Is FGSM Optimal or Necessary for L_∞ Adversarial Attack?

Chaoning Zhang* chaoningzhang@kaist.ac.kr
Adil Karjauv* mikolez@gmail.com
Philipp Benz* pbenz@kaist.ac.kr
Soomin Ham smham@kaist.ac.kr
Gyusang Cho gyusang.cho@kaist.ac.kr
Chan-Hyun Youn chyoun@kaist.ac.kr
In So Kweon iskweon77@kaist.ac.kr
Korea Advanced Institute of Science and Technology (KAIST)

Abstract

Due to its simplicity and efficiency, the fast gradient sign method (FGSM) has been widely used in L_∞ norm-bounded adversarial attack. Its iterative variant I-FGSM has become the de facto standard practice of performing a strong attack but suffers from a low transfer rate. Momentum-based iterative FGSM, i.e. *MI-FGSM, is the first technique for boosting the transferability of I-FGSM. In this work, we identify two drawbacks of MI-FGSM: inducing higher average pixel discrepancy (APD) to the image as well as making the current iteration update overly dependent on the historical gradients. They increase the perturbation visibility as well as limit the potential of even higher transferability. We revisit why momentum improves the transferability and attribute it to alleviating the unreliable sign directions for the small gradient values. This unreliable sign direction problem occurs because the sign operation in FGSM maps all positive and negative gradient values to 1 and -1, respectively while ignoring their actual values. To this end, we propose a new momentum-free iterative method that processes the gradient with a generalizable Cut&Norm operation instead of a sign operation. In a wide range of attack setups, our approach consistently outperforms existing MI-FGSM by a large margin for white-box and black-box attacks in both non-targeted and targeted settings.*

1. Introduction

Shortly after the discovery of adversarial examples [1, 7, 27], Goodfellow *et al*. has proposed the fast gradient sign method (FGSM), a surprisingly simple yet effective adversarial attack method. Specifically, they linearize the cost function around its current network parameter, obtaining an optimal max-norm constrained perturbation through one-step backpropagation [7]. [18] introduced an iterative variant of FGSM, *i.e*. I-FGSM, for achieving a stronger attack. Due to its simplicity and effectiveness, I-FGSM has become the de facto standard practice of performing L_∞ norm-bounded adversarial attacks. It is widely reported that I-FGSM has a lower transfer rate than FGSM, and the reason is often attributed to the over-fitting effect. Momentum-based approach [5], which won the first place in NIPS2017 Adversarial Attack competitions for both non-targeted and targeted settings, is one of the earliest approaches to improve the transferability of I-FGSM without relying on an ensemble of surrogate models [21, 28]. MI-FGSM has thus become the basic framework for a line of works that improve the transferability of adversarial examples, resulting in a family of I-FGSM based transferable attacks, such as MI-DI-FGSM [32] and MI-TI-FGSM [6].

In this work, we revisit why MI-FGSM improves the transferability and complementary to prior work attributing it to stabilizing the gradient updates, we come up with an intuitive explanation that links it to the sign operation in FGSM. Specifically, the sign operation maps all positive and negative gradient values to 1 and -1, respectively, while ignoring their actual values. This is equivalent to amplifying the gradient for the gradient values that are very close to zero. In other words, their sign directions are less reliable and the momentum approach alleviates this problem via aggregating all the historical gradient values. This comes at the cost of a higher APD to the images, thus increasing the visibility to the human eye. Note that higher APD itself might also improve the transferability, but this is not the intended effect. Another drawback we identify in MI-FGSM is that the contribution of the current gradient to the final gradient update direction gets smaller and smaller in the perturbation generation process. Note that these two drawbacks are momentum-inherent, so we intend to address to attempt a momentum-free iterative gradient method by challenging the long practice of adopting FGSM, *i.e*. sign operation, in this field.

