International Journal of Computer Vision (2023) 131:160–176
https://doi.org/10.1007/s11263-022-01701-w

Minimally Distorted Structured Adversarial Attacks

Ehsan Kazemi[1] · **Thomas Kerdreux**[2] · **Liqiang Wang**[1]

Received: 30 July 2021 / Accepted: 4 October 2022 / Published online: 10 October 2022

Abstract
White box adversarial perturbations are generated via iterative optimization algorithms most often by minimizing an adversarial loss on a ℓ_p neighborhood of the original image, the so-called distortion set. Constraining the adversarial search with different norms results in disparately structured adversarial examples. Here we explore several distortion sets with structure-enhancing algorithms. These new structures for adversarial examples might provide challenges for provable and empirical robust mechanisms. Because adversarial robustness is still an empirical field, defense mechanisms should also reasonably be evaluated against differently structured attacks. Besides, these structured adversarial perturbations may allow for larger distortions size than their ℓ_p counterpart while remaining imperceptible or perceptible as natural distortions of the image. We will demonstrate in this work that the proposed structured adversarial examples can significantly bring down the classification accuracy of adversarially trained classifiers while showing a low ℓ_2 distortion rate. For instance, on ImagNet dataset the structured attacks drop the accuracy of the adversarial model to near zero with only 50% of ℓ_2 distortion generated using white-box attacks like PGD. As a byproduct, our findings on structured adversarial examples can be used for adversarial regularization of models to make models more robust or improve their generalization performance on datasets that are structurally different.

Keywords Adversarial attacks · Blurriness · Group norm · Image classification

1 Introduction

Adversarial examples are inputs to machine learning classifiers designed to cause the model to misclassify the input images. These samples are searched in the vicinity of some samples in the test set, and typically in their norm-ball neighborhoods, the so-called *distortion set*. When replacing every test set sample with their corresponding adversarial examples, the accuracy of standardly trained classifiers drops to zero in the inverse correlation with the considered norm-ball radius. Thus, the lack of robustness of classifiers to adversarial samples challenges the security of some real-world systems and poses questions regarding the generalizing properties of neural classifiers (Schmidt et al., 2018; Stutz et al., 2019).

Thus far, there have been some successful studies on defense strategies against adversarial examples, though most of the attack and defense mechanisms considered ℓ_p neighborhoods. In particular, existing approaches for learning adversarially robust networks include methods that are both empirically robust via adversarial training (Goodfellow et al., 2015; Kurakin et al., 2016; Madry et al., 2017) and also certifiably robust with certified bounds (Wong & Kolter, 2017; Raghunathan et al., 2018; Zhang et al., 2019) and randomized smoothing (Cohen et al., 2019; Yang et al., 2020). Recently, there were some studies that outlined the inherent limitations of the ℓ_p balls (Sharif et al., 2018; Sen et al., 2019). While some recent papers (Xu et al., 2018; Wong et al., 2019) pointed out the benefits of other families of distortions sets, many classical norm families remained mostly unexplored in the adversarial setting. In this work, we consider white-box adversarial attacks on neural networks. In the white-box framework, the model and the in-place defenses are known to the attacker. Adversarial examples in this frame-

Communicated by Liwei Wang.

✉ Ehsan Kazemi
ehsan_kazemy@knights.ucf.edu

Thomas Kerdreux
thomaskerdreux@gmail.com

Liqiang Wang
lwang@cs.ucf.edu

[1] Department of Computer Science, University of Central Florida, Orlando, USA

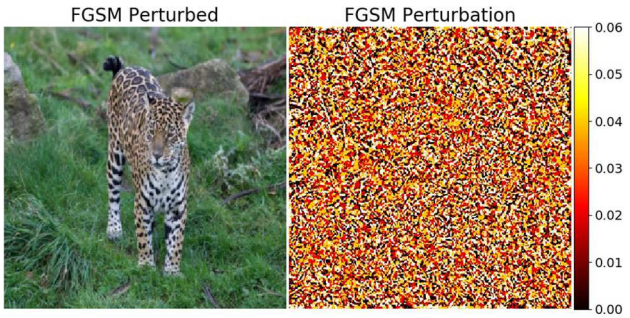
[2] Zuse Institute of Berlin, TU Universitat of Berlin, Berlin, Germany

work are typically crafted using optimization algorithms aiming to minimize constrained adversarial losses. In black-box attacks, the attacker can only make queries and observe the response of the model.

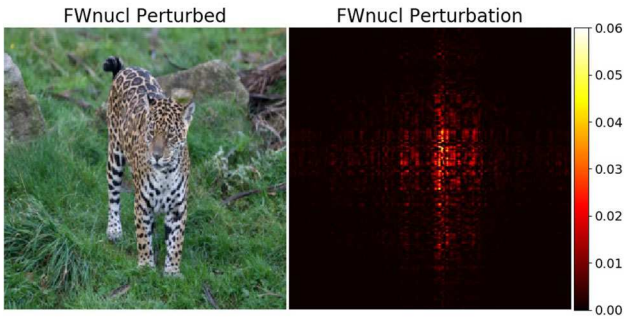
Although norms are equivalent in the image finite-dimensional space, the type of norm-ball influences the structure of the optimization algorithm iterations and the (local) minima to which they converge. As the studies on the robustness of neural model still remained empirical, it is hence necessary to explore the effect of particular structures in adversarial perturbation besides ℓ_p balls. For instance, Fig. 1 shows that the perturbation generated by the proposed attack (FWnucl) is more structured and targeted to the main objective in the image. Relatedly, important security concerns can be raised when some empirical defense mechanisms are vulnerable to certain patterns in the adversarial examples. Thus, providing a catalog of many structured attacks would cause the rapid development of robust machine learning algorithms due to an arms race between attack and defense mechanisms and can greatly expand the scope of adversarial defenses to new contexts. For instance, in Carlini et al. (2019), it is shown many defense mechanisms can be broken by stronger attacks while exhibiting robustness to the weaker attacks. Thus, finding more diverse attacks is important for evaluating defense strategies. In addition, as adversarial training uses attack methods as a form of regularization for training neural networks, the training process can be performed on the newly proposed adversaries to robustify models against discovered structured semantic perturbations. These sorts of training processes can better flatten the curvature of decision boundaries which can be potentially an important parameter to improve generalization performance in non-adversarial settings (Keskar et al. 2016).

Regarding generating the white-box adversarial samples, the radius of the convex balls is often considered sufficiently small to ensure that the added perturbations to the original samples are imperceptible. This imperceptibility requirement is pervasive in the literature, although it is not the only studied regime for adversarial examples (Gilmer et al. 2018). Arguably, the imperceptibility of the distortion does not play a crucial role in crafting adversaries, in particular when the ideal level of perturbation is aligned with the human perception in the sense that the perturbed image is labeled as the original image for a human observer. In fact, the imperceptible deformation regime of non-robust classifiers has received much attention because it highlights the gap between human perception and the processing done by machine learning systems to classify the non-perceptible class of perturbed samples (Gilmer et al. 2018).

In this work, we do not limit ourselves to the imperceptible regime of perturbation. Instead, we explore adversarial examples' structure leading to possibly perceptible deformations that would yet be considered as *non-suspicious* alteration



(a) FGSM



(b) FW with nuclear ball (FWnucl)

Fig. 1 The images correspond to two types of targeted attacks. Projected Gradient Descent (PGD) solve (1) constrained by a ℓ_∞ ball while FWnucl solves (1) constrained with a nuclear ball. The type of adversarial perturbations differs significantly in structure

of the image. In particular, we consider the trace norm ball (the nuclear ball), which is the convex relaxation of rank-1 matrices. Qualitatively, adding perturbation in this distortion set leads to blurring effects on the original image. This blurring effect could be further localized in a controlled way to specific semantic areas in the image by considering the group-nuclear ball distortion set, where the groups are defined on the specific semantic area of interest.

In the sequel, for the sake of simplicity of the presentation we focus on untargeted adversarial examples. Our approach is to use an auxiliary optimization problem to craft the adversarial perturbations. The optimization problem to generate untargeted adversarial attack for the original sample x^{ori} is formulated by

$$\begin{aligned} &\text{minimize } \mathcal{L}(x) = -L(f(x), y) \\ &\text{subject to } \left\| x - x^{ori} \right\| \leq \epsilon \end{aligned} \tag{1}$$

where L is an adversarial loss (e.g., cross entropy loss), f is the neural classifier and y is the label of the original sample x^{ori}. In this formulation, ϵ constrains the perturbation magnitude in particular norms.

Related Work.

Several recent research studies question the underlying reason for considering ℓ_p neighborhood as distortion sets and

propose alternative adversarial models. For instance, (Sharif et al., 2018) suggests that ℓ_p norms are neither the right metric for perceptually nor even content-preserving adversarial examples. In Sen et al. (2019) a behavioral study is conducted which shows that ℓ_p norms and some other metrics do not align with the human perception.

There are some recent works that consider adversarial perturbations beyond the ℓ_p distortion sets. In Engstrom et al. (2017) it is shown that simple rotation and translation can create efficient adversarial examples. Xu et al. (2018) consider group-lasso distortion sets which are optimized based on methods like ADMM. Liu et al. (2018) generate adversarial examples based on the geometry and physical rendering of the image. They notably suggest that *large pixel perturbations can be realistic if the perturbation is conducted in the physical parameter space (e.g., lighting)*. Wong et al. (2019) recently argue that robustness to Wasserstein perturbations of the original image is essentially an invariant that should typically exist in classifiers. Recently, (Wong & Kolter, 2020) investigate learning perturbation sets without optimization-based approaches and via applying conditional generative models.

