Selection of Source Images Heavily Influences the Effectiveness of Adversarial Attacks

Utku Ozbulak[1]
utku.ozbulak@ugent.be
Esla Timothy Anzaku[1]
eslatimothy.anzaku@ghent.ac.kr
Wesley De Neve[1]
wesley.deneve@ugent.be
Arnout Van Messem[2]
arnout.vanmessem@uliege.be

[1] Ghent University
Ghent, Belgium
Ghent University Global Campus
Incheon, South Korea

[2] University of Liège
Liège, Belgium

Abstract

Although the adoption rate of deep neural networks (DNNs) has tremendously increased in recent years, a solution for their vulnerability against adversarial examples has not yet been found. As a result, substantial research efforts are dedicated to fix this weakness, with many studies typically using a subset of source images to generate adversarial examples, treating every image in this subset as equal. We demonstrate that, in fact, not every source image is equally suited for this kind of assessment. To do so, we devise a large-scale model-to-model transferability scenario for which we meticulously analyze the properties of adversarial examples, generated from every suitable source image in ImageNet by making use of three of the most frequently deployed attacks. In this transferability scenario, which involves seven distinct DNN models, including the recently proposed vision transformers, we reveal that it is possible to have a difference of up to 12.5% in model-to-model transferability success, 1.01 in average L_2 perturbation, and 0.03 $(8/225)$ in average L_∞ perturbation when $1,000$ source images are sampled randomly among all suitable candidates. We then take one of the first steps in evaluating the robustness of images used to create adversarial examples, proposing a number of simple but effective methods to identify unsuitable source images, thus making it possible to mitigate extreme cases in experimentation and support high-quality benchmarking. In support of future research efforts, we make our code and the statistics for all evaluated source images as well as the list of identified fragile source images publicly available in https://github.com/utkuozbulak/imagenet-adversarial-image-evaluation.

1 Introduction

Thanks to recent advances in the field of deep neural networks, a wide range of problems that were once thought to be hard challenges found easy-to-adopt solutions [32, 55]. Indeed,

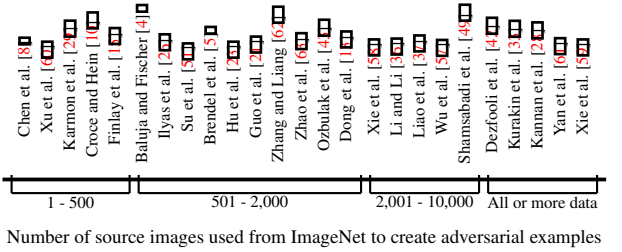


Figure 1: A number of studies that work with images taken from the ImageNet validation set, grouped based on the number of source images used for creating adversarial examples.

many deep learning libraries now come with built-in solutions and pre-trained models, further increasing the adoption rate of such networks in the area of computer vision [1, 27, 45]. In spite of receiving a large amount of research attention, a number of fundamental flaws of DNNs still remain unsolved. One of those flaws is their vulnerability to adversarial attacks, where small changes in inputs may lead to large changes in predictions [52].

Although adversarial attacks have been recognized to be a threat for all domains that make use of DNNs, the domain of vision in particular is said to be the one that suffers from adversarial attacks the most, since the perturbation is often invisible to the bare eye. Moreover, continuous deployment of DNNs for mission-critical tasks such as self-driving cars and medical diagnosis tools further amplify this threat since the adversarial examples are not easily detectable [9, 16, 39, 56].

In recent years, numerous adversarial defenses were proposed in order to prevent adversarial attacks or detect adversarial examples [19, 30, 36, 46]. Proposed defenses often claim a certain level of robustness against adversarial examples that have an amount of perturbation less than a selected norm [11]. Since the topic of adversariality is closely linked with security, reproducibility of newly proposed techniques is of utmost importance. As a result, there have been a number of impactful studies that analyze the correctness and reliability of newly proposed adversarial defenses [2, 3, 7, 53]. In this context, Carlini and Wagner [7], for instance, demonstrated that most of the defenses proposed for MNIST [34] do not even generalize to CIFAR [31]. This observation prompted research on the suitability of datasets for adversarial research [55], with Carlini and Wagner further suggesting that the usage of larger datasets such as ImageNet [48] may be necessary, given the lack of generalization of defenses proposed for smaller datasets [7].

Even though the results obtained with ImageNet are more convincing, working with ImageNet is much more challenging than, for example, working with MNIST or CIFAR. Indeed, not only does ImageNet contain more images than the other two, the images themselves are also larger. In addition, DNNs that achieve state-of-the-art results for ImageNet are also much bigger than their counterparts that achieve state-of-the-art results for MNIST or CIFAR, thus posing a challenge in terms of computational power needed. As a result, most of the studies that work with ImageNet only use a subset of images in order to create adversarial examples, unless that research is performed by a large industry lab that can afford the computational power (see Figure 1).

Although the studies of [51, 53] hinted that not all source images may be equally suitable for adversarial example creation, most of the studies that work with adversarial examples often randomly sample source images among the ones that are correctly classified. As such, any image that is correctly classified by the models of interest is thought to be suitable and equal in terms of model-to-model transferability and the required perturbation to achieve

adversariality. To the best of our knowledge, an in-depth analysis of source image suitability of adversarial examples in large-scale model-to-model scenarios has not been conducted yet. Hence, approaching the problem of adversarial examples from a different angle and following the directions of [17, 51, 53], instead of analyzing the effectiveness of attacks, the durability of defenses, or the robustness of models, our study focuses on the source images used to create adversarial examples, hereby investigating the impact of image selection on (1) the success of model-to-model adversarial transferability and (2) the required perturbation to achieve this transferability.

