CARBEN: Composite Adversarial Robustness Benchmark

Lei Hsiung[1*], **Yun-Yun Tsai**[2], **Pin-Yu Chen**[3] and **Tsung-Yi Ho**[1]
[1]National Tsing Hua University
[2]Columbia University
[3]IBM Research
hsiung@m109.nthu.edu.tw, yt2781@columbia.edu, pin-yu.chen@ibm.com, tyho@cs.nthu.edu.tw

Abstract

Prior literature on adversarial attack methods has mainly focused on attacking with and defending against a single threat model, e.g., perturbations bounded in Lp ball. However, multiple threat models can be combined into composite perturbations. One such approach, composite adversarial attack (CAA), not only expands the perturbable space of the image, but also may be overlooked by current modes of robustness evaluation. This paper demonstrates how CAA's attack order affects the resulting image, and provides real-time inferences of different models, which will facilitate users' configuration of the parameters of the attack level and their rapid evaluation of model prediction. A leaderboard to benchmark adversarial robustness against CAA is also introduced.

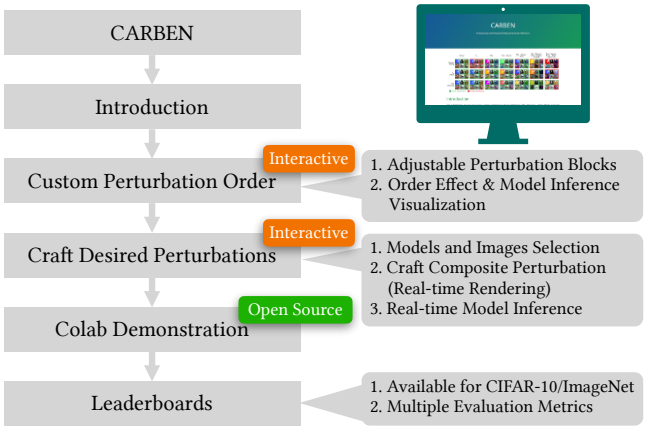


Figure 1: CARBEN overview. Browse on: hsiung.cc/CARBEN

1 Background

Deep neural networks (DNNs) have transformed computer vision and have been used in many fields, including aviation, climate forecasting, and medicine, among others. However, when a DNN encounters carefully crafted images, it can exhibit various vulnerabilities: for instance, lack of robustness in the face of *adversarial attack* [Szegedy *et al.*, 2013]. By utilizing adversarial attacks to optimize perturbations, one can intentionally derive a perturbed sample from a normal image, and make it imperceptible to human beings. In other words, despite the original and adversarial images looking very much alike to us, the latter can lead well-trained DNN models to make wrong predictions.

Previous studies have sought to create bounded perturbations in a metric manner [Goodfellow *et al.*, 2015; Chen *et al.*, 2018]. Most such work has focused on ℓ_p-norm perturbation (i.e., ℓ_1, ℓ_2, and ℓ_∞) and utilized gradient-based optimization – i.e., fast gradient sign method (FGSM), projected gradient descent (PGD), or C&W – to effectively generate the adversarial example. However, it is possible to extend adversarial perturbations beyond the ℓ_p-norm bounds. For instance, Laidlaw et al. generated their perturbation in a perceptual distance metric [Laidlaw *et al.*, 2021], and Hosseini and Poovendran proposed a method of modifying images semantically to create semantic adversarial examples [Hosseini and Poovendran, 2018]. Additionally, Mao et al. studied adversarial examples generated from multiple-threat models, and demonstrated that they were more effective than single-threat ones against DNN targets [Mao *et al.*, 2021].

Recently, Tsai et al. combined the ℓ_∞-norm and semantic perturbations (i.e., hue, saturation, rotation, brightness, and contrast), and proposed a novel approach – composite adversarial attack (CAA) – capable of generating unified adversarial examples [Tsai *et al.*, 2022]. The main differences between CAA and previously proposed perturbations are a) that CAA incorporates several threat models simultaneously, and b) that CAA's adversarial examples are semantically similar and/or natural-looking, but nevertheless result in large differences in ℓ_p-norm measures.