There exist some methods which solve the adversarial optimization problem on specific subspaces, which might lead to specifically structured adversarial examples. While a random subspace (Yan et al., 2019) does not necessarily induce a specific perturbation structure, projection on low-frequency domain (Guo et al., 2018) or onto the subspace generated by the top few singular vectors of the image (Yang et al., 2019, §3.4.) will induce structured adversarial examples. These approaches are leveraged to reduce the search space of adversarial perturbation for more efficient computational complexity. Finally, one can consider the problem of adversarial attack generation as an image processing task. A recent trend to various types of such algorithms are for instance conditional or unconditional generative models, style transfer algorithms, or image translation algorithms (Reed et al., 2016; Gatys et al., 2017; Risser et al., 2017; Lu et al., 2017).

In this paper, we particularly apply Frank–Wolfe methods to solve the adversarial optimization problem. These algorithms have shown a recent revival in constrained optimization problems for machine learning, where their success is notably due to their low-cost computational cost per iteration (Jaggi, 2013). It is known that Frank–Wolfe method exhibits linear convergence on polytopes (Guélat & Marcotte, 1986; Garber & Hazan, 2013a, b; Lacoste-Julien & Jaggi, 2013, 2015), on strongly convex set (Levitin & Polyak, 1966; Demyanov & Rubinov, 1970; Dunn, 1979; Garber & Hazan, 2015) or uniformly convex sets (Kerdreux & d'Aspremont, 2020). Frank–Wolfe algorithm has been extensively studied in convex setting for large scale nuclear norm regularization (Jaggi & Sulovskỳ, 2010) (Lee et al., 2010; Shalev-Shwartz et al., 2011; Harchaoui et al., 2012; Dudik et al., 2012; Allen-Zhu et al., 2017; Garber et al., 2018). Furthermore, many variations of Frank–Wolfe method exist (Freund et al., 2017; Cheung & Li, 2017) that leverage the facial properties to preserve structured solutions for non-polytope or strongly convex domains. A closer approach to this work is Chen et al. (2018), where the authors apply the zero-order Frank–Wolfe algorithm for solving adversarial problems in the black-box setting.

This work exploits an optimization method to generate adversarial attacks by imposing blurriness on the target images. Currently, the commonly-used packages for crafting adversarial samples, e.g., Foolbox (Rauber et al., 2017) apply spatial filters aiming to craft adversaries via blurring. In Guo et al. (2020) a method for forging visually natural motion-blurred adversarial examples is introduced where the misclassification capability is achieved by tuning the kernel weights. Their work is mainly inspired by the Gaussian blurring kernel, though using a learnable kernel. Nevertheless, to craft visually natural and plausible examples, the authors introduced a paradigm for the saliency-regularized adversarial kernel prediction and the predicted kernel is regularized to achieve natural visual effects. However, in our approach, the blurriness is generated using additive random noise-like perturbations.

Contribution.
Currently, the defense techniques and in particular the mechanisms which provide theoretical guaranties are designed for non-structured norms while structured norms are largely overlooked in the literature. This shortcoming may render previous defense algorithms less appealing when exposed to structured adversaries. We study some families of structured norms in the adversarial example setting. This is a pretense to more generally motivate the relevance of structured attacks (i.e. besides the ℓ_p distortion set), which are largely unexplored. It is also a versatile approach to producing specific modifications of the adversarial images, like (local) blurriness. We demonstrate in the experiments that the proposed structured adversaries generate samples that target the important parts of the image resulting in a lower number of perturbed elements from the original image, and therefore providing a lower perturbation magnitude which makes them undetectable (see Fig. 1). We also demonstrate an algorithm for the localized perturbations (blurriness) of the region of interest in the image using group norms.

2 Structured Distortion Sets

Here we describe some structured families of norms that to the best of our knowledge have not so far been explored in the context of adversarial attacks. To be more specific, we generate some specifically structured perturbations by solving the

adversarial problem (1), which provides the potential attacker a framework to derive adversarial alternation of the original test samples. In the sequel, we set the trace norm ball as the distortion set and design a framework to solve the optimization problem (1) based on conditional gradient algorithms. In the conditional gradient algorithm, in each iteration a Linear Minimization Oracle (LMO) is solved. More technically, for a direction d and a convex set $\mathcal{C}$, the LMO problem is defined as

$$\text{LMO}_{\mathcal{C}}(d) \in \underset{v \in \mathcal{C}}{\text{argmin}}\, d^T v. \tag{2}$$

The iterations of conditional gradient algorithms are then constructed as a (sparse) convex combination of the solutions to (2). These solution points can always be chosen as the vertices of $\mathcal{C}$. Hence, the specific structure of the solutions of the LMO is applied in the early iterations of the optimization problem. In the following section, we provide the mathematical formulation of the optimization problem.

2.1 Low-Rank Perturbation

We let $\|\cdot\|_{S1}$ denote the nuclear norm which is the sum of the matrix singular value, a.k.a. the trace norm or the 1-Schatten norm. The nuclear norm has been classically used to find low-rank solutions of convex optimization problems (Fazel et al., 2001; Candès & Recht, 2009) such as matrix completion. Here, we propose to simply consider nuclear balls as distortion sets when searching for adversarial examples in problem (1). We later explain the various potential benefits of using this structural distortion set. To our knowledge, the low-rank structure is leveraged in different aspects of some defense techniques (Yang et al., 2019) but it has never been acquired to craft adversarial attacks. As an empirical defense mechanism, (Langeberg et al., 2019) add a penalization in the training loss to enhance the low-rank structure of the convolutional layer filters. Yang et al. (2019) notably propose a pre-processing of the classifier outputs, which randomly removes some input pixels and further reconstructs them via matrix completion for denoising purposes.

More formally, with nuclear ball as a distortion set, the adversarial optimization problem (1) is reformulated as

$$\underset{\|x-x^{ori}\|_{S1} \leq \epsilon}{\text{argmin}} \quad \mathcal{L}(x) = -L(f(x), y). \tag{3}$$

This formulation is a particular example of the family of p-Schatten norms $\|\cdot\|_{Sp}$, i.e., the p-norm of the singular value vector with $p = 1$. These structured norms lead to differently structured adversarial examples. Given the lack of explicit mathematical translation across norms, these adversaries may end up defeating certified approaches in terms of ℓ_p neighborhoods. At this point, we solve the adversarial

Fig. 2 For a test image of CIFAR-10, we computed the various adversarial examples stemming from solving (1) on the nuclear ball with Frank–Wolfe algorithm. From left to right: original image, adversarial example with a nuclear radius of $\epsilon_{S1} = 5, 10, 20, 30$. Note that the adversarial examples are already miss-classified with $\epsilon_{S1} = 3$; here we increase the radius purposely to observe the perturbation on the initial image

problem (3) in the framework of conditional gradient methods. The analytical solution of LMO (2) for a nuclear ball of radius ρ is given by

$$\text{LMO}_{\|\cdot\|_{S1} \leq \rho}(M) \triangleq \rho\, U_1 V_1^T, \tag{4}$$

where U_1, V_1 are the first columns of matrices U and V in the SVD decomposition of matrix M given by USV^T. For q-Schatten norm (with $q > 1$) the LMO has also a closed-form solution involving the full singular decomposition (see e.g., Garber & Hazan, 2015, Lemma 7). Solving LMO involves computing the right and left singular vectors U_1 and V_1 which are associated with the largest singular value ρ. Lanczos algorithm can be used to calculate singular vectors corresponding to the largest singular value, where the solution is found using the Krylov subspace formed by the columns of matrix M. This demonstrates the computational efficiency of Frank–Wolfe methods as opposed to the other optimization approaches such as projected gradient descent, which requires the full SVD computation in each iteration. Qualitatively, adversarial perturbations in nuclear norm add a blurring effect to the original images, as for instance is depicted in Fig. 2. Thus, this can potentially pose a risk in some security scenarios, when such perturbations could be perceived as simple alterations of the image rather than a malware deformation of it, e.g., see Gilmer et al. (2018) for real-world scenarios.

2.2 Group Constraints

In this section, we demonstrate how to leverage weighted group norms in order to localize the low-rank perturbations. Group-norms are defined by a partition of the pixels' coordinates into groups. For instance, such a partition can be adapted from a segmentation of the sample image. These group-norms are a combination of two norms: a local one applied on vectors formed by each group of pixel values, and a global one applied on the vectors of the norms of all the groups. Here, we consider the nuclear norm as the local norm and the global ℓ_1 norm to induce sparsity at the group level.

Considering such norms provides some tools to substantially control the perturbations restricted to desirable parts to craft adversarially perturbed images.

Nuclear Group Norm.

Let $\mathcal{G}$ be an ensemble of groups of pixels' coordinates of the tensor image of (c, h, w), where each element $g \in \mathcal{G}$ is a set of pixel coordinates'. Then for $x \in \mathbb{R}^{c\times h\times w}$ we define $\mathcal{G}$-nuclear group-norm as

$$\|x\|_{\mathcal{G},1,p} = \left\| \|x[g]\|_{S(1),g\in\mathcal{G}} \right\|_p, \tag{5}$$

with $p \in [1, \infty[\cup\{\infty\}$ (see for instance Tomioka & Suzuki, 2013). When $\mathcal{G}$ is a partition of the pixels, $\|\cdot\|_{\mathcal{G},1,S(1)}$ is a norm. The nuclear group-norm allows to localize the blurring effect of the nuclear norm. Indeed, the LMO of $\mathcal{G}$-nuclear group-norm is given by

$$\text{LMO}_{\|\cdot\|_{\mathcal{G},1,S1}\leq\rho}(M) \triangleq \begin{cases} \rho\, U_1^{(g)}\big(V_1^{(g)}\big)^T \\ 0 \quad \text{otherwise} \end{cases}, \tag{6}$$

where $g^* = \underset{g\in\mathcal{G}}{\text{argmax}} \|M[g]\|_{S1}$ and the singular value decomposition of $M[g]$ for each group g is given by $U^{(g)}S^{(g)}\big(V^{(g)}\big)^T$. When solving (1) with such norms, each iteration of the conditional gradient will add to the adversarial perturbation a vertex of the form described by (6), i.e. a matrix of rank-one on the rectangle defined by the group of pixels in $g \in \mathcal{G}$. Note that the only modification for the approximate solution of group nuclear ball versus nuclear ball is the solution to LMO problem, and the rest of the conditional gradient method for both of the distortion sets is similar.