With the help of large-scale experiments, we demonstrate that, even when the most-studied adversarial attacks for benchmarking are used, model-to-model transferability successes of adversarial examples, as well as the amount of required perturbation to achieve this transferability, heavily depend on the source images used to create those adversarial examples. Moreover, we present a case study that shows how the experimental results obtained may lead to misleading conclusions when making use of certain subsets of source images.

2 Adversarial attacks

Assuming an M-class setting in which a data point and its categorical association are defined as $\boldsymbol{x} \in \mathbb{R}^k$ and $\boldsymbol{y} \in \mathbb{R}^M$, respectively, with $y_c = 1$ and $y_m = 0, \forall m \in \{0,\dots,M\}\backslash\{c\}$, let g be a classification function that maps inputs onto categorical predictions. In this setting, we define the output $g(\boldsymbol{\theta},\boldsymbol{x}) \in \mathbb{R}^M$ as the logits obtained by a prediction model/classifier using the parameters $\boldsymbol{\theta}$. The given data point is then classified into the category with the largest logit value: $G(\boldsymbol{\theta},\boldsymbol{x}) = \arg\max_t(g(\boldsymbol{\theta},\boldsymbol{x})_t)$. If $G(\boldsymbol{\theta},\boldsymbol{x}) = \arg\max_t(\boldsymbol{y}_t)$, then this classification is correct.

For the given setting, a perturbation Δ_x bounded by the L_p ball centered at $\boldsymbol{x}$ with a radius ε, formulated as $\mathcal{B}(\boldsymbol{x})^p_\varepsilon := \{\hat{\boldsymbol{x}} : ||\Delta_x||_p := ||\boldsymbol{x}-\hat{\boldsymbol{x}}||_p \leq \varepsilon\}$, is said to be an *adversarial perturbation* if $G(\boldsymbol{\theta},\boldsymbol{x}) \neq G(\boldsymbol{\theta},\hat{\boldsymbol{x}})$. In this case, $\hat{\boldsymbol{x}}$ is also said to be an adversarial example.

Since the discovery of adversarial examples, a plethora of attacks using a wide range of perturbation generation methods has been proposed [41, 47, 60]. Early research efforts in the field mostly made use of L-BFGS optimization [52], Fast Gradient Sign Method (FGSM) [18], and Iterative Fast Gradient Sign Method (IFGSM) [33]. However, Projected Gradient Descent (PGD) [38] , the Carlini & Wagner's Attack (CW) [6], and Momentum Iterative Fast Gradient Sign Method (MI-FGSM) [12] have taken the place of the aforementioned attacks in recent research efforts, thanks to the superior results obtained by the latter three. Following these findings, the study presented in this paper also uses these three attacks for examining the fragility of source images.

PGD can be seen as a generalization of FGSM and IFGSM. In particular, this attack aims at finding an adversarial example $\hat{\boldsymbol{x}}$ that satisfies $||\hat{\boldsymbol{x}}-\boldsymbol{x}||_\infty < \varepsilon$, where the perturbation is defined within an L_∞ ball centered at $\boldsymbol{x}$ with a radius ε. The adversarial example is iteratively generated as follows: $[\hat{\boldsymbol{x}}]_{n+1} = \Pi_\varepsilon\Big([\hat{\boldsymbol{x}}]_n - \alpha\ \text{sign}\big(\nabla_x J(g(\boldsymbol{\theta},[\hat{\boldsymbol{x}}]_n)_c)\big)\Big)$, with $[\hat{\boldsymbol{x}}]_1 = \boldsymbol{x}$, where the perturbation is calculated using the signature of the gradient of the cross-entropy loss, $\text{sign}(\nabla_x J(g(\boldsymbol{\theta},[\hat{\boldsymbol{x}}])_c))$, originating from the target class c. In this setting, α controls the exercised perturbation at each iteration and Π_ε is a function that controls the L_∞ limit imposed on the perturbation.

CW, on the other hand, is a complex attack that aims to find a perturbation within a small L_2 norm as follows: $\min_{\hat{\boldsymbol{x}}} f(\hat{\boldsymbol{x}},c) + ||\hat{\boldsymbol{x}}-\boldsymbol{x}||_2$, where f is a preferred loss function. Given [6, 51], we use the following loss: $f(\hat{\boldsymbol{x}},c) = \max_k\{\max_{c\neq k}\{g(\boldsymbol{\theta},\hat{\boldsymbol{x}})_c - g(\boldsymbol{\theta},\hat{\boldsymbol{x}})_k\} - \kappa\}$, selecting the target class with c and adjusting the confidence of the attack with κ.

The overall structure of MI-FGSM is similar to that of PGD and IFGSM. However, instead of adding perturbation directly to the image, it integrates the gradient of the cross-entropy loss into a variable that acts as a momentum term: $[\boldsymbol{\tau}]_{n+1} = \mu[\boldsymbol{\tau}]_n + \frac{J(g(\theta,[\hat{\boldsymbol{x}}]_n)_c)}{||J(g(\theta,[\hat{\boldsymbol{x}}]_n)_c)||_1}$, where μ is the multiplier for already-accumulated gradient in past iterations. Unlike the previous two attacks, the perturbation is generated from this momentum term $\boldsymbol{\tau}$ instead of the gradient itself, and iteratively added to the image as follows: $[\hat{\boldsymbol{x}}]_{n+1} = \Pi_\varepsilon\Big([\hat{\boldsymbol{x}}]_n - \alpha \operatorname{sign}(\boldsymbol{\tau})\Big)$, with $[\hat{\boldsymbol{x}}]_1 = \boldsymbol{x}$.

