Xue. Fingerprinting deep neural networks globally via universal adversarial perturbations, 2022.

[22] Deyan Petrov and Timothy M Hospedales. Measuring the transferability of adversarial examples. *arXiv preprint arXiv:1907.06291*, 2019.

[23] Maura Pintor, Fabio Roli, Wieland Brendel, and Battista Biggio. Fast minimum-norm adversarial attacks through adaptive norm constraints. In *NeurIPS*, 2021.

[24] Ali Rahmati, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Huaiyu Dai. Geoda: a geometric framework for black-box adversarial attacks. In *CVPR*, 2020.

[25] Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017.

[26] Hung-Jui Wang, Yu-Yu Wu, and Shang-Tse Chen. Enhancing targeted attack transferability via diversified weight pruning. *arXiv preprint arXiv:2208.08677*, 2022.

[27] Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *ICCV*, 2021.

[28] Xiaosen Wang, Jiadong Lin, Han Hu, Jingdong Wang, and Kun He. Boosting adversarial transferability through enhanced momentum. In *BMVC*, 2021.

[29] Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. Feature importance-aware transferable adversarial attacks. In *ICCV*, 2021.

[30] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

[31] Lei Wu and Zhanxing Zhu. Towards understanding and improving the transferability of adversarial examples in deep neural networks. In *ACML*, pages 837–850. PMLR, 2020.

[32] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *CVPR*, 2019.

[33] Yifeng Xiong, Jiadong Lin, Min Zhang, John E Hopcroft, and Kun He. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In *CVPR*, 2022.

[34] Liping Yuan, Xiaoqing Zheng, Yi Zhou, Cho-Jui Hsieh, and Kai-Wei Chang. On the transferability of adversarial attacks against neural text classifier. In *EMNLP*, 2021.

[35] Hanwei Zhang, Yannis Avrithis, Teddy Furon, and Laurent Amsaleg. Walking on the edge: Fast, low-distortion adversarial examples. *IEEE Trans. on IFS*, 2020.

[36] Jianping Zhang, Weibin Wu, Jen-tse Huang, Yizhan Huang, Wenxuan Wang, Yuxin Su, and Michael R Lyu. Improving adversarial transferability via neuron attribution-based attacks. In *CVPR*, 2022.

[37] Jingjing Zhao, Qingyue Hu, Gaoyang Liu, Xiaoqiang Ma, Fei Chen, and Mohammad Hassan. AFA: Adversarial Fingerprinting Authentication for deep neural networks. *Computer Communications*, 2020.

[38] Zhengyu Zhao, Hanwei Zhang, Renjue Li, Ronan Sicre, Laurent Amsaleg, and Michael Backes. Towards good practices in evaluating transfer adversarial attacks. *arXiv preprint arXiv:2211.09565*, 2022.

[39] Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable adversarial perturbations. In *ECCV*, 2018.

[40] Junhua Zou, Zhisong Pan, Junyang Qiu, Xin Liu, Ting Rui, and Wei Li. Improving the transferability of adversarial examples with resized-diverse-inputs, diversity-ensemble and region fitting. In *ECCV*, 2020.

A. Experimental Setup

In this study, we evaluated the transferability of adversarial attacks on a diverse set of 48 models trained for image classification on the ImageNet dataset with over one million annotated 224×224 images. The models were obtained from the Timm library [30], with the exception of $\texttt{ResNet50}_{\texttt{AdvTrain}}$, which was obtained from the GitHub repository of the original paper [2]. To ensure adequate representation, we randomly selected models from each architecture, with a minimum of three models per architecture. The only exception was the ReXNet architecture, which had two distinct models. The 48 selected models are:

- ConViT architecture: $\texttt{ConViT}_{\texttt{base}}$, $\texttt{ConViT}_{\texttt{small}}$, $\texttt{ConViT}_{\texttt{tiny}}$
- LeViT architecture: $\texttt{LeViT}_{192}$, $\texttt{LeViT}_{256}$, $\texttt{LeViT}_{128}$
- DenseNet architecture: $\texttt{DenseNet}_{169}$, $\texttt{DenseNet}_{121}$, $\texttt{DenseNet}_{161}$
- PiT architecture: $\texttt{PiT}_{\texttt{small}}$, $\texttt{PiT}_{\texttt{tight}}$, $\texttt{PiT}_{\texttt{tight-dist}}$, $\texttt{PiT}_{\texttt{small-dist}}$
- MobileNet architecture (V2): $\texttt{MobileNetV2}_{\texttt{110d}}$, $\texttt{MobileNetV2}_{100}$, $\texttt{MobileNetV2}_{\texttt{120d}}$
- CoaT architecture: $\texttt{CoatLite}_{\texttt{tiny}}$, $\texttt{CoatLite}_{\texttt{mini}}$, $\texttt{CoatLite}_{\texttt{small}}$
- xCiT architecture: $\texttt{xCiT}_{\texttt{medium}}$, $\texttt{xCiT}_{\texttt{nano}}$, $\texttt{xCiT}_{\texttt{small}}$
- Twins architecture: $\texttt{Twins}_{\texttt{small}}$, $\texttt{Twins}_{\texttt{large}}$, $\texttt{Twins}_{\texttt{base}}$,
- MixNet architecture: $\texttt{MixNet}_{\texttt{large}}$, $\texttt{MixNet}_{\texttt{small}}$, $\texttt{MixNet}_{\texttt{medium}}$, $\texttt{MixNet}_{\texttt{small-TensorFlow}}$, $\texttt{MixNet}_{\texttt{large-TensorFlow}}$, $\texttt{MixNet}_{\texttt{medium-TensorFlow}}$
- EfficientNet architecture: $\texttt{EfficientNetB0}$, $\texttt{EfficientNetB0}_{\texttt{AdvProp}}$, $\texttt{EfficientNetB0}_{\texttt{NS}}$
- ResNet architecture: $\texttt{ResNet}_{50}$, $\texttt{ResNet}_{\texttt{50d}}$, $\texttt{ResNet50}_{\texttt{AdvTrain}}$
- ResNetV2 architecture: $\texttt{ResNetV2}_{\texttt{50x1-dist}}$, $\texttt{ResNetV2}_{101}$, $\texttt{ResNetV2}_{50}$
- ReXNet architecture: $\texttt{RexNet}_{150}$, $\texttt{RexNet}_{130}$
- DPN architecture: $\texttt{DPN}_{92}$, $\texttt{DPN}_{107}$, $\texttt{DPN}_{\texttt{68b}}$
- DLA architecture: $\texttt{DLA}_{60}$, $\texttt{DLA}_{102}$, $\texttt{DLA}_{169}$

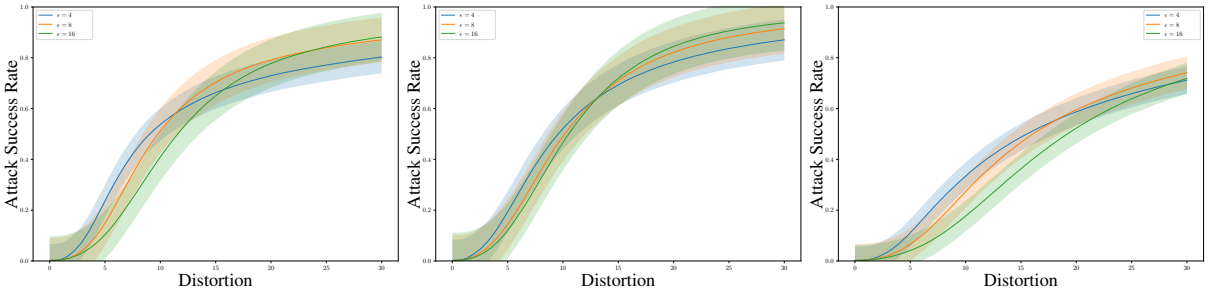
B. Epsilon Parameter

All transferable attacks share a common parameter ϵ. It controls the maximum perturbation norm added on a single pixel for the adversarial example built. Fig. 7 demonstrates the ASR obtained for various values of ϵ as a function of the perturbation norm. It shows that even if more freedom is given to the perturbation, in the sense that a larger maximum perturbation norm is allowed, the transferable directions remain consistent. Irrespective of the value of ϵ for a given attack, all scores for a given norm of the perturbation are similar.