Different Distortion Radius per Group.

When perturbing an image, modification in the pixel regions with high variance are typically harder to perceive than pixel modification in low variance regions. This knowledge was leveraged in Luo et al. (2018) or in the σ-map of (Croce & Hein, 2019, §2.2.) to craft more imperceptible adversaries. Weighted nuclear group norms allow to search adversarial perturbations with different distortion radius across the image. With some $w_g > 0$, the weighted nuclear group norm is defined as

$$\|x\|_{\mathcal{G},1,S(1),w} = \sum_{g\in\mathcal{G}} w_g \|x[g]\|_{S(1)}, \tag{7}$$

and the LMO for weighted nuclear group-norm is then obtained as

$$\text{LMO}_{\|\cdot\|_{\mathcal{G},1,S1}\leq\rho}(M) \triangleq \begin{cases} \frac{\rho}{w_{g^*}}\, U_1^{(g^*)}\big(V_1^{(g^*)}\big)^T & \text{on group of pixels } g^* \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $g^* = \underset{g\in\mathcal{G}}{\text{argmax}} \frac{1}{w_g}\|M[g]\|_{S1}$ and the singular value decomposition of $M[g]$ for each group g is given by $U^{(g)}S^{(g)}\big(V^{(g)}\big)^T$. In particular, this means that the solution corresponding to the group associated with g have a nuclear radius of $\frac{\rho}{w_g}$ and the weights w_g which allows to control the distortion in each group of pixels. The weights can be customized by the attacker to impose perturbation in desirable regions of the image. For instance, the weights can be chosen in inverse correlation with the variance of pixel regions to make the perturbations more targeted.

2.3 Structure Enhancing Algorithm for Adversarial Examples

We apply Frank–Wolfe algorithms (Frank & Wolfe, 1956), a.k.a. conditional gradient algorithms (Levitin & Polyak, 1966), for problem (1). Given the conditional gradient optimization framework, the algorithm 1 can iteratively find the adversarial perturbation to fool the network. For specific constraint structures such as the distortion set introduced earlier, conditional gradient algorithms naturally trade-off between the convergence accuracy and the structured solutions in the early iterations.

Algorithm 1 Vanilla Frank-Wolfe

Input: Original image $\boldsymbol{x}_0$
for $t = 0, \cdots, T$ **do**
  $s_t = \text{LMO}_{\mathcal{C}}\big(-\nabla\mathcal{L}(\boldsymbol{x}_t)\big)$.
  $\gamma_t = \text{LineSearch}(x_t, s_t - x_t)$
  $x_{t+1} = (1-\gamma_t)x_t + \gamma_t s_t$
end for

For almost all the distortion sets which we consider in this work, LMO has a closed-form solution. Note that the LMO has a low computational requirement as opposed to the projection-based approaches. In particular, LMO requires only computing the first singular vectors, while comparably projection steps demand the full SVD matrix to find the solution in each iteration. Provided the upper-bound for the Lipschitz constant L of the adversarial loss is known, we apply the short step size $\gamma_t = \text{clip}_{[0,1]}(\langle -\nabla f(x_t), s_t - x_t\rangle / L\,\|s_t - x_t\|^2)$ for the optimization method. These are the only parameters that should be tuned in the algorithm, which makes the method more versatile for many models as compared to attacks that require hyperparameter tuning such as CW attacks (Carlini & Wagner, 2017).

It is well-known that for non-convex, objective functions e.g. the adversarial losses, injecting noise might be useful to escape from local optimums. This noise could be added either via random starts or via randomized block-coordinate methods. Under some additional conditions, Kerdreux et al. (2018) proposes a version of Frank–Wolfe that solves linear

Table 1 MNIST and CIFAR-10 extensive white-box attack results

Network	Training model	Clean	Accuracy under attack			
			FWnucl 20 *	FWnucl 20 +	PGD 20	FGSM
MNIST						
LeNET	Madry	98.38	95.26	92.76	95.79	96.59
	ME-Net	99.24	97.63	75.41	74.88	46.18
SmallCNN	Madry	99.12	98.19	96.66	95.77	97.95
	ME-Net	99.42	89.56	78.65	76.84	54.09
CIFAR-10						
ResNet-18	Madry	81.25	44.28	3.06	49.95	55.91
	ME-Net	93.45	29.66	4.01	4.99	44.80
WideResNet	Madry	85.1	43.16	2.82	52.49	59.06
	ME-Net	95.27	40.09	16.04	12.73	59.33
ResNet-50	Madry	87.03	40.97	2.64	53.01	61.44
	ME-Net	92.09	47.66	17.81	9.14	58.51

FWnucl 20 *: FWnucl with $\epsilon_{S1} = 1$. FWnucl 20 +: FWnucl with $\epsilon_{S1} = 3$. On MNIST (resp. CIFAR-10) PGD and FGSM have a total perturbation scale of 76.5/255 (0.3) (resp. 8/255 (0.031)), and step size 2.55/255 (0.01) (resp. 2/255 (0.01)). PGD runs for 20 iterations. We reproduce the ME-Net and Madry defense with same training hyper-parameters

Table 2 ImageNet extensive white-box attack results on 4000 randomly selected images from validation dataset of ImageNet

Network	Training model	Clean	FWnucl						PGD					
			$\epsilon_{S1} = 1$		$\epsilon_{S1} = 3$		$\epsilon_{S1} = 5$		$\epsilon = 2/255$		$\epsilon = 4/255$		$\epsilon = 8/255$	
			Acc	ℓ_2	Acc	ℓ_2	Acc	ℓ_2	Acc	ℓ_2	Acc	ℓ_2	Acc	ℓ_2
ResNet-50	Standard	80.55	19.67	0.69	1.62	1.27	0.17	1.68	0.2	2.53	0.0	4.55	0.0	8.53
	Madry	50.02	38.3	1.45	16.8	3.82	6.62	5.80	42.07	2.97	34.52	5.90	18.9	11.69

FWnucl and PGD runs for 100 iterations. We reproduce Madry defense in ℓ_∞ norm with $\epsilon = 8/255$. In the table we report accuracy (Acc) versus the average of generated ℓ_2 distortion for each attack

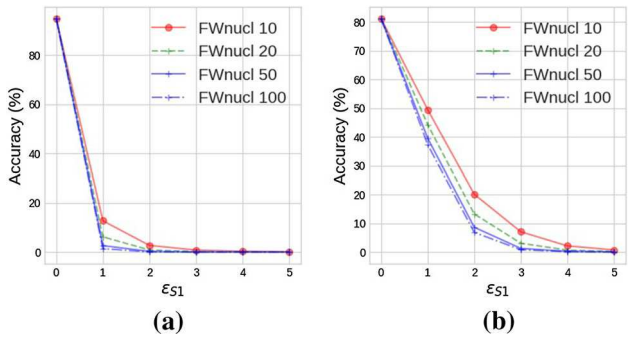


Fig. 3 Accuracy of standard model (left) and robust model of Madry (right) on ResNet-18 for CIFAR-10, versus the nuclear ball radius when varying the number of steps

minimization oracles on random subsets of the constraint sets. Here we consider subsampling the image channels, i.e., $\|x\|_{color,S1} = \sum_{c=1}^{3} \|x_c\|_{S1}$ where x_c is one of the image channels. Note that we did not impose the box constraints which demonstrate that the values of image elements should be inside the interval $[0, 1]^d$. To impose this restriction, we clamp the last iteration of the optimization process to satisfy box constraints. Although this approach does not guarantee the convergence to a saddle point but removes the need to compute the LMO over the intersection of two sets, which is non-trivial.

3 Numerical Experiments

This section aims at evaluating the adversarial accuracy of adversarial examples using Frank–Wolfe algorithms for the adversarial problem (1) with nuclear balls as distortion sets, which we refer to as FWnucl. The complementary results for Frank–Wolfe with group norms and random initialization are provided in the appendix.

Experiments Goal.

We tested the FWnucl white-box attack against two baselines of defenses for untargeted attacks. The first is Madry et al. (2017), the state-of-the-art defense against white-box attacks. It uses the training images augmented with adversarial perturbations to train the network. The second one (Yang et al., 2019) leverages matrix estimation techniques as a pre-processing step; each image is altered by randomly masking various proportions of the image pixels' and then

reconstructed using matrix estimation by the nuclear norm. For a given training image, this approach produces a group of images that are used during training, see Yang et al. (2019) for more details. This provides a non-differentiable defense technique, i.e. a method that cannot be straightforwardly optimized via back-propagation algorithms, and was reported to be robust against methods in Athalye et al. (2018) by circumventing the obfuscated gradients defenses. Qualitatively it leverages a structural difference between the low-rank structure of natural images and the adversarial perturbations that are not specifically designed to share the same structures. In addition, we evaluate our proposed attacks over more recently introduced robust models such as Adversarial Weight Perturbation (AWP) (Wu et al., 2020) and Learnable Boundary Guided Adversarial Training (LBGAT) (Cui et al., 2021). We also evaluate our method against a provably robust model trained with randomized smoothing (Cohen et al., 2019). In the randomized smoothing, a provably robust classifier is derived from the convolution of the base classifier with the isotropic Gaussian distribution of variance σ^2. This approach provides provable certified bounds in the L_2 norm for the smoothed classifier. We show that the structured attacks can bring down the accuracy of the model to the certified accuracy in almost all the smoothed models. Overall, a key motivation of our experiments is to propose adversarial examples with specific structures, serving at least as a sanity check for defense approaches.

Experiment Settings.
We assess the accuracy of networks in different scenarios over MNIST and CIFAR-10 testsets. For ImageNet we examine the neural models over randomly selected images from the ImageNet validation set that are correctly classified. For defense evaluation, for MNIST we use the LeNet model with two convolutional layers similar to Madry et al. (2017) and SmallCNN with four convolutional layers followed by three fully connected layers as in Carlini and Wagner (2017). For CIFAR-10 dataset we use ResNet-18 and its wide version WideResNet and ResNet-50. For the ImageNet dataset, we use ResNet-50 architecture.

