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: $\texttt{DLA}_{60}$, $\texttt{DLA}_{102}$, $\texttt{DLA}_{169}$

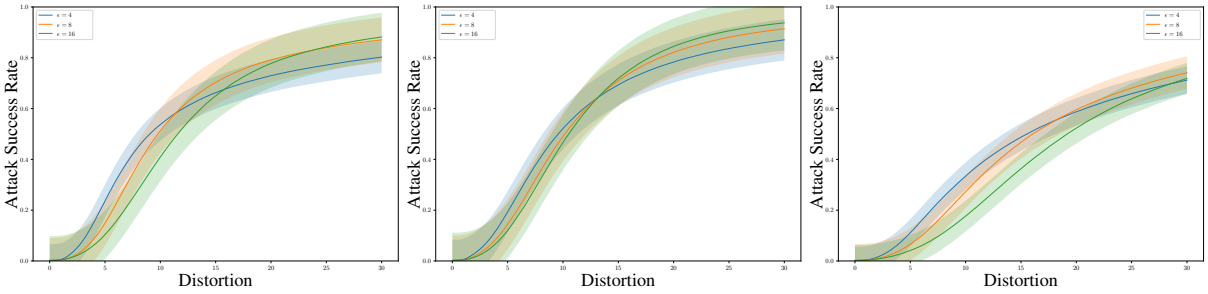
B. Epsilon Parameter

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

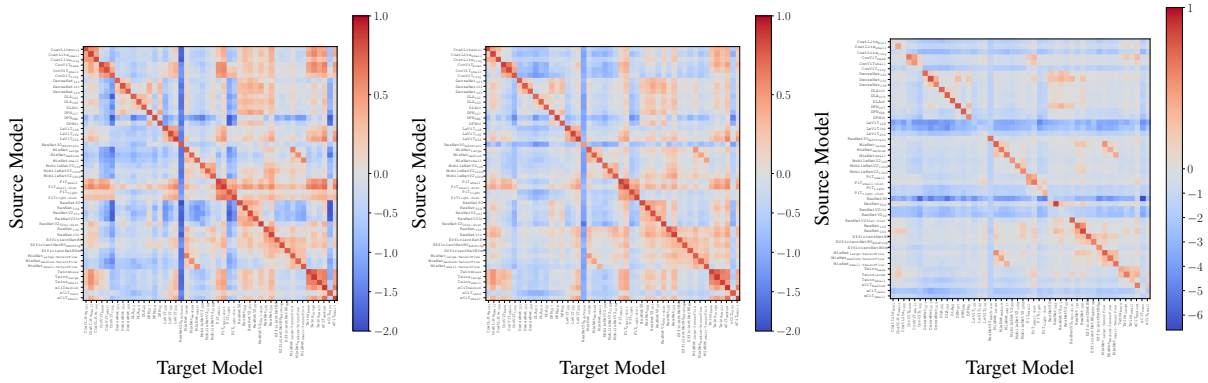
The 48 models considered in A are evaluated as both sources and targets in this study. For each possible pair of models, each source model is evaluated for its ability to transfer to each target model. This results in a total of $48^2 = 2304$ evaluations. The transferability is evaluated using the score 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

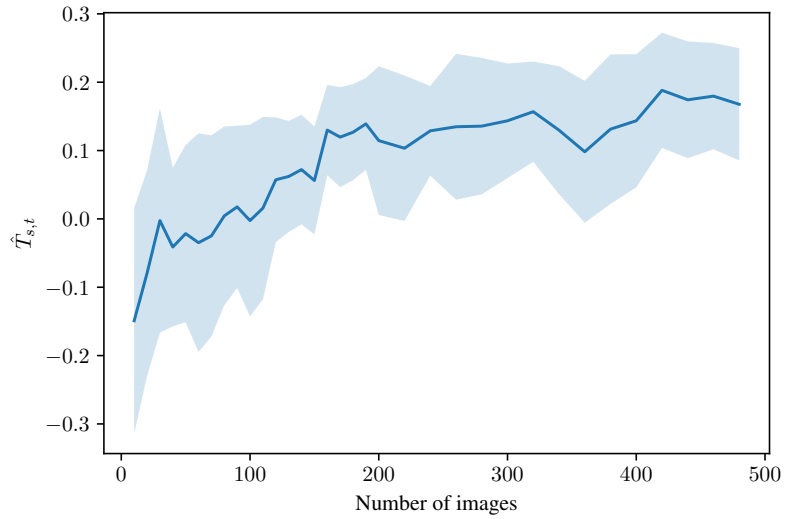


Figure 9: $\hat{T}_{s,t}$ function of the number of images used for FBI [18] to estimate the transferability between 45 sources and one target. Adversarial obtained with DI [32] and $\epsilon = 8$.

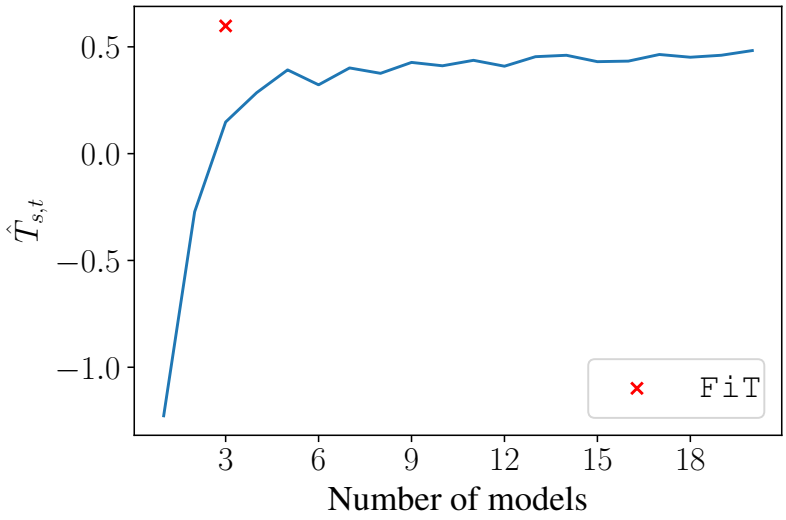


Figure 10: $\hat{T}_{s,t}$ function of the number of models used for ensemble-model to attack $\text{xCiT}_{\text{nano}}$. The models are randomly selected and added one by one and compared with FiT selecting the three best models for ensemble-model among the 20 models available.

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.