2022 年臺灣國際科學展覽會
優勝作品專輯

作品編號 **190045**

參展科別 電腦科學與資訊工程

作品名稱 **Limited Query Black-box Adversarial Attacks in the Real World**

得獎獎項 三等獎

國　　家 **United States**

就讀學校 **High School of Mathematics and Natural Sciences "Professor Emanuil Ivanov"**

指導教師

作者姓名 **Hristo Todorov**

關鍵詞 **machine learning, adversarial**

作者照片

LIMITED QUERY BLACK-BOX ADVERSARIAL ATTACKS IN THE REAL WORLD

Hristo Todorov
HSMS "Prof. Emanuil Ivanov"
Kyustendil, Bulgaria

ABSTRACT

We study the creation of physical adversarial examples, which are robust to real-world transformations, using a limited number of queries to the target black-box neural networks. We observe that robust models tend to be especially susceptible to foreground manipulations, which motivates our novel Foreground attack. We demonstrate that gradient priors are a useful signal for black-box attacks and therefore introduce an improved version of the popular SimBA. We also propose an algorithm for transferable attacks that selects the most similar surrogates to the target model. Our black-box attacks outperform state-of-the-art approaches they are based on and support our belief that the concept of model similarity could be leveraged to build strong attacks in a limited-information setting.

1 INTRODUCTION

With the further integration of contemporary machine learning techniques into a wide variety of real-world domains (Krizhevsky et al., 2012; Janai et al., 2017), the question about their security and reliability becomes central. It has been shown that all modern artificial neural network architectures are susceptible to adversarial examples (Szegedy et al., 2013) — altered inputs that are almost indistinguishable from natural data and yet classified incorrectly.

Adversarial attacks have been successfully executed in various fields, such as image classification (Szegedy et al., 2013), speech recognition (Carlini and Wagner, 2018), and reinforcement learning (Gleave et al., 2019). An adversarial example may be used either to make the model return any wrong output (untargeted attack) or even to shift its prediction to a desired state (targeted attack). As a result, one can obtain control over the output of a model by executing a targeted attack, which might be used for example as a tool for deceiving and penetrating facial recognition systems (Komkov and Petiushko, 2019) or forcing self-driving vehicles to recognize stop signs as speed limits (Eykholt et al., 2018). Therefore, the task of building robust and reliable models that are resistant to adversarial attacks is a major problem for society when considering their deployment to the real world.

The current state-of-the-art defenses against such malicious inputs are based on variations of adversarial training (Madry et al., 2017) (training a model on adversarial examples instead of regular data with the objective of classifying them correctly) and therefore efficient attack methods are vital for improving the defenses. Models with black-box access, which is the default setting for deployment in many domains, are particularly hard to attack due to the obscure amount of information they provide — in contrast to the white-box scenario, they do not allow access to the model architecture and weights and only provide their top predictions. Furthermore, multiple factors in the real world (e.g. weather conditions) endanger the successful execution of a given attack due to the extra perturbations they may apply to the adversarial examples. In our work, we consider the unification of black-box and physical attacks, which represents a realistic real-world scenario, and formulate our **main research question**, namely

> *Can we synthesize adversarial examples using only a very limited number of queries to the target black-box model, while maintaining robustness to physical transformations?*

A popular method for creating adversarial examples for black-box models relies on the fact that some artificial neural networks share a lot of common properties, and consequently knowledge obtained about one or several machine learning models could be integrated into an attack against another similar model (Liu et al., 2016). Our intuition is that the very same principle could be leveraged to solve our main problem. Thus, we pose the following questions:

- *What kind of similarities do machine learning models exhibit?*
- *How can we select models that resemble our target one the most?*
- *How can knowledge obtained from similar models be integrated into the attack methodology?*
- *What kinds of transformations are present in the real world and how can we model them exclusively digitally?*

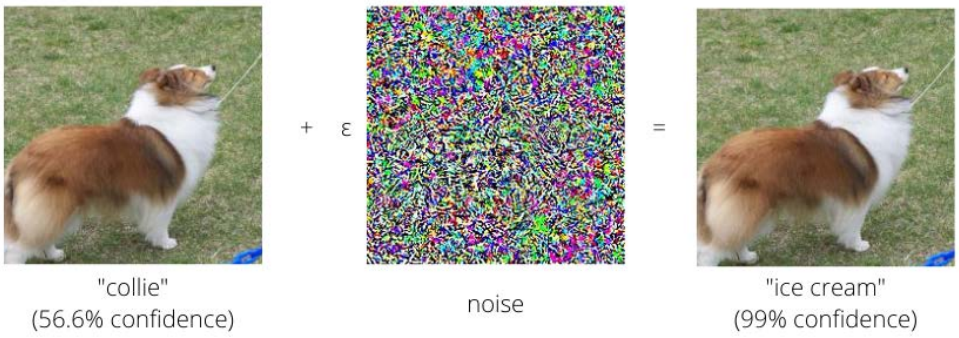


Figure 1: Creating an adversarial example: a picture of a Collie (dog breed), which is correctly classified by a given artificial neural network (trained on ImageNet in the current example), is insignificantly manipulated and then classified as ice cream by the same network

We make the following contributions:

- In our study of the impact of different discrete image regions (foreground and background) on model predictions, we observe that standard models are more sensitive to changes to the foreground rather than to the background. The same applies to robust models to an even greater extent.
- We propose a novel attack algorithm we call *Foreground Attack*. Our attack can achieve especially plausible results, despite constraining the perturbations only to the foreground of the given image, which supports our claim that region priors are useful for executing attacks with relatively little information from the target model, which is the case in the black-box scenario. We further point out that our Foreground Attack can be combined with adversarial training to increase the robustness of a model to physical adversarial examples by modeling the training data exclusively digitally.
- We introduce an ensemble-based transferable attack we call *Selective Transfer Attack* that replaces the manual process of picking appropriate surrogate models, which is vital for its success, with an automatic selection over a large set of available models.
- We present our *Ensemble Gradient-based SimBA*, which is based on one of the state-of-the-art black-box attacks (Guo et al., 2019). We show that gradient priors obtained from

a highly similar ensemble model are a stable and meaningful signal for selecting efficient perturbations.

2 Related Work

Physical adversarial examples. It has been shown that adversarial examples can transfer to the real world. Kurakin et al. (2016) were the first to demonstrate that property by creating printed photos that fool classifiers with a physical camera input. However, their photos are brittle to real-world transformations and therefore not suitable for a realistic attack. Athalye et al. (2017) proposed a method for synthesizing both 2D and 3D physical adversarial examples that are robust over a chosen distribution of transformations. Yet, generating such objects and putting them into a real-world setting is a rather computationally expensive process that is highly insufficient during the adversarial training of a model, which receives input from a physical camera. Other approaches use specially designed objects, such as printable stickers (Brown et al., 2017), which are placed onto or near a certain physical object in order to affect the prediction about it. Those objects are nevertheless usually quite noticeable and make the adversarial examples easily distinguishable from natural data.

Black-box attacks. Models with black-box access are especially hard to attack, due to the limited amount of information they provide. Papernot et al. (2016)'s model substitution method produced the first full black-box adversarial attack, but it is considered quite ineffective nowadays since training an entire model from scratch is a way too complex and slow task. Ensemble strategies (Liu et al., 2016; Hang et al., 2019) have proven to be an effective way for synthesizing adversarial examples that can transfer across a variety of models. A critical component of theirs is the surrogate model selection process, which is often coarse-grained and done manually. Guo et al. (2019)'s SimBA is currently one of the state-of-the-art black-box attacks. One of its main downsides is the uniform coordinate sampling process. Yang et al. (2020) propose a way to utilize gradient priors to improve SimBA's efficiency. We further develop their idea, by demonstrating that ensemble models selected through our model similarity evaluation technique provide more stable and meaningful priors.

3 Background

In this section, we introduce the concepts and notation needed to address the main research question and to describe the proposed algorithms precisely.

Adversarial attacks Suppose we have a target classifier $f(x, \theta)$ with parameters θ, and a natural input x with label y. We refer to the prediction of the model about the probability of x belonging to y as $f(x, \theta)_y$. If we denote the standard cross-entropy loss function by

$$l(p, y) = -\log\left(\frac{\exp(p_y)}{\sum_i \exp(p_i)}\right), \tag{1}$$

where $p = f(x, \theta)$, the process of generating an adversarial example can be expressed as the optimization problem

$$\max_{\delta \in \Delta} l(f(x + \delta, \theta), y), \tag{2}$$

where $\Delta = \{\delta : \|\delta\|_p \leq \epsilon\}$. The most popular perturbation sets are the l_2 and the l_∞ balls, due to the simplicity of projecting onto them.

We will consider Projected Gradient Descent (Madry et al., 2017) as a way of tackling the optimization problem in 2. If we refer to the gradient of the loss function with respect to a given image as $\nabla_x l$, then the adversarial perturbation δ can be iteratively updated as

$$\delta : \mathcal{P}(\delta + \sigma * \nabla_\delta l(f(x + \delta, \theta), y)), \tag{3}$$

where σ is the chosen step size and $\mathcal{P}$ is the projection over the ball of interest.

Adversarial training The given model architecture can increase its robustness by replacing the regular training objective

$$\min_{\theta} l(x, y)$$

with its adversarial training counterpart (Madry et al., 2017), viz.

$$\min_{\theta} \max_{\delta \in \Delta} l(f(x + \delta, \theta), y). \tag{4}$$

Note that the robustness of a given model is relative to a chosen l_p ball with a small radius ϵ, because a large radius (e.g. 1 for a l_∞ ball) would mean that the image may be perturbed to an extent that it is either no longer recognizable even to humans or it portrays an entirely different concept.

