Measuring Robustness with Black-Box Adversarial Attack using Reinforcement Learning

Soumyendu Sarkar[†*], **Sajad Mousavi**[†], **Ashwin Ramesh Babu**[†], **Vineet Gundecha**,
Sahand Ghorbanpour, Alexander Shmakov
Hewlett Packard Enterprise
soumyendu.sarkar, ashwin.ramesh-babu, sajad.mousavi, vineet.
gundecha,sahand.ghorbanpour, alexander.shmakov @hpe.com

Abstract

A measure of robustness against naturally occurring distortions is key to the trustworthiness, safety, and success of machine learning models on deployment. We investigate an adversarial black-box attack that adds minimum Gaussian noise distortions to input images to make deep learning models misclassify. We used a Reinforcement Learning (RL) agent as a smart hacker to explore the input images to add minimum distortions to the most sensitive regions to induce misclassification. The agent employs a smart policy also to remove noises introduced earlier, which has less impact on the trained model at a given state. This novel approach is equivalent to doing a deep tree search to add noises without an exhaustive search, leading to faster and optimal convergence. Also, this adversarial attack method effectively measures the robustness of image classification models with the misclassification inducing minimum L_2 distortion of Gaussian noise similar to many naturally occurring distortions. Furthermore, the proposed black-box L_2 adversarial attack tool beats state-of-the-art competitors in terms of the average number of queries by a significant margin with a 100% success rate while maintaining a very competitive L_2 score, despite limiting distortions to Gaussian noise. For the ImageNet dataset, the average number of queries achieved by the proposed method for ResNet-50, Inception-V3, and VGG-16 models are 42%, 32%, and 31% better than the state-of-the-art "Square-Attack" approach while maintaining a competitive L_2. **Demo:** https://tinyurl.com/pzrca5fj

1 Introduction

Research on adversarial attacks has shown Deep learning models suffer from a vulnerability where small distortions could lead to wrong predictions. However, naturally occurring distortions that affect the inputs are of greater concern in safety-critical applications such as self-driving cars, facial recognition, and image-based authorization Joos et al. [2022]Ozdag et al. [2019]. Measuring robustness, is a key to discovering vulnerabilities of poorly trained models.

Even though there are both White box attacks Szegedy et al. [2013]Goodfellow et al. [2014] and Black box attacks Andriushchenko et al. [2020]Su et al. [2019], visibility into the models is not practical in many real-world applications for IP concerns and support issues. On the contrary, black box attacks suffer from inefficiency and require many queries to create the adversarial sample that could break the evaluated model.

The motivation for using RL is for the policy to learn the optimum heuristic based on the data and exploration, unlike the other State-of-the-art adversarial attacks. The main contribution of the work can be summarized as follows.

[*] Corresponding author

[†] These authors contributed equally

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

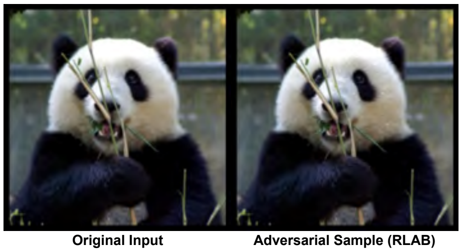


Figure 1: An example of adversarial perturbations driven by the learnt policy of RLAB agent. The image "x" classified as **Panda**, an adversarial sample generated with RLAB(ours) "$x + \delta$" has been classified as **dolphin**.

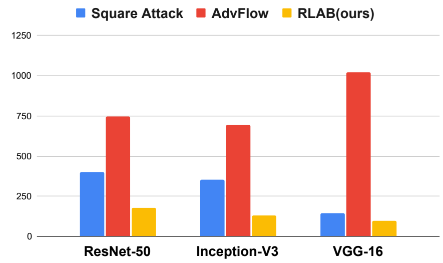


Figure 2: Average number of queries in untargeted L_2-attacks for ImageNet datasets of 3 CNN models for black-box attacks. RLAB outperforms all other attacks by a large margin

1. A novel Reinforcement Learning agent, that makes black-box adversarial attacks faster and beats the state-of-the-art un-targeted black-box L_2 attack models on average number of queries by wide margins with 100% success rate a very competitive L_2-norm.
2. A high-performance adversarial attack agent that limits the distortions to Gaussian noise, which is one of the naturally occurring real-life distortions, unlike most adversarial attacks.