Intuitively, the actual gradient values at different pix-

*Equal contribution

els have implications on update direction and magnitude of the perturbation weight at the corresponding pixels; however, with the sign operation, FGSM only takes the direction into account while discarding the pixel-wise magnitude difference. One non-trivial challenge of reflecting the gradient values into the perturbation update with the L_∞ constraint taken into account is that the gradient values vary within an extensive range. To this end, we propose a simple yet intuitive method coined as Cut&Norm. Specifically, to make the perturbation more compatible with L_∞ constraint, we cut the perturbation gradient value to make its absolute value lower than a certain threshold and then normalize the resulted gradient values by its mean absolute value. Empirically, we find that keeping more original values is beneficial for boosting performance. One extreme case is to keep all gradient values proportional to their values, resulting in a more simple yet effective variant of our Cut&Norm gradient method, and we adopt it as our final approach. Contrary to the FGSM that discards all values, our final approach keeps all of them and automatically addresses the unreliable sign direction concern of FGSM. Naturally, the momentum is no longer necessary in our approach and might even decrease the performance due to its drawback of being overly dependent on historical gradient values. Overall, our approach outperforms existing MI-FGSM by a large margin in a wide range of attack setups. Our finding challenges the long-time established practice of adopting FGSM in L_∞ norm-bounded transferable adversarial attack.

2. Related works

Pioneering works in the field show the phenomenon of transferability on a wide range of white-box attacks, such as FGSM [7], I-FGSM [18], PGD [22], C&W [2]. This intriguing phenomenon has been partially attributed to the hypothesis of linear nature of modern DNNs [7], and this hypothesis has also been recently supported by the finding that backpropagating linearly [9] or relying more on the skip connections [31] improves transferability. [14] has also attributed transferability to similar non-robust features between models. Pixel interaction has also been recently found to provide a unified perspective on transferability [30]. On the other hand, towards improving the transferability, one line of work extended the I-FGSM attack [5, 6, 32], while another line of work has attempted to improve the transferability through fine-tuning adversarial examples on the intermediate features [10, 20]. Multiple works have also attempted to improve the transferability in the targeted setting through designing a new Poincaré ball distance loss. Optimizing the loss on the feature space has also been investigated in [15–17]. We highlight that they are all based on I-FGSM, and momentum is also by default adopted to improve the transferability. In contrast to them, our work investigates a new momentum-free iterative gradient processing method that is not based on FGSM. Since our approach just replaces the FGSM with a new gradient processing method, it can be directly plugged into the above methods. Recently, the transferability has also been exploited in [4, 11, 33] to achieve more query-efficient black-box attack [3, 8, 12, 13, 23, 24, 26, 29]. We refer the reader to the supplementary material for a detailed description of white-box and black-box attacks.

3. Methodology

Average Pixel Discrepancy (APD). In the field of transfer-based black-box attacks, the L_∞ norm is commonly adopted to make an adversarial example look natural. However, the $L_\infty \leq \epsilon$ constraint does not necessarily guarantee that different attacks have equivalent perturbation visibility. To illustrate this, we can consider two extreme cases of adversarial perturbations, one that only modifies a single pixel with either ϵ or $-\epsilon$ and the other one that changes all pixels with either ϵ or $-\epsilon$. Consequently, the induced change of the latter case would be more noticeable than the former one, despite both cases resulting in $L_\infty = \epsilon$. The reason can easily be attributed to the fact that the L_∞ constraint only limits the maximum change of each pixel without measuring their average change. To this end, we introduce an auxiliary metric termed average pixel discrepancy (APD), which is formally defined as the average of absolute changes that perturbation brings to the sample on all pixels. Under the $L_\infty \leq \epsilon$ constraint, the APD is guaranteed to be in the range of $[0, \epsilon]$. Within this range, however, a higher APD induces higher perturbation visibility. We refer the reader to the supplementary for an in-depth discussion of the APD.