4 METHODOLOGY

Gradient analysis. The influence of different image regions on model predictions is nonuniform and our intuition is that the foreground (rather than the background) encodes more vital information associated with the real label of a given image (e.g. if an image is labeled as "horse", then its muzzle is more directly associated with the class label than the grass surrounding it), even though it has been demonstrated that the background is also quite a helpful signal (Xiao et al., 2020). The gradient $\nabla_x l$ provides useful information about the relative impact of each value of the image and many successful white-box attacks are built on top of that property (Goodfellow et al., 2014; Madry et al., 2017). We analyze the gradients of the loss functions of models with respect to a variety of annotated images to validate our hypothesis about the importance of the foreground. Let us denote an annotation mask by s, so that

$$s_{i,j,k} = \begin{cases} 1, & \text{if } (i, j, k) \text{ is inside the foreground of } x \\ 0, & \text{otherwise} \end{cases}, \tag{5}$$

where $i \in [3]$ (number of channels), $j \in [W]$, $k \in [H]$, and W and H are respectively the width and the height of x. We calculate the gradient $\nabla_x l$ and then its mean ($\mu \in \mathbb{R}^3$) and standard deviation ($\sigma \in \mathbb{R}^3$) channel-wise. We standardize the gradient by updating

$$\nabla_x l_{i,j,k} = \frac{\nabla_x l_{i,j,k} - \mu_i}{\sigma_i} \tag{6}$$

Finally, we segment the foreground $\psi = s * |\nabla_x l|$ and the background $\hat{\psi} = |\nabla_x l| - \psi$ from the absolute value of the standardized gradient and compare their mean values to gain insight about their relative importance.

Foreground Attack In order to further support our claim that region priors are an important signal for adversarial attacks, we develop a novel technique we call Foreground Attack. It is motivated by our findings that the foreground of an image affects model predictions a lot more than its background (see Figure 5). In our Foreground Attack, we just use an annotation mask s to constrain the adversarial noise δ to the foreground of the image x, namely

$$\hat{x} = x + s * \delta, \tag{7}$$

where $\hat{x}$ is the perturbed image. Our technique is universal and can be combined with any other kind of attack because all kinds of image manipulations can be represented as additive noises (by just subtracting the original version of the image from the transformed one). In our experiments, we integrate it into the optimization procedure of PGD.

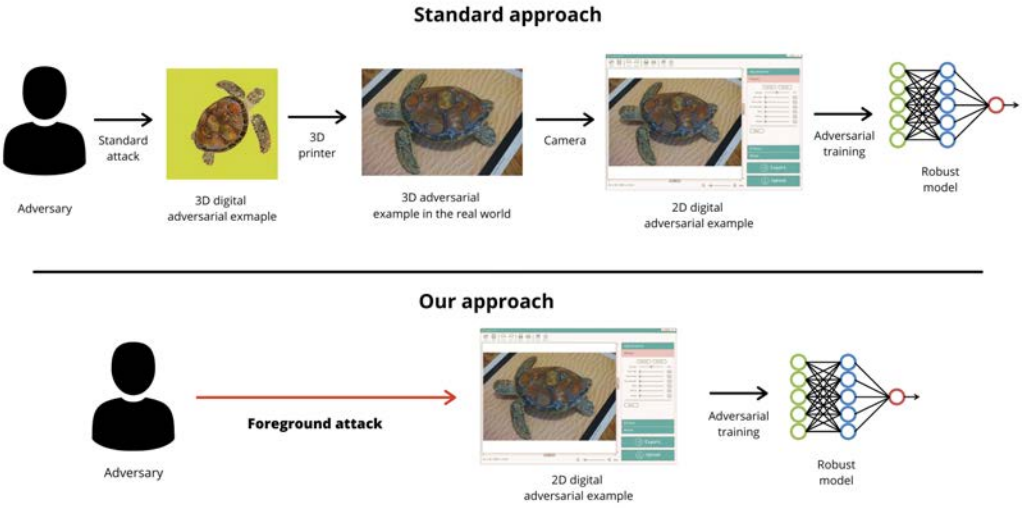


Figure 2: Comparison between a standard approach for building defenses against physical attacks and one employing our Foreground Attack (the photo of the adversarial turtle is from Athalye et al. (2017)'s paper)

Systems that collect input data from physical cameras (e.g. self-driving vehicles, facial recognition systems) are highly susceptible to physical adversarial examples. As such systems are consistently gaining more popularity, building defenses that are robust against such physical attacks becomes a vital task. In a realistic setting, the adversary cannot carefully apply perturbations to all objects that are going to be captured by the camera (e.g. the sky may be captured, and it is hard to apply permanent physical manipulations to it). Usually, such attacks are executed by placing small objects near the items that have to be classified (e.g. placing a tiny sticker on a stop sign (Brown et al., 2017)). Building defenses against such attacks is not straightforward, because creating physical adversarial objects is a slow and costly process, and the execution of such attacks has to be integrated into the adversarial training procedure. However, our Foreground Attack can apply perturbations directly to the foregrounds of the taken photos (which are passed to the model for classification), thus making the chosen objects in the photos look modified. We constrain the perturbations in the photos to specific areas, as real-world physical attacks are constrained too. Therefore, our Foreground Attack could be used for building defenses against realistic attacks, as the final inputs to the target models are just 2D images (see Figure 4).

Selective Transfer Attack. We consider a set of n pre-trained models $G = \{g_1, ..., g_n\}$, where each model architecture g_i is parameterized by different parameters θ_i, Our goal is to select the most similar ones to our target black-box model $f(x)$ (the parameters θ are unknown). As a metric of similarity between a given surrogate and our target model, we consider the sum of the cross-entropy losses between the respective predictions of the available models for a given image over a set of transformations (e.g. Gaussian noise), with lower metric values corresponding to higher degrees of similarity. We evaluate all our models from G on the given metric and choose k models with the lowest corresponding values, which form the set of surrogates $S = \{g_1, ..., g_k\}$.