2 Related Works

White-box attacks showed great results with one of the initial works from Goodfellow et. al in their work Goodfellow et al. [2014] introducing Fast Gradient Sign Method (FGSM) based attack. Other incremental works based on gradients-based distortion that could flip the model Kurakin et al. [2016a]Kurakin et al. [2016b]Dong et al. [2018]. DeepFool by moosavi et al. Moosavi-Dezfooli et al. [2016] proposed a simple yet effective approach to add perturbations to the input to fool the models.

In **Black-box attacks**, there is only partial visibility to no visibility into the model. In a partially visible black-box attack, information about the loss function, the prediction probabilities, or top-K sorted labels could be available based on which the attack is executed in a query access approach. Some of the most popular black-box attack in recent times that has been acknowledged by the research community include Square attack Andriushchenko et al. [2020], SimBA Guo et al. [2019a], and LeBA Yang et al. [2020], which achieved significant results in breaking Convolutional Neural Network based models. Other significant works include Guo et al. Guo et al. [2019a], Andriushckenko et al. Andriushchenko et al. [2020], EigenBA Zhou et al. [2022], Pixle Pomponi et al. [2022], Querynet Chen et al. [2021], advFlow Mohaghegh Dolatabadi et al. [2020], and CG attack Feng et al. [2022] producing state-of-the-art results.

RL has solved many complex problems Sarkar et al. [2022a], Sarkar et al. [2022b], Sarkar et al. [2021]. But so far little has been done with **Reinforcement Learning** for adversarial attacks and work done by Yang et al. Yang et al. [2020](Patch Attack) is rudimentary and applies RL to attack CNN models by superimposing unnatural textured patches on the input image and the results were very poor. Unlike the previous approach, our RL agent uses a comprehensive state representation that captures the model's sensitivity to various image regions and implements a patch-based process with natural distortions. This enables our approach to significantly outperform state-of-the-art adversarial attacks, in terms of minimum distortion measured by L2-norm, query efficiency, and success rate.

3 Proposed Reinforcement Learning Method

Our approach **RLAB**, Reinforcement Learning for Adversarial Black-box Attack, as represented in figure 3 is based on selective Gaussian noise distortion to specific fixed-size square patches in the image with the images split into multiple patches of size $n \times n$. The main idea of the algorithm is to first find the sensitivity of the change in classification probability of ground truth P_{GT} with changes in distortion and then use Reinforcement Learning to make the complex decision on selecting the patches to which we add or remove distortion.

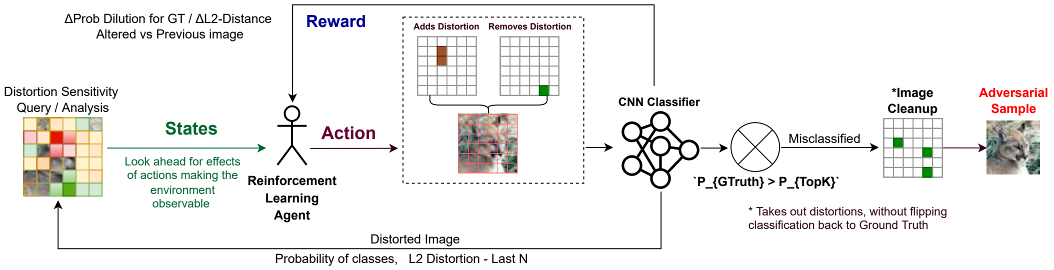


Figure 3: Reinforcement Learning agent for proposed method (RLAB).

In our proposed approach, we limit all distortions to Gaussian noise as its a commonly encountered and naturally occurring distortion. During the image sensitivity analysis, we generated a fixed number "k" of noise masks of size $n \times n$ sampled from a normal distribution as represented in the equation 1.

$$NoiseMask(n \times n) = NormalDistribution(0, Noise\_level) \tag{1}$$

At every step during the training and validation, one mask is randomly chosen from the generated noise masks and applied across all image patches to evaluate the drift in the ground truth classification probability P_{GT}. In our experiments, we got the best results for a $noise\_level = 0.005$. A lower noise level helps more granular addition of noise in successive steps to specific regions that create maximum drift with the P_{GT}, while keeping L_2 to a near minimum.