We report the adversarial accuracy of FWnucl along with those of classical attack methods like Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015), and Projected Gradient Descent (PGD) (Madry et al., 2017) to solve adversarial problem (1) using ℓ_∞ ball as the distortion set. FGSM generates adversarial examples with a single gradient step, while PGD is a more powerful adversary that performs a multi-step variant of FGSM. In addition, we compare our attacks with the adversarial attack Auto-attack (Croce & Hein, 2020) which is an ensemble of parameter-free attacks consisting of a stepsize free version of PGD (APGD-CE), a stepsize free PGD with a novel loss (APGD-DLR), Fast Adaptive Boundary Attack (FAB) with the goal of finding minimal ℓ_p norm perturbations, and Square Attack which is a query-efficient black-box attack. While this ensemble of attacks leverage non-standard mechanisms to optimize the threat model, our goal is to compare the proposed approach with the fundamental basic blocks of optimization approaches in adversarial attacks to show that the proposed method is introducing novel adversarial structures which are structurally different from the ℓ_p norm structures.

Empirical Results.
In Table 1 we report accuracy as the percentage of adversarial examples that are classified correctly. We repeated the experiments several times to insure the results are general. These numerical experiments demonstrate that the attack success rates for FWnucl are comparable to the classical ones in an imperceptibility regime while also retaining specific structures in the perturbation. Note that FGSM for ME-Net provides a better success rate (lower adversarial accuracy) compared to PGD which indicates the gradient masking generated by ME-Net over MNIST dataset. Table 1 also shows that FWnucl with $\epsilon = 3$ significantly performs better than other attacks. We attribute this difference to the fact that FWnucl has a tendency to induce low-rank solutions, leading to global structure perturbation in images without any ℓ_p norm restrictions. This key characteristic of FWnucl makes it orthogonal to the existing adversarial attacks. FWnucl is specifically designed to iterate over solutions that lie on low-dimensional faces of the feasible set, as low-dimensional faces of the feasible region contain desirable well-structured low-rank matrices.

In Table 2 we provided the adversarial accuracy for standardly and adversarially trained models over ImageNet dataset. The results show that the attacks created by PGD show at least 50% increase in ℓ_2 norm distortion compared with FWnucl. Note that enlarging the radius of the norm ball for PGD attack significantly increases the distortion while for FWnucl the increase in distortion rate is not fierce per increasing the nuclear ball radius. It confirms our earlier intuition that FWnucl is designed to selectively add distortion to pixels which are important for the label predictions. For the adversarially trained model, the robust accuracy for FWnucl is significantly lower that the counterparts from PGD. It indicates that FWnucl generates patterns that the robust models may not be robust to them.

Figure 3 summarizes the results for FWnucl with varying ϵ_{S1} for standard and robust model on CIFAR-10. The figure shows FWnucl algorithm noticeably drops the accuracy rate by increasing the radius ϵ_{S1}. The performance of different FWnucl methods is slightly different, as the higher number of FWnucl steps may gain better performance.

We also extend our evaluations to models trained adversarially using PGD_2 method, i.e., PGD with ℓ_2 norm, for MNIST, CIFAR-10 and ImageNet datasets in Table 3. The

Table 3 MNIST, CIFAR-10 and ImageNet extensive white-box attack results for models adversarially trained with PGD_2

Network	Clean	FWnucl 20			PGD_2 20			Auto-attack-ℓ_2		
		$\epsilon_{S1} = 0.5$	$\epsilon_{S1} = 1.0$	$\epsilon_{S1} = 1.5$	$\epsilon = 0.5$	$\epsilon = 1.0$	$\epsilon = 1.5$	$\epsilon = 0.5$	$\epsilon = 1.0$	$\epsilon = 1.5$
MNIST										
LeNET	99.31	98.65	97.64	94.39	98.35	96.41	91.45	98.24	95.68	87.14
SmallCNN	99.19	98.57	97.4	94.61	98.37	96.53	92.57	98.16	95.46	87.44
CIFAR-10										
ResNet-18	89.97	67.22	33.31	11.54	69.38	41.01	18.86	66.63	33.37	9.57
WideResNet	90.64	68.21	34.54	11.78	70.6	43.02	19.46	68.27	35.18	9.83
ResNet-50	90.83	68.95	35.55	12.8	71.78	44.77	21.36	69.24	36.16	11.17
ImageNet										
ResNet-50	63.68	60.91	54.24	57.9	61.01	57.99	54.97	58.89	53.74	49.33

The radius for adversarial examples for training in ℓ_2-norm for MNIST is 1.5 and for CIFAR-10 and ImageNet is 0.5. The results for ImageNet is for 10,000 randomly selected images from validation dataset of ImageNet

Table 4 Adversarial accuracy on CIFAR-10 and CIFAR-100 datasets using WideResNet (WRN-34-10) under nuclear norm threat model with FWnucl and ℓ_∞ threat models with Auto-attack for adversarially trained models using AWP (Wu et al., 2020) and LBGAT (Cui et al., 2021)

Network	Training model	Clean	FWnucl						Auto-attack-ℓ_∞					
			$\epsilon_{S1} = 1$		$\epsilon_{S1} = 3$		$\epsilon_{S1} = 5$		$\epsilon = 2/255$		$\epsilon = 4/255$		$\epsilon = 8/255$	
			Acc	ℓ_2	Acc	ℓ_2	Acc	ℓ_2	Acc	ℓ_2	Acc	ℓ_2	Acc	ℓ_2
CIFAR-10														
WRN-34-10	AT-AWP	85.57	36.46	1.49	1.68	3.54	0.14	5.14	79.41	0.03	72.12	0.11	53.90	0.54
	TRADES-AWP	85.36	38.90	1.48	2.98	3.67	0.32	5.48	79.65	0.02	72.74	0.11	56.19	0.49
	LBGAT	88.22	33.94	1.43	2.01	3.40	0.26	5.00	81.72	0.03	73.31	0.14	52.21	0.61
CIFAR-100														
WRN-34-10	AT-AWP	60.38	42.67	0.93	13.57	2.52	4.99	3.84	51.73	0.04	43.58	0.14	28.84	0.53
	TRADES-AWP	60.17	42.16	0.94	15.98	2.53	6.63	3.78	50.82	0.04	42.50	0.15	28.79	0.53
	LBGAT	70.25	44.50	0.90	12.23	2.17	4.91	3.04	58.53	0.05	46.78	0.20	26.72	0.73

Table 5 Adversarial accuracy and the average norm of generated perturbations on CIFAR-10 dataset using WideResNet (WRN-34-10) under nuclear norm threat model with FWnucl and ℓ_∞ threat model with Auto-attack for robust models AWP (Wu et al., 2020) and LBGAT (Cui et al., 2021)

Training model	Metric	FWnucl			Auto-attack-ℓ_∞			
		$\epsilon_{S1} = 1$	$\epsilon_{S1} = 3$	$\epsilon_{S1} = 5$	$\epsilon = 4/255$	$\epsilon = 8/255$	$\epsilon = 16/255$	$\epsilon = 32/255$
CIFAR-10-WRN-34-10								
Clean Acc 85.7	Acc	36.46	1.68	0.14	72.12	53.90	16.29	0.32
AT-AWP	ℓ_2	1.49	3.54	5.14	0.11	0.54	2.32	5.49
	$\|\cdot\|_{S1}$	1.00	2.93	4.79	0.26	1.22	5.21	12.41
	SSIM	0.9732	0.8925	0.8078	0.9977	0.9819	0.8836	0.6740
Clean Acc 88.22	Acc	33.94	2.01	0.26	73.31	52.21	13.81	0.25
LBGAT	ℓ_2	1.43	3.40	5.00	0.13	0.61	2.47	5.58
	$\|\cdot\|_{S1}$	1.00	2.91	4.73	0.30	1.40	5.57	12.51
	SSIM	0.9739	0.8982	0.8174	0.9972	0.9781	0.8754	0.6766

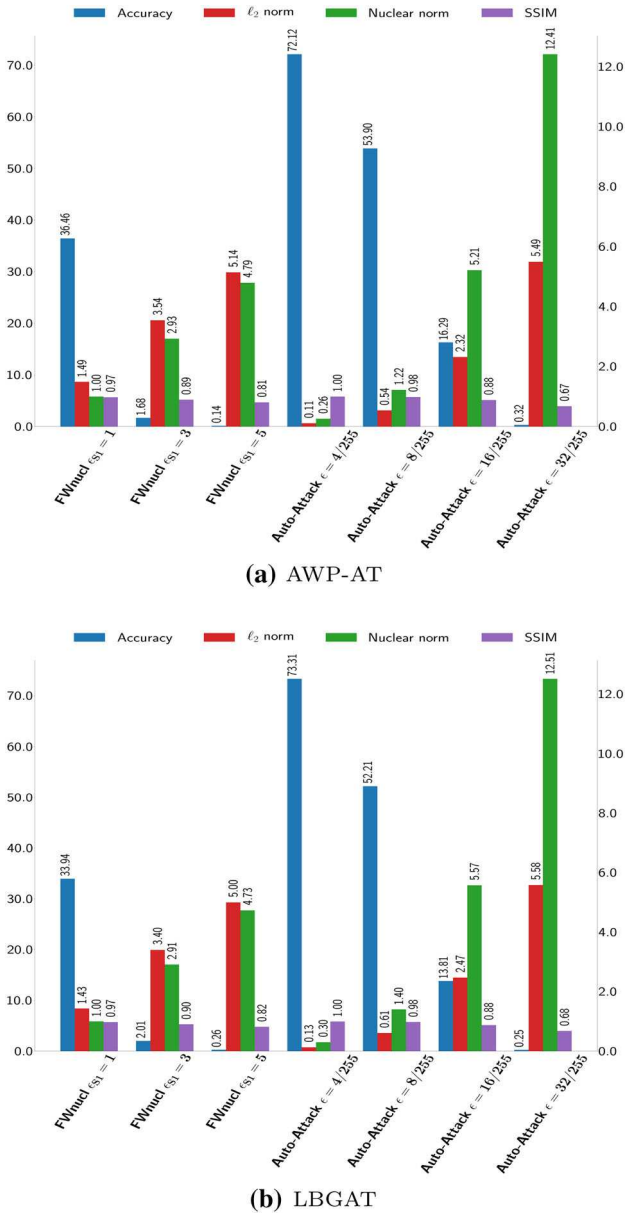
results show that these models fail when they are attacked by the FWnucl attack. This demonstrates that FWnucl could symmetrically fail the ℓ_∞ and ℓ_2 adversarially trained models.