Algorithm 1 Surrogate models selection

Input: number of iterations t, target model $f(x)$, image x, set of n pretrained models G, number of models to select k, standard deviation σ (Gaussian distribution)
Output: set of k models
$s := 0_n$
for $i = 1$ to t **do**
    Sample noise δ with the same size as x from the Gaussian distribution $\mathcal{N}(0, \sigma^2, I)$
    Compute the current model prediction $p = f(x + \delta)$
    Evaluate the predicted label $\hat{y} = \underset{y}{\operatorname{argmax}}\, p_y$
    for $j = 1$ to n **do**
        Compute the cross-entropy loss $l(g_j(x + \delta, \theta_j), \hat{y})$ and add it to s_j
    end for
end for
return the k models $\{g_1, .., g_k\}$ from G with the lowest respective values at s

After the selection process, we optimize the image x using PGD (shown in 3) using the loss

$$l_e(x, y) = \frac{\sum_{i=1}^{k} l(g_i(x, \theta_i), y)}{n} \tag{8}$$

over S. Our approach queries the target model only about the arbitrary set of transformations of the given image.

The selected surrogates do not exhibit equally high degrees of similarity to the target model, so the ensemble loss l_e could be estimated proportionally to the relative similarity scores s, viz.

$$l_e(x, y) = \sum_{i=1}^{k} ((g_i(x, \theta_i), y) * w_{k-i+1}), \tag{9}$$

where

$$w = \frac{s}{|s|_1}, \tag{10}$$

and s contains the sorted in increasing order relative similarity scores only for the selected surrogates (corresponding to the indices with the lowest respective values). The model with index i is assigned weight w_{k-i+1} because we reverse the order of the weights so that the model with the lowest score has the highest weight (as lower scores correspond to higher degrees of similarity to the target one).

Ensemble Gradient-based SimBA It has been demonstrated that black-box attacks can be executed without any priors by observing the change of the target model's predictions over a distribution of perturbations of the given image x. These observations can be leveraged by estimating $\nabla_x l$ (Wierstra et al., 2014) and then combined with a standard gradient-based optimization method, such as PGD, to craft adversarial examples. Interestingly, more efficient results can be achieved by using a much simpler approach, which directly applies perturbations that maximize $l(f(x + \delta), f(x)$ (Guo et al., 2019; Andriushchenko et al., 2020).

One such approach is Guo et al. (2019)' Simple Black-box Attack (SimBA), which randomly perturbs individual coordinates of the chosen image and keeps only the perturbations that optimize its objective. However, one of our main goals is to minimize the number of queries to the target model, and therefore an approach without any priors is an inferior choice.

Yang et al. (2020) further improved this idea by leveraging gradient priors obtained from a surrogate model. Instead of sampling coordinates uniformly, they consider a surrogate model $f_s(x, \theta_s)$ and sample coordinate-wise perturbation with probabilities proportional to $|\nabla_x l(f_s(x+\delta, \theta_s), y)|$, where y is the correct label of x and $|\cdot|$ denotes absolute value.

However, we believe that by randomly choosing a surrogate model to guide the sampling procedure it is possible that the surrogate model could provide misguiding instructions if it has learned sub-

stantially different representations of the training data in comparison to our target model. We instead point out that we can use the selection procedure described in Algorithm 1 to sample surrogates that are highly similar to our target one, which would induce a higher correlation with their respective gradients with respect to the input. Furthermore, it has been shown that ensemble methods (Liu et al., 2016) provide more robust predictions. We, therefore, augment Yang et al. (2020)'s approach by replacing the single surrogate model with an ensemble one, selected using our model similarity metric.

Algorithm 2 Standard SimBA

Input: target model $f(x)$, original image x, original label y,
Output: adversarial example x
Parameters: step size σ, number of iterations n, dimensions of x: W, H, C
$p \leftarrow f(x)_y$
for $i = 1$ to n **do**
    Sample uniformly $w \in [W]$, $h \in [H]$, $c \in [C]$
    $x_{w,h,c} \leftarrow x_{w,h,c} + \sigma$
    $p' \leftarrow f(x)_y$
    if $p' < p$ **then**
        $p \leftarrow p'$
    else
        $x_{w,h,c} \leftarrow x_{w,h,c} - 2 * \sigma$
        $p' \leftarrow f(x)_y$
        if $p' < p$ **then**
            $p \leftarrow p'$
        end if
    end if
end for
return x

Algorithm 3 Ensemble Gradient-based SimBA

Input: target model $f(x)$, original image x, original label y
Output: adversarial example x
Parameters: step size σ, number of iterations n, dimensions of x: W, H, C (width, height, number of channels)
Select a set surrogate models f_s using Algorithm 1
$p \leftarrow f(x)_y$
for $i = 1$ to n **do**
    Estimate $g := \nabla_x l_s(x, y)$ (see Algorithm 1
    Sample coordinate indices (w, h, c) from x with probabilities proportional to $|g|$
    $x_{w,h,c} \leftarrow x_{w,h,c} + \sigma$
    $p' \leftarrow f(x)_y$
    if $p' < p$ **then**
        $p \leftarrow p'$
    else
        $x_{w,h,c} \leftarrow x_{w,h,c} - 2 * \sigma$
        $p' \leftarrow f(x)_y$
        if $p' < p$ **then**
            $p \leftarrow p'$
        end if
    end if
end for
return x

