Author Response

A. Responses to Reviewer #2

Question 1: How does the method perform against other API services? Are there limitations to applying your method to other API services?

For Google Cloud Vision and Imagga API services, we only measured the average query time in Table 1 (Line 182), but did not perform a benchmark because it would cost $1800 to run our experiments against Google Cloud Vision and Imagga. As a result, we designed DeepAPI to reduce the cost, and DeepAPI supports the same models used in prior research. Using our DeepAPI deployed on Azure costs less than $300 for all experiments in this paper.

The querying cost is one of the reasons why prior research tested black-box attacks against local models and thus overestimated the success rate of their methods.

Except the query budget, there is no limitation to apply concurrent queries to other API services since load balancing has become a fundamental component for cloud services to reduce latency (supported by major cloud service providers, AWS, Azure, Google Cloud, etc). Besides, the query limit for each user is 1,800 requests per minute (Google Cloud Vision), and our distributed black-box attacks (1,068 requests/min) did not reach the limit.

Question 2: Is there any performance difference when attacking a image in two distributed settings?

For a single image, only vertical distribution can accelerate the attack because horizontal distribution aims to accelerate attacking a batch of images. Thus, horizontal distribution reduces to a non-distributed attack for a single image.

Question 3: How were the 100 images in the experiments sampled? Is there any bias in the sampling process?

We used the FiftyOne ImageNet Sample Dataset that contains 1,000 images, one randomly chosen from each class of the validation split of the ImageNet 2012 dataset.

For the benchmark, we only tested 100 images, one randomly chosen from each class of the 1000 classes because our experiments on DeepAPI took 120 hours for 100 images in total (see Fig. 6). A benchmark on 1,000 images could take over 50 days depending on network connectivity.

Question 4: What is the attack performance like in the targeted setting?

Since the original paper (Square [1] and Bandits [2] Attack) did not provide experimental results on targeted settings, we only tested untargeted distributed attacks to compare with prior research.

B. Responses to Reviewer #3

Question 1: Is the meaning of the symbol in line 222 the same as that in line 208?

The symbol δ in line 208 represents the final perturbation added to the input image (see line 212), while the δ in line 222 represents the exploration step for gradient estimation.

We used the same notation as the original papers to make it easier to compare our distributed attacks with non-distributed attacks. We hope our first step towards online black-box attacks could bring black-box attacks closer to being a practical threat and facilitate more future research on this more realistic scenario.

C. Responses to Reviewer #1

Question 1: This paper seems to violate the anonymous author guidelines: Demo youtube link in the abstract which is not anonymous.

The reviewer refers to a Youtube link in the abstract. However, no such Youtube link exists in the paper.

D. Responses to Reviewer #5

Question 1: This paper provides a new framework in the area of Blackbox Adversarial Attack research that considers the time it takes to launch an attack. The paper's focus is realistic and interesting. However, the paper fails to anonymize the code repository, which violates the submission policy of BMVC 2023 since it is possible to identify the authors.

We included the code repository because the open-source cloud service and the new framework for online black-box attacks are two contributions of this paper.

PyPI does not allow anonymous packages in order to prevent abuse and ensures that package authors are accountable for the code they distribute. Thus, we used our authenticated account for the source code.

Our intention was to provide access to the code and demonstrate our contributions to the open-source community to make it easier to benchmark online black-box attacks for future research.

References

[1] 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 (ECCV)*, pages 484–501, 2020. 1

[2] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. *arXiv preprint arXiv:1807.07978*, 2018. 1