A Few Seconds Can Change Everything: Fast Decision-based Attacks against DNNs

Ningping Mou[†‡], **Baolin Zheng**[†§], **Qian Wang**[*‡], **Yunjie Ge**[‡], **Binqing Guo**[‡]
Wuhan University, Wuhan, China
{ningpingmou, baolinzheng, qianwang, yunjiege, binqingguo}@whu.edu.cn

Abstract

Previous researches have demonstrated deep learning models' vulnerabilities to decision-based adversarial attacks, which craft adversarial examples based solely on information from output decisions (top-1 labels). However, existing decision-based attacks have two major limitations, *i.e.*, expensive query cost and being easy to detect. To bridge the gap and enlarge real threats to commercial applications, we propose a novel and efficient decision-based attack against black-box models, dubbed FastDrop, which only requires a few queries and work well under strong defenses. The crux of the innovation is that, unlike existing adversarial attacks that rely on gradient estimation and additive noise, FastDrop generates adversarial examples by dropping information in the frequency domain. Extensive experiments on three datasets demonstrate that FastDrop can escape the detection of the state-of-the-art (SOTA) black-box defenses and reduce the number of queries by 13∼133× under the same level of perturbations compared with the SOTA attacks. FastDrop only needs 10∼20 queries to conduct an attack against various black-box models within 1s. Besides, on commercial vision APIs provided by Baidu and Tencent, FastDrop achieves an attack success rate (ASR) of 100% with 10 queries on average, which poses a real and severe threat to real-world applications.

1 Introduction

Recently, commercial deep learning models have suffered from adversarial attacks, which pose a severe threat to widely used DNN-based applications, especially in those security-sensitive areas such as automatic driving and intelligent health. Attackers can craft indistinguishing adversarial examples by adding small perturbations to mislead the models, even when the models' internal information is not accessible, as shown in Figure 1. Furthermore, to pose a real threat to these commercial models, attackers can merely utilize the exposed decisions (top-1 labels) from victim models to perform more practical attacks, *i.e.*, decision-based attacks. More specifically, the SOTA decision-based attacks, including GeoDA [Rahmati *et al.*, 2020] and HSJA [Chen *et al.*, 2020], have been proposed to explore the black-box space in the real-world scenario via a large number of queries.

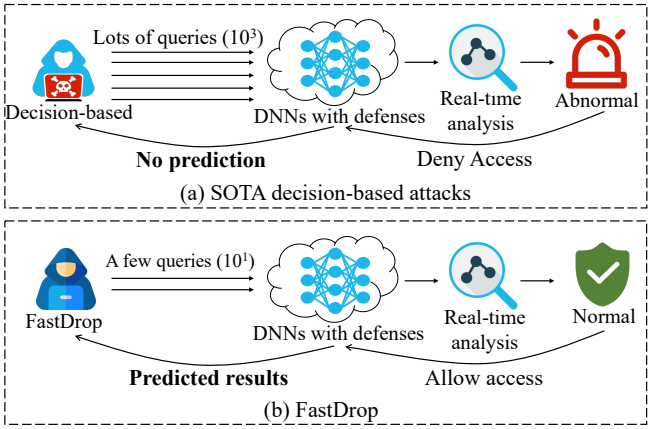
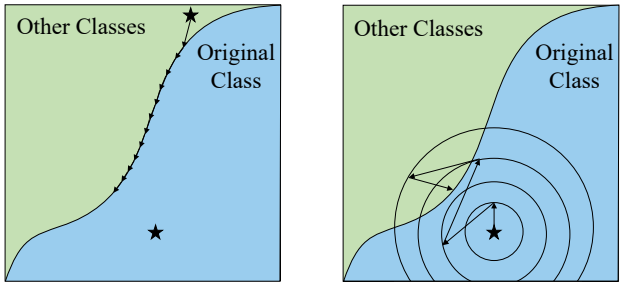


Figure 1: Comparison between other SOTA decision-based attacks and our FastDrop against DNNs with strong defenses.

Despite the excellent attack performance of prior attacks, they have two major weaknesses that limit their practicality against commercial applications, as shown in Figure 1 (a): 1) the expensive query cost. The SOTA attacks, such as GeoDA and HSJA, still require 10^3∼10^5 queries to generate an adversarial example. Specifically, previous works adopt the rejection sampling strategy [Brendel *et al.*, 2018] and gradient estimation [Chen *et al.*, 2020; Rahmati *et al.*, 2020] to carefully and gradually move the solution from a remote initial adversarial example to the vicinity of the original example along the decision boundary, as shown in Figure 2 (a). However, lots of queries hinder the effectiveness and timeliness

*Qian Wang is the corresponding author.
†The first two authors contributed equally to this work.
‡The Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, Hubei, China. This work was partially supported by the National Key R&D Program of China (2020AAA0107701), and the NSFC under Grants U20B2049 and U21B2018.
§School of Computer Science, Wuhan University, Wuhan 430072, Hubei, China.



(a) SOTA decision-based attacks (b) FastDrop

Figure 2: (a) SOTA decision-based attacks move along the decision boundary between adversarial and benign examples. (b) FastDrop moves with large steps from benign examples.

of attacks since the attacker needs to spend time and money on queries. 2) easy to detect. A large number of queries not only increase the cost of time and resources, but also raise the potential risk of exposure. A recent study, Blacklight [Li *et al.*, 2020], showed that query-based adversarial attacks are easy to detect since such attacks produce too many similar queries. Based on the similarity, Blacklight can detect adversarial examples with high rates in the early stage, and defend current SOTA decision-based attacks with the rate of nearly 100%. These limitations motivate us to design a more effective, practical, and robust attack as shown in Figure 1 (b).