Robustness to real-world transformations. During the preparation of a physical attack (e.g. a sticker placed on a stop sign in order to fool the computer vision system of a self-driving vehicle (Eykholt et al., 2018)) additional real-world factors, such as weather change or a different viewpoint, should be considered because they alter the captured by a camera adversarial example, which might no longer be able to fool the target model. In order to circumvent this problem, we first develop a framework for transformations, which digitally resemble real-world phenomena (light adjustment, noise addition, blurring, rotation, and translation), and then use the technique Expectation Over Transformation (EOT) (Athalye et al., 2017) to optimize the objective of our examples to maximize the cross-entropy loss over a chosen distribution of transformations T, created with our transformation framework, which can be expressed as the optimization problem

$$\underset{\delta}{\operatorname{argmin}} \mathbb{E}_{t \sim T} \log(f(t(x + \delta))_y). \tag{11}$$

5 EXPERIMENTS & RESULTS

Setup. Every method mentioned in the paper is implemented from scratch in Python. We evaluate our methods on randomly selected samples from ImageNet (Deng et al., 2009) and COCO (employing both images and annotation masks) datasets.

Gradient analysis. For each image in a given COCO class, we compute respectively the segmented foreground and background mean values. Then, we compute the final foreground and background mean values for the whole class (by averaging the results for the different images), and finally plot and analyze those values.

Our empirical evidence (see Figure 5) supports our claim that foregrounds indeed have a broader impact on model predictions than backgrounds, which is a possible explanation for why our Foreground Attack is successful despite the highly constrained perturbation space.

We find a more interesting correlation when we run the same experiment, replacing the regular model with a robust one (adversarially trained). We observe that robust models are even more sensitive to the foreground (see Figure 6), which implies that they may learn better representations that align with humans' (humans also receive a more useful signal from the foreground when recognizing an object) and is another argument for Salman et al. (2020)'s claim that robust models transfer better (the whole idea of transfer learning is to use already obtained meaningful representations).

Foreground Attack. We compare a PGD attack with our Foreground Attack (which too uses PGD as an optimization algorithm). We however note that it can be integrated with any kind of optimization technique because all kinds of perturbations can be represented as additive noises. We evaluate both standard white-box attacks and black-box ensemble transferable attacks (the surrogate models are manually selected from the list of available ones).

Type	ϵ (l_∞)	Queries	Transferable	Success rate
Regular attack	0.031	50	False	100%
	0.063	128	True	100%
Foreground Attack (our)	0.031	50	False	96%
	0.063	128	True	84%

Table 1: Comparison between regular PGD attack and Foreground PGD attack (standard ResNet50 trained on ImageNet used for evaluation)

Our results suggest that the foreground of the image is indeed a useful signal for executing attacks. Our Foreground Attack has high success rates both during white-box and black-box scenarios (see Table 1), even though the perturbation space is quite constrained — the images used are mainly photos of airplanes, in which the foregrounds usually cover relatively small parts of the photos.

Selective Transfer Attack and Expectation Over Transformation We evaluate our Selective Transfer Attack and compare it to a standard ensemble-based transferable black-box attack. Both of the approaches have the additional objective of producing robust adversarial examples, which they optimize via EOT.

Our Selective Transfer Attack outperforms the standard approach for synthesizing transferable adversarial examples even when using fewer surrogate models (see Table 2), which means that our attack is both more efficient and computationally optimal, and also implies that evaluating similarities between models over a distribution of input transformations is a sufficient metric for selecting surrogates that closely resemble the target. Our results for the standard transfer attack may however vary, because the surrogates are chosen uniformly, thus repeating the same experiment might hold different results (the chosen surrogates may be more or less relevant). Selective Transfer Attack is also not purely deterministic (it samples Gaussian noise), but is more consistent when querying the target over a large distribution of transformations.

Type	ϵ	Surrogates	Queries	Success rate
Standard transfer attack	0.016 (l_∞)	5	50	58%
	0.016 (l_∞)	10	50	75%
Selective Transfer Attack	0.0156 (l_∞)	5	50	87%
	0.0156 (l_∞)	10	50	93%

Table 2: Comparison between standard transferable black-box attack with random surrogate sampling and and with our Selective Transfer Attack (both utilizing EOT)

Ensemble Gradient-based SimBA We compare our ensemble gradient-based variation of SimBA with the original approach. It is important to point out that our ensemble gradient-based sampling approach can be integrated with any attack that uses coordinate-wise sampling (such as SquareAttack (Andriushchenko et al., 2020)).

Type	ϵ	Queries	Success rate
Original SimBA	0.126 (l_∞)	1400	66%
Gradient-based SimBA (our)	0.05 (l_∞)	1400	66%

Table 3: Comparison between the original SimBA and our ensemble gradient-based approach (evaluated on ResNet50 as target model and Inception-v4 as surrogate)

Our results support the claim that gradient priors, obtained via a well-selected ensemble model, are indeed a useful signal, as our Ensemble Gradient-based SimBA is able to achieve the same success rate as the original approach when constrained by a l_∞ ball with a much smaller radius (see Table 3). While ensemble-based transferable attacks do not hold any guarantees that the generated adversarial example will be able to fool the target model, our Ensemble Gradient-based SimBA's focusing on only fooling the target model (which in turn requires a higher number of queries).

