GenGradAttack: Efficient and Robust Targeted Adversarial Attacks Using Genetic Algorithms and Gradient-Based Fine-Tuning

Naman Agarwal[a] and James Pope[b]
Intelligent Systems Laboratory, School of Engineering Mathematics and Technology, University of Bristol, Bristol, U.K.

Keywords: Adversarial Machine Learning, Privacy-Preserving Image Classification, Genetic Algorithms, Gradient-Based Fine-Tuning, Black-Box Attack.

Abstract: Adversarial attacks pose a critical threat to the reliability of machine learning models, potentially undermining trust in practical applications. As machine learning models find deployment in vital domains like autonomous vehicles, healthcare, and finance, they become susceptible to adversarial examples—crafted inputs that induce erroneous high-confidence predictions. These attacks fall into two main categories: white-box, with full knowledge of model architecture, and black-box, with limited or no access to internal details. This paper introduces a novel approach for targeted adversarial attacks in black-box scenarios. By combining genetic algorithms and gradient-based fine-tuning, our method efficiently explores input space for perturbations without requiring access to internal model details. Subsequently, gradient-based fine-tuning optimizes these perturbations, aligning them with the target model's decision boundary. This dual strategy aims to evolve perturbations that effectively mislead target models while minimizing queries, ensuring stealthy attacks. Results demonstrate the efficacy of *GenGradAttack*, achieving a remarkable *95.06%* Adversarial Success Rate (ASR) on MNIST with a median query count of *556*. In contrast, conventional GenAttack achieved 100% ASR but required significantly more queries. When applied to InceptionV3 and Ens4AdvInceptionV3 on ImageNet, *GenGradAttack* outperformed GenAttack with *100%* and *96%* ASR, respectively, and fewer median queries. These results highlight the efficiency and effectiveness of our approach in generating adversarial examples with reduced query counts, advancing our understanding of adversarial vulnerabilities in practical contexts.

1 INTRODUCTION

Adversarial attacks in machine learning pose a serious threat to the reliability and security of deployed systems, especially in critical applications like autonomous vehicles, medical diagnosis, and financial systems. As machine learning models become integral to these domains, addressing their susceptibility to carefully crafted adversarial perturbations is crucial for safe real-world deployment.

Black-box attacks, simulating realistic conditions where adversaries lack direct access to internal model workings, are particularly relevant in cloud-based APIs or third-party model scenarios. This thesis explores targeted black-box attacks using a hybrid approach that combines genetic algorithms (GA) with gradient-based optimization. This complexity necessitates innovative methodologies to achieve high attack success rates while minimizing visual disruptions, especially for high-dimensional models like those in ImageNet.

Our method employs a genetic algorithm inspired by evolution principles, guiding the search process to produce potent adversarial scenarios through crossover, mutation, and selection processes. To further refine perturbations, we integrate gradient-based optimization techniques to iteratively update the noise pattern, addressing challenges in existing methods. The combination of genetic algorithms and gradient-based optimization aims to generate potent and inconspicuous black-box adversarial perturbations. By minimizing queries, we create visually imperceptible adversarial samples, enhancing attack potency and stealth.

In summary, our objectives are:

1. **Effectiveness of Genetic Algorithms:** Investigate the potency of genetic algorithms in generating black-box adversarial perturbations, leveraging their ability to explore vast search spaces and discover optimized solutions. Our findings demonstrate their effectiveness in consistently deceiving target models.
2. **Combining Optimization Techniques:** Enhance attack efficiency by combining gradient-based op-

[a] https://orcid.org/0009-0007-4899-7494
[b] https://orcid.org/0000-0003-2656-363X

Agarwal, N. and Pope, J.
GenGradAttack: Efficient and Robust Targeted Adversarial Attacks Using Genetic Algorithms and Gradient-Based Fine-Tuning.
DOI: 10.5220/0012314700003636

In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 3*, pages 202-209
ISBN: 978-989-758-680-4; ISSN: 2184-433X

timization with genetic algorithms. This hybrid approach achieves faster convergence, higher success rates, and query-efficient attacks.