3.1. Cut&Norm Gradient Method Based Momentum-free Iterative Approach

Intuitively, the unstable sign issue in the vanilla I-FGSM lies in the fact that sign operation is equivalent to *amplifying* the values close to zero while diminishing the gradient values far from zero. To avoid *amplifying* the values close to zero, we propose a new gradient method termed Cut&Norm, which is shown in Algorithm 1. It has the objective to keep a portion of the input gradient values proportional to original input gradient values while limiting the remaining values to a minimum or maximum value. Before detailing the algorithm, we point out two heuristic observations of the input gradients. First, the input gradient values are almost symmetric and centered around a mean value of close to 0. Second, the input gradient values cover a relatively wide range of values, with most of them concentrating near zero. Due to the widespread of input gradient values, it is non-trivial to set an appropriate threshold to cut off the gradient values; hence we choose the threshold with the percentile. Specifically, after calculating the absolute val-

ues of the input gradients, we calculate the p-th percentile of these values, resulting in the cut threshold value t_{cut} (See line 5). For example, considering $p = 20\%$ the 20-th percentile threshold value t_{cut} indicates the gradient value for which 20% of all absolute input gradients are smaller. Given the threshold value, all values below $-t_{cut}$ and above t_{cut} are set to $-t_{cut}$ and t_{cut}, respectively, which represents the *cut*-operation of our proposed algorithm. This is followed by the *norm*-operation, which linearly scales the cut gradients by the mean of the absolute of the cut gradient values (line 7). This operation serves the purpose of giving a fixed APD of the updated perturbation at each iteration. Note that the APD at each iteration is 1, which is somewhat equivalent to FGSM mapping all values to 1 or -1. It allows a fair comparison with the FGSM since they make the APD of the updated perturbation at each iteration the same.

Our Cut&Norm still resembles the philosophy of FGSM, and an astute reader can quickly find out that FGSM is a special case of our proposed method when the t_{cut} is set to a very small value. Empirically, we find that in general, the attack performance increases with a larger t_{cut}, suggesting the vanilla FGSM is not an optimal mapping function. Somewhat surprisingly, we empirically find that removing the cut, *i.e.* setting the t_{cut} threshold to an infinitely large value, results in the best performance in most cases, see the ablation study on cut ratio in the supplementary. Motivated by this observation, our proposed generalizable Cut&Norm approach can be simplified to a special variant that requires no need to cut the perturbation values, and we adopt it as our final approach. Contrary to FGSM, our final gradient method without the cut operation in essence is just a linear mapping function that fully keeps the gradients proportional to their original values. We note that this aligns well with the practice of updating the model weight proportional to their gradient values in modern network training. It is worth mentioning that the updated perturbation after each iteration is still clipped to be in the range of $[-\epsilon, \epsilon]$ to make it fulfill the L_∞ constraint.

Algorithm 1: Cut&Norm Algorithm

Input: Data $\mathcal{X}$, Classifier $\hat{C}$, Loss function $\mathcal{L}$, mini-batch size m, Number of iterations I, perturbation magnitude ϵ
Output: Perturbation vector δ
$\delta_1 \leftarrow 0$ ▷ Initialize perturbation
for $t = 1, \ldots, T$ **do**
  $g \leftarrow \nabla_v \mathcal{L}(x + \delta_t, y)$ ▷ Calculate input gradient
  $g^{abs} = |g|$
  $t_{cut} = \texttt{Percentile}\,(g^{abs}, p)$ ▷ Percentile
  $g_{cut} = \texttt{Clip}\,(g, -t_{cut}, t_{cut})$ ▷ Cut-operation
  $g_{cut}^{norm} = g_{cut} \,/\, \text{mean}(|g_{cut}|)$ ▷ Norm-operation
  $\delta_{t+1} = \delta_t + \alpha \cdot g_{cut}^{norm}$
  $\delta_{t+1} = \texttt{Clip}\,(\delta_{t+1}, -\epsilon, \epsilon)$
end

4. Experiments

4.1. Experimental Setup

Dataset and Networks. We generate transferable adversarial examples on a ImageNet-compatible dataset consisting of 1,000 images. Notably, this dataset has also been widely used in [5, 6, 32]. Following [19], we study 9 state-of-the-art pre-trained models on ImageNet [25]. Out of them, we have 6 normally trained models, namely Inception-v3 (Inc-v3), Inception-v4 (Inc-v4), Inception-Resnet-v2 (IncResv2), and Resnet-v2-50, 101, 152 (Res-50, 101, 152) and three adversarially trained networks, namely Inc-v3$_{ens3}$, Inc-v3$_{ens4}$, and IncRes-v2$_{ens}$.

