

1 Additional Related Works

2 1.1 Image-based Adversarial Attacks

3 Image-based adversarial attacks have been extensively studied, focusing primarily on neural network
4 classifiers. Prominent attack methods include the Fast Gradient Sign Method (FGSM) attack Goodfel-
5 low et al. [2015], which perturbs the pixels in the direction of the gradient’s sign; Projected Gradient
6 Descent (PGD) attack Madry et al. [2019], which iteratively optimizes adversarial perturbations using
7 gradient information; Square Attack [Andriushchenko et al., 2020], a query-efficient gradient-free
8 approach; and Simultaneous Perturbation Stochastic Approximation (SPSA) Uesato et al. [2018],
9 which estimates gradients through randomized sampling.

10 Beyond these, a diverse array of attacks has been developed, targeting various threat models and
11 settings. Optimization-based white-box attacks include l_1 -APGD [Croce and Hein, 2021], which
12 adapts projected gradient descent for l_1 -norm constraints with adaptive step sizes, and the Carlini
13 & Wagner (C&W) attack [Carlini and Wagner, 2016], which formulates adversarial example gener-
14 ation as an optimization problem under various norms. DeepFool [Moosavi-Dezfooli et al., 2016]
15 iteratively approximates the decision boundary to find minimal perturbations. Universal Adversarial
16 Perturbations [Moosavi-Dezfooli et al., 2017] generate image-agnostic perturbations effective across
17 multiple inputs and models.

18 Physical and localized attacks such as the Adversarial Patch [Brown et al., 2017] create robust,
19 universal patches that induce misclassification under real-world conditions. Auto-PGD [Croce and
20 Hein, 2020] automates step-size selection in PGD for improved convergence and attack success.

21 Numerous black-box attacks have also emerged. Bandits [Ilyas et al., 2018b] and NES [Ilyas et al.,
22 2018a] leverage gradient estimation and evolutionary strategies, respectively, for query-efficient
23 adversarial example generation. Boundary Attack [Brendel et al., 2018] is a decision-based method
24 that refines large initial perturbations to minimize their magnitude. Other approaches include One
25 Pixel Attack [Su et al., 2019] (differential evolution on a single pixel), ZOO [Chen et al., 2017] (zeroth-
26 order coordinate-wise gradient estimation), GenAttack [Alzantot et al., 2018] (genetic algorithms),
27 Parsimonious Black-Box Attack [Tashiro et al., 2019], NATTACK [Li et al., 2019], and Saliency
28 Attack [Li et al., 2022], which focus on query efficiency, minimal perturbation, or targeting salient
29 regions.

30 These approaches typically aim to induce misclassification or alter model output minimally and
31 undetectably. Our work expands upon these methodologies by leveraging semantic manipulation
32 through text-guided diffusion models, aiming to influence model decisions at a deeper semantic level.

References

- Moustafa Alzantot, Yash Sharma, Supriyo Chakraborty, Huan Zhang, Cho-Jui Hsieh, and Mani Srivastava. Genattack: Practical black-box attacks with gradient-free optimization. *arXiv preprint arXiv:1805.11090*, 2018. URL <https://arxiv.org/abs/1805.11090>.
- Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search, 2020. URL <https://arxiv.org/abs/1912.00049>.
- Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *International Conference on Learning Representations*, 2018. URL <https://arxiv.org/abs/1712.04248>.
- Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017. URL <https://arxiv.org/abs/1712.09665>.
- Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. *arXiv preprint arXiv:1608.04644*, 2016. URL <https://arxiv.org/abs/1608.04644>.
- Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. 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, 2017. URL <https://arxiv.org/abs/1708.03999>.
- Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *arXiv preprint arXiv:2003.01690*, 2020. URL <https://arxiv.org/abs/2003.01690>.
- Francesco Croce and Matthias Hein. On adaptive attacks and robust training for ℓ_1 -adversarial robustness. *arXiv preprint arXiv:2103.01208*, 2021. URL <https://arxiv.org/abs/2103.01208>.
- Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. URL <https://arxiv.org/abs/1412.6572>.
- Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. *arXiv preprint arXiv:1804.08598*, 2018a. URL <https://arxiv.org/abs/1804.08598>.
- Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Bandits for black-box adversarial attacks. *arXiv preprint arXiv:1807.07978*, 2018b. URL <https://arxiv.org/abs/1807.07978>.
- Xinyun Li, Yinpeng Li, Xiaoliang Wang, Baoyuan Li, and Yisen Wang. Saliency-based black-box adversarial attacks via regionally masked perturbations. *arXiv preprint arXiv:2206.01898*, 2022. URL <https://arxiv.org/abs/2206.01898>.
- Yinpeng Li, Xiaoliang Wang, Baoyuan Li, Tong Zhang, and Boqing Gong. Nattack: Learning the distributions of adversarial examples for black-box attacks. *arXiv preprint arXiv:1905.00441*, 2019. URL <https://arxiv.org/abs/1905.00441>.
- Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019. URL <https://arxiv.org/abs/1706.06083>.
- 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. URL <https://arxiv.org/abs/1511.04599>.
- Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017. URL <https://arxiv.org/abs/1610.08401>.

- 80 Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep
81 neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019. URL
82 <https://arxiv.org/abs/1710.08864>.
- 83 Yusuke Tashiro, Yang Song, and Stefano Ermon. Parsimonious black-box adversarial attacks via
84 efficient combinatorial optimization. *arXiv preprint arXiv:1905.06635*, 2019. URL [https:](https://arxiv.org/abs/1905.06635)
85 [//arxiv.org/abs/1905.06635](https://arxiv.org/abs/1905.06635).
- 86 Jonathan Uesato, Brendan O’Donoghue, Aaron van den Oord, and Pushmeet Kohli. Adversarial risk
87 and the dangers of evaluating against weak attacks, 2018. URL [https://arxiv.org/abs/1802.](https://arxiv.org/abs/1802.05666)
88 [05666](https://arxiv.org/abs/1802.05666).