Copy and Paste: A Simple But Effective Initialization Method for Black-Box Adversarial Attacks

Thomas Brunner[1,2] Frederik Diehl[1,2] Alois Knoll[2]
[1] fortiss GmbH [2] Technical University of Munich
{brunner,diehl}@fortiss.org knoll@in.tum.de

Abstract

Many optimization methods for generating black-box adversarial examples have been proposed, but the aspect of initializing said optimizers has not been considered in much detail. We show that the choice of starting points is indeed crucial, and that the performance of state-of-the-art attacks depends on it. First, we discuss desirable properties of starting points for attacking image classifiers, and how they can be chosen to increase query efficiency. Notably, we find that simply copying small patches from other images is a valid strategy. We then present an evaluation on ImageNet that clearly demonstrates the effectiveness of this method: Our initialization scheme reduces the number of queries required for a state-of-the-art Boundary Attack by 81%, significantly outperforming previous results reported for targeted black-box adversarial examples.

1. Introduction

Black-box adversarial attacks describe a scenario in which an attacker has access only to the input and output of a machine learning model, but no specific knowledge about its parameters or architecture [11]. Intuitively, it may seem improbable for an attacker to simply guess the vulnerabilities of a model, but recent work has demonstrated that it is indeed possible to create custom-tailored adversarial examples for any black box when allowed to query the model a large number of times [1, 8, 9, 15, 5].

Naturally, the practicality of such attacks depends on the number of queries required and much work has gone into designing efficient optimization strategies. Gradient estimation techniques have been popular [8, 5, 15] for models that provide real-valued output (*e.g.* softmax activations), and sophisticated sampling strategies have been proposed that drastically reduce the number of iterations [9]. The same has happened in the much harder label-only setting, where models only output a single discrete value (*e.g.* the top-1 class label). Early success in this setting was sparked by the Boundary Attack [1], which essentially performs a random walk along the decision boundary, and several variants have

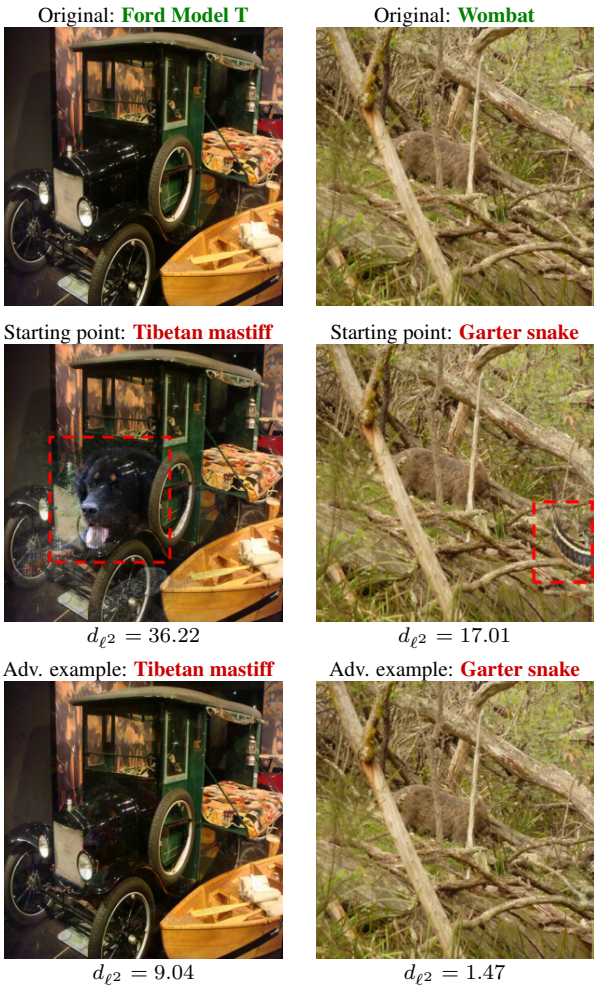

Figure 1. Starting points for adversarial attacks, synthesized by our method. (Top) Original images. (Middle) Starting points obtained by inserting small patches from images of the target class. (Bottom) Adversarial examples obtained after refining the images with a targeted black-box attack. Less than 1000 model queries are needed to render the perturbation virtually imperceptible, while the adversarial label is retained.

been proposed that improve the efficiency of this search procedure [7, 4]. Another family of successful black-box attacks exploits the fact that machine learning models often share vulnerabilities. They train surrogate models, perform white-box attacks and find adversarial examples that can be transferred to the model under attack [10, 14, 11].