C. Transferability Dependences

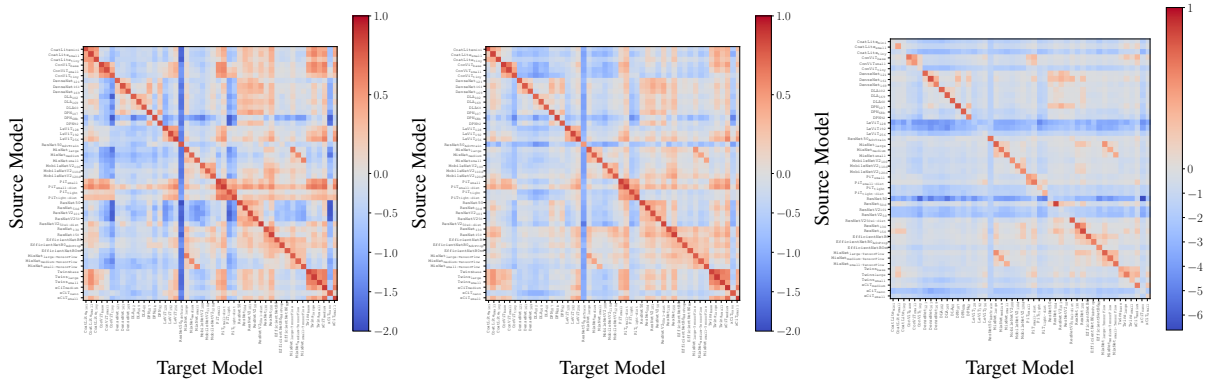
The 48 models considered in A are evaluated as both sources and targets in this study. For each possible pair of models, each source model is evaluated for its ability to transfer to each target model. This results in a total of $48^2 = 2304$ evaluations. The transferability is evaluated using the score defined in 3.2 and their matrices for the attacks DI [32], TAIG [9], and DWP [26] are presented in 8. Each matrix exhibits a similar structure, with models that have high transferability values appearing in each matrix. However, the values achieved are different for each attack. The DI [32] and TAIG [9] attacks achieve higher values than DWP [26], indicating that these attacks create better quality transferable examples.

[2] https://github.com/MadryLab/robustness



(a) DI [32] (b) TAIG [9] (c) DWP [26]

Figure 7: Attack Success Rate function of the perturbation norm for different values of ϵ.



(a) DI [32] (b) TAIG [9] (c) DWP [26]

Figure 8: Transferability score $\hat{T}_{s,t}$ matrix of 48 sources and 48 targets listed in A for DI [32], TAIG [9] and DWP [26].

D. Results

D.1. Fingerprinting

Transferability can be divided into three components: the attack, the model, and the attacked image. To estimate transferability, the FiT measure defined in 3.2 first estimate the similarity between the source and the target models. In a defensive scenario, fingerprinting methods have been proposed to estimate model similarity without accessing one of the models. These methods do not modify the model during training but instead take an already trained model and find images that are its signatures. They usually generate adversarial examples specially designed for this model [2, 15, 21]. FBI [18] is the only method using benign images to assess the similarity of two models by measuring the independence between the two models using mutual information. All fingerprinting methods are sensitive to the number of images used for fingerprinting. More images lead to more accurate similarity scores, but they also have a cost. In the scenario considered here, the number of images submitted must be minimized. Figure 9 shows the $\hat{T}_{s,t}$ function of the number of images used for FBI [18]. Increasing the number of images submitted provides a better estimation of the transferability. The score reaches a plateau at 200 images submitted.