3.1 Computational efficiency

We further limited the number of mask "k" to 10 for computational efficiency and caching considerations. The impact of choosing different noise masks is the same as the difference in change in L_2 is negligible with one mask over the other.

For RL problems like these and board games the most effective moves or actions are figured out through a Deep Tree Search (DTS). DTS is computationally expensive, and unlike a board game, in this problem, we can always replay the earlier moves when we realize that we have made a less optimized move a few steps back. In RLAB this is done by removing distortions from some patches and adding distortions to some other patches considering the state of the modified image at any given step.

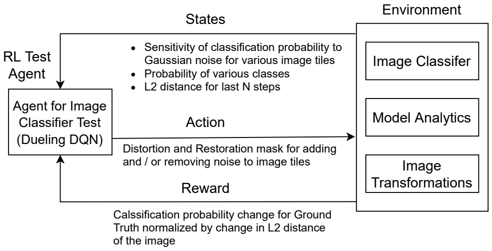
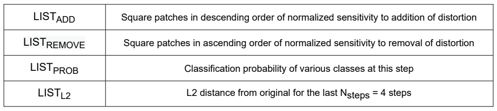

States
Environment
RL Test
Agent
Sensitivity of classification probability to Gaussian noise for various image tiles
Probability of various classes
L2 distance for last N steps
Image Classifer
Agent for Image
Classifier Test
(Dueling DQN)
Action
Model Analytics
Distortion and Restoration mask for adding
and / or removing noise to image tiles
Image
Transformations
Reward
Calssification probability change for Ground
Truth normalized by change in L2 distance
of the image

Figure 4: Reinforcement Learning agent



$LIST_{ADD}$	Square patches in descending order of normalized sensitivity to addition of distortion
$LIST_{REMOVE}$	Square patches in ascending order of normalized sensitivity to removal of distortion
$LIST_{PROB}$	Classification probability of various classes at this step
$LIST_{L2}$	L2 distance from original for the last N_{steps} = 4 steps

Figure 5: RL States

3.2 Reinforcement Learning

The decision of which patches to choose for adding or removing distortion has multiple dependencies and needs to be adaptive for the most efficient generation of adversarial examples. Mapping this adversarial sample generation as a Reinforcement Learning (RL) problem requires defining the states, actions, and rewards. The state-space is constructed such that the environment becomes observable in a way it enables the RL agent to learn the optimum policy to take actions while maximizing the reward. We used the Dueling DQN Reinforcement Learning (RL) based agent in RLAB. Figure 3 represents the overall flow of the proposed method. Figure 4 represents the steps involved in adding and removing distortion by the RL agent.

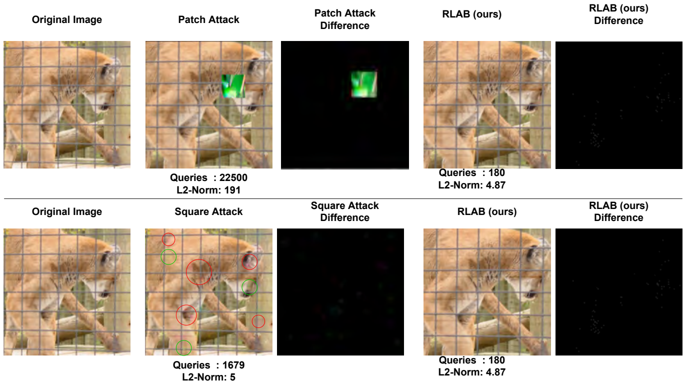


Figure 6: RLAB's distortion comparison with Patch Attack Yang et al. [2020] and SOTA Square Attack Andriushchenko et al. [2020]. **Unlike RLAB other attacks show unnatural color patches**

Table 1: Comparing L_2 and average queries of the proposed method with competitors on the ResNet-50 model trained on IMAGENET dataset.