Parameters. It is conventional practice for works in the field to conduct the same hyperparameter setting. For example, the maximum perturbation in [5, 6, 32] is all set to 16 out of 255 to make the perturbation invisible. Following [19], we set the step size $\alpha = \epsilon/T$ where the total iteration number T is 10. Following [5], for the momentum in MI-FGSM, the decay factor μ is set to 1, and the probability p of applying the stochastic input diversity in DI-FGSM is set to 0.7. The kernel length in TI-FGSM is set to 5 for normally trained models, while it is set to 15 for attacking adversarially trained models. For the Po-Trip loss, we follow [19] to set λ to 0.01 and margin γ to 0.007.

4.2. Transferable Targeted Attack

We mainly investigate the more challenging targeted setting. For a comprehensive comparison between our proposed method and FGSM, we report the attack success rate for the averaged attack success rate on each individual white-box model in the ensemble training as well as the attack success rate for the hold-out black-box model.

Attacking Standard Models. We first demonstrate the results when adversarial examples are crafted on an ensemble of models and are evaluated on standard models. Similar to [5, 19] we realize an ensemble of models through $l(x) = \sum_{k=1}^{K} w_k l_k(x)$, where K indicates the number of models in the ensemble, l_k refers to the logits of model k for input sample x, while l indicates the resulting ensemble logit values. The logit values of each model are weighted with w_k, for which we choose an equal weighting $w_k = 1/K$. We generate the transferable adversarial examples on an ensemble of 3 models and evaluate them in the white-box scenario, meaning on the ensemble network and the black-box setting, referring to a hold-out network. Note that the hold-out network is indicated as the column heading

Table 1. The ASR/APD of targeted FGSM-based attack and our method. We study four models—Inc-v3, Inc-v4, Res-152, and IncRes-v2, and adversarial examples are crafted via an ensemble of three of them. In each column, "-" denotes the hold-out model.