We also examine our proposed method over the recently introduced robust models AWP (Wu et al., 2020) and LBGAT (Cui et al., 2021) over CIFAR-10 and CIFAR-100 datasets. In Table 4 we compare our attacks versus more novel adversarial training approaches and compare the performance with more powerful attacks such as Auto-attack. The results show that the robust models are vulnerable to the plausible modifications to the images which are generated using FWnucl and can provide a better success rate compared to Auto-attack. Auto-attack produces the adversarial with a lower ℓ_2 norm as it always exploits a budget for the pixel-wise modification. The results show that the generated adversarial attacks with a bounded nuclear norm distortion within the perceptual ball can easily fail the recent robust model though with a higher distortion in the ℓ_p norm compared to Auto-attack. Our attacks, while might not satisfy the ℓ_∞ threshold for the pixel-wise modifications, introduce perturbations that are not perceptible and can change the classifier decision.

Imperceptibility nuclear threshold.
The attacks such as FWnucl, PGD, and Auto-attack are constrained with perturbation limitation on particular norms which characterize the distortion set for each adversarial attack. FWnucl is constrained to the nuclear ball with a given radius and therefore it might increase beyond the standard perturbation limit used to craft adversaries with PGD or Auto-attack constrained over ℓ_∞ or ℓ_2 distortion sets. Conversely, as PGD or Auto-attack are constrained over either ℓ_2 or ℓ_∞ balls, the forged adversaries with these threat models may increase beyond the limitation for the perturbation limit for the nuclear norm in FWnucl.

As demonstrated in Sharif et al. (2018), ℓ_p norms do not capture the perceptual quality of images and are unsuitable to assess the quality of adversarial images. Therefore, we compute the structural similarity index (SSIM) measure (Wang et al., 2004) as advocated in the literature (Hameed & Gyorgy, 2021; Gragnaniello et al., 2021) to measure the perceptual similarity between the original image and the adversarial images. In Table 5 we listed the achieved adversarial accuracy, the average norm of generated distortions, and SSIM values for AT-AWP and LBGAT models in Table 4 over CIFAR-10 dataset. The results for Table 5 are also visualized in Fig. 4.

Table 5 shows that the images generated by FWnucl and Auto-attack with different perturbation limits excluding Auto-attack with $\epsilon = 32/255$ are equally good when the requirement is to achieve a minimum SSIM value up to 0.8, which guarantees high-quality images. The results also show that FWnucl with $\epsilon_{S1} = 3$ generated adversarial



(a) AWP-AT

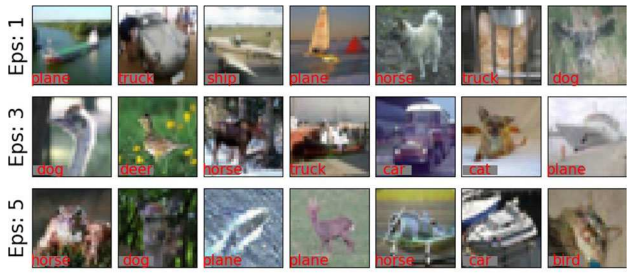
(b) LBGAT

Fig. 4 Performance of adversarial attacks on CIFAR-10 with adversarial training, listed in Table 5

images of better perceptual quality compared with Auto-attack with $\epsilon = 16/256$ whereas providing lower adversarial accuracy. It is noted from Table 5 that the nuclear norm for Auto-attack with the perturbation limits $\epsilon = 16/255$ and $\epsilon = 32/255$ for AWP model are 5.21 and 12.41, respectively, which are significantly beyond the perturbation limit of nuclear norm used for FWnucl with $\epsilon_{S1} = 3$. This indicates that the images crafted with Auto-attack using $\epsilon = 16/256$ and $\epsilon = 32/255$ are heavily distorted in the nuclear norm. This also shows that FWnucl crafts adversarial perturbations which are structurally different from adversaries generated based on ℓ_p-norm constraints.

Table 6 Adversarial accuracy and the average norm of generated perturbations of successful adversaries on 1000 randomly selected images from ImageNet validation dataset using ResNet-50 under nuclear norm threat model with FWnucl and ℓ_∞ threat models with PGD and Auto-attack for ℓ_∞ adversarially trained model

Network	Metric	FWnucl			PGD				Auto-attack-ℓ_∞			
		$\epsilon_{S1}=1$	$\epsilon_{S1}=3$	$\epsilon_{S1}=5$	$\epsilon=4/255$	$\epsilon=8/255$	$\epsilon=16/255$	$\epsilon=32/255$	$\epsilon=4/255$	$\epsilon=8/255$	$\epsilon=16/255$	$\epsilon=32/255$
ImageNet-Clean Acc 66.66												
ResNet-50	Acc	51.58	24.90	11.01	50.00	30.65	8.03	1.19	46.23	25.19	2.87	0.10
	ℓ_2	1.45	3.62	5.31	5.89	11.57	21.42	35.49	5.94	11.76	23.11	44.27
	$\|\cdot\|_{S1}$	1.0	2.91	4.86	30.56	59.07	109.17	214.02	31.05	60.95	120.40	237.50
	SSIM	0.99	0.99	0.98	0.96	0.90	0.75	0.46	0.96	0.89	0.72	0.46



(a)

Fig. 5 FWnucl adversarial examples for the CIFAR-10 dataset for different radii. The fooling label is shown on the image

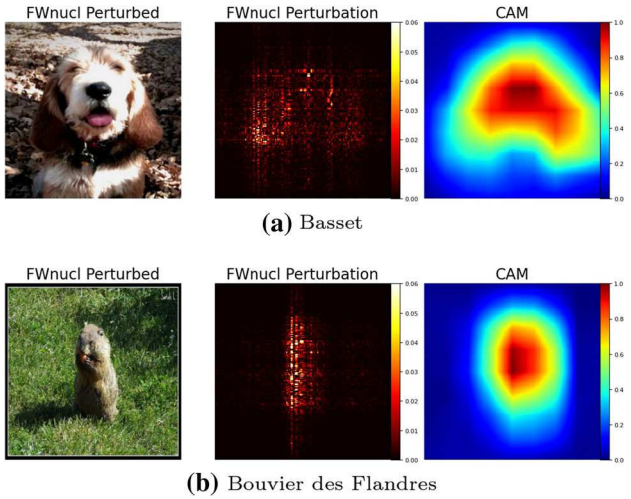


Fig. 6 The images display some structural pattern of FWnucl perturbations for the ImageNet dataset on DenseNet121 architecture, standardly trained. Observe that the adversarial perturbed pixels are accumulated on the areas containing semantic information about the image. FWnucl is conducted with $\epsilon_{S1} = 5$ and 20 iterations

In Table 6 we also compare our approach with ℓ_∞ threat models for successful adversaries over distortion sets with large radii on 1000 randomly selected images from ImageNet validation dataset. We see from Table 6 that for FWnucl with $\epsilon_{S1} = 3$ the adversarial accuracy is 23% lower than the adversarial accuracy for Auto-attack with $\epsilon = 8/256$ while the corresponding SSIM value is 9% higher for FWnucl with $\epsilon_{S1} = 3$. In addition, this table again shows that the perturbation generated by ℓ_∞ attack models over ImageNet are highly distorted in the nuclear norm level providing an average nuclear norm of 60.95 for Auto-attack with $\epsilon = 8/255$, which is by orders of magnitude larger than the maximum nuclear norm used for FWnucl with $\epsilon_{S1} = 5$. The results indicate that our FWnucl is able to make the most trade-offs between the SSIM value and the distortion rate in nuclear norm level with the adversarial accuracy.

We illustrate in Fig. 5 some adversarial examples generated by FWnucl, for three different values of ϵ_{S1}. On CIFAR-10, we qualitatively observed that with $\epsilon_{S1} = 1$, all adversarial examples are perceptually identical to the original images. The imperceptibility threshold exclusively depends on the dataset as the dataset becomes more complex, the tolerance of imperceptibility to nuclear ball radius values ϵ_{S1} increases; on ImageNet we realized the imperceptibility threshold is $\epsilon_{S1} = 10$.

We interpret the adversarial perturbations using class activation maps (CAM) (Zhou et al., 2016) which localizes class-specific discriminate semantics in the input images to the neural model. Adversarial images are generated to suppress the most discriminative regions related to the original labels and promote the discriminatory semantics in regard to the adversarial labels. We have shown the heat-map for perturbation and the CAM visualization for the corresponding images in Fig. 6. The figure shows that the perturbations are tightly projected to the most discriminatory image regions (i.e., body, head), which are localized by CAM with respect to the original label. While the noise generated by PGD attack exhibits abrupt changes in pixel intensities (see Fig. 1), the perturbation from FWnucl has continuous variations in pixel values. It is seen from the same figure that the conventional ℓ_p norm constrained methods e.g., FGSM, PGD do not encourage any structure and tends to generate perturbations even for pixels that might not be crucial for the label predictions, e.g., the background. However, FWnucl only focuses on important regions of the image which might induce a dramatic shift in the predictions. Therefore, the FWnucl attack significantly reduces the number of perturbed elements in the image. For instance, the number of non-zero pixel coordinates for PGD and FGSM on ImageNet is respectively almost $11x$ and $14x$ larger than the number of non-zero pixel intensities for FWnucl with $\epsilon_{S1} = 1$.