Attack	**Avg Queries**	**Avg L_2**	**Success Rate**
Boundary Attack Brendel et al. [2017]	5000	24.67	100
Q-Fool Chen et al. [2020]	5000	7.52	-
HopSkipJumpAttack Rahmati et al. [2020]	1000	11.76	-
Bandits Ilyas et al. [2018a]	5251	5	80.5
SimBA-DCT Guo et al. [2019a]	1665	**3.98**	98.6
Square Andriushchenko et al. [2020]	401	5	99.8
EigenBA Zhou et al. [2022]	518	3.6	98
querynet Chen et al. [2021]	-	5	-
RLAB (ours)	**178**	4.03	**100%**

4 Results

We evaluate on image classification datasets ILSVRC2012 Russakovsky et al. [2015]. 80 percent of the validation set was used to train our RL agents, and 20 percent of the validation set was used for evaluation. We performed our attacks on three major Convolution-based Neural Network architectures: ResNet, Inception-V3, and VGG-16. We used **three metrics** to evaluate the performance of our approach. L_2 distance, the average number of queries to make a model miss-classify a correctly classified sample, and the average success rate.

As shown in table 1 Our proposed approach **RLAB outperforms competitor state-of-the-art black-box adversarial attacks for the average number of queries**. For validation, we had an overall average L_2 of 4.03 with the values of pixels ranging between 0 and 1. More results in appendix.

5 Conclusion

RLAB outperforms the state-of-the-art adversarial attacks in query efficiency by a significant margin and achieves a highly competitive L_2-norm indicative of very low distortion with 100% success rate for miss-classification. But as RLAB only uses Gaussian noise, the distortions are similar to real-life deployment. This makes it valuable for a more appropriate test for non-malicious distortions and an effective measure of robustness. Also, Reinforcement Learning was very effective in learning the policy to make the complex decision of choosing the square patches for changing distortion and making RLAB adaptive. This is by far the best RL implementation of this type of Black-Box adversarial attack, and this same approach is applicable for a wide variety of adversarial attack agents beyond image classifiers. This RL design may provide good insights to future work.

We made several simplifications in RLAB to limit the computation and we can relax some of those constraints in future work to explore for even better results. Also, RLAB can generate distortions to the test data set to retrain the model and enhance its robustness.

References

Sander Joos, Tim Van hamme, Davy Preuveneers, and Wouter Joosen. Adversarial robustness is not enough: Practical limitations for securing facial authentication. In *Proceedings of the 2022 ACM on International Workshop on Security and Privacy Analytics*, pages 2–12, 2022.

Mesut Ozdag, Sunny Raj, Steven Lawrence Fernandes, Alvaro Velasquez, Laura Pullum, and Sumit Kumar Jha. On the susceptibility of deep neural networks to natural perturbations. In *AISafety@ IJCAI*, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pages 484–501. Springer, 2020.

Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. Adversarial examples in the physical world, 2016a.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016b.

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*, pages 9185–9193, 2018.

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*, pages 2574–2582, 2016.

Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. Simple black-box adversarial attacks. In *International Conference on Machine Learning*, pages 2484–2493. PMLR, 2019a.

Jiancheng Yang, Yangzhou Jiang, Xiaoyang Huang, Bingbing Ni, and Chenglong Zhao. Learning black-box attackers with transferable priors and query feedback. *Advances in Neural Information Processing Systems*, 33:12288–12299, 2020.

Linjun Zhou, Peng Cui, Xingxuan Zhang, Yinan Jiang, and Shiqiang Yang. Adversarial eigen attack on black-box models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15254–15262, 2022.

Jary Pomponi, Simone Scardapane, and Aurelio Uncini. Pixle: a fast and effective black-box attack based on rearranging pixels. *arXiv preprint arXiv:2202.02236*, 2022.

Sizhe Chen, Zhehao Huang, Qinghua Tao, and Xiaolin Huang. Querynet: Attack by multi-identity surrogates. *arXiv e-prints*, pages arXiv–2105, 2021.

Hadi Mohaghegh Dolatabadi, Sarah Erfani, and Christopher Leckie. Advflow: Inconspicuous black-box adversarial attacks using normalizing flows. *Advances in Neural Information Processing Systems*, 33:15871–15884, 2020.

Yan Feng, Baoyuan Wu, Yanbo Fan, Li Liu, Zhifeng Li, and Shu-Tao Xia. Boosting black-box attack with partially transferred conditional adversarial distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15095–15104, 2022.