3. **Generic and Transferable Attacks:** Develop techniques for generic and transferable attacks, successfully deceiving various models beyond the initially targeted one.
4. **Query-Efficient Attacks:** Minimize target model queries to increase the challenge of detecting and defending against adversarial attacks. Our approach significantly reduces the interactions required for perturbation, enhancing attack stealth.

2 BACKGROUND

Adversarial Attacks

Szegedy et al. (Szegedy et al., 2013) first exposed neural network vulnerability to carefully crafted adversarial examples, revealing ”shortcut learning” reliance on non-robust features. Techniques like Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2014) and Projected Gradient Descent (PGD) (Madry et al., 2017) expanded attack methods, with recent focus on pixel-wise prediction tasks (Agnihotri and Keuper, 2023).

Black-Box Attacks

Papernot et al. (Papernot et al., 2017) introduced transferability in black-box attacks, using substitute modeling and decision-based querying. Zeroth Order Optimisation (ZOO) (Chen et al., 2017) relied solely on the target model's input-output interface. Notable approaches include SimBA (Guo et al., 2019), lightweight attacks on shallow layers (Sun et al., 2022), query-efficient decision-based patch attacks (Chen et al., 2023), and deep reinforcement learning (Kang et al., 2023).

Black-Box Adversarial Attacks Using Genetic Algorithm

GenAttack (Alzantot et al., 2019) pioneered genetic algorithms for black-box attacks, followed by POBA-GA (Chen et al., 2019) and localized scratch attacks (Jere et al., 2019). Evolution-based methods like MF-GA (Wu et al., 2021) and attentional mechanisms in PICA (Wang et al., 2021) improved perturbation optimization. This study fills a gap by combining genetic algorithms with gradient-based fine-tuning, exploring their potential for more efficient and effective black-box attacks.

3 DESIGN

The considered attack framework assumes zero information about the network's architecture, parameters, or training data. The attacker, having no access to model features, operates solely through querying the model as a black-box function:

$$f : \mathbb{R}^d \to [0,1]^K \tag{1}$$

Here, d is the number of input features, and K is the number of classes. The attacker's objective is a precisely targeted assault, aiming to discover a perturbed instance x_{adv} for an innocuous input instance x. The perturbed instance aligns with the attacker's target prediction t, selected from the label set 1....K, while minimizing the L_p distance:

$$\arg\max_{c\in\{1..K\}} f(x_{\text{adv}})_c = t \text{ such that } \|x - x_{\text{adv}}\|_p \leq \delta \tag{2}$$

The L_p distance is often chosen as L_2 or L_∞ norm, following established methodologies in black-box attacks (Chen et al., 2017), (Papernot et al., 2017).

3.1 Genetic Algorithm and Gradient-Based Optimization

GenGradAttack builds upon the GenAttack algorithm (Alzantot et al., 2019) by introducing gradient-based fine-tuning. This hybrid approach combines natural selection within a population-based search with precision from gradient-based optimization.

GenGradAttack iteratively engages a population, denoted as P, in which candidate solutions evolve through crossover, mutation, and selection processes. Crossover integrates genetic information from two parents to produce offspring, simulating genetic recombination. Finite difference approximation estimates gradients when the true gradient is unknown, common in black-box optimization. Mutation introduces small random changes, promoting population variety and preventing local optima entrapment. This approach enhances solution quality, evolving perturbations that mislead the target model while minimizing queries, and aligning with the attack objective.