It is evident that considerable effort has gone into the design of sophisticated optimization procedures. Yet surprisingly, the question of how to initialize them has not been discussed in detail. We consider this to be a rather important gap in current literature, since costly optimization procedures can often be sped up by a smart choice of starting points. Our contribution is as follows:

- We discuss beneficial properties of starting points and how they can improve the efficiency of iterative black-box attacks.
- As a proof of concept, we propose a simple copy-and-paste scheme, in which patches from images of the adversarial class are added to the image under attack.
- We use this strategy to initialize a state-of-the-art attack and evaluate it against an ImageNet classifier. Our initialization reduces the number of queries by 81% when compared with previous results, and thus forms a new state of the art in query-efficient black-box attacks. The source code for repeating our experiment is publicly available [1].

In this work, we exclusively focus on the targeted label-only black-box setting, where an attacker must change the classification to a specific label and does not have access to gradients or confidence scores. This is one of the hardest settings currently considered [1, 4, 6, 8] and therefore our results should be valid for easier settings as well.

2. Initialization strategies

Currently, black-box adversarial attacks start with either (a) the original image or (b) an example of the target class. In the case of (a), the attack tries to take steps into directions that lead to an adversarial region. This is considered very hard and is typically approached by estimating gradients [8, 5] or transferring them from a surrogate [10]. This approach can be unreliable and many queries must be spent before the adversarial region is found [4]. In order to improve reliability, other attacks [1, 8, 4] employ (b), where the starting point already has the desired class label but the distance to the original image is high. As a result, the attack needs to travel a great distance through the input space, requiring many steps until it arrives at an example that is reasonably close to the original.

[1] https://github.com/ttbrunner/blackbox_starting_points

Both strategies have potential for improvement. For (a), Tramèr et al. [14] propose adding small random perturbations to the input, which they find to increase the overall success rate. In this work, we focus on (b) – recent black-box attacks have achieved impressive results using this method [4, 8], and at the same time it seems very easy to improve. Surely some images of the target class would be better suited than others, and the number of required queries could be reduced by choosing them in a systematic manner.

2.1. Criteria for suitable starting points

In image classification, an adversarial example is considered successful if has low distance to the original image (*e.g.* measured by some norm of the perturbation) and is at the same time classified as an adversarial label. We assume two properties to be beneficial for attack efficiency:

Starting points should be close. Intuitively, it makes sense to pick points that are already close to that goal. The optimization procedure would then merely refine them, requiring less iterations to arrive at an adversarial example or produce a better one in the same number of steps. The most straightforward approach is to search a large data set (*e.g.* ImageNet) for images of the target class and then pick the one with the lowest distance to the original.

Starting points should reduce dimensionality. Optimization often suffers from very large search spaces. An ImageNet example at a resolution of 299 x 299 has 268,203 dimensions. It is therefore desirable to reduce this search space and to concentrate only on specific dimensions. Notably, Brunner et al. [4] demonstrate that attacks gain efficiency by concentrating only on pixels that differ between the current image and the original, ignoring the rest. A suitable starting point should facilitate this, ideally by not replacing the original entirely but only small regions of it. The attack can then be limited to these regions.

3. Copy-pasting adversarial features

In order to test our assumptions, we propose a simple strategy that segments images of the target class into small patches and then inserts them into the image under attack. It is geared towards attacks targeting the ℓ^2 norm, but similar strategies could be applied to improve ℓ^∞ attacks. This work should be understood as a proof of concept that offers many opportunities for refinement. Nevertheless, our evaluation in Section 4 shows that this simple approach already delivers a large boost in efficiency.