To bridge the gap, we first propose a novel decision-based attack, namely FastDrop, to craft adversarial examples by dropping information in the frequency domain. As high-frequency information contributes to the classification of DNNs [Wang *et al.*, 2020a], dropping high-frequency information could cause the model to misclassify the images. Besides, compared with adding high-frequency noise, the search space of dropping information is finite. Searching adversarial examples in the smaller space may lead to fewer steps to craft an adversarial example. In this way, only a few steps can cause the misclassification of DNNs. Therefore, we craft adversarial examples by dropping high-frequency information with acceptable queries. Our attack is composed of two strategies: Orderly Frequency Dropping and Progressive Dual Backtrack. Orderly Frequency Dropping drops high-frequency information incrementally to push the example close to the decision boundary, as shown in Figure 2 (b). Progressive Dual Backtrack explores the complex dependency among high-frequency information to craft adversarial examples and minimize perturbations.

Extensive experiments demonstrate the superiority of FastDrop compared with the SOTA decision-based attacks. It can reduce the number of querie (NoQ) by 13$\sim$133$\times$ in around 0.9s on ImageNet when perturbations of the same level is considered. Besides, even under the strong black-box defense, *i.e.*, Blacklight, FastDrop could achieve an ASR of around 90%, while that of other SOTA attacks is nearly 0. Moreover, FastDrop can also attack commercial vision APIs with an ASR of 100% in around 7s. In summary, this paper makes the following contributions.

- We propose a new strategy, namely FastDrop, to conduct untargeted attacks under the decision-based scenario. To the best of our knowledge, this is the first attempt to generate adversarial examples by dropping information in the black-box setting.
- FastDrop is of superior efficiency compared with the SOTA decision-based attacks, which is validated by extensive experiments on three datasets and six models. With the same level of perturbations, FastDrop can reduce the NoQ by 13$\sim$133$\times$ compared with the SOTA attacks and achieve an ASR of 100% on ImageNet.
- FastDrop is robust to various defenses and effective against commercial vision APIs. It can escape the detection of the black-box defense, *i.e.*, Blacklight, with a success rate of 90% while other SOTA decision-based attacks fail. Besides, for traditional defenses, FastDrop is as effective as the SOTA ones. Furthermore, FastDrop can attack Baidu GOSR* and Tencent DL† with an ASR of 100% using around 10 queries.

2 Related Work

2.1 Adversarial Attacks

White-box attacks [Goodfellow *et al.*, 2015; Carlini and Wagner, 2017] are the early attempts of adversarial attacks. To overcome the constraint of black-box access, score-based attacks [Chen *et al.*, 2017; Bhagoji *et al.*, 2018; Guo *et al.*, 2019] are proposed. Then decision-based attacks [Brendel *et al.*, 2018; Cheng *et al.*, 2019] consider a more challenging and practical scenario where the target model only returns predicted labels without per-class probabilities, among which HSJA [Chen *et al.*, 2020] and GeoDA [Rahmati *et al.*, 2020] are the SOTA methods.

There is a very recent work, *i.e.*, AdvDrop [Duan *et al.*, 2021], considering the concept of dropping information to craft adversarial examples. But its white-box precondition makes it not practical in the real world. Besides, to reduce the NoQ, Shukla *et al.* [2021] proposed a Bayesian optimization-based attack (BOA) to search for adversarial examples in the structured low-dimensional subspace. BOA could craft adversarial examples within limited queries, but its ASR is relatively low, around 65%, whereas the ASR of HSJA and GeoDA are both near 100%. Conversely, FastDrop can generate adversarial examples with an ASR of 100% only using only a few (10$\sim$20) queries.

2.2 Adversarial Defenses

To mitigate the threat of adversarial attacks, many defensive strategies have been proposed, such as adversarial training [Madry *et al.*, 2018], feature squeezing [Xu *et al.*, 2018], JPEG compression [Shin and Song, 2017], and TWIS [Hu *et al.*, 2019]. However, these methods are not robust when facing strong black-box attacks. To effectively defend black-box attacks in real time, Li *et al.* [2020] proposed Blacklight, which is based on an insight that queries of an attack are similar while normal queries share little similarities. Based on this, it can detect adversarial examples of the SOTA attacks like HSJA and GeoDA with a rate of nearly 100% in the early

*https://cloud.baidu.com/product/imagerecognition/general
†https://cloud.tencent.com/product/tiia

stage. Empirically, Blacklight is probably the most effective defense against black-box attacks. However, FastDrop could bypass it and successfully conduct attacks with an ASR of around 90%.

3 Our Insight

Previous decision-based attacks consider incremental search in the spatial domain, which needs a lot of queries. But when the NoQ is limited, attackers should optimize the examples with fewer queries. Besides, due to the general goal of invisible perturbations [Carlini and Wagner, 2017; Brendel *et al.*, 2018], the perturbations should also be small.