6 Conclusion

In this paper, we study the problem of synthesizing adversarial examples against black-box models using only a limited number of queries, with the additional objective for these examples to be robust to physical transformations to be exploitable in a realistic setting (e.g. attacks against self-driving cars or facial recognition systems). We observe that models, especially robust ones, are more susceptible to changes in the foregrounds of the images. This motivates our Foreground Attack, which is able to achieve satisfactory results despite the highly constrained perturbation space and serves as another argument for our claim that the different regions of a given image do not equally contribute to model predictions. Additionally, we point out that our Foreground Attack could be used to synthesize photos in which certain objects appear adversarial, and consequently to build defenses against physical adversarial examples exclusively digitally. We also introduce a novel method for surrogate model selection via a transformation-based similarity metric and demonstrate that the choice of surrogate models is vital for the success of ensemble-based attacks. Finally, we propose a new black-box attack, which outperforms the state-of-the-art method it is based on, and show that carefully selected surrogate models can provide useful signals for the attack.

There is clearly further work to be done on leveraging techniques such as meta-tailoring (Alet et al., 2020) to assess whether training over small amounts of data can further increase the resemblance between the surrogates and the target model. In order to verify the generality of our techniques,

evaluation on a wider variety of models should be performed, including real-world computer vision systems such as Google Cloud's Vision API[1] or Clarifai[2].

REFERENCES

Alet, F., Kawaguchi, K., Bauza, M., Kuru, N. G., Lozano-Perez, T., and Kaelbling, L. P. (2020). Tailoring: encoding inductive biases by optimizing unsupervised objectives at prediction time.

Andriushchenko, M., Croce, F., Flammarion, N., and Hein, M. (2020). Square attack: a query-efficient black-box adversarial attack via random search.

Athalye, A., Engstrom, L., Ilyas, A., and Kwok, K. (2017). Synthesizing robust adversarial examples.

Brown, T. B., Mané, D., Roy, A., Abadi, M., and Gilmer, J. (2017). Adversarial patch.

Carlini, N. and Wagner, D. (2018). Audio adversarial examples: Targeted attacks on speech-to-text.

Deng, J., Dong, W., Socher, R., Li, L., Kai Li, and Li Fei-Fei (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. (2018). Robust physical-world attacks on deep learning models.

Gleave, A., Dennis, M., Wild, C., Kant, N., Levine, S., and Russell, S. (2019). Adversarial policies: Attacking deep reinforcement learning.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples.

Guo, C., Gardner, J. R., You, Y., Wilson, A. G., and Weinberger, K. Q. (2019). Simple black-box adversarial attacks.

Hang, J., Han, K., Chen, H., and Li, Y. (2019). Ensemble adversarial black-box attacks against deep learning system. *Pattern Recognition*, 101:107184.

Janai, J., Güney, F., Behl, A., and Geiger, A. (2017). Computer vision for autonomous vehicles: Problems, datasets and state of the art.

Komkov, S. and Petiushko, A. (2019). Advhat: Real-world adversarial attack on arcface face id system.

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25.

Kurakin, A., Goodfellow, I., and Bengio, S. (2016). Adversarial examples in the physical world.

Liu, Y., Chen, X., Liu, C., and Song, D. (2016). Delving into transferable adversarial examples and black-box attacks.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. (2016). Practical black-box attacks against machine learning.

Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., and Madry, A. (2020). Do adversarially robust imagenet models transfer better?

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks.

[1] https://cloud.google.com/vision

[2] https://www.clarifai.com/

Wierstra, D., Schaul, T., Glasmachers, T., Sun, Y., Peters, J., and Schmidhuber, J. (2014). Natural evolution strategies. *Journal of Machine Learning Research*, 15(27):949–980.

Xiao, K., Engstrom, L., Ilyas, A., and Madry, A. (2020). Noise or signal: The role of image backgrounds in object recognition.

Yang, J., Jiang, Y., Huang, X., Ni, B., and Zhao, C. (2020). Learning black-box attackers with transferable priors and query feedback.