Various defense strategies have also recently been proposed. For example, adversarial training (AT) is one of the most efficient ways to defend against adversarial attacks. However, recent results showed that ℓ_∞-robust models (e.g., [Madry *et al.*, 2018]) might become fragile when they encounter composite perturbations [Tsai *et al.*, 2022]. Motivated by this limitation in ℓ_∞-centric robustness, generalized adversarial training (GAT) overcomes this weakness and

*Contact Author

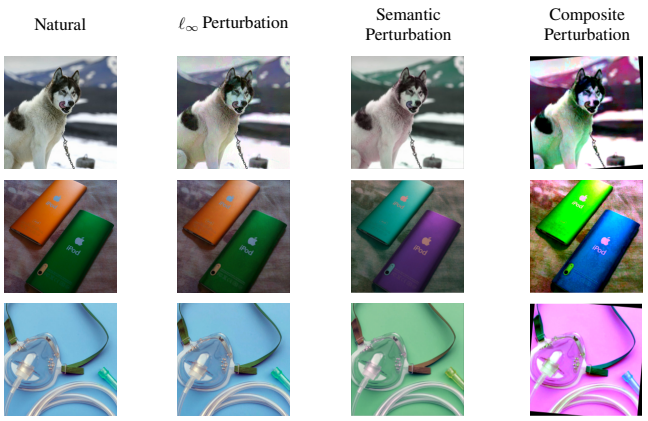
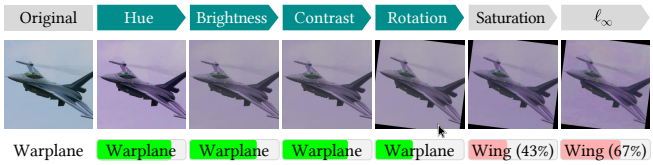
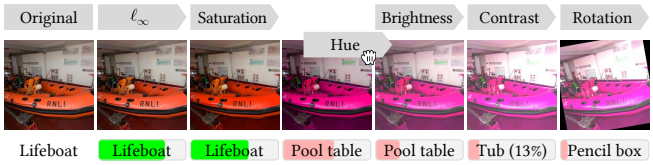


Figure 2: Examples of different perturbation types



(a) Perturbation block highlight. If the cursor hovers over the image below, the perturbation blocks are highlighted, indicating which perturbations are included in the image.



(b) Adjustment of perturbation order. Users can change the perturbation order by exchanging perturbation blocks.

Figure 3: Perturbed images, with each row representing a series of consecutive perturbations in a different specified attack order. Several samples and the inference results from an ℓ_∞-robust model are presented, and the confidence bar is marked in green (red) if the prediction is correct (incorrect).

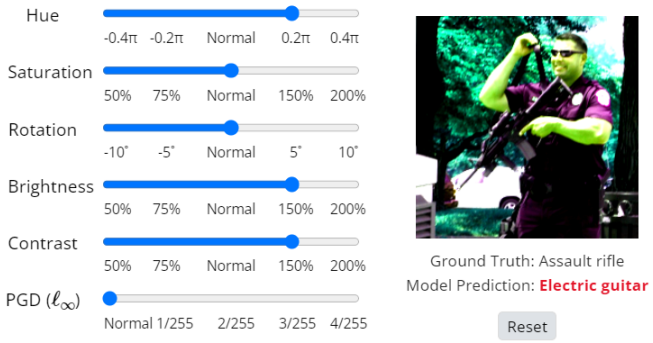


Figure 4: CARBEN's interactive perturbation panel. Users can change the attack level and render the perturbed image on the canvas at right. Real-time model inference about the resulting instance is also provided, and marked in green (red) if it gives the correct (incorrect) prediction. The model used here is naturally trained model from *torchvision* package.

shows the robustness against a variety of composite perturbations [Tsai *et al.*, 2022].