Algorithm 1 outlines the operation of GenGradAttack, aiming to generate an adversarial example x_{adv} from the original example x_{orig} and the attacker's chosen target label t, with the constraint $\|x_{\text{orig}} - x_{\text{adv}}\|_\infty \leq \delta_{\text{max}}$.

```
Input: Original example $x_{\text{orig}}$, target label $t$,
       maximum $L_\infty$ distance $\delta_{\max}$,
       mutation-range $\alpha$, mutation
       probability $\rho$, population size $N$,
       sampling temperature $\tau$
for $i = 1, \ldots, N$ in population do
    $P_0^{(i)} \leftarrow x_{\text{orig}} + U(-\delta_{\max}, \delta_{\max})$;
end
for $d = 1, 2, \ldots, G$ generations do
    for $i = 1, \ldots, N$ in population do
        $F_{d-1}^{(i)} = \text{ComputeFitness}(P_{d-1}^{(i)})$;
    end
    Find the elite member;
    $x_{\text{adv}} = P_{d-1}^{\text{argmax}}(F_{d-1})$;
    if $argmax_c f(x_{adv})_c = t$ then
        Return: $x_{\text{adv}}$    Found successful
          attack;
    end
    $P_d^1 = \{x_{\text{adv}}\}$;
    Compute selection probabilities;
    probs $= \text{Softmax}\left(\frac{F_{d-1}}{\tau}\right)$;
    for $i = 2, \ldots, N$ in population do
        Sample $parent1$ from $P_{d-1}$ according
          to probs;
        Sample $parent2$ from $P_{d-1}$ according
          to probs;
        $child = \text{Crossover}(parent1, parent2)$;
        ------------------------------------------------
          Finite Difference Approximation for
          Gradients;
        $gradients =$
          $\text{ComputeGradients}(child_{\text{mut}})$;
        $normalized\_gradients =$
          $\frac{gradients - \text{mean}(gradients)}{\text{std}(gradients) + \varepsilon}$;
        $child_{\text{mut}} = child_{\text{mut}} + \text{learning\_rate} \cdot$
          $normalized\_gradients$;
        Fine-tuning step using
          fine_tune_child;
        $child_{\text{mut}} = \text{fine\_tune\_child}(child_{\text{mut}})$;
        ------------------------------------------------
        Apply mutations and clipping;
        $child_{\text{mut}} = child + \text{Bernoulli}(\rho) \cdot$
          $U(-\alpha\delta_{\max}, \alpha\delta_{\max})$;
        $child_{\text{mut}} = \Pi_{\delta_{\max}}(child_{\text{mut}}, x_{\text{orig}})$;
        Add mutated child to next generation;
        $P_d^{(i)} = \{child_{\text{mut}}\}$;
    end
    Adaptively update $\alpha, \rho$ parameters;
    $\rho, \alpha = \text{UpdateParameters}(\rho, \alpha)$;
end
```

Algorithm 1: Genetic Attack Algorithm with Fine-tuning using Gradients Approximation.

Fitness Function: The ”ComputeFitness” function assesses each population member’s quality, incorporating the assigned output score for the target class label and considering the reduction in probability for other classes. The fitness function is defined as:

$$ComputeFitness(x) = \log f(x)_t - \log \sum_{j=0, j\neq t}^{j=k} f(x)_c. \tag{3}$$

Selection: Members of the population are ranked based on fitness, and Softmax is applied to determine the selection probability. Random parent pairings are then chosen, including the elite members using the elitism technique (Bhandari et al., 1996).

Crossover: Parents engage in a mating process, and offspring attributes are selected from either parent1 or parent2 based on selection probabilities $(p, 1-p)$, with p determined by fitness.

Gradient Approximation: In the black-box setting, gradients are approximated using central finite difference approximations to construct a loss function. The central difference approximation is given by:

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h}. \tag{4}$$

Mutation: A mutation process, guided by a probability factor ρ, introduces random noise within the interval $(-\alpha\delta_{\max}, \alpha\delta_{\max})$. The noise is applied to attributes derived from the outcomes of the crossover operation, ensuring pixel values stay within the allowable L_∞ distance from the original example x_{orig}.

4 RESULTS

GenGradAttack’s performance is evaluated against state-of-the-art image classification models, including MNIST and ImageNet. We utilize models identical to ZOO (Chen et al., 2017) and GenAttack (Alzantot et al., 2019) for each dataset. The MNIST model achieves 99.5% accuracy, and for ImageNet, Inception-v3 is used with 79.1% top-1 accuracy and 95.2% top-5 accuracy.

The evaluation compares GenGradAttack, ZOO, and GenAttack in terms of median queries, attack success rate (ASR), and runtime. Notably, the runtime and query count specifically apply to successful attacks, where a single query assesses the target model’s output for a given input image.