Soumyendu Sarkar, Vineet Gundecha, Alexander Shmakov, Sahand Ghorbanpour, Ashwin Ramesh Babu, Paolo Faraboschi, Mathieu Cocho, Alexandre Pichard, and Jonathan Fievez. Multi-agent reinforcement learning controller to maximize energy efficiency for multi-generator industrial wave energy converter. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11): 12135–12144, 2022a.

Soumyendu Sarkar, Vineet Gundecha, Sahand Ghorbanpour, Alexander Shmakov, Ashwin Ramesh Babu, Alexandre Pichard, and Mathieu Cocho. Skip training for multi-agent reinforcement learning controller for industrial wave energy converters. pages 212–219, 2022b.

Soumyendu Sarkar, Vineet Gundecha, Alexander Shmakov, Sahand Ghorbanpour, Ashwin Ramesh Babu, Paolo Faraboschi, Mathieu Cocho, Alexandre Pichard, and Jonathan Fievez. Multi-objective reinforcement learning controller for multi-generator industrial wave energy converter. In *NeurIPs Tackling Climate Change with Machine Learning Workshop*, 2021.

Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017.

Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)*, pages 1277–1294. IEEE, 2020.

Ali Rahmati, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Huaiyu Dai. Geoda: a geometric framework for black-box adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8446–8455, 2020.

Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. *arXiv preprint arXiv:1807.07978*, 2018a.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning*, pages 2137–2146. PMLR, 2018b.

Yiwen Guo, Ziang Yan, and Changshui Zhang. Subspace attack: Exploiting promising subspaces for query-efficient black-box attacks. *Advances in Neural Information Processing Systems*, 32, 2019b.

Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. *Advances in neural information processing systems*, 32, 2019.

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019.

