Rebuttal for 3893: Black-Box Ripper: Copying black-box models using generative evolutionary algorithms

We thank reviewers for their useful comments and insights. The reviewers appreciated our idea as intuitive and original (R1, R3), our method as easy to grasp (R1), significant and highly relevant to NeurIPS (R3), our results as showing the superiority of our method (R1, R3) and our paper as well-written and clear (R1, R2, R3). Furthermore, R3 appreciated our nice review of existing approaches. We next address concerns raised by the reviewers.

R1: The evolutionary algorithm might collapse to certain regions of the output space. **Answer:** We have analyzed the pixel-level variability of 1000 random images generated by GAN and 1000 random images optimized by our evolutionary strategy, the resulting means of the normalized pixel-level standard deviations being 0.26 and 0.27, respectively. This demonstrates that the proposed evolutionary strategy does not collapse to certain regions.

R1: Did the authors look at mode collapse of GAN or blurry effect of VAE? **Answer:** The GANs used in our experiments did not suffer from mode collapse. However, for the second set of experiments, the considered VAE does indeed output blurry images. Nonetheless, aspects such as the quality of the images generated by GAN or VAE are not relevant to our approach, as long as we are able to achieve state-of-the-art performance levels in stealing the functionality of black-box models, as shown in Tables 1, 2, 3 from our paper.

R2: The experimental parts are weak, only small data sets being employed. **Answer:** First, we considered the same data sets as Addepalli et al. [1], our main competitor, their paper presenting results on such data sets being published at AAAI 2020. Aside from the experimental setup used by Addepalli et al. [1], in which the images have low resolution, we have performed additional experiments using larger images from CelebA-HQ, ImageNet Cats and Dogs and 10 Monkey Species. In the experiments presented in Table 3 (see paper), the teacher and student models take images of 224×224 pixels as input. We thus believe that our experiments cover a wider range than those of Addepalli et al. [1].

R2: Concerns about the proposed method in practical applications with many classes. **Answer:** The experiments performed on higher-resolution image shows that our method works well when confronted with a large multi-class latent space, considering that ImageNet Cats and Dogs contains 143 classes. Regarding our results on 90 and 40 classes from CIFAR-100, we note that our accuracy rates are comparable to those reported by Addepalli et al. [1]. However, our method is significantly more general, as their method only works in white-box scenarios, requiring complete access to the network to perform back-prop. Our method works in a black-box setup, requiring no knowledge of the internal structure or parameters of the model.

R2: Is the GAN used in the paper same as in the baseline [1]? **Answer:** The GANs have equivalent architectures, but the models are not identical, mainly because Addepalli et al. [1] train the GAN along with the student, whereas we only use a pre-trained GAN.

R3: The main weakness is the lack statistics over several independent runs. **Answer:** We note that Addepalli et al. [1] typically reported results of independent runs, with one exception. They performed 5 runs for a single scenario, in which 10 classes from CIFAR-100 are used as proxy. We hereby present results for 5 runs of our approach for the same case in Table 1, demonstrating the stability and superiority of our approach over that of Addepalli et al. [1]. Upon acceptance, we will include mean and standard deviations over 5 runs for all our experiments in the final paper.

Table 1: Accuracy rates (in %) for 5 runs on CIFAR-10 as true data set and 10 classes of CIFAR-100 as proxy.

Method	Run 1	Run 2	Run 3	Run 4	Run 5	Mean ± Std. Dev.
DeGAN [1]	66.9	74.6	72.6	76.6	71.6	72.6 ± 3.26
Black-Box Ripper (Ours)	78.2	78.0	77.4	77.9	78.2	77.9 ± 0.29

R3: Wouldn't it help to perform adaptive mutations? **Answer:** In a set of preliminary experiments, we tried using CMA-ES, without observing any performance improvements. Hence, we decided to stick to the most straight forward method that already outperforms existing approaches.

R3: Authors only evolve for 10 generations. **Answer:** In a set of preliminary experiments, we tested with up to 50 generations, but we observed that the evolved exemplars typically converge in 10 generations or even less. To reduce the computational time, we decided to keep only 10 generations, which seems sufficient (see Tables 1, 2, 3 in paper).

R5: The detailed description of the evolutionary algorithm used in this paper is not clear, such as what kind of coding method is used. **Answer:** The coding method is straight forward: our data samples are input noise vectors from the latent space of the generator. Upon acceptance, we will strive to further improve the clarity of our presentation.

R5: Why do you choose this kind of evolutionary algorithm? **Answer:** We tried the simplest evolutionary algorithm, which already seems capable of surpassing glass-box approaches. Further tweaking the evolutionary algorithm can only bring improvements, putting an even greater gap between our results and those of the state-of-the-art methods.

R5: Paper is not innovative enough. **Answer:** Model stealing has never been studied or proven possible with evolutionary algorithms. We show that our evolutionary strategy surpasses other, more relaxed, glass-box state-of-the-art methods. Furthermore, the originality of our idea is appreciated by R1, R3.

[1] Addepalli et al., DeGAN: Data-Enriching GAN for Retrieving Representative Samples from a Trained Classifier. AAAI 2020.