To systematically track the progress of adversarial robustness, [Croce *et al.*, 2021] created a leaderboard and benchmarks for 120+ state-of-the-art models' performance on an image-classification task conducted under ℓ_p-threat and common corruptions. However, their approach did not cover evaluation of robustness against semantic attacks or composite perturbations, and these absences could potentially have led to bias in their interpretation and ranking of DNN models. To bridge this gap, familiarize other researchers with the concept of composite adversarial robustness, and ultimately, create more trustworthy AI, we developed a browser-based composite perturbation generation demo, CARBEN (composite adversarial robustness benchmark).

As shown in Figure 1, CARBEN is a web application featuring *interactive* and *real-time* perturbation panels in which its users render the perturbed image and can see the models' predictions in real time. Figure 2 presents examples of several perturbation types. CARBEN also includes an interactive section that lets users generate images in any attack combination. We have also created a leaderboard that tracks the accuracy of robustness against CAA.

2 System Design

The main purpose of CARBEN is to visualize the mechanism of generating composite perturbations, and thereby help its users to understand how models change their predictions or confidences when under adversarial attack. In this demonstration, users can adjust the attack parameters manually, to explore the effects such adjustments will have, and gain valuable hands-on experience of generating desired examples. Once users understand the concept of CAA and attack ordering, they can use our Google Colab demonstration and the automated CAA to generate a set of composite adversarial examples for robustness evaluation, and report their model-performance results on our leaderboard.

2.1 Composite Perturbations in Custom Order

To demonstrate how attack order can affect the outcomes of composite perturbations, we designed CARBEN to allow its users to explore order effects. Figure 3 shows some perturbation examples in different orders of CAA. We pre-generated the images with all attack combinations and orders, and one can move the arrow-shaped attack blocks to the desired position to see the images in that selected order. Furthermore, we also presented the highest confidence scores for each instance, measured with an ℓ_∞-robust model. This visualization also illustrates why previous models are more susceptible to semantic perturbations and lead to erroneous predictions when it comes to semantic and compositional scenarios.

2.2 Perturbation Panel

Because each attack component has its own perturbation level (or attack parameter), CARBEN also includes a perturbation panel, as shown in Figure 4, in which users can create composite perturbations on images. More specifically, a user can

slide the bar to specify the attack level, and if it is set to *Normal*, the attack will be disabled. However, it should be noted that for purposes of this CARBEN feature, the attack order is fixed as: $\ell_\infty \rightarrow$ Hue $\rightarrow$ Saturation $\rightarrow$ Brightness $\rightarrow$ Contrast $\rightarrow$ Rotation.

In addition, we provide several ImageNet examples and real-time model prediction on the user-generated image, including a standard training model, ℓ_∞-robust model, and GAT. Potentially, CARBEN could also be extended to support uploading images from user's phone or computer.

2.3 Colab Demonstration and Leaderboard

Colab Demonstration. The opensource code for generating a composite adversarial example with an automatically optimized attack order is currently available on GitHub.[1] We provide a step-by-step tutorial on Colab to guide users on how to execute this CAA in the notebook and see the results. The CIFAR-10 dataset was used for the experiment in our Colab notebook. Because CAA is implemented based on gradient optimization, computational cost and computing time will increase significantly as the number of enabled attacks increases. Therefore, we recommend using NVIDIA 2080Ti or above GPU to get better efficiency.

Benchmark and Leaderboard. When we compared the robustness rankings of the top 10 models on RobustBench (CIFAR-10, ℓ_∞) [Croce *et al.*, 2021], the results show that rankings between Auto-Attack and CAA (Full attacks) have a low correlation, suggesting that only considering perturbations in ℓ_p ball for robustness evaluation is biased and incomplete. Statistically, the Spearman's rank correlation coefficients between Auto-Attack and CAA (Semantic attacks and Full attacks) are as follows: 0.16 for semantic attacks, and 0.38 for full attacks.

To provide a complete robustness evaluation, we sought to offer several metrics for measuring the model performance. Therefore, we maintain leaderboards to facilitate tracking and benchmarking adversarial robustness progress across literature. Inspired from RobustBench, CARBEN's leaderboard focuses on tracking the robustness of model's robust accuracy not merely AutoAttack but also two CAAs (Semantic/Full attacks). Figure 5 shows the top three models on our leaderboard, as evaluated using CIFAR-10 and ImageNet datasets. In our leaderboard, models are ranked according to their *Full Attacks robust accuracy*; their architectures and papers are also listed.