How can we achieve these two seemingly contradictory goals at the same time? Recently, some studies [Guo *et al.*, 2019; Madry *et al.*, 2018; Xu *et al.*, 2018; Wang *et al.*, 2020b] considered perturbations in the high-frequency domain. This gives us an insight that we can manipulate adversarial examples in the high-frequency domain. Besides, Wang *et al.* [2020a] holds that high-frequency information contributes to the high accuracy of DNNs. Since the high-frequency information of images is important to the function of DNNs, if we drop some high-frequency information of benign examples, DNNs may make wrong predictions. Unlike adding perturbations, the scale of dropping information is limited, meaning that a few steps can drop all the information. In this way, we could constrain the queries to a limited number. Moreover, compared with low-frequency information, the change of high-frequency information does not affect the visual effect of examples too much, meaning that the perturbations are not too large. Therefore, we consider dropping high-frequency information as our solution.

4 FastDrop: Attack Strategy

4.1 Threat Model

Previous decision-based attacks [Brendel *et al.*, 2018; Chen *et al.*, 2020; Rahmati *et al.*, 2020] considered the black-box scenario where the attacker can only get predicted labels from the target DNN. Given an input x, the model returns only the predicted label y. However, with the intensifying arms race of adversarial attacks, strong defenses are proposed to detect malicious queries and limit the queries of suspectable users. We consider a more practical attack scenario of this kind, where the attacker can only use a limited number of queries, *e.g.*, 100. The attacker's goal is to craft an adversarial example x^* with a few queries based on the returned labels and escape the defenses. In this paper, we focus on untargeted attacks, meaning that the attacker's goal is to change the predicted label of an adversarial example to any incorrect label. We leave targeted attacks as our future work.

4.2 Overview

As shown in Figure 3, FastDrop consists of two phases: Orderly Frequency Dropping and Progressive Dual Backtrack.

- *Orderly Frequency Dropping(OFD).* We gradually set the blocks of the frequency spectrum to zero in order of importance. The loss of frequency information means that the useful information of the example to DNNs is less, thus leading to the misclassification of DNNs.

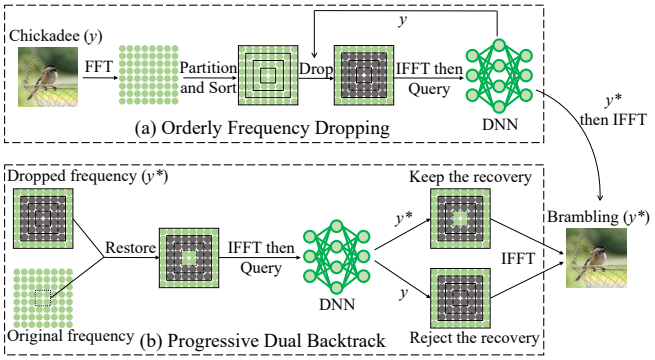


Figure 3: A simple illustration of two steps to generate an adversarial example by FastDrop. (a) OFD: FFT on a benign input and drop some blocks of the frequency spectrum in order. (b) PDB: restore some blocks while keeping the example adversarial.

- *Progressive Dual Backtrack(PDB).* We gradually restore the dropped blocks of the frequency spectrum in the reverse order of OFD while keeping the example adversarial. Restored blocks mean more similarities between the adversarial example and the original one, thus leading to smaller perturbations.

Besides, if the query limitation is not too restricted, we could adopt binary search between the adversarial example and the original one to reduce perturbations further. With a few additional queries, the perturbations will be smaller. The analysis is shown in Appendix B.2.

4.3 Orderly Frequency Dropping

In general, modifications of the structure will cause a larger influence on visual effect than that of color, to which human eyes are more sensitive [Chen *et al.*, 2021]. There are three mainstream transformations from the spatial domain to the frequency domain, *i.e.*, Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT), and Discrete Wavelet Transform (DWT). If we modify the frequency spectrum of DCT or DWT, the structural information will be affected. But the frequency spectrum of DFT is composed of the amplitude spectrum and the phase spectrum. If we only modify the amplitude spectrum of DFT without changing the phase spectrum, the structural information of the image will not be changed. Therefore, crafting adversarial examples by modifying the amplitude spectrum of DFT tends to cause smaller perturbations. We choose DFT and adopt a fast implementation of it, namely FFT, as it can effectively distinguish between high and low frequency. Formally, one-dimensional DFT and Inverse-DFT (IDFT) are given by

$$X(k) = DFT[x(n)] = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}kn}, \qquad (1)$$

$$x(n) = IDFT[X(k)] = \frac{1}{N} \sum_{n=0}^{N-1} X(k) e^{j\frac{2\pi}{N}kn}, \qquad (2)$$

where $x(n)$ is a spatial domain signal, and $X(k)$ is the corresponding frequency domain signal that could be divided into the amplitude spectrum and the phase spectrum.

Algorithm 1 Orderly Frequency Dropping

Input: Original input (x, y), target model f
Output: Adversarial input (x^*, y^*)

$F_{amplitude}, F_{phrase} \leftarrow FFT(x)$.
$\{b_1, b_2, ..., b_n\} \leftarrow$ split $F_{amplitude}$ into blocks.
$\{s_1, s_2, ..., s_n\} \leftarrow$Sort($\{b_1, b_2, ..., b_n\}$).
for $i = 1 : n$ **do**
  $s_i \leftarrow 0$.
  $x' \leftarrow IFFT(F_{amplitude}, F_{phrase})$.
  if $f(x') \neq y$ **then**
    $x^* \leftarrow x'$.
    break.
  end if
end for
return x^*

We firstly perform FFT on the input image to get the corresponding spectrums. Then the amplitude spectrum is split into centrosymmetric border-shaped blocks, as shown in Figure 3. As the amplitude spectrum of FFT is centrosymmetric, modifying the centrosymmetric block will not influence other blocks, which is therefore controllable. Specifically, the width of the border in each block is 1, that is, a 224×224 spectrum will be split into 112 blocks. For each block, we conduct global average pooling to get a value for sorting. In the frequency domain, small value is often corresponding to high-frequency information. Therefore, the blocks with minimal values are first set to zero. Then the modified amplitude spectrum and the phase spectrum are transformed to the spatial domain, outputting an image with small perturbations. After that, the image is used to query the model. If the prediction is different from its original label, we successfully craft an adversarial example. Otherwise, we will repeat the aforementioned steps and set more blocks to zero to drop information. The process is depicted in Algorithm 1. In practice, as the top sorted blocks contain too little information, we set more blocks to zero at each iteration in the early stage.