Adversaries with Group Nuclear Norm—FWnucl-group. In FWnucl-group we deal with sparse attacks and we want to modify the smallest number of pixels to change the model decision on the class label. We compare FWnucl-group with state-of-the-art sparse attacks Structured Attack (StrAttack) (Xu et al., 2018), Coordinate Search (CS) (Croce & Hein, 2019) with $\ell_0 + \sigma$ version and Auto-attack with ℓ_2 and ℓ_∞ norms. The attack StrAttack uses group lasso and it is shown that the generated attacks enjoy localized sparse patterns. The attack CS computes the perturbations in direct correlation to the variance of neighboring pixels. We conduct the experiments using the open-source codes from the papers. We run the attacks on 1000 samples of the test sets for ℓ_2-norm robust models. In Tables 7 and 8 we report the test accuracy of each method for MNIST and CIFAR-10 datasets, respectively. The test accuracy is the fraction of classified samples that can correctly be classified, and we listed the value of perturbation in different norms. We report the statistics of the attacks in Tables 7 and 8 only for successful attacks. The tables also report the number of pixel changes for each image

Table 7 Comparison of the FWnucl and FWnucl-group with sparse attacks Coordinate Search (CS) and Structured Attack (StrAttack) and with Auto-attack with ℓ_2 and ℓ_∞ norms for 1000 randomly selected examples of MNIST dataset on ℓ_2 adversarially trained models

Metric	FWnucl	FWnucl-group			StrAttack	CS	Auto-attack-ℓ_2		Auto-attack-ℓ_∞	
	$\epsilon_{S1}=1$	$\epsilon_{S1}=1$	$\epsilon_{S1}=3$	$\epsilon_{S1}=5$			$\epsilon=0.5$	$\epsilon=1.0$	$\epsilon=0.1$	$\epsilon=0.2$
Net-Clean Acc 99.5										
Acc	96.3	97	77	44	98	90.7	97.7	94.4	94.7	43.7
ℓ_2	0.90	0.88	2.34	3.08	0.28	1.79	0.5	1.0	2.08	3.98
$\|\cdot\|_{S1}$	1.26	1.07	3.01	4.63	0.76	5.20	1.32	2.55	7.60	14.30
ℓ_∞	0.37	0.43	0.91	0.97	0.09	0.48	0.16	0.32	0.1	0.2
pixels	465.6	54.76	70.8	90.2	462.2	32.03	320.3	326.7	560	556.4
IS	0.94	0.9339	0.9134	0.9492	0.9076	0.8682	0.9182	0.9127	0.9287	0.9810
SmallCNN-Clean Acc 99.4										
Acc	96	96.9	84.8	72.7	98.8	91.1	97.1	93.6	94.6	46.2
ℓ_2	0.91	0.86	2.19	2.80	0.14	1.70	0.5	1.0	2.03	4.14
$\|\cdot\|_{S1}$	1.27	1.09	2.95	4.29	0.36	4.81	1.30	2.47	14.26	7.50
ℓ_∞	0.36	0.41	0.86	0.94	0.05	0.48	0.17	0.34	0.1	0.2
pixels	458.08	60	79.02	94.23	334.33	29.93	224.4	230.5	587.8	696.2
IS	0.9422	0.9633	0.9048	0.9302	0.9642	0.9120	0.9577	0.9074	0.9553	0.9886

Table 8 Comparison of the FWnucl and FWnucl-group with sparse attacks Coordinate Search (CS) and Structured Attack (StrAttack) and with Auto-attack with ℓ_2 and ℓ_∞ norms for 1000 randomly selected samples of CIFAR-10 dataset on ℓ_2 adversarially trained models

Metric	FWnucl	FWnucl-group			StrAttack	CS	Auto-attack-ℓ_2		Auto-attack-ℓ_∞	
	$\epsilon_{S1}=1$	$\epsilon_{S1}=1$	$\epsilon_{S1}=3$	$\epsilon_{S1}=5$			$\epsilon=0.5$	$\epsilon=1.0$	$\epsilon=4/255$	$\epsilon=8/255$
ResNet-18-Clean Acc 90										
Acc	32.6	74.5	34.6	17.6	57.8	47.9	65.7	33.7	60.1	24.2
ℓ_2	1.64	0.82	2.02	2.70	44.87	52.49	0.5	1.0	0.85	1.68
$\|\cdot\|_{S1}$	1.01	0.56	1.61	2.58	60.61	71.23	1.05	2.09	2.09	4.10
ℓ_∞	0.24	0.20	0.42	0.49	2.05	2.17	0.05	0.12	0.015	0.031
pixels	1022.5	148.2	225.44	297.46	1011.11	906.26	1022.46	1022.12	1022.56	1022.08
IS	0.9983	0.9653	0.9978	0.9992	0.9955	0.9932	0.9843	0.9976	0.9864	0.9982
ResNet-50-Clean Acc 91.5										
Acc	35.8	77	37.3	19.6	56.0	49.9	70.3	36.8	66.2	29.6
ℓ_2	1.64	0.85	2.02	2.72	44.31	52.90	0.5	1.0	0.85	1.69
$\|\cdot\|_{S1}$	1.01	0.57	1.65	2.57	59.66	71.95	1.05	2.07	2.07	4.08
ℓ_∞	0.24	0.21	0.41	0.50	2.02	2.17	0.06	0.14	0.015	0.031
pixels	1022.72	137.91	214.62	282.03	1011.65	908.60	1021.44	1022.35	1022.45	1022.77
IS	0.9976	0.9684	0.9972	0.9990	0.9949	0.9919	0.9812	0.9972	0.9827	0.9980
WideResNet-Clean Acc 91.3										
Acc	35	75.62	38.19	18.3	54.3	49.7	67.9	36.3	62.9	28.5
ℓ_2	1.65	0.86	2.03	2.73	44.12	52.73	0.5	1.0	0.85	1.69
$\|\cdot\|_{S1}$	1.01	0.56	1.63	2.57	58.97	71.23	1.06	2.08	2.09	4.12
ℓ_∞	0.25	0.21	0.41	0.50	2.00	2.17	0.06	0.12	0.015	0.031
pixels	1022.63	136.75	204.02	271.92	1007.25	905.05	1022.76	1022.1	1023.02	1022.31
IS	0.9975	0.9737	0.9977	0.9990	0.9948	0.9930	0.9817	0.9976	0.9872	0.9983

where MNIST samples have 784 pixels and CIFAR-10 samples have 1024 pixels. FWnucl-group significantly reduces the number of pixel changes for FWnucl, while showing a higher success rate compared to the other sparse attacks. It is shown that FWnucl-group can produce sparse attacks that outperform the other attacks in terms of sparsity. For instance, for MNIST dataset the group norm reduces the number of pixels for FWnucl-group to the 8th fraction of that of FWnucl while decreasing the rate of success only by 0.7%. The results from these tables show that Auto-attack perturbs almost all the pixels for the successful adversaries and generates large distortion in terms of nuclear norm for Auto-attack-ℓ_∞.

We leverage the adversarial saliency map (ASM) (Papernot et al., 2016) to evaluate the effect of structured adversarial perturbation generated by FWnucl on image classification. For this purpose, we introduce the IS metric based on ASM score. For any d-dimensional image $x \in \mathbb{R}^d$, we let $\mathrm{ASM}(x,t) \in \mathbb{R}^d$ denote the adversarial saliency score corresponding to label t. In particular, the i-th element of $\mathrm{ASM}(x,t)$ shows how much the classification score with respect to label t increases and how much with respect to the original label decreases if a modification is applied to the i-th pixel value. We define a Boolean map $\mathbf{B}_{\mathrm{ASM}}$ corresponding to ASM score by

$$\mathbf{B}_{\mathrm{ASM}}[i] = \begin{cases} 1 & \text{if } \mathrm{ASM}(x,t)[i] \geq \nu \\ 0 & \text{otherwise} \end{cases}$$

where ν is the threshold on the pixel value. We then define the IS metric using $\mathrm{IS}(\delta) = \frac{\mathbf{B} \circ \delta}{\|\delta\|}$ where $\circ$ is the element-wise product. The logic behind this definition is that if $\mathrm{IS}(\delta)$ is close to 1 then the perturbation is applied mostly for the pixels which are critical for class prediction by the model, and by contrast the IS scores close to zero can not be interpreted based on ASM scores. We also listed IS scores in in Tables 7 and 8 where ν equals to the 30 percentile of ASM scores across the entire samples. In Fig. 7 we have shown the IS metric for the attacks in Table 8 versus ν, being different percentiles of ASM score across the samples of CIFAR-10 dataset. The figure shows that FWnucl-group has higher IS scores compared to the other sparse attacks and compared to FWnucl.

In Fig. 8 we display the adversarial images and the corresponding perturbation generated by the nuclear norm versus group nuclear norm $\|\cdot\|_{\mathcal{G},1,S(1),w}$, where the weights w are calculated based on the local variance of each group. The figures show that FWnucl-group creates perturbations that are more targeted and are localized to groups of pixels around the objects which are important for the classifier to make the prediction.