Given that adversarial examples are trivial to generate in white-box cases [2, 7], and given the recent focus on the importance of adversarial evaluation in black-box scenarios [26, 54], our study mainly focuses on analyzing the properties of adversarial examples that achieve model-to-model transferability. In this context, an adversarial example created by a model is said to achieve model-to-model adversarial transferability if it is also incorrectly classified by another model, provided that the source image used to create the adversarial example is initially correctly classified by both models.

3 Experimental setup

Models – In this study, we use five different deep learning architectures that see frequent use in the literature. The considered models are: AlexNet [32], SqueezeNet [25], VGG-16 [50], ResNet-50 [22], and DenseNet-121 [24]. In addition to these models, we also include two recently proposed vision transformer models that achieve state-of-the-art results on ImageNet [14]: Vision Transformer Base/16 – 224 (ViT-B) and Vision Transformer Large/16 – 224 (ViT-L). From here on, each model will be denoted by its set of parameters $\theta_i, i \in \{1,\dots,7\}$, and multiple models will be denoted by $\Theta = \{\theta_1,\dots,\theta_7\}$.

Data – We follow the approach used by previous studies on adversariality, leveraging the images in the ImageNet validation set for generating adversarial examples. In this paper, these unperturbed images are referred to as *source images*. Further adopting previously used methods, we only rely on images that are correctly classified by all selected models in order to conduct trustworthy experiments on adversarial transferability, thus ensuring $G(\theta_i,\boldsymbol{x}) = \arg\max_t(\boldsymbol{y}_t), \forall i \in \{1,\dots,7\}$. By doing so, we filter out images that are hard to correctly classify for at least one of our models, thus limiting the hypothesis space and allowing us to perform a best-case analysis. After this filtering operation, we are left with a set of 19,025 source images, which approximately corresponds to 38% of the ImageNet validation set. We will refer to this set of 19,025 source images as:

$$\mathbb{S} = \{\boldsymbol{x} \mid G(\theta_i,\boldsymbol{x}) = \arg\max_t(\boldsymbol{y}_t); i \in \{1,\dots,7\}\}. \tag{1}$$

Adversarial perturbation – Although the methods used to identify perturbation in images are not a perfect match for how humans perceive noise, L_p norms (with $p \in \{0,2,\infty\}$) are commonly used since the early days of research on adversarial examples [6, 17, 44]. We adopt both L_2 and L_∞ norms for measuring the added perturbation. In terms of the used L_∞ perturbation budget, another large-scale study on adversarial transferability [51] uses $\varepsilon_{[0,1]} \in \{0.1, 0.2, 0.3\}$, which approximately corresponds to $\varepsilon_{[0,255]} \in \{25, 45, 67\}$ in discretized settings. We observed that using $\varepsilon_{[0,255]} \geq 45$ leads to adversarial examples that come with large perturbation budgets. In light of this observation, we limit the perturbation on an L_∞ ball to 38 (i.e., $\varepsilon_{[0,255]} = 38$, $\varepsilon_{[0,1]} = 0.15$), thus ensuring that the perturbation is not excessive. Further details on the calculation of L_p norms and the employed attacks, as

well as a comparison of perturbation visibility, can be found in the supplementary material (Section A).

For PGD and MI-FGSM, we perform the attack with 50 iterations and allow a perturbation budget of $\varepsilon_{[0,1]} = 0.15$. For MI-FGSM, we follow the work of [12] and use $\mu = 1$. For CW, we use $\kappa = 20$ (as suggested by the authors of the attack). Using a randomly selected class that differs from the true class of the source image, we perform the aforementioned attacks on source images. In order to avoid cases where the image/target class combination is challenging, if an attack does not succeed within the allocated number of iterations, we select another class, and we perform the attack on the same image up to five times. At each iteration of the adversarial attack, we analyze whether or not the prediction for the image changed for the other six models (i.e., evaluating the non-targeted transferability) and then save the adversarial examples with the smallest perturbation. By doing so, we aim at finding the least-required perturbation, as exercised by all three attacks, that is sufficient to convert a source image into an adversarial one.

Non-adversarial perturbation – In addition to the adversarial attacks, we also make use of commonly used image distortion techniques in order to measure the robustness of source images. For this analysis, we employ (1) uniform noise, (2) Gaussian noise, and (3) change in contrast to create "adversarial examples", where all of these additive types of noise respect the L_∞ limit put in place for the adversarial attacks. Details on the usage of these operations can be found in the supplementary material (Section B).

4 Methodology for the source image analysis

In this section, we explain the notation and methodology used for the analysis of source images in adversarial scenarios. We denote by $\hat{\boldsymbol{x}}^{(\mathrm{A}):i\to j}$ an adversarial example that is created through the addition of adversarial perturbation with the attack $(\mathrm{A}) \in \{\mathrm{PGD}, \mathrm{CW}, \mathrm{MI\text{-}FGSM}\}$, calculated from the model θ_i, but that is misclassified by model θ_j, thus achieving adversarial transferability. We then denote the set of all adversarial examples that achieve adversarial transferability, created through the usage of source image $\boldsymbol{x}$, as follows:

$$\widehat{\mathcal{X}}^{(\mathrm{A})} := \{\hat{\boldsymbol{x}}^{(\mathrm{A}):i\to j} \,|\, i,j = 1,\ldots,7\,; i \neq j\}\,. \tag{2}$$

We measure the added perturbation with $L_{\{2,\infty\}}$ norms. Moreover, we denote the least amount of perturbation required to convert a source image into an adversarial example for a particular target model (j), regardless of which other model it is generated from, by:

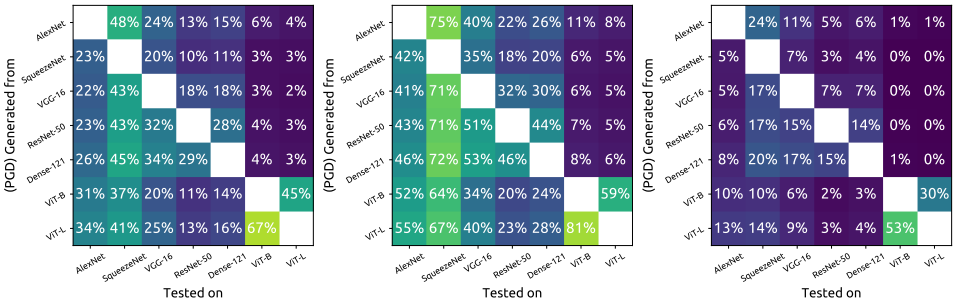
$$d_p(\theta_j, \widehat{\mathcal{X}}^{(\mathrm{A})}) = \min_{i\in\{1,\ldots,7\}\setminus\{j\}} ||\boldsymbol{x} - \hat{\boldsymbol{x}}^{(\mathrm{A}):i\to j}||_p\,, \tag{3}$$

where p denotes the selected norm. We also measure the minimum amount of perturbation required to convert a source image into an adversarial example for any model as follows:

$$D_p(\Theta, \widehat{\mathcal{X}}^{(\mathrm{A})}) = \min_{j\in\{1,\ldots,7\}} d_p(\theta_j, \widehat{\mathcal{X}}^{(\mathrm{A})})\,. \tag{4}$$

Another important benchmark is the transferability count of adversarial examples created from individual source images. Since we have seven models, and since we are only interested in model-to-model transferability, we count the successful model-to-model transfers for adversarial examples generated from a source image $\boldsymbol{x}$ and the attack A as follows:

$$T(\Theta, \widehat{\mathcal{X}}^{(\mathrm{A})}, \boldsymbol{y}) = \sum_{i,j=1, i\neq j}^{7} \mathbb{1}_{\{G(\theta_j, \hat{\boldsymbol{x}}^{(A):i\to j}) \neq \arg\max_t(\boldsymbol{y}_t)\}}\,. \tag{5}$$



(a) All images ($\mathbb{S}$) (b) Fragile images ($\mathbb{S}_f$) (c) Hard images ($\mathbb{S}_h$)

Figure 2: Proportion of source images in (left) $\mathbb{S}$, (center) $\mathbb{S}_f$, and (right) $\mathbb{S}_h$ that achieved (untargeted) adversarial transferability with the usage of PGD.

For each source image and attack, this (untargeted) transferability count $T(\Theta, \widehat{\mathcal{X}}^{(\mathrm{A})}, \boldsymbol{y})$ can take a value between 0 and 42. In this context, having zero model-to-model transferability means that none of the adversarial examples generated from a particular source image achieved adversarial transferability and 42 means that the adversarial examples created from a source image achieved adversarial transferability in all model-to-model scenarios.

5 Experimental results

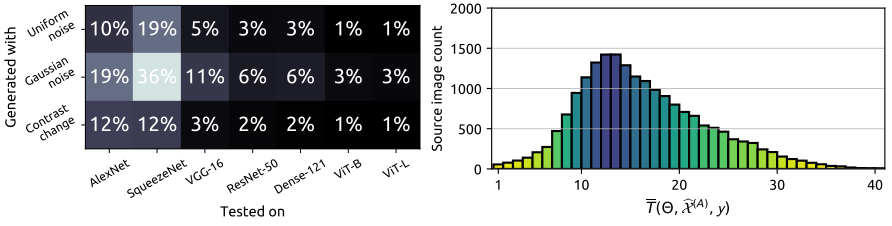
Through the methodology described above, we successfully created 825,005 adversarial examples that achieve adversarial transferability for at least one model-to-model scenario (excluding white-box cases). Specifically, 173,542, 115,688, and 535,775 adversarial examples were produced with PGD, CW, and MI-FGSM, respectively. In the remainder of this paper, we provide and discuss experimental results for these 825,005 adversarial examples, as well as for the 19,025 source images used to obtain them.

Since MI-FGSM is able to create a large number of adversarial examples that achieve model-to-model transferability compared to the other two attacks, experimental results obtained through the usage of all adversarial examples may be skewed towards adversarial examples created with MI-FGSM. For the sake of precise experimentation, when we inspect all adversarial examples for an experiment, we provide the same experiment in the supplementary material using adversarial examples created with individual attacks.

5.1 Model-to-model transferability

In Figure 2a, we show the model-to-model transferability success ratio of adversarial examples generated with PGD. Specifically, we provide details for the source and target models of all 173,542 adversarial examples that achieved (untargeted) adversarial transferability.

In order to answer the question of whether or not adversarial transferability success can be influenced by source image selection, let us continue with an unusual experiment. In Figure 3a, we show the number of source images that had their predictions changed for the models listed on the x-axis through the application of the non-adversarial perturbations listed on the y-axis. Surprisingly, relying on common noise generation methods that do not require any special setup, we observe that a large portion of source images have their classification changed in a limited L_∞ ball setting. Specifically, 9,615 unique source images, corresponding to approximately 50% of the source images ($\mathbb{S}$), become "adversarial examples" for at least one model through the introduction of non-adversarial noise.