We utilize the code from the authors of ZOO and

GenAttack to configure our evaluation[1][2]. The experiments extend to a comparison with the C&W white box attack (Carlini and Wagner, 2017) for better understanding. Additionally, we analyze the efficacy of GenGradAttack against ensemble adversarial training (Tramèr et al., 2020) using models provided by the authors[3].

In MNIST studies, GenGradAttack is limited to a maximum of 100,000 queries, with hyperparameters set as follows: $\rho = 5 \times 10^{-2}$ (mutation probability), N = 6 (population size), and $\alpha = 1.0$ (step size). For ImageNet, given the larger image size, a maximum of 1,000,000 queries is allowed. We set $\delta_{max} = 0.3$ for MNIST and $\delta_{max} = 0.05$ for ImageNet, aligning with ZOO's mean L_∞ distortion in successful instances. This setup ensures meaningful comparisons with GenAttack (Alzantot et al., 2019).

5 CRITICAL ASSESSMENT OF USE CASES

This section provides a comprehensive evaluation of the presented use cases, focusing on their strengths and vulnerabilities. Through systematic analysis, we aim to illuminate the real-world implications of these use cases, offering a nuanced understanding of their significance in the broader research domain.

5.1 Use Case 1: Attacking MNIST Images

In the assessment of different attack methods on the MNIST dataset with an L_∞ perturbation of 0.30, Table 1 summarizes the performance of GenGradAttack.

Table 1: ASR and queries for different attack methods on the MNIST dataset with L_∞ = 0.30 perturbation. The bold row represents the results obtained using our proposed method, **GenGradAttack**.

Dataset	*Attack Method*	*ASR*	*Queries*
MNIST	C&W	100%	–
(L_∞ = 0.30)	ZOO	97%	2.1M
	GenAttack	94.25%	996
	GenGradAttack	**95.06%**	**556**

GenGradAttack achieves an Adversarial Success Rate (ASR) of 95.06%, requiring a median query count of **556** for success. Comparatively, C&W achieves a 100% ASR (white-box attack), ZOO attains a 97% ASR with 2,118,514 queries, and GenAttack achieves a 94.25% ASR with 996 queries. GenGradAttack is **3810** times more efficient than ZOO and **2** times more efficient than GenAttack. This suggests that GenGradAttack is promising for generating efficient adversarial examples on MNIST, balancing effectiveness and query efficiency.

Figure 1 visually presents the adversarial examples generated using GenGradAttack. The subtle perturbations introduced in the images lead to misclassifications, highlighting the effectiveness of GenGradAttack in compromising the model's classification accuracy.

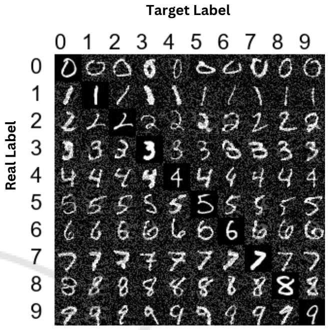

Figure 1: MNIST adversarial examples.

5.2 Use Case 2: Attacking ImageNet Images

Table 2 presents the results of experiments on Ens4AdvInceptionV3 and InceptionV3 models using different attack methods. Figure 2 shows some of the adversarial examples generated using GenGradAttack on the ImageNet dataset against the InceptionV3 model. The presented visual comparison reveals a remarkable observation, as there is a seemingly imperceptible divergence between the images on the left and the images on the right, yet the classifier's classification decision dramatically contrasts the inherent nature of the original image.

GenGradAttack demonstrates exceptional performance, achieving a 96% ASR on Ens4AdvInceptionV3 and a 100% ASR on InceptionV3. Notably, it requires a median query count of **12,623** for Ens4AdvInceptionV3 and **7,254** for InceptionV3, showcasing its efficiency. Comparatively, ZOO requires 3.5M queries with a 6% ASR on Ens4AdvInceptionV3 and 2.6M queries with an 18% ASR on InceptionV3. GenAttack Basic achieves a 93% ASR with 164K queries on Ens4AdvInceptionV3

[1] https://github.com/huanzhang12/ZOO-Attack

[2] https://github.com/nesl/adversarial_genattack

[3] https://github.com/tensorflow/models/tree/archive/research/adv_imagenet_models

Table 2: ASR and the number of queries for different attack methods on Ens4AdvInceptionV3 and InceptionV3 on ImageNet dataset.