A Appendix

A.1 Results

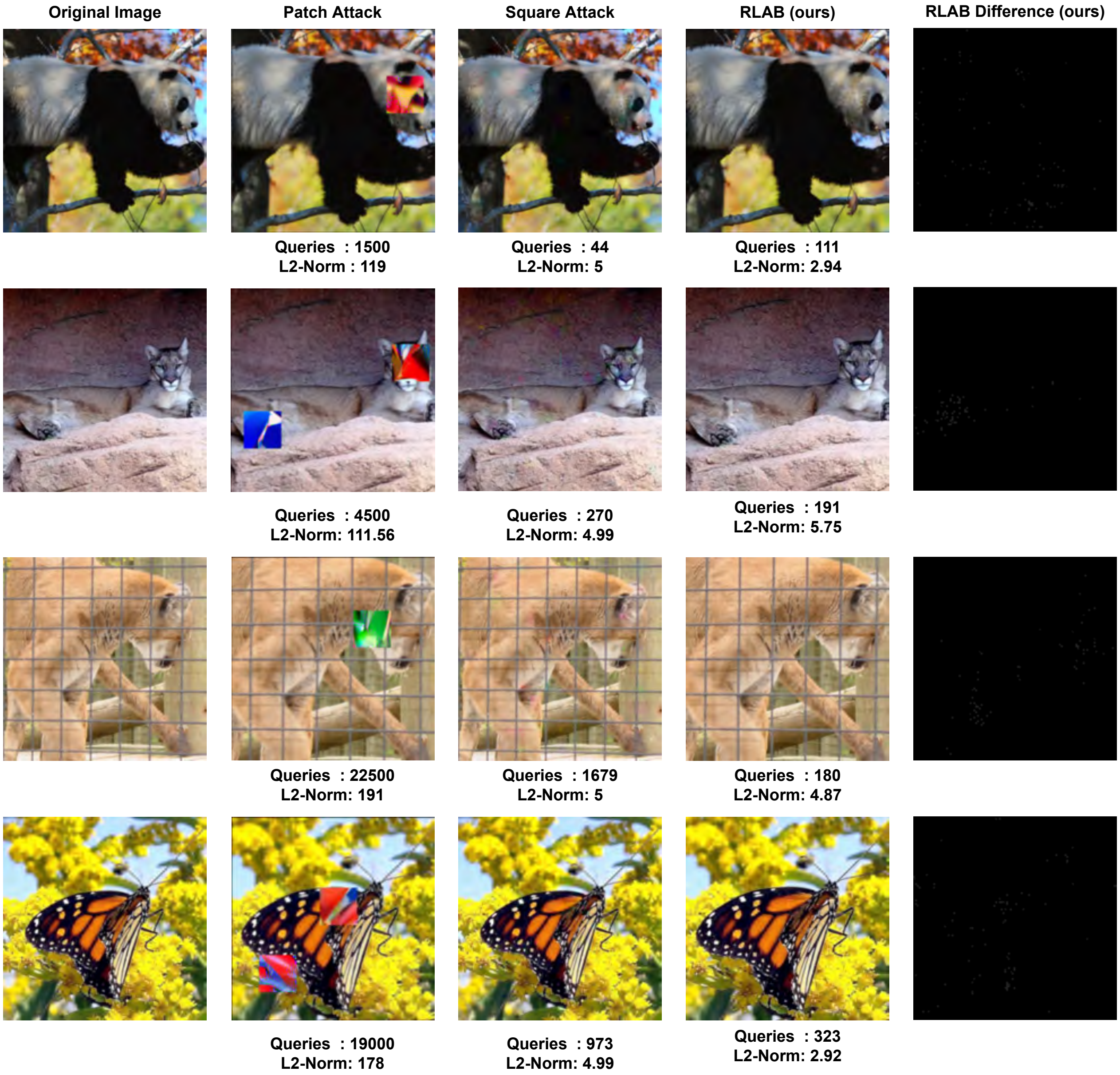


Figure 7: Comparing performance of our approach with competitors

A.1.1 Evaluation on Imagenet

Table 2 aggregates the results of the proposed method compared to other state-of-the-art black-box algorithms. The competitors' results were generated with the best parameters described in their paper. Average Success Rate (ASR) and Average Query (AVG.Q) are calculated for each victim model. It can be observed from the results that, for the Average Success Rate, our proposed approach achieves 100 percent accuracy even for the maximum query value set to 3500. At the same time, our competitors have experimented with a maximum query set to 10000. Similarly, our proposed approach outperforms competitors for Inception-v3 and ResNet-50 for the average number of queries. Furthermore, Table 1 shows the comparison of different L_2 values for competitor approaches that published their L_2 values. It can be observed that our proposed approach achieves competitive L_2 values while having an extremely low average query rate and 100 % success rate. Figure 7 shows the comparison of RLAB with the competition RL method and Square attack.

Table 2: Performance comparison of our approach with State-of-the-art methods. Average number of queries (AVG.Q) and Success Rate (ASR) evaluated on victim models including ResNet-50, Inception-V3 and VGG-16 with ImageNet dataset. All performance is reported using the official codes, under L_2 norm and maximum queries of 10,000 except of the proposed method with a maximum queries of 3500.

Method	ResNet-50		Inception-v3		VGG-16	
	ASR %	AVG.Q	ASR %	AVG.Q	ASR %	AVG.Q
NES (2018) Ilyas et al. [2018b]	82.7	1632.4	88.2	1726.2	84.8	1119.6
Bandits$_{TD}$ (2018) Ilyas et al. [2018a]	93.0	765.3	97.7	836.1	91.1	275.9
Subspace (2019) Guo et al. [2019b]	94.4	1078.7	96.6	1035.8	96.2	1085.8
P-RGF$_D$ (2019) Cheng et al. [2019]	99.3	270.5	99	637.4	99.8	393.1
TIMI (2019) Dong et al. [2019]	68.6	-	49	-	51.3	-
LeBA Yang et al. [2020]	99.9	178.7	99.4	243.8	99.9	145.5
SimBA (2019) Guo et al. [2019a]	98.6	873.9	99.2	874.5	99.9	423.3
Square Attack (2020) Andriushchenko et al. [2020]	99.8	401	99.4	351.9	**100**	142.3
querynet (2021) Chen et al. [2021]	-	-	-	518	-	-
AdvFlow (2021) Mohaghegh Dolatabadi et al. [2020]	96.7	746	99.3	694	95.5	1022
EigenBA (2022) Zhou et al. [2022]	98.6	518	95.7	968	-	-
Pixle (2022) Pomponi et al. [2022]	98	341	-	-	99	519
CG-Attack (2022) Feng et al. [2022]	97.3	210	**100**	139	99.4	**77**
Patch Attack Yang et al. [2020]	-	983	-	-	-	-
RLAB(ours)	**100**	**178**	**100**	**132**	**100**	98