(a) Non-adversarial noise and transferability (b) Source images that attained transferability

Figure 3: (left) Number (proportion) of source images that became "adversarial examples" through the addition of non-adversarial noise and (right) histogram of transferability count of source images of source images and their transferability count according to $\overline{T}(\Theta, \widehat{\mathcal{X}}^{(A)}, \boldsymbol{y})$.

Combining the two experiments (Figure 2a and Figure 3a) that have been discussed thus far, let us divide $\mathbb{S}$ into two sets $\mathbb{S}_f$ and $\mathbb{S}_h$, where the former contains *fragile source images* that had, at least once and for any model, their prediction changed with the application of non-adversarial noise (9,615 source images) and where the latter contains the remaining images (9,410 source images), with $\mathbb{S} = \mathbb{S}_f \cup \mathbb{S}_h$. According to this separation, we provide Figure 2b and Figure 2c, where we show the model-to-model transferability of the adversarial examples originating from the source images in $\mathbb{S}_f$ and $\mathbb{S}_h$, respectively. As can be seen, even though we use a similar number of source images taken from the same dataset for both $\mathbb{S}_f$ and $\mathbb{S}_h$, we obtain outcomes that are completely different in terms of adversarial transferability success. We present detailed versions of all transferability matrices, as well as the results obtained for CW and MI-FGSM, in the supplementary material (Section C).

The reason for the large discrepancy between the results presented in Figure 2b and Figure 2c is the fragility of a subset of the source images. Compared to the other, non-fragile images, these fragile source images have their predictions easily changed for a large number of models, even when other conditions are held the same (e.g., attacks and models). In order to lay bare the fragility of these source images, we perform an aggregate analysis of their average transferability per attack, leading to a histogram of $\overline{T}(\Theta, \widehat{\mathcal{X}}^{(A)}, \boldsymbol{y})$ for all source images in $\mathbb{S}$, as shown in Figure 3b. This histogram illustrates that, with one of the employed attacks, a large portion of the adversarial examples achieve adversarial transferability between 10 to 20 times. However, an intriguing observation can be made for the leftmost and the rightmost side of this figure, where 585 source images achieve adversarial transferability less than 5 times and where 1,743 sources images achieve transferability more than 25 times. These images that, through the added perturbation, do not easily become adversarial examples, as well as the fragile source images, which easily change predictions between models and which achieve unnaturally high model-to-model transferability, will be our main focus for the remainder of this paper.

5.2 Adversarial perturbation

Another important aspect of model-to-model adversarial transferability is how easy a source image becomes an adversarial example, since the robustness of adversarial defenses, as well as recently proposed models, are certified under an L_p norm perturbation. In Figure 4, we perform a correlation analysis between $\overline{T}(\Theta, \widehat{\mathcal{X}}^{(A)}, \boldsymbol{y})$ and the minimum required L_p perturbation to achieve adversarial transferability $D_{\{2,\infty\}}(\Theta, \widehat{\mathcal{X}}^{(A)})$. In this context, we observe a

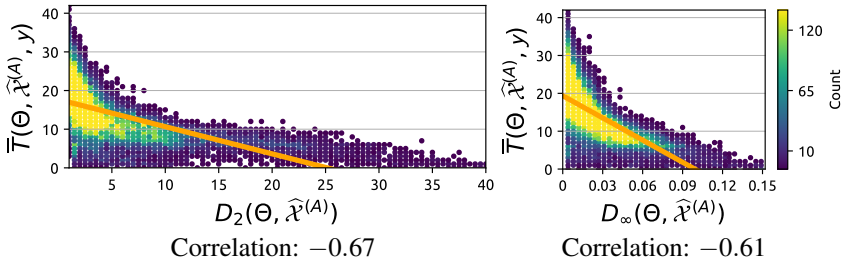


Figure 4: Scatter plot of $D_p(\Theta, \widehat{\mathcal{X}}^{(A)})$, the minimum amount of perturbation required for each source image, against average adversarial transferability count $\overline{T}(\Theta, \widehat{\mathcal{X}}^{(A)}, \boldsymbol{y})$, for $p = 2$ (left) and $p = \infty$ (right). The regression line is shown in orange.

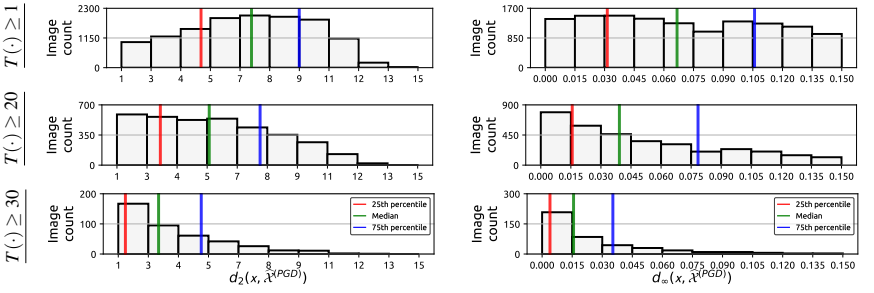


Figure 5: Source images that achieved adversarial transferability to ViT-B are selected based on transferability count, with $T(\Theta, \widehat{\mathcal{X}}^{(\mathrm{PGD})}, \boldsymbol{y}) \geq \{1, 20, 30\}$. The minimum amount of perturbation required for creating adversarial examples from these source images is histogrammed, measuring the perturbation using $d_p(\boldsymbol{x}, \widehat{\mathcal{X}}^{(\mathrm{PGD})})$, with $p \in \{2, \infty\}$. The median perturbation, as well as the 25th and the 75th percentile, are provided in order to improve interpretability.

mild negative correlation between added noise and transferability count, where the adversarial examples originating from source images that achieve higher transferability counts are also the ones that require less perturbation. These results hint that the fragile images we have identified do not only achieve high adversarial transferability, but that they also do so with smaller perturbation budgets.