D.2. Ensemble model attack

When attackers have access to multiple models, they can perform an ensemble-model attack to generate transferable adversarial examples. This approach has been shown to offer better transferability than the best single-model attack. However, existing methods for performing ensemble-model attacks have only been evaluated with a limited number of source models, typically with a maximum of three models. In this paper, a high number of models is used to build large ensemble-model attacks in the scenario described in the experimental setup in 4.1. At each step of the attack, a model is randomly selected from the available sources and added to the ensemble-model. To build transferable adversarial examples, the logits of the models are averaged together, as proposed in [28]. Transferability is computed for ensemble-model attacks of up to 20

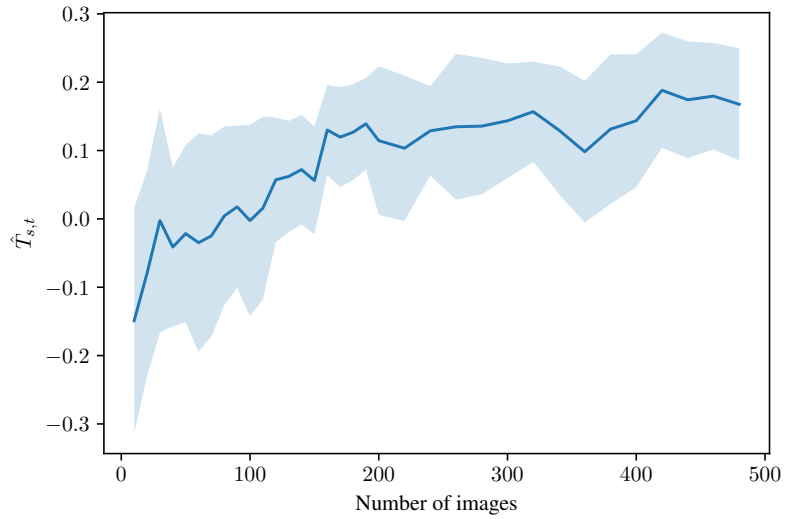


Figure 9: $\hat{T}_{s,t}$ function of the number of images used for FBI [18] to estimate the transferability between 45 sources and one target. Adversarial obtained with DI [32] and $\epsilon = 8$.

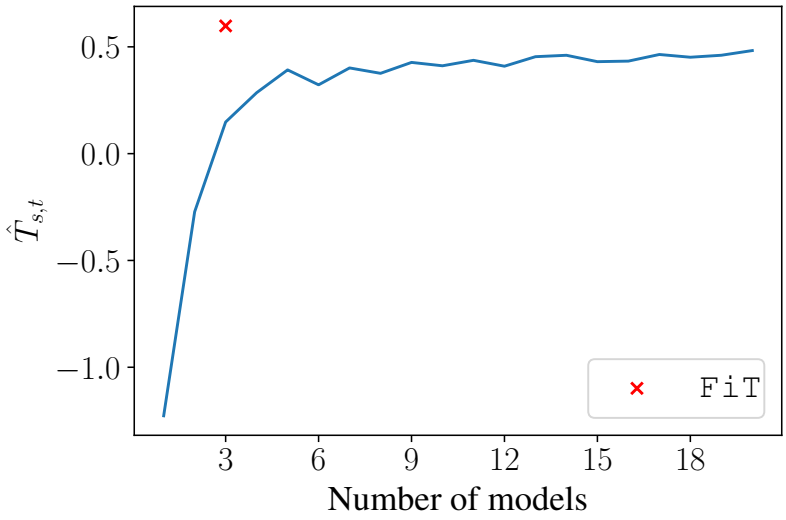


Figure 10: $\hat{T}_{s,t}$ function of the number of models used for ensemble-model to attack $\text{xCiT}_{\text{nano}}$. The models are randomly selected and added one by one and compared with FiT selecting the three best models for ensemble-model among the 20 models available.

models. Figure 10 shows the FiT score as a function of the ensemble-model size and compares the results with FiT scores obtained by selecting the three best models for the ensemble-model among the 20 models available. Ensemble-model attacks demonstrate significant improvements when only a few models are considered, but beyond 5 models, the improvements become negligible. Additionally, the FiT score for the ensemble-model attack with 20 models was lower than that of the ensemble-model attack with only three models, which were carefully selected using FiT. These findings suggest that the quality of the selected models is more crucial than the quantity of models for effective ensemble-model attacks.