4.4 Progressive Dual Backtrack

Simply dropping information may result in too large perturbations, which is not acceptable for the requirement of invisibility. Besides, for some images, even after dropping all the information of the amplitude spectrum, though the probability is quite small, they are still not adversarial.

To solve these two problems, we propose Progressive Dual Backtrack, which is composed of two consecutive backtracks with the same procedure. A basic operation of the backtrack is shown in Figure 3 (b). For the amplitude spectrum of an image processed by OFD, some blocks are set to zero. We restore the values of these blocks according to the reversed order in OFD. After one block is restored, we check whether the corresponding image is adversarial. If not, we set this block to zero again. Otherwise, we keep the change and try to restore the next zero-value block, which was set to zero just before the previous restored block in OFD. After the first backtrack, we can restore a lot of blocks while keeping the image adversarial. In this way, the adversarial example and the original one are more similar, thus reducing perturbations. During this process, those images that cannot become adversarial by simply dropping information in OFD, although very few, will become adversarial, as the result of some unexplained dependencies. Besides, we conduct the backtrack once again, which could restore some extra blocks, further reducing perturbations.

5 Implementation and Evaluation

In this section, we conduct extensive experiments to validate the efficiency, robustness, and practicality of FastDrop.

5.1 Setup

Datasets and models. We use commonly-used ImageNet [Deng *et al.*, 2009], Flowers-102 [Nilsback and Zisserman, 2008], and STL-10 [Coates *et al.*, 2011] as our datasets. For each dataset, we randomly select 1000 images evenly from each class in our experiments. And two different models are used for each dataset to validate the generality of our method. For ImageNet, ResNet50 [He *et al.*, 2016] and MobileNetV3 [Howard *et al.*, 2019] are used. We use ResNet34 and VGG19 [Simonyan and Zisserman, 2015] for Flowers-102 and use ResNet18 and VGG16 for STL-10. As for small images like CIFAR10, we evaluate them in Appendix C.

Baselines. To show the superiority of FastDrop, we compare it with three SOTA attacks, including HSJA, GeoDA, and BOA. Specifically, GeoDA-F and GeoDA-S mean GeoDA with full space and GeoDA with subspace, respectively. BOA-100 and BOA-1000 represent BOAs with a budget of 100/1000 queries.

Defense mechanisms. We evaluate the performance of FastDrop against many defenses, including Blacklight, adversarial training, feature squeezing, pixel deflection, JPEG compression and TWIS.

Metrics. For efficiency, we set the median of perturbations, namely l_2, to the same level and compare ASR and NoQ. For robustness, we adopt the detection rate of defenses or ASR under defenses as evaluation metrics. For practicality, the metrics include ASR, l_2 and NoQ. Unless otherwise emphasized, l_2 refers to the median of perturbations measured by l_2-$norm$.

5.2 Efficiency Evaluation

FastDrop is compared with SOTA attacks including HSJA, GeoDA and BOA. To make fair comparisons, we carefully tune the hyper-parameters of these methods to achieve better results, and show the least queries under the l_2 constraint. The results are shown in Table 1. FastDrop could achieve l_2 of around 15.0 with only about 13 queries and an ASR of 100% on ImageNet. And the results of ResNet50 and MobileNetV3 are similar. To achieve the same level of l_2, HSJA and GeoDA spend at least 169.8 queries on average, which is 13.48 times as much as that of FastDrop. For GeoDA-F on ResNet50, the result is even much more surprising, which needs 1817.7 queries, 132.68 times as much as that of FastDrop. Results of Flowers-102 and STL-10 show similar phenomena. Therefore, we conclude that FastDrop is much more query-efficient than HSJA and GeoDA.

Attack	Metric	ImageNet		Flowers-102		STL-10	
		ResNet50	MobileNetV3	ResNet34	VGG19	ResNet18	VGG16
HSJA	ASR(%)	100.0	100.0	99.0	100.0	100.0	100.0
	l_2	15.1	15.4	18.8	15.7	11.6	12.0
	NoQ	1245.0	512.0	317.0	317.0	317.0	154.0
GeoDA-F	ASR(%)	100.0	100.0	100.0	100.0	100.0	99.2
	l_2	15.3	14.7	17.6	14.6	10.4	10.7
	NoQ	1817.7	582.6	363.8	360.3	482.2	272.9
GeoDA-S	ASR(%)	100.0	100.0	100.0	100.0	100.0	99.1
	l_2	15.0	**14.4**	18.3	17.4	**10.3**	**9.8**
	NoQ	453.3	169.8	218.0	127.5	179.8	273.0
BOA-1000	ASR(%)	67.4	60.8	59.8	71.1	73.3	79.7
	l_2	20.0	20.0	20.0	20.0	12.0	12.0
	NoQ	359.8	429.1	437.9	319.8	305.5	239.6
BOA-100	ASR(%)	59.6	52.9	51.5	64.6	64.6	72.4
	l_2	20.0	20.0	20.0	20.0	12.0	12.0
	NoQ	66.4	76.3	76.1	59.7	65.2	56.9
FastDrop	ASR(%)	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
	l_2	**15.0**	15.4	**15.9**	**14.0**	11.0	10.4
	NoQ	**13.7**	**12.6**	**14.7**	**13.0**	**20.6**	**13.1**

Table 1: Fundamental results of SOTA decision-based attacks and FastDrop on three datasets and six models. ASR: attack success rate. l_2: the median of perturbations measured by l_2-$norm$. NoQ: the number of queries needed to conduct an attack.