It is important to characterize the type of deformation that arise with radii above the applied thresholds as the imper-

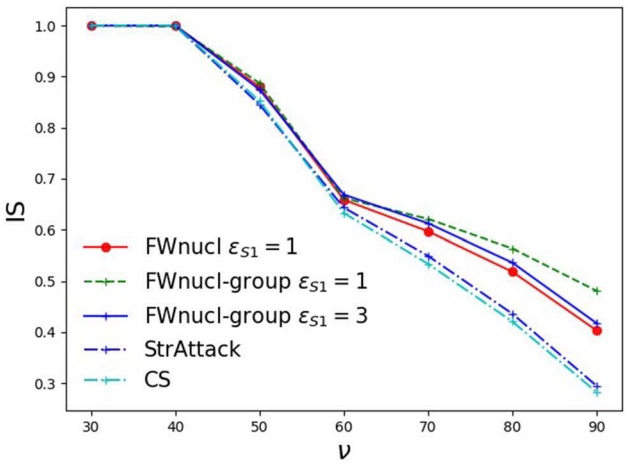
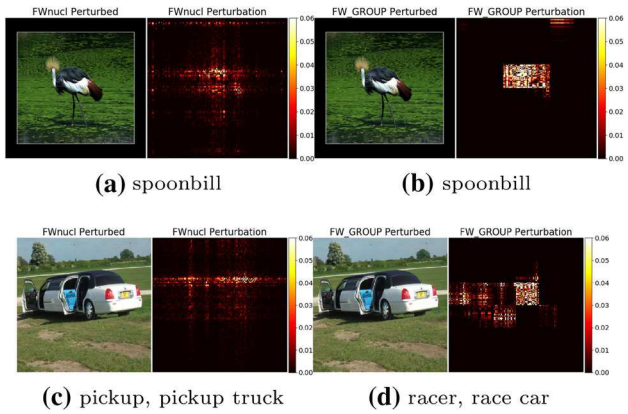


Fig. 7 IS metric vs ν, computed from the 30th percentile to the 90th percentile of ASM scores for CIFAR-10 dataset



(a) spoonbill **(b)** spoonbill

(c) pickup, pickup truck **(d)** racer, race car

Fig. 8 The images display some structural patterns of FWnucl (left) versus FWnucl-group (right) perturbations for the ImageNet dataset on DenseNet121 architecture for the nuclear ball of radius 5. Observe that the blurriness effect and perturbed pixels for images crafted by FWnucl-group are localized and restricted to some specific groups of pixels

Table 9 Evaluation of FWnucl against certifiably robust classifiers trained by randomized smoothing over CIFAR-10 dataset

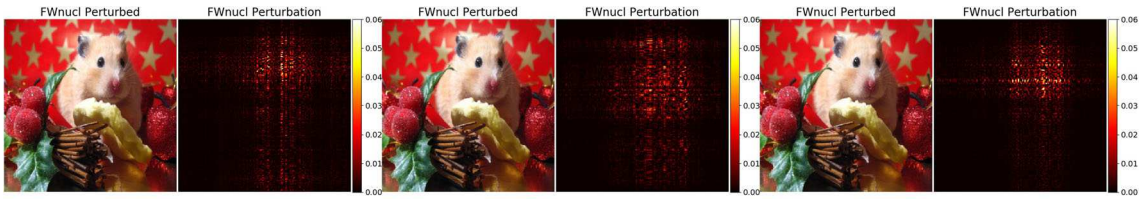
Network	Certified-l_2 radius FWnucl-l_{S_1} radius	0.25 0.25	0.5 0.5	0.75 0.75
ResNet-18	Certified	52.26	39.94	27.49
	Adversarial	48.66	38.37	28.28
WideResNet	Certified	55.9	28.97	16.83
	Adversarial	53.89	41.65	30.75
ResNet-50	Certified	50.04	37.11	24.77
	Adversarial	49.77	38.38	27.95

Certified (accuracy) denotes the certifiable accuracy of smoothed classifier (Cohen et al., 2019). Adversarial (accuracy) is the accuracy under FWnucl attack with the perturbation specified in the table

Table 10 Fooling rates of FWnucl adversarial perturbations between several models for 4000 samples from ImageNet

Source model/target model	ResNet-18	DenseNet121	GoogLeNet
ResNet-18	100	18.15	12.91
DenseNet121	16.56	99.30	11.74
GoogLeNet	15.03	12.37	99.40

The row indicates the source model and the column indicates the target model



(a) ResNet-18 **(b)** DenseNet121 **(c)** GoogLeNet

Fig. 9 General layout of the FWnucl perturbations for ImageNet across three different architectures

Table 11 Adversarial training using FWnucl norm for the MNIST dataset

Clean	FWnucl		PGD_2		Auto-attack-ℓ_2		PGD		Auto-attack-ℓ_∞	
	$\epsilon_{S1}=1$	$\epsilon_{S1}=3$	$\epsilon=0.5$	$\epsilon=1$	$\epsilon=0.5$	$\epsilon=1$	$\epsilon=0.1$	$\epsilon=0.2$	$\epsilon=0.1$	$\epsilon=0.2$
LeNet—FWnucl $\epsilon_{S1}=1.5$										
99.20	96.85	74.82	97.33	93.76	96.99	90.76	94.08	79.85	90.82	22.97
LeNet—FWnucl $\epsilon_{S1}=3$										
99.1	97.26	77.5	97.54	93.94	96.77	92.24	95.15	74.01	92.77	46.96
LeNet—PGD_2 $\epsilon=1.5$										
99.31	97.64	55.68	98.35	96.41	98.24	95.68	96.7	64.66	95.95	50.48
LeNet—PGD $\epsilon=0.3$										
98.38	95.08	86.04	95.16	93.38	93.91	82.94	97.41	95.96	96.91	93.67
SmallCNN—FWnucl $\epsilon_{S1}=1.5$										
99.23	97.84	90.57	98.21	96.82	97.19	92.68	97.12	94.06	95.23	55.79
SmallCNN—FWnucl $\epsilon_{S1}=3$										
98.81	97.32	83.22	97.6	94.66	97.19	89.84	95.49	73.99	92.88	47.15
SmallCNN—PGD_2 $\epsilon=1.5$										
99.19	97.4	71.2	98.37	96.53	98.16	95.46	96.83	69.24	96.09	52.05
SmallCNN—PGD $\epsilon=0.3$										
99.12	98.09	93.73	98.1	97.24	97.62	91.92	98.57	97.6	98.3	96.51

We used 20 iterations for adversarial training with FWnucl and the threat models FWnucl, PGD, PGD_2 are using 20 iterations with random initialization to converge

Table 12 Adversarial training using FWnucl norm for the CIFAR-10 dataset

Clean	FWnucl		PGD_2		Auto-attack-ℓ_2		PGD		Auto-attack-ℓ_∞	
	$\epsilon_{S1}=1$	$\epsilon_{S1}=1.5$	$\epsilon=0.5$	$\epsilon=1$	$\epsilon=0.5$	$\epsilon=1$	$\epsilon=2/255$	$\epsilon=4/255$	$\epsilon=2/255$	$\epsilon=4/255$
ResNet-18—FWnucl $\epsilon_{S1}=1.5$										
79.59	42.84	28.22	63.76	47.87	54.57	28.91	66.02	51.01	65.72	50.1
ResNet-18—PGD_2 $\epsilon=0.5$										
89.97	33.31	11.54	69.38	41.01	66.63	33.37	78.58	62.02	78.08	61.34
ResNet-18—PGD $\epsilon=8/255$										
81.25	38.04	19.14	63.56	41.05	58.38	27.36	74.97	67.68	73.06	65.12

We used 10 iterations for adversarial training with FWnucl and the threat models FWnucl, PGD, PGD_2 are using 10 iterations with random initialization to converge

ceptibility regimes are not the only restricted to the existing scenarios for generating adversarial examples. In particular accuracy of robust networks quickly drops to zero in the regimes above the seemingly imperceptible regions, see Fig. 3 and appendix for adversarial accuracy with various values for ϵ_{S1}. In the nuclear ball case, as the radius ϵ_{S1} of the nuclear ball increases, the perturbation becomes perceptible with a blurring effect. Structure in the adversarial examples can be leveraged to create specific perceptible deformation effects that look natural to humans.

Provably robust models with certifiable bounds.
We also evaluate the performance of the proposed adversary against the robust models with provable certified bounds. Table 9 shows that the accuracy of certified classifiers trained with randomized smoothing with standard deviation $\sigma = 0.5$. The results show that FWnucl can bring down the accuracy of the certifiably robust classifier up to the certified accuracy provided by provable defense methods for ResNet-18 and ResNet-50 models. It is also observed from Table 9 that there exists a gap between the certified and adversarial accuracy of WideResNet model.

Transferability.
In Table 10 we investigate the transferability of FWnucl adversarial examples over different architectures for ImageNet. This table shows that there should be some similar structural patterns between independent architectures that FWnucl employs, but the adversaries are mainly network dependent. In Fig. 9, we illustrate how the adversarial nuclear structure varies from one network to another for the same image; in particular, the perturbation continuously concentrates around the important regions of the image with however varying layouts and the patterns of perturbation for each network.

Adversarial Training.
In order to enhance the robustness of neural models to structured attacks, we adopt adversarial training using FWnucl adversarial attacks. We train models on MNIST and CIFAR-10 datasets with the architectures detailed in Tables 11 and 12 for MNIST and CIFAR-10 datasets, respectively, and we reported the robust accuracy with FWnucl and threat model with ℓ_2 and ℓ_∞ norms and compare the results with ℓ_p norm adversarially trained models for $p = 2, \infty$. From the tables, we see that the models adversarially trained with FWnucl show better to competitive performance versus FWnucl threat model compared with the other adversarially trained models, but nevertheless, they show competitive performance versus ℓ_p norm threat models for $p = 2, \infty$. The adversarial accuracy of our model over CIFAR-10 is almost $2.5x$ and $1.5x$ higher compared to the model trained with respectively ℓ_2 and ℓ_∞ norms against FWnucl adversarial attacks with $\epsilon_{S1} = 1.5$. This shows that adversarial training using FWnucl can reduce the success rate of our proposed nuclear attacks. In addition, the clean accuracy of robust models in Table 11 is showing that training using the augmented examples generated by FWnucl does not decrease the clean accuracy of the models significantly.

4 Conclusion

We consider adversarial attacks beyond the ℓ_p distortion set. Our proposed structured attacks allow an attacker to design imperceptible adversarial examples with specific characteristics, like localized blurriness. Furthermore, in the imperceptible regime, some defensive techniques may rely on a lack of certain structured patterns in the adversarial perturbations. Evaluating robustness against various structured adversarial examples then seems to be a reasonable defense sanity check. Our method is a competitor to the methods designed to craft sparse and targeted perturbations while maintaining success rates similar to powerful attacks like PGD.