	Attacks	-Inc-v3	-Inc-v4	-Res-152	-IncRes	Avg.
Ensemble White-box	I	95.0 / 3.3	88.0 / 3.2	92.7 / 3.3	88.9 / 3.2	91.2 / 3.2
	MI	94.1 / 9.9	89.8 / 9.9	93.6 / 9.7	90.6 / 9.8	92.0 / 9.8
	Ours	**97.4** / 3.9	**95.1** / 3.8	**97.9** / 3.9	**94.4** / 3.8	**96.2** / 3.8
	DI	79.1 / 3.3	75.5 / 3.2	83.5 / 3.2	78.1 / 3.2	79.0 / 3.2
	MI-DI	74.8 / 10.1	74.5 / 10.1	80.4 / 10.1	76.9 / 10.1	76.7 / 10.1
	Ours	**91.4** / 4.0	**89.4** / 3.9	**95.3** / 3.9	**89.3** / 3.9	**91.4** / 3.9
	TI	93.6 / 3.4	87.5 / 3.3	92.7 / 3.3	88.3 / 3.2	90.5 / 3.3
	MI-TI	93.7 / 9.8	89.3 / 9.8	93.5 / 9.6	90.4 / 9.7	91.7 / 9.7
	Ours	**97.5** / 3.9	**94.9** / 3.9	**97.9** / 3.9	**94.4** / 3.8	**96.2** / 3.9
	TI-DI	78.7 / 3.3	75.5 / 3.3	83.8 / 3.3	78.4 / 3.3	79.1 / 3.3
	MI-TI-DI	75.1 / 10.1	74.3 / 10.1	79.9 / 10.1	76.2 / 10.1	76.4 / 10.1
	Ours	**91.6** / 4.0	**90.2** / 3.9	**95.2** / 4.0	**89.0** / 3.9	**91.5** / 4.0
	TI-DI-Po	78.6 / 3.4	74.2 / 3.3	79.2 / 3.3	76.6 / 3.3	77.2 / 3.3
	MI-TI-DI-Po	79.6 / 10.2	76.7 / 10.1	79.9 / 10.1	79.3 / 10.1	78.9 / 10.1
	Ours	**90.1** / 4.0	**88.0** / 3.9	**92.1** / 3.9	**88.7** / 3.9	**89.7** / 3.9
Hold-out Black-box	I	1.7 / 3.3	1.0 / 3.2	0.0 / 3.3	0.3 / 3.2	0.8 / 3.2
	MI	6.6 / 9.9	3.0 / 9.9	**1.5** / 9.7	3.0 / 9.8	3.5 / 9.8
	Ours	**7.8** / 3.9	**5.4** / 3.8	1.4 / 3.9	**5.2** / 3.8	**5.0** / 3.8
	DI	14.4 / 3.3	13.9 / 3.2	2.7 / 3.2	8.2 / 3.2	9.8 / 3.2
	MI-DI	25.2 / 10.1	23.9 / 10.1	**13.1** / 10.1	21.9 / 10.1	21.0 / 10.1
	Ours	**30.9** / 4.0	**31.1** / 3.9	9.3 / 3.9	**26.5** / 3.9	**24.4** / 3.9
	TI	2.0 / 3.4	1.8 / 3.3	0.1 / 3.3	0.4 / 3.2	1.1 / 3.3
	MI-TI	7.5 / 9.8	5.4 / 9.8	1.4 / 9.6	4.2 / 9.7	4.6 / 9.7
	Ours	**9.7** / 3.9	**7.9** / 3.9	**2.8** / 3.9	**6.0** / 3.8	**6.6** / 3.9
	TI-DI	17.1 / 3.3	16.8 / 3.3	3.3 / 3.3	9.9 / 3.3	11.8 / 3.3
	MI-TI-DI	27.0 / 10.1	27.0 / 10.1	**14.1** / 10.1	22.2 / 10.1	22.6 / 10.1
	Ours	**36.8** / 4.0	**34.6** / 3.9	11.7 / 4.0	**31.0** / 3.9	**28.5** / 4.0
	TI-DI-Po	21.9 / 3.4	20.1 / 3.3	5.2 / 3.3	13.5 / 3.3	15.2 / 3.3
	MI-TI-DI-Po	34.1 / 10.2	34.4 / 10.1	**17.6** / 10.1	30.7 / 10.1	29.2 / 10.1
	Ours	**45.1** / 4.0	**44.0** / 3.9	15.4 / 3.9	**36.3** / 3.9	**35.2** / 3.9

with a " – ". In each scenario, we report the transferability results comparing ours with I-FGSM, MI-FGSM in a wide range of setups, such as including, DI [32], TI [6], Po loss [19]. The results are available in Table 1. We observe that in all attack setups, MI-FGSM outperforms I-FGSM by a significant margin on the hold-out black models but with a much higher APD, which is expected with our aforementioned relevant discussion. Our approach outperfoms I-FGSM by a even larger margin, for examples improving the ASR from 0.8% to 5.0% in the vanilla attack setup and our APD is only slightly higher than that of I-FGSM. Compared with MI-FGSM, ours has a higher ASR with less than half APD. Overall, the results suggest that our approach is our momentum-free iterative is superior than existing momentum-free I-FGSM and momemtum-based MI-FGSM. It is also worth mentioning that our approach also achieves stronger attack on the white-box models.

Attacking Robust Models. Among various defenses against adversarial examples, adversarial training is arguably the most widely used. Since adversarial models can drastically reduce the effectiveness of adversarial examples, we further test the transferability of adversarial examples crafted on adversarially trained (robust) models. To craft the adversarial examples, we use an ensemble of two robust models and evaluate their transferability for the white-box and black scenario as for the standard models above. The results are presented in Table 2. Overall, the trend mirrors that for attacking standard models. Our momentum-free iterative approach consistently our-performs MI-FGSM in all setups by a significant margin. For example, in the strong transferable setup combining DI and TI, our momentum-free iterative approach improves the performance from 17.7% to 40.2%, yet staying less visible with a smaller APD.