Dataset	Model	HSJA	GeoDA-F	GeoDA-S	BOA-1000	BOA-100	FastDrop
ImageNet	ResNet50	100.0	100.0	100.0	50.8	50.8	**4.0**
	MobileNetV3	100.0	100.0	99.9	51.5	51.5	**4.2**
Flowers-102	ResNet34	100.0	100.0	99.9	62.8	62.8	**1.6**
	VGG19	100.0	100.0	99.8	50.4	50.4	**1.9**
STL-10	ResNet18	99.9	100.0	100.0	48.6	48.6	**15.3**
	VGG16	99.5	100.0	100.0	40.2	40.2	**14.9**

Table 2: Performance of SOTA decision-based attacks and FastDrop under the defense of Blacklight. The metric is detection rate (DT): the proportion of black-box attacks that are detected before the attacks are completed.

Besides, BOA is conducted following the same settings in [Shukla *et al.*, 2021], where a maximal query budget of 1000 is configured on BOA (BOA-1000). Furthermore, as the NoQ of successful attacks is mostly under 100, to make fair comparisons, we also set the query budget of BOA to 100 (BOA-100). As shown in Table 1, although the average NoQ of BOA-100 is much less than that of HSJA and GeoDA, it is still several times more than that of FastDrop. Besides, the ASR of BOA-1000 and BOA-100 is less than 80%, whereas FastDrop achieves an ASR of 100%. Therefore, FastDrop outperforms BOA regarding both ASR and NoQ. Moreover, when conducting experiments of ResNet50 on a GeForce RTX 3080, BOA-1000 needs 207.15s to finish an attack of an image, while FastDrop only needs 0.90s. Thus, FastDrop is much more efficient than BOA. Furthermore, we validate our settings of FastDrop are effective in Appendix A and Appendix B.

5.3 Robustness Evaluation

In this section, we evaluate the performance of FastDrop when facing defenses on ResNet50. Black-box defense named Blacklight [Shukla *et al.*, 2021] and other traditional defenses are considered to comprehensively validate the robustness of FastDrop.

Black-box Defense

The emerging black-box defenses, such as Blacklight, showed that query-based black-box adversarial attacks are easy to detect. Blacklight is a strong defense to detect abnormal queries of current SOTA decision-based attacks with nearly 100% precision. Although those attacks could optimize the examples gradually based on the returned labels, if they are detected, the process will be stopped, preventing the attacks from being effective. We adopt the same settings of Section 5.2, and deploy Blacklight during the process of the attacks. The results are shown in Table 2. For HSJA and

Metric	ImageNet		Flowers-102		STL-10	
	Baidu GOSR	Tencent DL	Baidu GOSR	Tencent DL	Baidu GOSR	Tencent DL
ASR(%)	100.0	99.9	100.0	100.0	100.0	99.9
l_2	8.6	12.2	9.3	20.1	5.0	6.5
NoQ	5.1	18.5	7.3	15.7	3.7	6.7

Table 3: Performance of FastDrop on commercial vision APIs. Baidu GOSR: General Object and Scene Recognition Service of Baidu Intellignet Cloud. Tencent DL: Detect Label Service of Tencent Cloud.

GeoDA, nearly 100% of attacks are detected, meaning that they could not work when facing Blacklight. For BOA-1000 and BOA-100, the detection rates are the same. Although the detection rate of BOA is lower than 100%, it is still more than 40%. In comparison, the detection rate of FastDrop is nearly zero on ImageNet and Flowers-102. Therefore, FastDrop is of relatively higher robustness when facing Blacklight.

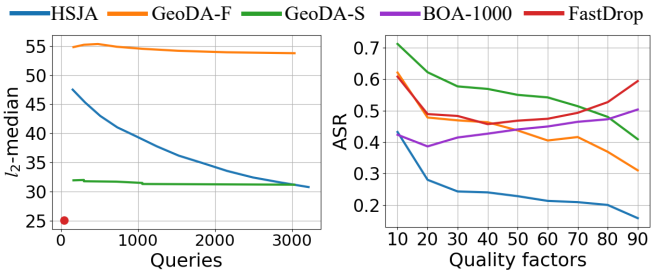
Traditional Defenses

We evaluate FastDrop under some traditional effective defenses, including adversarial training [Madry *et al.*, 2018], feature squeezing [Xu *et al.*, 2018], pixel deflection[Prakash *et al.*, 2018], JPEG compression[Shin and Song, 2017], and TWIS[Hu *et al.*, 2019]. We conduct same experiments on HSJA, GeoDA, and BOA to make fair comparisons. As BOA-1000 is stronger than BOA-100, we only show the result of BOA-1000.