	Ens4AdvInceptionV3		InceptionV3	
	Queries	ASR	Queries	ASR
GenGradAttack	**12,623**	**96%**	**7,254**	**100%**
ZOO	3.5M	6%	2.6M	18%
GenAttack Basic	164K	93%	97,493	100%
GenAttack (with adaptive parameter)	21,329	95%	11,201	100%
C&W	-	100%	-	100%

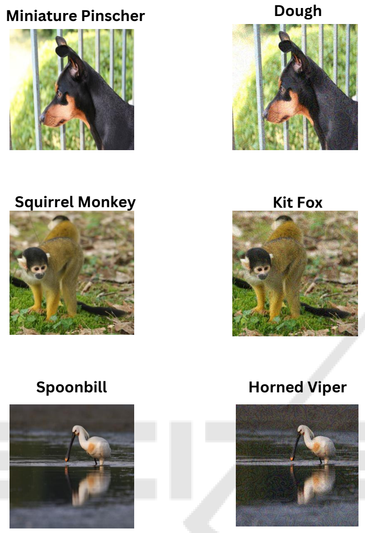

Figure 2: ImageNet adversarial example. Left figure: real label, right figure: target label.

6 DISCUSSION

The discussion below provides a comprehensive analysis of the obtained results from two distinct use cases: one involving the MNIST dataset subjected to $L_\infty = 0.30$ perturbations and the other involving the InceptionV3 and Ens4AdvInceptionV3 models. The results highlight the efficacy and performance of various attack methods, as summarized in Table 1 and Table 2.

6.1 ImageNet Models Use Case

GenGradAttack showcases remarkable query efficiency, but the visual quality and interpretability of the generated adversarial examples are crucial considerations. The attack method aims to transform an input image from its original class to a target class, making the extent of distortion introduced and the number of queries vital factors.

In scenarios where the target class is significantly different from the original class, the resulting adversarial image might exhibit noticeable and disruptive alterations. Figure 3 illustrates an example where an image of a ”German Shepherd” was transformed into an adversarial example classified as a ”Miniature Pinscher.” The adversarial image shows minimal geometric changes, indicating success in achieving a class transformation with minimal visual distortion.

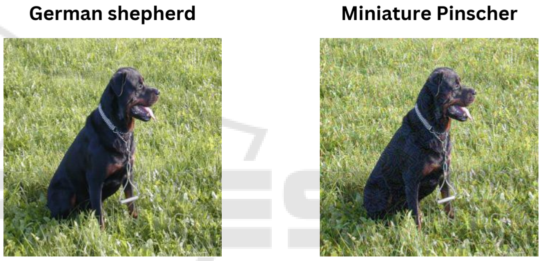

Figure 3: ImageNet Adversarial example with a target class similar to the original class. Left figure: real label, right figure: target label.

Table 3 presents key metrics for this transformation process, highlighting the query count, attack duration, and L_2 distance. The efficiency of GenGradAttack is evident in the low query count, short attack duration, and subtle alterations in the adversarial image.

Table 3: Key Metrics for the attack process on similar classes using ImageNet dataset.

Query Count	**Attack Duration**	**Average L2 Distance**
96.0	366.76 seconds	16.59

On the contrary, figure 4 illustrates an example where an image of a ”Junco” was transformed into an adversarial example classified as a ”Home Theater.” The adversarial image exhibits noticeable geometric changes, emphasizing the challenges in converting images between semantically distant classes.

Table 4 presents key metrics for this transforma-

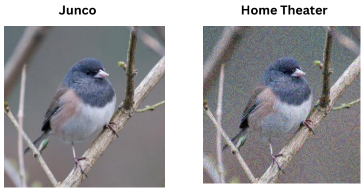

Figure 4: ImageNet Adversarial example with a target class very different from the original class. Left figure: real label, right figure: target label.

tion process, indicating a higher query count, longer attack duration, and a larger L_2 distance. The increased computational effort and visual alterations highlight the difficulty in manipulating predictions across dissimilar classes.