Table 2. The ASR/APD of targeted FGSM-based attack and our method. We study nine models—Inc-v3, Inc-v4, Res-152, Res-101, Res-50, IncRes-v2, Inc-v3$_{ens3}$, Inc-v3$_{ens4}$ and IncRes-v2$_{ens}$, and adversarial examples are crafted via an ensemble of eight of them. In each column, "-" denote the hold-out model.

	Attacks	-Inc-v3$_{ens3}$	-Inc-v3$_{ens4}$	-IncRes-v2$_{ens}$	Avg.
Ensemble White-box	I	81.9 / 3.4	81.2 / 3.4	79.4 / 3.4	80.8 / 3.4
	MI	84.0 / 9.8	83.3 / 10.0	82.6 / 9.9	83.3 / 9.9
	Ours	**92.1** / 3.9	**92.4** / 4.0	**91.3** / 4.0	**91.9** / 4.0
	DI	64.4 / 3.3	62.5 / 3.3	62.4 / 3.3	63.1 / 3.3
	MI-DI	61.8 / 10.1	60.0 / 10.1	62.0 / 10.1	61.3 / 10.1
	Ours	**83.4** / 4.0	**82.7** / 4.0	**83.0** / 4.0	**83.0** / 4.0
	TI	61.8 / 4.3	61.4 / 4.3	63.3 / 4.3	62.2 / 4.3
	MI-TI	56.5 / 9.6	56.5 / 9.6	57.8 / 9.6	56.9 / 9.6
	Ours	**82.7** / 4.6	**82.2** / 4.6	**83.4** / 4.6	**82.8** / 4.6
	TI-DI	43.1 / 4.0	43.6 / 4.0	44.9 / 4.1	43.9 / 4.0
	MI-TI-DI	35.1 / 10.1	35.7 / 10.2	37.4 / 10.2	36.1 / 10.2
	Ours	**72.8** / 4.6	**73.4** / 4.6	**75.1** / 4.7	**73.8** / 4.6
	TI-DI-Po	47.3 / 4.1	47.6 / 4.1	48.9 / 4.1	47.9 / 4.1
	MI-TI-DI-Po	43.8 / 10.2	44.0 / 10.2	46.3 / 10.3	44.7 / 10.2
	Ours	**74.0** / 4.6	**74.5** / 4.6	**76.2** / 4.6	**74.9** / 4.6
Hold-out Black-box	I	0.0 / 3.4	0.0 / 3.4	0.0 / 3.4	0.0 / 3.4
	MI	0.0 / 9.8	0.0 / 10.0	0.0 / 9.9	0.0 / 9.9
	Ours	**0.0** / 3.9	**0.1** / 4.0	**0.1** / 4.0	**0.1** / 4.0
	DI	0.1 / 3.3	0.3 / 3.3	0.2 / 3.3	0.2 / 3.3
	MI-DI	0.9 / 10.1	1.0 / 10.1	0.8 / 10.1	0.9 / 10.1
	Ours	**1.3** / 4.0	**1.6** / 4.0	**2.0** / 4.0	**1.6** / 4.0
	TI	5.5 / 4.3	5.0 / 4.3	3.1 / 4.3	4.5 / 4.3
	MI-TI	14.4 / 9.6	12.3 / 9.6	10.5 / 9.6	12.4 / 9.6
	Ours	**24.0** / 4.6	**20.4** / 4.6	**18.3** / 4.6	**20.9** / 4.6
	TI-DI	12.2 / 4.0	11.7 / 4.0	9.5 / 4.1	11.1 / 4.0
	MI-TI-DI	14.8 / 10.1	13.7 / 10.2	13.7 / 10.2	14.1 / 10.2
	Ours	**38.2** / 4.6	**37.7** / 4.6	**35.3** / 4.7	**37.1** / 4.6
	TI-DI-Po	12.9 / 4.1	12.3 / 4.1	11.6 / 4.1	12.3 / 4.1
	MI-TI-DI-Po	18.6 / 10.2	19.1 / 10.2	17.7 / 10.3	18.5 / 10.2
	Ours	**41.2** / 4.6	**39.2** / 4.6	**39.9** / 4.6	**40.1** / 4.6