In order to solidify these observations regarding perturbation and transferability, we part from an aggregate analysis to a more granular one and investigate the perturbations of adversarial examples that achieve transferability for each model individually. In Figure 5, we provide for ViT-B the smallest required $L_{\{2,\infty\}}$ perturbation for source images progressively filtered with $T(\Theta, \widehat{\mathcal{X}}, \boldsymbol{y}) \geq \{1, 20, 30\}$. Note that, as $T(\Theta, \widehat{\mathcal{X}}, \boldsymbol{y})$ increases, the distribution of the perturbation shifts towards zero, thus confirming our previous observations. These results indicate that source images that achieve high transferability counts are, most likely, also the ones that require less perturbation. Similar results, as available in the supplementary material (Section D and Section E), can be observed for the other models.

6 Source image suitability

Our experiments indicate that, while a certain portion of images never becomes adversarial, another portion of images can be easily turned into adversarial examples using relatively small perturbation budgets. Given the importance of research reproducibility, this leads to

Table 1: Correlation coefficients between various estimates of errors in source image predictions and properties of adversarial examples (transferability and perturbation) created from those source images are given for PGD, CW, and MI-FGSM.

Error measurement	PGD			CW			MI-FGSM		
	$T(\cdot)$	$d_2(\cdot)$	$d_\infty(\cdot)$	$T(\cdot)$	$d_2(\cdot)$	$d_\infty(\cdot)$	$T(\cdot)$	$d_2(\cdot)$	$d_\infty(\cdot)$
$\mathrm{Q}(P(\theta,\boldsymbol{x}))$	0.58	$\mathbf{-0.64}$	$\mathbf{-0.58}$	**0.57**	$\mathbf{-0.59}$	$\mathbf{-0.66}$	0.42	−0.54	−0.54
$1-\max(P(\theta,\boldsymbol{x}))$	**0.61**	−0.60	−0.57	**0.57**	−0.54	−0.63	**0.43**	$\mathbf{-0.58}$	$\mathbf{-0.57}$
$\mathrm{MSE}(P(\theta,\boldsymbol{x}),\boldsymbol{y})$	0.56	−0.57	−0.53	0.56	−0.51	−0.61	0.37	−0.51	−0.53
$\mathrm{WD}(P(\theta,\boldsymbol{x}),\boldsymbol{y})$	0.33	−0.35	−0.37	0.33	−0.32	−0.37	0.29	−0.38	−0.38

the question of how much variance can be observed when randomly sampling source images. In order to answer this question, we randomly sample $1,000$ source images from $\mathbb{S}$ (since this number seems to be the most commonly selected number in Figure 1), subsequently inspecting the transferability success and $L_{\{2,\infty\}}$ perturbation norms of the adversarial examples generated for the individual model-to-model transferability cases. We perform the aforementioned routine $10,000$ times. As a result, we calculate the lowest, the highest, and the average transferability, as well as the $L_{\{2,\infty\}}$ perturbations. Overall, we observe that, while the average case closely matches the usage of all available source images, it is possible to have differences between the lowest and the highest case of up to 12.5% in transferability, 1.01 in L_2 norm perturbation, and 0.03 (i.e., $8/255$) in L_∞ norm perturbation. These results indicate that, even when random sampling is used, it is indeed possible to have conflicting results, depending on the source images selected.

In Section 5.1, we demonstrated that one way to identify source images that are fragile to adversarial attacks is to perform a large-scale analysis of model-to-model transferability using all possible source images. However, such an approach is not scalable, unless an abundance of computational power is available, thus forcing us to investigate alternate methods for the identification of these atypical source images. An important piece of information we have for each source image is the vector of prediction probabilities obtained through the softmax function, $P(\theta,\boldsymbol{x}) = [e^{g(\theta,\boldsymbol{x})_c} / \sum_{k=1}^{M} e^{g(\theta,\boldsymbol{x})_k}]_{c\in\{1,\ldots,M\}}$. The softmax output in conjunction with various error quantification methods has seen a significant use in recent research efforts on measuring the robustness and calibrated nature of DNNs [20]. Relying on the knowledge obtained from these studies, we use the following error quantification methods for evaluating the suitability of source images: the error made for the correct class, as calculated by (1) $1-\max(P(\theta,\boldsymbol{x}))$, (2) mean squared error (MSE), (3) Wasserstein distance (WD), and (4) the ratio of probabilities (Q) (that is, the second-largest to the largest one). Details on the way the different errors are calculated can be found in the supplementary material (Section F).

In Table 1, we provide the correlation between (a) the error measurement for the prediction of source images and (b) the properties of adversarial examples originating from those images (i.e., transferability and perturbation). Even though we use a large number of data points for this analysis, we still find a moderate correlation between multiple error estimates and adversarial properties. In particular, the simple approach of $\mathrm{Q}(\cdot)$ has the largest correlation when it comes to estimating perturbations, while having a comparably large correlation with transferability. Based on Table 1, we observe that, when $P(\theta,\boldsymbol{x})$ for a source image has its second-largest prediction closer to the largest one, adversarial examples originating from that source image are more likely to achieve adversarial transferability while requiring less perturbation. This leads to the question whether or not these error estimates can be used to identify fragile images, thus alleviating the need for large-scale experimentation. To answer

this question, we devise the following experimental procedure.