3.1. Segmentation by saliency

The most important pixels for classification are those that contain salient features. They typically concentrate in small regions (*e.g.* nose and eyes of an animal, see Figure 2), whereas the rest of an image matters little for the predicted class label. The unimportant regions can safely be removed

#	Attack	Initialization	Success rate vs number of queries						Median queries until success
			500	1000	2500	5000	10000	15000	
1	BBA [4]	Closest image	0.04	0.09	0.25	0.46	0.72	0.80	5485
2	BBA [4]	Copy and paste (ours)	0.29	0.38	0.63	0.75	0.88	0.90	1541
3	BBA (tuned)	**Copy and paste (ours)**	**0.32**	**0.49**	**0.74**	**0.86**	**0.94**	**0.96**	**1028**

Table 1. Comparison of initialization strategies for a targeted label-only Boundary Attack on ImageNet. All numbers include queries made during initialization. Run 1 is the biased Boundary Attack (BBA) as implemented by Brunner et al. [4]. Run 2 replaces their starting points with ours, but otherwise performs the same attack. In run 3, we tune some of the attack hyperparameters to take better advantage of our starting points, resulting in an even larger performance gain.

– compare the starting points in Figure 3 (left), where removing the background significantly lowers the ℓ^2-distance to the original but retains the adversarial label. To do this, we can apply any segmentation method of our choice. In our implementation, we use a surrogate model to construct saliency maps for images of the target class and then blend these pixels into the original image.

Saliency maps are model-specific, and therefore our initialization could be interpreted as a transfer attack that is not guaranteed to generalize across models. To address this concern, we apply heavy smoothing and amplification to the map. This results in contiguous patches that cover the salient regions and are therefore likely to contain the core motif of an image (see Figure 2). We expect this method to generalize well across models, but in practice it can also be replaced by any other segmentation technique available to the attacker.

3.2. Placing adversarial patches

It is apparent that the inserted pixels constitute small adversarial patches that change the classification of an entire image. This effect has been described by Brown et al. [3], who construct small patches that are very salient but also strikingly visible to human observers. It would be possible to use their patches as starting points, however their strong visibility is contrary to our goals – attacks in our setting aim to reduce the perturbation as much as possible in an attempt make the patches completely imperceptible to humans.

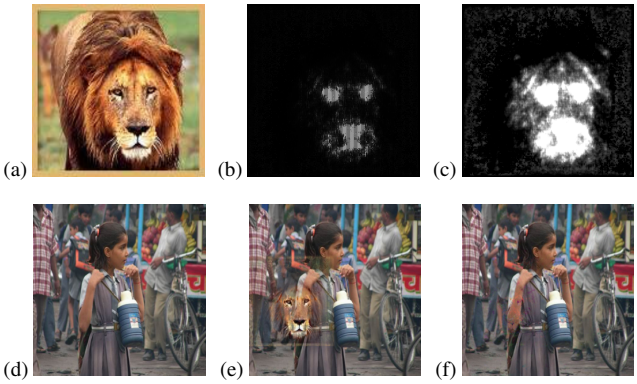

Figure 2. Segmentation of starting points. (a) An image of the target class, label "lion". (b) Saliency mask on a surrogate. (c) Smoothened and amplified mask to improve reliability. (d) Original image under attack, label "water bottle". (e) Starting point after the patch is inserted. (f) Final adversarial example.

We use a simple brute-force method for placement. Salient patches are extracted from multiple images of the target class, randomly scaled and translated, and then blended over the original image. We create 50 such candidates and rank them by distance to the original image. The candidates are then tested against the black box and the first image to be adversarially classified is chosen as the final starting point.

We typically find success within the first 10 candidates, but in the case that all are unsuccessful on the black-box model, an attacker may wish to fine-tune the procedure, increase patch size or fall back to full-sized images. In our evaluation, we are able to synthesize valid starting points for all examples.

4. Evaluation

To demonstrate the speedup provided by our initialization method, we apply it to a black-box attack against a pretrained Inception-v3 ImageNet classifier [13].