Table 4: Key Metrics for the attack process on different classes using ImageNet dataset.

Query Count	Attack Duration	Average L2 Distance
5986	21359.35 seconds	32.64

These examples illustrate the trade-off between successful class transformation and visual coherency. GenGradAttack achieves its goal of manipulating model predictions, but the generated images may exhibit varying levels of visual distortion that impact real-world interpretability.

6.2 MNIST Dataset Use Case

The MNIST dataset serves as a platform to explore GenGradAttack's subtleties, especially in small-sized images. Results highlight the method's efficiency in small domains, with a low query count, demonstrating its agility in manipulating model predictions for MNIST digits.

However, challenges arise in scenarios where the target class is significantly distinct from the source class. Figure 5 shows an example of transforming the digit ”3” into a ”7,” resulting in visible alterations and challenges in achieving a successful class conversion.

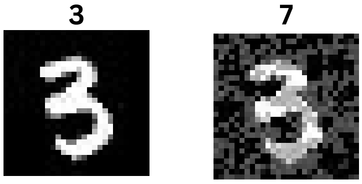

Figure 5: MNIST Adversarial example with a target class very different from the original class. Left figure: real label, right figure: target label.

Table 5 presents key metrics for this transformation process, emphasizing the increased query count and attack duration, as well as a larger L_2 distance. These challenges underscore the impact of semantic gaps between classes, leading to more pronounced visual discrepancies.

Table 5: Key Metrics for the attack process on different classes using MNIST dataset.

Query Count	Attack Duration	Average L2 Distance
42956	359.01 seconds	6.28

In contrast, figure 6 illustrates an example where the digit ”2” was transformed into a ”3” with subtle modifications and minimal visual distortion.

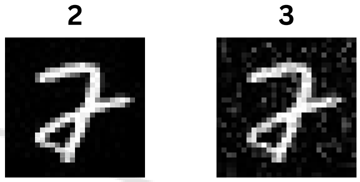

Figure 6: MNIST Adversarial example with a target class similar to the original class. Left figure: real label, right figure: target label.

Table 6 presents key metrics for this transformation process, showcasing the method's efficiency in achieving a seamless conversion between similar classes.

Table 6: Key Metrics for the attack process on similar classes using MNIST dataset.

Query Count	Attack Duration	Average L2 Distance
26	8.10 seconds	2.68

It's crucial to note that MNIST images are considerably smaller, contributing to quicker computation times. The examples underline the trade-off between successful class transformation and visual coherency, emphasizing the varying interpretability of generated adversarial images.

6.3 Hyper-Parameter Selection in Genetic Algorithms

Genetic algorithms are traditionally sensitive to the choice of hyper-parameter values, such as population size, mutation rate, and others. In this section, we discuss the impact of these choices on query efficiency.

6.3.1 Population Size

The population size plays a crucial role in balancing exploration and exploitation within the search space. A larger population size enhances diversity and improves exploration, leading to better search space coverage in fewer iterations. However, this advantage comes with a trade-off, as evaluating each population member incurs a query cost. Figure 7 illustrates this trade-off by showcasing the mean number of queries and iterations until success across various population sizes on a dataset of 20 images. Based on this experiment, we advocate for a relatively small population size of six, striking a balance between convergence speed and the total number of queries expended.

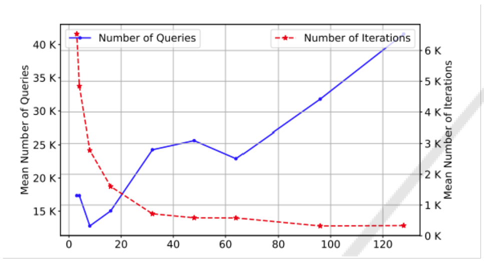

Figure 7: Effect of population size selection on both the speed of convergence and the number of queries.

6.3.2 Mutation Rate

The mutation rate, denoted as ρ, significantly influences algorithm performance. Experimentally, we explored different mutation rate strategies and found that a fixed mutation rate outperformed other approaches. The fixed mutation rate effectively balances exploration and exploitation, contributing to the algorithm's overall success without the need for adaptive adjustments.