In order to approximate the adversarial properties of source images, we group source images according to the $Q(P(\theta,\boldsymbol{x}))$-value obtained. Specifically, we sort $\mathbb{S}$ according to $Q(P(\theta,\boldsymbol{x}))$ and create subsets based on certain percentiles of $Q(\cdot)$. Doing so, we observe the results for the same experimental routine described above (i.e., $1,000$ source images sampled $10,000$ times), but with a small difference: only the source images that have $Q(\cdot)$ larger than the 75th and 90th percentile ($\mathbb{S}_{Q>\{75,90\}}$), as well as source images that have $Q(\cdot)$ smaller than the 10th and 25th percentile ($\mathbb{S}_{Q<\{10,25\}}$), are selected.

We observe that source images with lower error estimates, as measured through $Q(P(\theta,\boldsymbol{x}))$, are harder to convert to adversarial examples, whereas the ones with higher $Q(P(\theta,\boldsymbol{x}))$ estimates are easier to convert. Furthermore, the required amount of perturbation for creating adversarial examples also differs greatly between the lower and the upper end of $Q(\cdot)$, with source images having a lower $Q(\cdot)$ requiring more perturbation, and vice versa. These results indicate that error estimates based on the prediction of source images can be used as a proxy for the properties of adversarial examples originating from these source images.

Finally, we measure the adversarial properties obtained with source images filtered from both ends, with $\mathbb{S}\setminus(\mathbb{S}_{Q<P}\cup\mathbb{S}_{Q>100-P})$. Using this approach, overall, we are able to reduce the difference between the highest and the lowest transferability from 12.5% to 7.6%, the difference in L_2 norm perturbation from 1.01 to 0.71, and the difference in L_∞ norm perturbation from 0.03 to 0.01, thus pointing to a more stable experimentation that is closer to the average case. Moreover, when we filter the same number of images from both ends (e.g., $\mathbb{S}\setminus(\mathbb{S}_{Q<10}\cup\mathbb{S}_{Q>90})$), the average transferability goes down slightly compared to using all available source images, while the average amount of required perturbation goes up slightly. These results indicate that the usage of $Q(\cdot)$ is more reliable in identifying fragile (easy) source images than hard source images. Consequently, we believe that the way error measurements are performed can be further improved, for instance through the usage of more complex analysis that takes into account the categories of source images.

Comprehensive experimental results for the experiments detailed in this section are provided in the supplementary material (See Table I to Table VI).

7 Conclusions and outlook

With the help of large-scale experiments, we exposed the fragility of a subset of source images to adversariality, with the adversarial examples created from these fragile images achieving high transferability rates for relatively small perturbation budgets. We then took one of the first steps to identify unusual source images that are either very hard or very easy to convert to adversarial examples, with the goal of supporting high-quality experimentation.

Given the security concerns associated with adversarial examples, an important item for future work is to evaluate how the observations made in this paper extend to adversarial defenses. In particular, we believe that the fragile images we have identified, given the properties discussed in this paper, can easily be leveraged to circumvent adversarial defenses.

We noted that a large number of adversarial examples are misclassified into categories that are semantically close to the categories of their source image counterparts, thus achieving untargeted adversarial transferability. In the supplementary material (Section G), we provide a number of qualitative examples of such cases. In that regard, we believe a detailed investigation of this topic, involving the semantic relationships between different categories, is also a promising item for future work, and where this future work item could make use of the hierarchies available in the WordNet database [40].

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015. URL https://www.tensorflow.org/.

[2] Anish Athalye and Nicholas Carlini. On the Robustness of the CVPR 2018 White-box Adversarial Example Defenses. *CoRR*, abs/1804.03286, 2018.

[3] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated Gradients Give A False Sense Of Security: Circumventing Defenses To Adversarial Examples. *International Conference on Machine Learning*, 2018.

[4] Shumeet Baluja and Ian Fischer. Adversarial Transformation Networks: Learning to Generate Adversarial Examples. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[5] Wieland Brendel, Jonas Rauber, Matthias Kümmerer, Ivan Ustyuzhaninov, and Matthias Bethge. Accurate, Reliable and Fast Robustness Evaluation. In *Advances in Neural Information Processing Systems*, 2019.

[6] Nicholas Carlini and David A. Wagner. Towards Evaluating The Robustness of Neural Networks. *2017 IEEE Symposium on Security and Privacy*, 2017.

[7] Nicholas Carlini and David A. Wagner. Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods. *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017.

[8] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Ead: Elastic-Net Attacks to Deep Neural Networks via Adversarial Examples. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[9] Alesia Chernikova, Alina Oprea, Cristina Nita-Rotaru, and BaekGyu Kim. Are Self-Driving Cars Secure? Evasion Attacks Against Deep Neural Networks For Steering Angle Prediction. *IEEE Security and Privacy Workshops*, 2019.

[10] Francesco Croce and Matthias Hein. Sparse and Imperceivable Adversarial Attacks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[11] Francesco Croce and Matthias Hein. Provable Robustness Against All Adversarial l_p-perturbations for $p \geq 1$. In *International Conference on Learning Representations*, 2020.

[12] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting Adversarial Attacks with Momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[13] Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. Benchmarking Adversarial Robustness on Image Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

[15] Chris Finlay, Aram-Alexandre Pooladian, and Adam Oberman. The LogBarrier Adversarial Attack: Making Effective use of Decision Boundary Information. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[16] Samuel G Finlayson, Isaac S Kohane, and Andrew L Beam. Adversarial Attacks Against Medical Deep Learning Systems. *Science*, 2019.

[17] Justin Gilmer, Ryan P Adams, Ian Goodfellow, David Andersen, and George E Dahl. Motivating The Rules Of The Game For Adversarial Example Research. *CoRR*, abs/1807.06732, 2018.