We use a ResNet50 model pretrained with adversarial training from Madry *et al.* [2018] as the target model [‡]. The results are shown in Figure 4 (a). As BOA-1000 only achieves an ASR of 23.6%, we do not show it in Figure 4 (a). Specifically, all of these attacks achieve 100% ASR except for BOA-1000. Particularly, FastDrop only spends 43.4 queries to achieve l_2 of 25.1, whereas HSJA spends 3216 queries to get that of only 30.8. For GeoDA-F and GeoDA-S, the case is even worse, for they reduce perturbations too slowly. And after 3000 queries, l_2 is still high and close to that of the beginning. As for BOA-1000, it needs 773.3 queries on average to achieve an ASR of 23.6%. Therefore, FastDrop is far more effective compared with HSJA, GeoDA, and BOA-1000 when facing adversarial training.

The results of feature squeezing (FS), pixel deflection (PD), and TWIS are shown in Table 4. Specifically, we adopt two defenses in feature squeezing, namely Bit-4 and MF-3. Under different defenses, FastDrop can achieve performance close to the best. Besides, we also consider JPEG compression which could remove perturbations partly by DCT transformation and quantization. We set different quality factors of JPEG compression to comprehensively exhibit the robustness of FastDrop, shown in Figure 4 (b). Although FastDrop does not always perform best, it is relatively more stable. For HSJA and GeoDA, the ASR is larger with the decreased quality factor, the reason behind which is that smaller quality factors may cause more corruption instead of recovering the correct label. FastDrop shares the same phenomenon when the quality factor is less than 40, but its ASR increases when

[‡]https://github.com/MadryLab/robustness



(a) Robust model (b) JPEG compression

Figure 4: Performance of SOTA decision-based attacks and FastDrop under (a) Adversarial training and (b) JPEG compression.

Attack	No defense	FS		PD	TWIS
		Bit-4	MF-3		
HSJA	100.0	99.7	98.5	15.2	89.5
GeoDA-F	100.0	100.0	98.3	18.7	85.5
GeoDA-S	100.0	100.0	99.5	47.4	89.8
BOA	67.4	64.1	65.9	67.7	83.7
FastDrop	100.0	99.5	90.1	44.8	96.7

Table 4: ASR(%) of SOTA decision-based attacks and FastDrop under feature squeezing, pixel deflection and TWIS.

the factor increases beyond 40, which demonstrates that FastDrop is more robust and not easy to be corrupted when the quality factor is not too small. Based on the above results, we draw a conclusion that FastDrop is as strong as other SOTA decision-based attacks when facing traditional defenses.

5.4 Practicality Evaluation

In this section, we evaluate the performance of FastDrop against commercial vision APIs, including Baidu GOSR and Tencent DL, to validate its efficiency and practicality against real-world applications. We only use the top-1 hard labels to conduct attacks. Results are shown in Table 3, which demonstrate the extreme vulnerability of the commercial vision APIs, especially Baidu GOSR, to FastDrop. Specifically, FastDrop only requires several (about 10) queries to successfully perform an attack against commercial vision APIs at the seconds time-consuming level (around 7s), which poses a severe threat to real-world applications. Furthermore, we also conduct experiments of HSJA and Geoda-S to further validate the superiority of FastDrop. To achieve the same level of perturbations, HSJA and GeoDA-S require an average of 77×

and 62× as many queries as FastDrop, respectively. And the time consumed are 357.6s and 706.1s, unbearable and much longer than that of FastDrop. These results demonstrate the superior efficiency of FastDrop.

6 Conclusion and Future Work

In this paper, we propose a novel untargeted decision-based attack, namely FastDrop. By dropping high-frequency information, it can craft an adversarial example with only 10^1 queries and achieve an ASR of 100%. Compared with SOTA decision-based attacks, it greatly reduces the query budget. Besides, extensive studies validate the robustness of FastDrop against defenses, especially the SOTA black-box defense. Furthermore, it can attack commercial vision APIs with an ASR of 100% and a NoQ of around 10. Consequently, FastDrop is a strong attack against black-box DNNs and poses a severe threat to commercial vision APIs.

As for our future work, we would like to solve the problem of targeted attacks under secure-sensitive settings. We are also interested in crafting more invisible adversarial examples.

A Hyperparameter Analysis

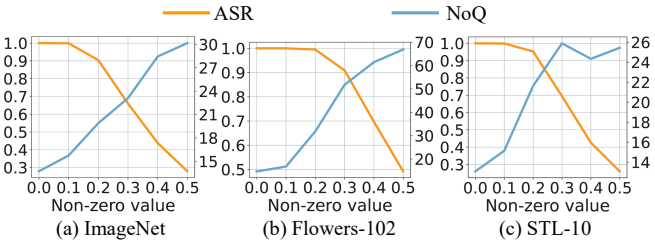
To validate our settings are optimal, we analyze the hyperparameters of FastDrop. After three aspects of analysis including order of the sorted blocks, non-zero modification and threshold of OFD, we demonstrate that our settings could achieve the best results.

A.1 Order of the Sorted Blocks

The order of sorted blocks plays an important role in optimizing adversarial examples. If FastDrop is combined with an unreasonable order, it may drop too many blocks but still can not generate adversarial examples. We test three different orders and report the results in Table 5. To comprehensively analyze the effect of the order, we also show the mean of perturbations. Specifically, O_1 refers to the reversed order of the default, that is, the blocks that affect the visual effect most will be first dropped. Besides, O_2 and O_3 are based on the location of the amplitude spectrum. O_2 refers to the order from the outer to the inner, which is the reversed version of O_3. The energy is mainly located at low-frequency areas that are composed of the outer blocks, meaning that the outer blocks have larger values in the amplitude spectrum. Therefore, O_1 and O_2 are similar, while O_3 and the default are similar, which explains the close results in Table 5. Based on the results of NoQ, l_2-median and l_2-mean, we can draw a conclusion that the default order is a better choice.