Setting. We choose a query-limited targeted label-only setting, which is currently considered to be one of the most difficult scenarios [1] to attack. We do not consider untargeted attacks, as ImageNet contains 1000 classes: An attacker could simply substitute one dog breed for another and thus create an "adversarial" example.

Data set. We randomly pick 1000 images from the ImageNet validation set and resize them to 299 x 299 x 3. For each image, we pick a random class label as the adversarial target. Our initialization method chooses images of the target class from the ImageNet validation set and uses a pretrained Inception-ResNet [12] to extract salient regions.

Attack method. We perform a Boundary Attack [1], which is one of the simplest ℓ^2 attacks and at the same time has recently been shown to be one of the most query-efficient [2, 4] in our setting. We use the publicly available implementation of Brunner et al. [4], who perform biased sampling to obtain state-of-the-art performance. We replace their initialization (closest image of the target class)

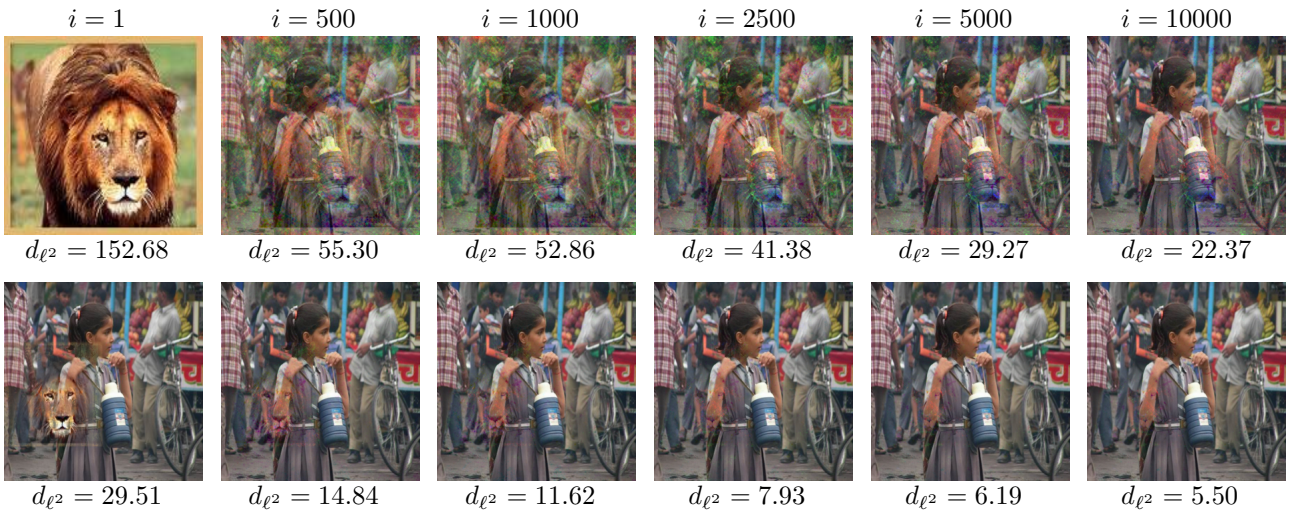

Figure 3. Number of queries vs. perturbation magnitude for a Boundary Attack. (Top) Initialized with a full-sized image of the target class. (Bottom) Initialized with our method. All images are classified as the adversarial label, "lion".

with our own and activate their low-frequency and regional masking biases. Low-frequency perturbations have been shown to improve efficiency [7, 9], and the masking bias should be able to take full advantage of a low-dimensional search space. For the sake of simplicity we deactivate the surrogate bias, and use our surrogate model only for extracting the (smoothened) saliency maps.

Success criteria. We measure attack efficiency in number of queries until success. We define success as $d_{\ell^2} < 25.89$, which is the same threshold as used in the original attack [4]. This number roughly corresponds to a worst-case ℓ^∞ perturbation of 0.05 in our resolution, which is a threshold also used by other black-box attacks on ImageNet [8, 9].