[18] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. *International Conference on Learning Representations*, 2015.

[19] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. On The (Statistical) Detection Of Adversarial Examples. *CoRR*, abs/1702.06280, 2017.

[20] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning*, 2017.

[21] Chuan Guo, Jacob R Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Q Weinberger. Simple Black-box Adversarial Attacks. *International Conference on Machine Learning*, 2019.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning For Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[23] Shengyuan Hu, Tao Yu, Chuan Guo, Wei-Lun Chao, and Kilian Q Weinberger. A New Defense Against Adversarial Images: Turning a Weakness Into a Strength. In *Advances in Neural Information Processing Systems*, 2019.

[24] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[25] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *CoRR*, abs/1602.07360, 2016.

[26] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box Adversarial Attacks with Limited Queries and Information. *International Conference on Machine Learning*, 2018.

[27] Eric Jones, Travis Oliphant, and Pearu Peterson. Scipy: Open Source Scientific Tools For Python, 2001.

[28] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial Logit Pairing. *CoRR*, abs/1803.06373, 2018.

[29] Danny Karmon, Daniel Zoran, and Yoav Goldberg. Lavan: Localized and Visible Adversarial Noise. *International Conference on Machine Learning*, 2018.

[30] Jinkyu Koo, Michael Roth, and Saurabh Bagchi. Hawkeye: Adversarial Example Detector For Deep Neural Networks. *CoRR*, abs/1909.09938, 2019.

[31] Alex Krizhevsky and Geoffrey Hinton. Learning Multiple Layers Of Features From Tiny Images. Technical report, Citeseer, 2009.

[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, 2012.

[33] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial Examples In The Physical World. *Workshop Track, International Conference on Learning Representations*, 2016.

[34] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-Based Learning Applied To Document Recognition. *Proceedings of the IEEE*, 1998.

[35] Eden Levy, Yael Mathov, Ziv Katzir, Asaf Shabtai, and Yuval Elovici. Not all datasets are born equal: On heterogeneous data and adversarial examples. *CoRR*, abs/2010.03180, 2020.

[36] Xin Li and Fuxin Li. Adversarial Examples Detection In Deep Networks With Convolutional Filter Statistics. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[37] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense Against Adversarial Attacks Using High-level Representation Guided Denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[38] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant To Adversarial Attacks. *International Conference on Learning Representations*, 2018.

[39] Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal Adversarial Perturbations Against Semantic Image Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[40] George A Miller. *WordNet: An Electronic Lexical Database*. MIT press, 1998.

[41] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A Simple And Accurate Method To Fool Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[42] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal Adversarial Perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[43] Utku Ozbulak, Manvel Gasparyan, Wesley De Neve, and Arnout Van Messem. Perturbation Analysis of Gradient-based Adversarial Atacks. *Pattern Recognition Letters*, 2020.

[44] Nicolas Papernot, Patrick D. McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation As A Defense To Adversarial Perturbations Against Deep Neural Networks. *IEEE Symposium on Security and Privacy*, 2016.

[45] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic Differentiation in PyTorch, 2017. URL https://pytorch.org/.

[46] Kevin Roth, Yannic Kilcher, and Thomas Hofmann. The Odds Are Odd: A Statistical Test For Detecting Adversarial Examples. In *International Conference on Machine Learning*, 2019.

[47] A. Rozsa, E. M. Rudd, and T. E. Boult. Adversarial Diversity and Hard Positive Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016.

[48] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 2015.

[49] Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla, and Andrea Cavallaro. Colorfool: Semantic Adversarial Colorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[50] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks For Large-Scale Image Recognition. *International Conference on Learning Representations*, 2015.

[51] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is Robustness the Cost of Accuracy?–A Comprehensive Study on the Robustness of 18 Deep Image Classification Models. In *Proceedings of the European Conference on Computer Vision*, 2018.

[52] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing Properties Of Neural Networks. *International Conference on Learning Representations*, 2014.

[53] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On Adaptive Attacks to Adversarial Example Defenses. *Advances in Neural Information Processing Systems*, 2020.

[54] Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based Zeroth Order Optimization Method for Attacking Black-Box Neural Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[55] Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, Łukasz and Polosukhin, Illia. Attention is All you Need. In *Advances in Neural Information Processing Systems*, 2017.

[56] Ke Wang, Guangyu Wang, Ning Chen, and Ting Chen. How Robust Is Your Automatic Diagnosis Model? In *IEEE International Conference on Bioinformatics and Biomedicine*, 2019.

[57] Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip Connections Matter: On the Transferability of Adversarial Examples Generated with ResNets. In *International Conference on Learning Representations*, 2020.

[58] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating Adversarial Effects Through Randomization. *International Conference on Learning Representations*, 2018.

[59] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial Examples Improve Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[60] Kaidi Xu, Sijia Liu, Pu Zhao, Pin-Yu Chen, Huan Zhang, Quanfu Fan, Deniz Erdogmus, Yanzhi Wang, and Xue Lin. Structured Adversarial Attack: Towards General Implementation and Better Interpretability. *International Conference on Learning Representations*, 2019.

[61] Ziang Yan, Yiwen Guo, and Changshui Zhang. Deep Defense: Training DNNs with Improved Adversarial Robustness. In *Advances in Neural Information Processing Systems*, 2018.

[62] Yuchen Zhang and Percy Liang. Defending against Whitebox Adversarial Attacks via Randomized Discretization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

[63] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. Towards Large yet Imperceptible Adversarial Image Perturbations with Perceptual Color Distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.