References

Allen-Zhu, Z., Hazan, E., & Hu, W., et al. (2017). Linear convergence of a frank-wolfe type algorithm over trace-norm balls. In *Advances in neural information processing systems* (pp. 6191–6200).
Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. arXiv:1802.00420
Candès, E. J., & Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics, 9*(6), 717.
Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *2017 IEEE symposium on security and privacy (SP)* (pp. 39–57). IEEE.
Carlini, N., Athalye, A., & Papernot, N., et al. (2019). On evaluating adversarial robustness. arXiv:1902.06705
Chen, J., Yi, J., & Gu, Q. (2018). A frank-wolfe framework for efficient and effective adversarial attacks. arXiv:1811.10828
Cheung, E., & Li, Y. (2017). Projection free rank-drop steps. arXiv:1704.04285
Cohen, J. M., Rosenfeld, E., & Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. arXiv:1902.02918
Croce, F., & Hein, M. (2019). Sparse and imperceivable adversarial attacks. *In Proceedings of the IEEE international conference on computer vision* (pp. 4724–4732).
Croce, F., & Hein, M. (2020). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning, PMLR* (pp. 2206–2216).
Cui, J., Liu, S., & Wang, L., et al. (2021). Learnable boundary guided adversarial training. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 15721–15730).
Demyanov, V. F., & Rubinov, A. M. (1970). Approximate methods in optimization problems. In *Modern analytic and computational methods in science and mathematics*.
Dudik, M., Harchaoui, Z., & Malick, J. (2012). Lifted coordinate descent for learning with trace-norm regularization. In *Artificial intelligence and statistics* (pp. 327–336).

Dunn, J. C. (1979). Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals. *SIAM Journal on Control and Optimization, 17*(2), 187–211.

Engstrom, L., Tran, B., & Tsipras, D., et al. (2017). A rotation and a translation suffice: Fooling cnns with simple transformations. arXiv:1712.02779

Fazel, M., Hindi, H., & Boyd, S. P. (2001). A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American control conference (Cat. No. 01CH37148)* (pp. 4734–4739). IEEE.

Frank, M., & Wolfe, P. (1956). An algorithm for quadratic programming. *Naval Research Logistics Quarterly, 3*(1–2), 95–110.

Freund, R. M., Grigas, P., & Mazumder, R. (2017). An extended frank-wolfe method with "in-face" directions, and its application to low-rank matrix completion. *SIAM Journal on Optimization, 27*(1), 319–346.

Garber, D., & Hazan, E. (2013a). A linearly convergent conditional gradient algorithm with applications to online and stochastic optimization. arXiv:1301.4666

Garber, D., & Hazan, E. (2013b). Playing non-linear games with linear oracles. In *2013 IEEE 54th annual symposium on foundations of computer science* (pp. 420–428). IEEE.

Garber, D., & Hazan, E. (2015). Faster rates for the frank-wolfe method over strongly-convex sets. In *32nd International conference on machine learning, ICML 2015*.

Garber, D., Sabach, S., & Kaplan, A. (2018). Fast generalized conditional gradient method with applications to matrix recovery problems. arXiv:1802.05581

Gatys, L. A., Ecker, A. S., & Bethge, M., et al. (2017). Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3985–3993).

Gilmer, J., Adams, R.P., & Goodfellow, I., et al. (2018). Motivating the rules of the game for adversarial example research. arXiv:1807.06732

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *International conference on learning representations*.

Gragnaniello, D., Marra, F., Verdoliva, L., et al. (2021). Perceptual quality-preserving black-box attack against deep learning image classifiers. *Pattern Recognition Letters, 147,* 142–149.

Guélat, J., & Marcotte, P. (1986). Some comments on Wolfe's 'away step'. *Mathematical Programming*.

Guo, C., Frank, J. S., & Weinberger, K. Q. (2018). Low frequency adversarial perturbation. arXiv:1809.08758

Guo, Q., Juefei-Xu, F., & Xie, X., et al. (2020). Watch out! motion is blurring the vision of your deep neural networks. arXiv:2002.03500

Hameed, M. Z., & Gyorgy, A. (2021). Perceptually constrained adversarial attacks. arXiv:2102.07140

Harchaoui, Z., Douze, M., & Paulin, M., et al. (2012). Large-scale image classification with trace-norm regularization. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 3386–3393). IEEE.

Jaggi, M. (2013). Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th international conference on machine learning, CONF* (pp. 427–435).

Jaggi, M., & Sulovskỳ, M. (2010). A simple algorithm for nuclear norm regularized problems. In *ICML*.

Kerdreux, T., & d'Aspremont, A. (2020). Frank-wolfe on uniformly convex sets. arXiv:2004.11053

Kerdreux, T., Pedregosa, F., & d'Aspremont, A. (2018). Frank-wolfe with subsampling oracle. arXiv:1803.07348

Keskar, N. S., Mudigere, D., & Nocedal, J., et al. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. arXiv:1609.04836

Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial examples in the physical world. arXiv:1607.02533

Lacoste-Julien, S., & Jaggi, M. (2013). An affine invariant linear convergence analysis for frank-wolfe algorithms. arXiv:1312.7864

Lacoste-Julien, S., & Jaggi, M. (2015). On the global linear convergence of Frank–Wolfe optimization variants. In Cortes, C., Lawrence, N. D., Lee, D. D., et al (Eds.). *Advances in neural information processing systems* (Vol. 28, pp. 496–504). Curran Associates, Inc.

Langeberg, P., Balda, E. R., & Behboodi, A., et al. (2019). On the effect of low-rank weights on adversarial robustness of neural networks. arXiv:1901.10371

Lee, J.D., Recht, B., & Srebro, N., et al. (2010). Practical large-scale optimization for max-norm regularization. In *Advances in neural information processing systems* (pp. 1297–1305).

Levitin, E. S., & Polyak, B. T. (1966). Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics, 6*(5), 1–50.

Liu, H. T. D., Tao, M., & Li, C. L., et al. (2018). Beyond pixel norm-balls: Parametric adversaries using an analytically differentiable renderer. In *International conference on learning representations*.

Lu, M., Zhao, H., & Yao, A., et al. (2017). Decoder network over lightweight reconstructed feature for fast semantic style transfer. In *Proceedings of the IEEE international conference on computer vision* (pp. 2469–2477).

Luo, B., Liu, Y., & Wei, L., et al. (2018). Towards imperceptible and robust adversarial example attacks against neural networks. In *Thirty-second AAAI conference on artificial intelligence*.

Madry, A., Makelov, A., & Schmidt, L., et al. (2017). Towards deep learning models resistant to adversarial attacks. arXiv:1706.06083

Papernot, N., McDaniel, P., & Jha, S., et al. (2016). The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)* (pp. 372–387). IEEE.

Raghunathan, A., Steinhardt, J., & Liang, P. (2018). Certified defenses against adversarial examples. arXiv:1801.09344

Rauber, J., Brendel, W., & Bethge, M. (2017). Foolbox: A python toolbox to benchmark the robustness of machine learning models. In *Reliable machine learning in the wild workshop, 34th international conference on machine learning*. http://arxiv.org/abs/1707.04131

Reed, S.E., Akata, Z., & Mohan, S., et al. (2016). Learning what and where to draw. In: Advances in neural information processing systems (pp. 217–225).

Risser, E., Wilmot, P., & Barnes, C. (2017). Stable and controllable neural texture synthesis and style transfer using histogram losses. arXiv:1701.08893

Schmidt, L., Santurkar, S., & Tsipras, D., et al. (2018). Adversarially robust generalization requires more data. In *Advances in neural information processing systems* (pp. 5014–5026).

Sen, A., Zhu, X., & Marshall, L., et al. (2019). Should adversarial attacks use pixel p-norm? arXiv:1906.02439

Shalev-Shwartz, S., Gonen, A., & Shamir, O. (2011). Large-scale convex minimization with a low-rank constraint. arXiv:1106.1622

Sharif, M., Bauer, L., & Reiter, M. K. (2018). On the suitability of lp-norms for creating and preventing adversarial examples. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 1605–1613).

Stutz, D., Hein, M., & Schiele, B. (2019). Disentangling adversarial robustness and generalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6976–6987)

Tomioka, R., & Suzuki, T. (2013). Convex tensor decomposition via structured schatten norm regularization. In *Advances in neural information processing systems* (pp. 1331–1339).

Wang, Z., Bovik, A. C., Sheikh, H. R., et al. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing, 13*(4), 600–612.

Wong, E., & Kolter, J. Z. (2017). Provable defenses against adversarial examples via the convex outer adversarial polytope. arXiv:1711.00851

Wong, E., & Kolter, J. Z. (2020). Learning perturbation sets for robust machine learning. arXiv:2007.08450

Wong, E., Schmidt, F. R., & Kolter, J. Z. (2019). Wasserstein adversarial examples via projected sinkhorn iterations. arXiv:1902.07906

Wu, D., Xia, S. T., & Wang, Y. (2020). Adversarial weight perturbation helps robust generalization. arXiv:2004.05884

Xu, K., Liu, S., & Zhao, P., et al. (2018). Structured adversarial attack: Towards general implementation and better interpretability. arXiv:1808.01664

Yan, Z., Guo, Y., & Zhang, C. (2019). Subspace attack: Exploiting promising subspaces for query-efficient black-box attacks. arXiv:1906.04392

Yang, G., Duan, T., & Hu, E., et al. (2020). Randomized smoothing of all shapes and sizes. arXiv:2002.08118

Yang, Y., Zhang, G., & Katabi, D., et al. (2019). Me-net: Towards effective adversarial robustness with matrix estimation. arXiv:1905.11971

Zhang, H., Chen, H., & Xiao, C., et al. (2019). Towards stable and efficient training of verifiably robust neural networks. arXiv:1906.06316

Zhou, B., Khosla, A., & Lapedriza. A., et al. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2921–2929).