7 CONCLUSIONS

In this study, we introduced **GenGradAttack**, a pioneering approach that seamlessly integrates genetic algorithms and gradient-based optimization for black-box adversarial attacks. Our results showcase the impressive efficacy of **GenGradAttack**, achieving notable Adversarial Success Rates (ASR) with reduced query counts. Notably, on the MNIST dataset, we attained a **95.06%** ASR with a median query count of **556**, outperforming conventional GenAttack.

The success of **GenGradAttack** stems from its ability to evolve perturbations that effectively mislead the target model, demonstrating the potency of genetic algorithms in generating adversarial perturbations. Moreover, the combination of gradient-based optimization with genetic algorithms leads to faster convergence, higher ASRs, and query-efficient attacks.

While our achievements are significant, this research lays the groundwork for future exploration. Further analysis, including extensive experimentation and the incorporation of adaptive learning rate strategies, holds the potential to enhance the attack's effectiveness. Delving into factors influencing transferability could yield more universally effective adversarial perturbations.

In summary, our research advances the landscape of adversarial black-box attacks, providing a robust tool for evaluating the vulnerabilities of machine-learning models. We anticipate that this work will inspire continued exploration in the dynamic realm of adversarial attacks and defenses.

ACKNOWLEDGEMENTS

This work was supported, in part, by the Engineering and Physical Sciences Research Council [grant number EP/X036871/1] and Horizon Europe [grant number HORIZON-MISS-2022-CIT-01-01].

REFERENCES

Agnihotri, S. and Keuper, M. (2023). Cospgd: a unified white-box adversarial attack for pixel-wise prediction tasks.

Alzantot, M., Sharma, Y., Chakraborty, S., Zhang, H., Hsieh, C.-J., and Srivastava, M. B. (2019). Genattack: Practical black-box attacks with gradient-free optimization. In *Proceedings of the genetic and evolutionary computation conference*, pages 1111–1119.

Bhandari, D., Murthy, C., and Pal, S. K. (1996). Genetic algorithm with elitist model and its convergence. *International journal of pattern recognition and artificial intelligence*, 10(06):731–747.

Carlini, N. and Wagner, D. (2017). Towards evaluating the robustness of neural networks.

Chen, J., Su, M., Shen, S., Xiong, H., and Zheng, H. (2019). Poba-ga: Perturbation optimized black-box adversarial attacks via genetic algorithm. *Computers & Security*, 85:89–106.

Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C.-J. (2017). 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*, pages 15–26.

Chen, Z., Li, B., Wu, S., Ding, S., and Zhang, W. (2023). Query-efficient decision-based black-box patch attack.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Guo, C., Gardner, J., You, Y., Wilson, A. G., and Weinberger, K. (2019). Simple black-box adversarial attacks. In *International Conference on Machine Learning*, pages 2484–2493. PMLR.

Jere, M., Rossi, L., Hitaj, B., Ciocarlie, G., Boracchi, G., and Koushanfar, F. (2019). Scratch that! an evolution-based adversarial attack against neural networks. *arXiv preprint arXiv:1912.02316*.

Kang, X., Song, B., Guo, J., Qin, H., Du, X., and Guizani, M. (2023). Black-box attacks on image classification model with advantage actor-critic algorithm in latent space. *Information Sciences*, 624:624–638.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. (2017). Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519.

Sun, C., Zhang, Y., Chaoqun, W., Wang, Q., Li, Y., Liu, T., Han, B., and Tian, X. (2022). Towards lightweight black-box attack against deep neural networks. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 19319–19331. Curran Associates, Inc.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. (2020). Ensemble adversarial training: Attacks and defenses.

Wang, J., Yin, Z., Tang, J., Jiang, J., and Luo, B. (2021). Pica: A pixel correlation-based attentional black-box adversarial attack. *arXiv preprint arXiv:2101.07538*.

Wu, C., Luo, W., Zhou, N., Xu, P., and Zhu, T. (2021). Genetic algorithm with multiple fitness functions for generating adversarial examples. In *2021 IEEE Congress on Evolutionary Computation (CEC)*, pages 1792–1799.