A.2 Non-zero Modification

After understanding the principles of FastDrop, one may wonder that directly setting blocks to zero may drop too much information and results in large perturbations. To validate the superiority of this strategy, we compare it with other settings: when dropping information, we multiply the values in the selected blocks and a decimal, namely non-zero value, from 0.1 to 0.5. The results are shown in Figure 5. The default is best with the smallest NoQ and the largest ASR. Besides, we also test the setting where we drop the block by multiplying each value in it with a random decimal. However, this setting needs around 114 queries to achieve an ASR of 68.5%, which is much worse than the default. Based on the above analysis, we validate the superiority of the default setting.

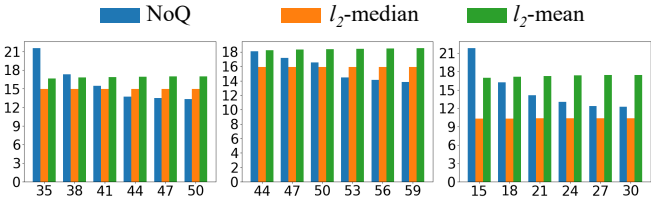


(a) ImageNet (b) Flowers-102 (c) STL-10

Figure 5: Analysis of non-zero modification.

A.3 Threshold of OFD

The threshold of OFD is an important hyper-parameter that decides whether to conduct PDB or not. It is the upper bound of perturbations in OFD. More specifically, it is a hyper-parameter predefined by balancing between the number of queries and perturbations. When the perturbations of the output of OFD are larger than the threshold, we perform PDB to further reduce the perturbations with additional queries. In our previous experiments, we set it to 44, 53 and 24 on three datasets respectively. In this section, we change it with the step of 3 around the default and report the results in Figure 6. Based on the results in Figure 6, we can find that the thresholds smaller than the default cause more queries, while the larger ones causes quite similar results to those of the default. Therefore, the default is the smallest threshold that leads to the best results.



(a) ImageNet (b) Flowers-102 (c) STL-10

Figure 6: Analysis of the threshold of OFD.

B Ablation Study

In this section, we evaluate the effectiveness of PDB and binary search. Specifically, binary search is not a component of FastDrop. We do an ablation study of it to validate its effectiveness in reducing perturbations.

B.1 Effectiveness of PDB

From previous experiments, we know that OFD plays a relatively more important role than PDB. Therefore, we abandon PDB and keep other settings the same to explore the effect of PDB. The results are shown in Table 6. Compared with

Order	ImageNet(ResNet50)			Flowers-102(ResNet34)			STL-10(VGG16)		
	NoQ	l_2-median	l_2-mean	NoQ	l_2-median	l_2-mean	NoQ	l_2-median	l_2-mean
O1	63.4	164.6	163.7	64.3	139.6	138.3	49.9	68.7	68.3
O2	63.4	164.6	163.5	64.3	139.6	138.5	49.8	68.7	68.3
O3	13.8	15.1	16.9	14.9	16.0	18.6	**13.0**	10.7	19.0
default	**13.7**	**14.9**	**16.9**	**14.7**	**15.9**	**18.4**	13.1	**10.4**	**17.4**

Table 5: Analysis of the order of the sorted blocks.

Dataset	Model	ASR(%)	NoQ
ImageNet	ReNet50	98.4	10.9
Flowers-102	ReNet34	98.8	12.9
STL-10	VGG16	85.8	4.7

Table 6: Performance of FastDrop without PDB.

Table 1, three metrics of the results are still acceptable, validating the effectiveness of OFD. Besides, the ASR in Table 6 is obviously lower than that of Table 1, especially for STL-10. This phenomenon demonstrates that when OFD can not generate an adversarial example for an image, the following PDB could make it with only a few queries. The unexplained dependency among blocks is explored by PDB, contributing to the final adversarial examples. In a nutshell, our method can work without PDB, but both OFD and PDB are necessary to achieve an ASR of 100%. In addition, we also validate that two backtracks is an appropriate setting. For ResNet50, the perturbations of the examples optimized by two backtracks are on average 38% smaller than those optimized by one backtrack. Besides, using more backtracks does not further reduce perturbations and cost more queries. Therefore, two backtracks is a better choice.

B.2 Ablation of Binary Search

In Section 4.2, we mention that binary search between an original example and the corresponding adversarial one in the frequency domain can reduce perturbations. To validate the function of binary search in reducing perturbations, we test it on ImageNet. With extra 9 queries, the binary search could reduce l_2 from 15.0 to 13.4 on ImageNet.

C Experiments on Small Images

After browsing the experiments of this paper, one may wonder if our method could achieve good performance on small images like those in CIFAR10, which is composed of RGB images of 32×32×3. In this section, we answer this question by testing FastDrop on a ResNet18 model with 94.8% top-1 accuracy on CIFAR10. We randomly select 1000 images evenly from each class as before. The results in table 7 demonstrate the superior performance of FastDrop. On CIFAR10, it still achieves much better performance regarding NoQ. One may doubt that the ASR of FastDrop is not 100%. Indeed, if the attacker wants to achieve an ASR of 100%, FastDrop may not be the best choice. But in most cases, an ASR of 97.5% is enough due to the small NoQ.