Results. Table 1 shows that our synthesized starting points greatly improve query efficiency, especially in the early stages of the attack. Simply replacing the initialization method reduces the median number of queries from 5485 to 1541, which is a reduction by 72%. Interestingly, some of the starting points synthesized by our method are already below the ℓ^2 threshold (such as the snake in Figure 1, which was obtained in a single query).

We perform another run with modified hyperparameters to take better advantage of our starting points (exact values are provided in our source code). This run achieves a median of 1028 queries, which in total amounts to a **reduction by 81%** and, to the best of our knowledge, significantly outperforms the previous state of the art in targeted black-box attacks. Figure 3 shows a side-by-side comparison of attack progress for a single example.

5. Conclusion

We have shown that the performance of iterative black-box attacks greatly depends on their initialization. In a proof of concept, we have demonstrated how starting points can be synthesized by adding small patches to an image. Indeed, some of the images crafted by our method could be considered adversarial from the start, rendering the actual attack largely obsolete.

This seems to indicate that the limits of current benchmarks for evaluating black-box attacks are being reached, and that subsequent improvements might not add to their applicability in the real world. A better definition of "adversarial" is needed, and metrics such as ℓ^2 (and, for that matter, ℓ^∞) should be replaced by measures based on human cognition. On high-resolution images, robustness should not be benchmarked by performing top-1 classification – an image that visibly contains both a car and a dog (Figure 1) should not be considered adversarial. Future benchmarks could address this problem by including multi-object detection, or at least top-k classification.

Still, our discovery provides some pointers for future work on real-world attacks. For example, an attacker may wish to create an adversarial patch as suggested by Brown et al. [3], and then use an iterative attack to camouflage it in the environment. It should also be interesting to see if adaptive attacks like the Boundary Attack can be modified to maintain a certain distance from the decision boundary, which would make the resulting examples robust in noisy scenarios and against non-deterministic classifiers. We consider this a prerequisite for finally bringing this family of attacks to the real world.

Our work shows that it is much easier to create black-box adversarial examples for high-resolution image classifiers than previously thought, and we hope this discovery provides a stepping stone towards a more realistic evaluation of black-box attacks in the future.

References

[1] W. Brendel, J. Rauber, and M. Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *International Conference on Learning Representations*, 2018. 1, 2, 3

[2] W. Brendel, J. Rauber, A. Kurakin, N. Papernot, B. Veliqi, M. Salathé, S. P. Mohanty, and M. Bethge. Adversarial vision challenge. *arXiv preprint arXiv:1808.01976*, 2018. 3

[3] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer. Adversarial patch. *arXiv preprint arXiv:1812.09803*, 2018. 3, 4

[4] T. Brunner, F. Diehl, M. Truong Le, and A. Knoll. Guessing smart: Biased sampling for efficient black-box adversarial attacks. *arXiv preprint arXiv:1812.09803*, 2018. 2, 3, 4

[5] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec '17, pages 15–26, 2017. 1, 2

[6] M. Cheng, T. Le, P.-Y. Chen, J. Yi, H. Zhang, and C.-J. Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. In *International Conference on Learning Representations*, 2019. 2

[7] C. Guo, J. S. Frank, and K. Q. Weinberger. Low frequency adversarial perturbation. *arXiv preprint arXiv:1809.08758*, 2018. 2, 4

[8] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin. Black-box adversarial attacks with limited queries and information. In *Proceedings of the 35th International Conference on Machine Learning, ICML*, 2018. 1, 2, 4

[9] A. Ilyas, L. Engstrom, and A. Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. *arXiv preprint arXiv:1807.07978*, 2018. 1, 4

[10] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. *International Conference on Learning Representations*, 2018. 2

[11] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ASIA CCS '17, pages 506–519, 2017. 1, 2

[12] C. Szegedy, S. Ioffe, and V. Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2016. 3

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *arXiv preprint arXiv:1512.00567*, 2015. 3

[14] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018. 2

[15] C. Tu, P. Ting, P. Chen, S. Liu, H. Zhang, J. Yi, C. Hsieh, and S. Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. *arXiv preprint arXiv:1805.11770*, 2018. 1