A PERTURBATION SETS

(a) $\epsilon = \frac{2}{255}$

(b) $\epsilon = \frac{8}{255}$

Figure 3: Adversarial examples, constrained by l_∞ balls with different radii

B ADVERSARIAL TRAINING

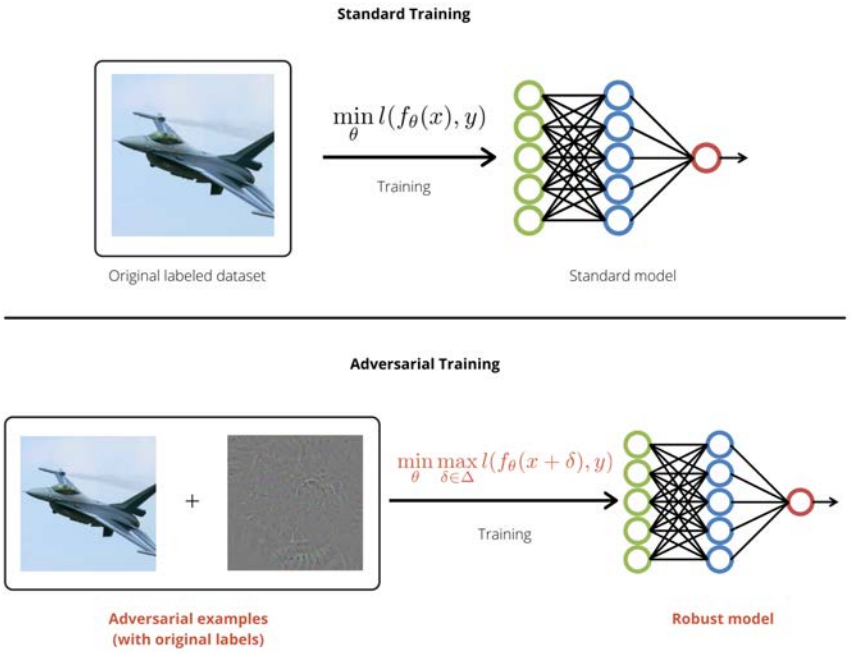


Figure 4: Comparison between the objectives of normal training and of adversarial training

C GRADIENT ANALYSIS RESULTS FIGURES

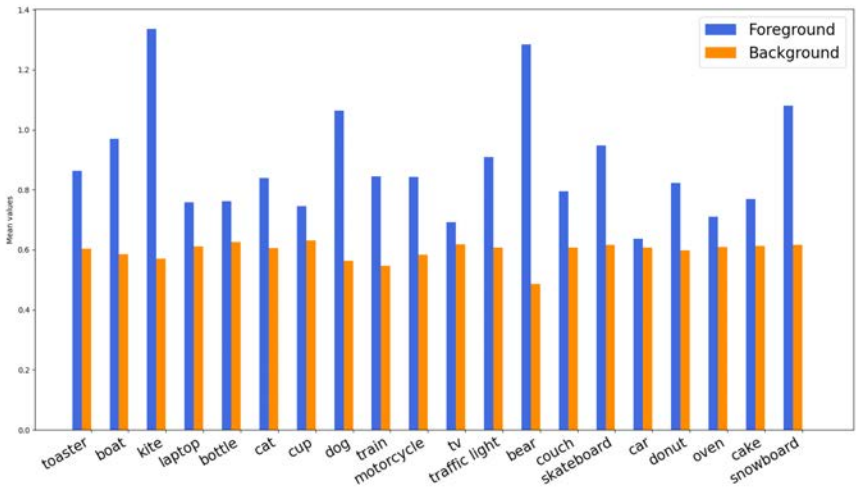


Figure 5: Comparison between foreground and background mean values (obtained via a regular ResNet50 trained on ImageNet) for 20 arbitrary COCO classes

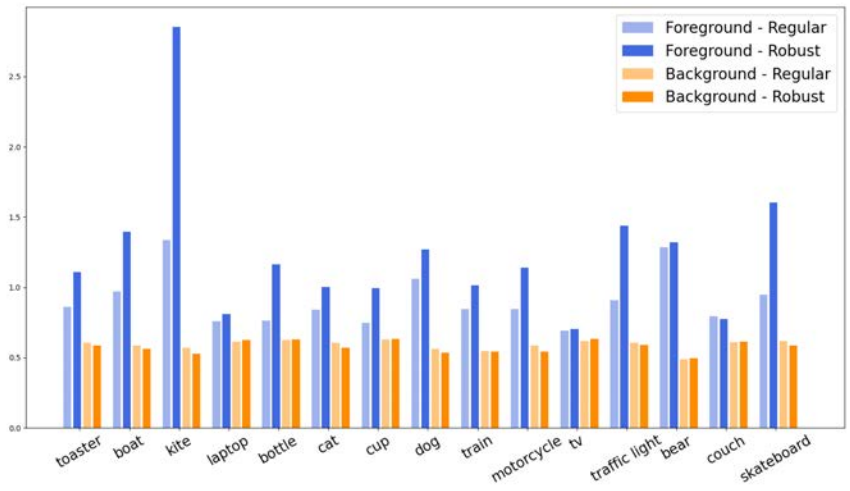


Figure 6: Comparison between foreground and background mean values of a regular and of a robust ResNet50 (both trained on ImageNet) for 20 arbitrary COCO classes

D Implementation Libraries

We have utilized the following Python libraries:

- PyTorch [3]
- NumPy [4]
- matplotlib [5]
- robustness [6]
- pandas [7]
- pretrainedmodels [8]