Attack	ASR(%)	l_2	NoQ
HSJA	100.0	2.4	694.0
GeoDA-F	100.0	2.7	87.2
GeoDA-S	100.0	3.0	156.6
FastDrop	97.5	**2.3**	**16.0**

Table 7: Fundamental results of SOTA decision-based attacks and FastDrop on CIFAR10.

D Visualization of FastDrop

We exhibit some randomly selected images to show the common results of FastDrop. As shown in Figure 7, adversarial examples generated by FastDrop are visually similar to benign images.

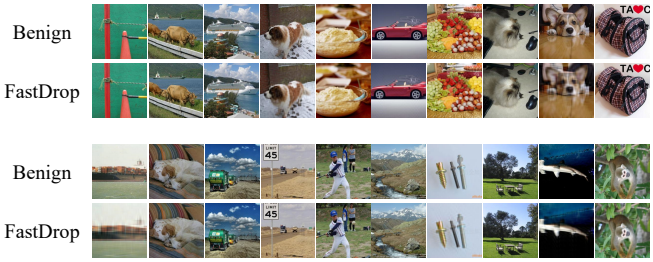


Figure 7: Common results of FastDrop.

References

[Bhagoji *et al.*, 2018] Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. Practical black-box attacks on deep neural networks using efficient query mechanisms. In *Proc. of ECCV*, pages 158–174, 2018.

[Brendel *et al.*, 2018] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *Proc. of ICLR*, 2018.

[Carlini and Wagner, 2017] Nicholas Carlini and David A. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proc. of ACM Workshop on Artificial Intelligence and Security*, pages 3–14, 2017.

[Chen *et al.*, 2017] 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 *Proc. of ACM Workshop on Artificial Intelligence and Security*, pages 15–26, 2017.

[Chen *et al.*, 2020] Jianbo Chen, Michael I. Jordan, and Martin J. Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *Proc. of IEEE S&P*, pages 1277–1294, 2020.

[Chen *et al.*, 2021] Guangyao Chen, Peixi Peng, Li Ma, Jia Li, Lin Du, and Yonghong Tian. Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In *Proc. of IEEE ICCV*, pages 458–467, 2021.

[Cheng *et al.*, 2019] Minhao Cheng, Thong Le, Pin-Yu Chen, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. In *Proc. of ICLR*, 2019.

[Coates *et al.*, 2011] Adam Coates, Andrew Y. Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proc. of AISTATS*, pages 215–223, 2011.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. of IEEE CVPR*, pages 248–255, 2009.

[Duan *et al.*, 2021] Ranjie Duan, Yuefeng Chen, Dantong Niu, Yun Yang, A. Kai Qin, and Yuan He. Advdrop: Adversarial attack to dnns by dropping information. In *Proc. of IEEE ICCV*, pages 7486–7495, 2021.

[Goodfellow *et al.*, 2015] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proc. of ICLR*, 2015.

[Guo *et al.*, 2019] Chuan Guo, Jacob R. Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Q. Weinberger. Simple black-box adversarial attacks. In *Proc. of ACM ICML*, pages 2484–2493, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of IEEE CVPR*, pages 770–778, 2016.

[Howard *et al.*, 2019] Andrew Howard, Ruoming Pang, Hartwig Adam, Quoc V. Le, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, and Yukun Zhu. Searching for mobilenetv3. In *Proc. of IEEE ICCV*, pages 1314–1324, 2019.

[Hu *et al.*, 2019] Shengyuan Hu, Tao Yu, Chuan Guo, Wei-Lun Chao, and Kilian Q. Weinberger. A new defense against adversarial images: Turning a weakness into a strength. In *Proc. of NeurIPS*, pages 1633–1644, 2019.

[Li *et al.*, 2020] Huiying Li, Shawn Shan, Emily Wenger, Jiayun Zhang, Haitao Zheng, and Ben Y. Zhao. Blacklight: Defending black-box adversarial attacks on deep neural networks. *arXiv preprint arXiv:2006.14042*, 2020.

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

[Nilsback and Zisserman, 2008] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proc. of IEEE ICVGIP*, pages 722–729, 2008.

[Prakash *et al.*, 2018] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James A. Storer. Deflecting adversarial attacks with pixel deflection. In *Proc. of IEEE CVPR*, pages 8571–8580, 2018.

[Rahmati *et al.*, 2020] Ali Rahmati, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Huaiyu Dai. Geoda: A geometric framework for black-box adversarial attacks. In *Proc. of IEEE CVPR*, pages 8443–8452, 2020.

[Shin and Song, 2017] Richard Shin and Dawn Song. Jpeg-resistant adversarial images. In *Proc. of NIPS Workshop on Machine Learning and Computer Security*, 2017.

[Shukla *et al.*, 2021] Satya Narayan Shukla, Anit Kumar Sahu, Devin Willmott, and J. Zico Kolter. Simple and efficient hard label black-box adversarial attacks in low query budget regimes. In *Proc. of ACM SIGKDD*, pages 1461–1469, 2021.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. of ICLR*, 2015.

[Wang *et al.*, 2020a] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P. Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proc. of IEEE CVPR*, pages 8681–8691, 2020.

[Wang *et al.*, 2020b] Zifan Wang, Yilin Yang, Ankit Shrivastava, Varun Rawal, and Zihao Ding. Towards frequency-based explanation for robust CNN. *arXiv preprint arXiv:2005.03141*, 2020.

[Xu *et al.*, 2018] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Proc. of NDSS*, 2018.