5. Conclusion

We identify that I-FGSM has the gradient sign unreliability issue and momentum in MI-FGSM alleviates it with a stabilization effect via aggregating all historical gradient values. MI-FGSM improves the transferability at the cost of higher APD to the image, making the perturbation update at the current iteration being overly dependent on the historical gradient values. Inspired by this, our work attempts a momentum-free iterative approach based on our proposed new gradient method termed Cut&Norm. Our momentum-free approach outperforms the existing momentum-based approach by a large margin in a wide range of attack setups while staying less visible with less than half APD. Our proposed simple gradient method can be easily applied to any FGSM-based attack for improving the performance.

References

[1] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, 2013. 1

[2] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Symposium on Security and Privacy (SP)*, 2017. 2

[3] 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 *ACM workshop on artificial intelligence and security*, 2017. 2

[4] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. *NeurIPS*, 2019. 2

[5] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018. 1, 2, 3

[6] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *CVPR*, 2019. 1, 2, 3, 4

[7] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 1, 2

[8] Chuan Guo, Jacob R Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Q Weinberger. Simple black-box adversarial attacks. *ICML*, 2019. 2

[9] Yiwen Guo, Qizhang Li, and Hao Chen. Backpropagating linearly improves transferability of adversarial examples. *arXiv preprint arXiv:2012.03528*, 2020. 2

[10] Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *ICCV*, 2019. 2

[11] Zhichao Huang and Tong Zhang. Black-box adversarial attack with transferable model-based embedding. *ICLR*, 2020. 2

[12] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *ICML*, 2018. 2

[13] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. *ICLR*, 2019. 2

[14] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *NeurIPS*, 2019. 2

[15] Nathan Inkawhich, Kevin J Liang, Lawrence Carin, and Yiran Chen. Transferable perturbations of deep feature distributions. *ICLR*, 2020. 2

[16] Nathan Inkawhich, Kevin J Liang, Binghui Wang, Matthew Inkawhich, Lawrence Carin, and Yiran Chen. Perturbing across the feature hierarchy to improve standard and strict blackbox attack transferability. *NeurIPS*, 2020. 2

[17] Nathan Inkawhich, Wei Wen, Hai Helen Li, and Yiran Chen. Feature space perturbations yield more transferable adversarial examples. In *CVPR*, 2019. 2

[18] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *ICLR*, 2017. 1, 2

[19] Maosen Li, Cheng Deng, Tengjiao Li, Junchi Yan, Xinbo Gao, and Heng Huang. Towards transferable targeted attack. In *CVPR*, 2020. 3, 4

[20] Qizhang Li, Yiwen Guo, and Hao Chen. Yet another intermediate-level attack. In *ECCV*, 2020. 2

[21] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *ICLR*, 2017. 1

[22] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 2

[23] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *ACM on Asia conference on computer and communications security*, 2017. 2

[24] Ali Rahmati, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Huaiyu Dai. Geoda: a geometric framework for black-box adversarial attacks. In *CVPR*, 2020. 2

[25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 3

[26] Yucheng Shi, Siyu Wang, and Yahong Han. Curls & whey: Boosting black-box adversarial attacks. In *CVPR*, 2019. 2

[27] 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. 1

[28] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *ICLR*, 2018. 1

[29] Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *AAAI*, 2019. 2

[30] Xin Wang, Jie Ren, Shuyun Lin, Xiangming Zhu, Yisen Wang, and Quanshi Zhang. A unified approach to interpreting and boosting adversarial transferability. *ICLR*, 2021. 2

[31] Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip connections matter: On the transferability of adversarial examples generated with resnets. *arXiv preprint arXiv:2002.05990*, 2020. 2

[32] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *CVPR*, 2019. 1, 2, 3, 4

[33] Ziang Yan, Yiwen Guo, and Changshui Zhang. Subspace attack: Exploiting promising subspaces for query-efficient black-box attacks. *NeurIPS*, 2019. 2