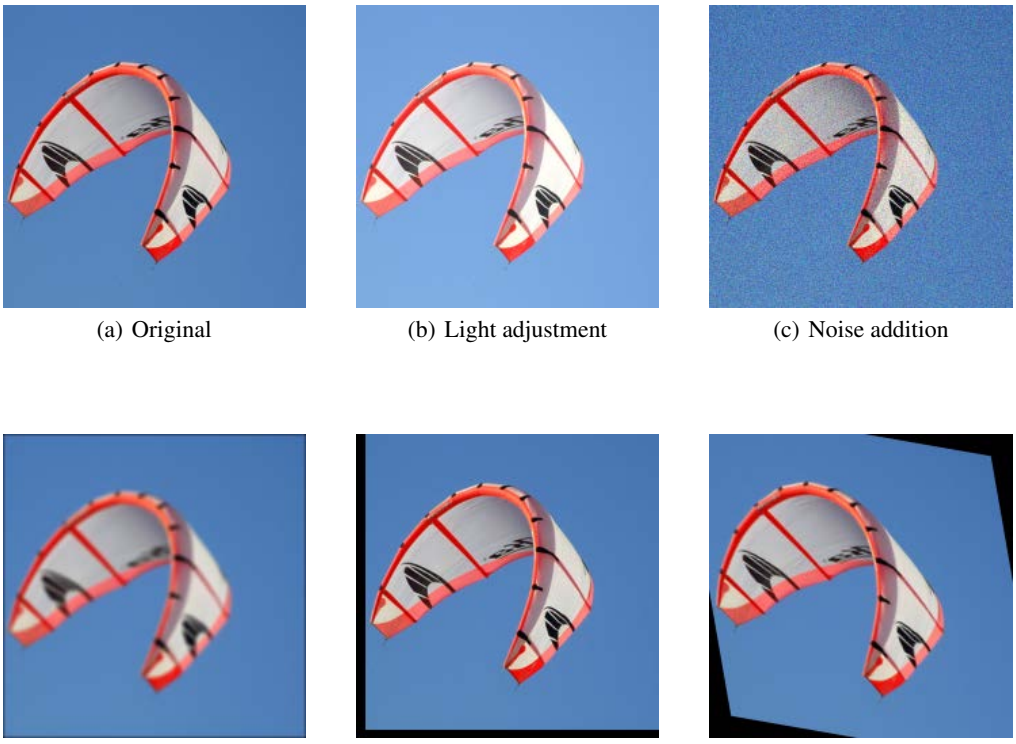
E Available Surrogates for Selection

The models are loaded using the pretrainedmodels library[8]

- mobilenet_v2
- shufflenet_v2_x0_5
- squeezenet1_1
- mnasnet1_0
- googlenet
- resnet18
- resnet34
- resnet50
- resnet101
- resnet152
- fbresnet152
- bninception
- resnext101_32x4d
- resnext101_64x4d
- alexnet
- densenet121
- densenet169
- densenet201
- densenet161
- vgg11
- vgg11_bn
- vgg13
- vgg13_bn
- vgg16
- vgg16_bn
- vgg19_bn
- vgg19
- nasnetamobile
- dpn68
- dpn68b
- dpn92
- dpn98
- dpn131
- dpn107
- xception
- senet154
- se_resnet50
- se_resnet101
- se_resnet152
- se_resnext50_32x4d
- se_resnext101_32x4d
- cafferesnet101

[3] https://github.com/pytorch/pytorch
[4] https://github.com/numpy/numpy
[5] https://github.com/matplotlib/matplotlib
[6] https://github.com/MadryLab/robustness
[7] https://github.com/pandas-dev/pandas
[8] https://github.com/Cadene/pretrained-models.pytorch

F EXPECTATION OVER TRANSFORMATION



(a) Original (b) Light adjustment (c) Noise addition

(d) Blur (e) Translation (f) Rotation

Figure 7: Types of transformation implemented in our framework

Transformation	Minimal value of σ	Maximal value of σ
Light adjustment	-0.1	0.1
Noise addition	0.0	0.05
Translation	-10	10
Blur	0	8
Translation	-10	10
Rotation	-35	35

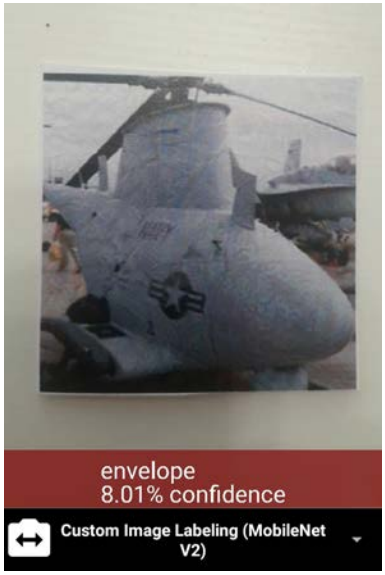
Table 4: Minimal and maximal values of the parameter σ for each kind of transformation

G SAMPLE ADVERSARIAL EXAMPLES

We display natural data on the left and their corresponding adversarial examples on the right. The digital images are labeled with the classes predicted by a standard ResNet50, whereas the physical are evaluated using MobileNetV2.

G.1 PHYSICAL

(a) Original



(b) Adversarial

(c) Original

(d) Adversarial

(e) Original

(f) Adversarial

G.2 DIGITAL

airliner

cleaver, meat cleaver, chopper

airliner

envelope

warplane, military plane

syringe

wing

hook, claw

warplane, military plane

paper towel

airliner

hair slide

【評語】190045

This project studies limited query black-box adversarial attacks in the real world. Three algorithms are presented in the project report and experimental results are discussed in depth. The overall quality of this project report is quite good.