In this leaderboard, we focus on "white-box" scenarios in which the attacker has all knowledge of the models. We have provided similar entries to those in the RobustBench leaderboard, and hereby solicit model submissions to compete against composite perturbations in our leaderboard.

Rank	Method	Clean	AutoAttack	Semantic Attacks	Full Attacks	Arch.
1	Towards Compositional Adversarial Robustness: Generalizing Adversarial Training to Composite Semantic Perturbations	85.37%	41.92%	70.33%	21.70%	WideResNet 34-10
2	Improving Robustness using Generated Data	85.64%	56.85%	22.21%	6.56%	WideResNet 34-20
3	Improving Robustness using Generated Data *It uses additional 100M synthetic images in training.*	88.74%	66.10%	17.37%	4.88%	WideResNet 70-16

(a) CIFAR-10

Rank	Method	Clean	AutoAttack	Semantic Attacks	Full Attacks	Arch.
1	Towards Compositional Adversarial Robustness: Generalizing Adversarial Training to Composite Semantic Perturbations	59.96%	20.94%	36.21%	11.65%	ResNet 50
2	Towards Deep Learning Models Resistant to Adversarial Attacks *Robustness library*	62.42%	28.94%	8.95%	3.13%	ResNet 50
3	Do Adversarially Robust ImageNet Models Transfer Better?	68.41%	38.14%	9.82%	1.26%	WideResNet 50-2

(b) ImageNet

Figure 5: The current top three entries on our leaderboard for model accuracy and robustness evaluation

3 Potential Impacts

CARBEN is believed to be the first demo aimed at explaining composite perturbations, and demonstrates the effects on model prediction of altering the attack order. As such, it can help its users gain valuable insights into composite adversarial robustness, and thereby accelerate research devoted to the design of robust models.

Looking beyond conventional ℓ_p-norm perturbations, our demo focuses on semantic adversarial attacks and composite perturbations, which are closer to real-life scenarios demanding robustness. For example, a brightness attack may happen during adjustments to a camera's aperture, and a hue-and-saturation attack may occur when the camera lens is fitted with a filter or mask. Our CARBEN demo and leaderboard offer unprecedentedly realistic and comprehensive robustness assessment. Our web-based demonstration and hands-on interaction can also help to build awareness of AI and robustness, and thus serve as an education toolkit for trustworthiness in AI.

4 Conclusion

This demonstration enables CARBEN users to gain familiarity with CAAs. Its design features interactive sections, and provides Colab tutorials for both manual and automated composite-attack generation. Our robustness evaluation and leaderboard are believed to be the first attempts to benchmark model performance against complex and realistic threats beyond small-norm and single-type perturbations.

Acknowledgments

This work is supported by Ministry of Science and Technology, Taiwan (MOST 111-2218-E-005-006-MBK).

[1] https://github.com/twweeb/composite-adv

References

[Chen *et al.*, 2018] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Ead: Elastic-net attacks to deep neural networks via adversarial examples. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.

[Croce *et al.*, 2021] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[Goodfellow *et al.*, 2015] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[Hosseini and Poovendran, 2018] Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018.

[Laidlaw *et al.*, 2021] Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. In *International Conference on Learning Representations*, 2021.

[Madry *et al.*, 2018] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[Mao *et al.*, 2021] Xiaofeng Mao, Yuefeng Chen, Shuhui Wang, Hang Su, Yuan He, and Hui Xue. Composite adversarial attacks. *Association for the Advancement of Artificial Intelligence (AAAI)*, 2021.

[Szegedy *et al.*, 2013] 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.

[Tsai *et al.*, 2022] Yun-Yun Tsai, Lei Hsiung, Pin-Yu Chen, and Tsung-Yi Ho. Towards compositional adversarial robustness: Generalizing adversarial training to composite semantic perturbations. *arXiv preprint arXiv:2202.04235*, 2022.