International Journal of Latest Trends in Engineering and Technology
Vol.(20)Issue(4), pp.001-012
DOI: http://dx.doi.org/10.21172/1.204.01
e-ISSN:2278-621X

AN EFFICIENT BLACK-BOX ADVERSARIAL ATTACK VIA TENSOR SINGULAR VALUE DECOMPOSITION

Qipeng Chen[1], Jianting Cao[1,2*]

Abstract- Machine learning (ML) models are playing an increasingly important role in daily life, such as automatic speech recognition, image classification, self-driving technology, etc. However, they are vulnerable to adversarial attacks such as convolution neural networks (CNN). Unlike the white-box attack, the black-box attack is practical but query-consuming to construct the adversarial images. In this paper, our method utilizes the following simple iterative principle: we decompose the original image by Tensor Singular Value decomposition (t-SVD), the noise tensor is randomly picked from the pre-specified set and then either add or subtract it to the Singular value tensor which is a rectangular diagonal data and its size is same as the original image but with much fewer value, therefore our method significantly reduces the query cost. From the experiment result, we demonstrate the efficacy and efficiency of the proposed method by fooling some widely used neural networks including Google Cloud Vision API.

Keywords- Black-box attack, Tensor Singular value decomposition, Adversarial attack, Untargeted and targeted attack

I. INTRODUCTION

Machine learning (ML) plays an essential role in our daily life and ML classifiers are used in many fields to finish the work of classification. For instance, a credit card fraud detector is a classifier taking the user's credit card transactions as inputs and identify which transactions are performed by the user and which are not. However, the safety of the model becomes an important topic for consideration. Adversarial attacks is to add a small perturbation to the input to misclassify the result and it is proved that the output of neural networks can be affected by small perturbation [1] [2]. There are two kinds of adversarial attacks, the white-box attacks and the black-box attacks. The white-box technology requires the attacker to know complete information about the target model, but there is no such restriction on black-box technology and it modified the perturbation according to the output of the previous query [3]. It seems that the output of most image classification models can be changed by white-box attacks [4] and the result indicates that after learning by ML classifiers, these image data are going to be close to decision boundaries. The white-box attack is an effective method to attack the target model because the attacker possesses the model's information, including its parameter values setting and training methods, etc. The white-box attack can be guided effectively with gradient descent [1] [5] and tends to have high query efficiency than black-box attack (the search for successful ResNet/ImageNet attacks require on the order of 10^4-$10^{\wedge 5}$ queries). But in most scenarios, it is impossible to acquire the information of the model. Hence black-box attack is more applicable for attackers [6] [7]. The number of queries is a vital indicator of the

[1] *Graduate School of Engineering, Saitama Institute of Technology, Japan*
[2] *Tensor Learning Unit, RIKEN Center for Advanced Intelligence Project (AIP), Japan*

efficiency of the attack algorithm. A low number of queries means less money and time cost for adversarial attacks. It is necessary to propose a query-efficient black-box attack method.

In order to improve the query efficiency, we propose a method that changes the objective of the adversarial perturbation attacks from the original image pixel data to another form with a smaller amount of data. Vector and matrix can be regarded as tensors, one-dimensional tensor is the vector, and the matrix is a two-dimensional tensor. If the dimension of data is greater than or equal to 3, it is considered a high-dimensional tensor. Videos, color pictures, and Magnetic Resonance Imaging (MRI) images are high-dimensional tensors. Preserve the original structure of high-dimensional tensor can obtain more spatial information from data processing by tensor method [8]. Tensor singular value decomposition [9] is one of the essential tensor methods and it is utilized to decompose the image data and it is an important tool to analyze data [10] [11], we can obtain low-rank (high value) parts and high-rank parts of the image. Some attack methods have been confirmed that the perturbation is roughly concentrated in the high-rank part and these attack methods can be easily defended by low-rank assumptions [12] [13]. In the proposed method, the perturbation is added to both the high-rank part and the low-rank part.

In this paper, we propose a simple and effective black-box attack method. Firstly, the original image is divided into two orthogonal tensors and one rectangular diagonal tensor by Tensor Singular Value decomposition (t-SVD). The noise tensor is added into the rectangular diagonal tensor to construct image perturbation. In order to improve the efficiency of the proposed method, we don’t have to pay too much attention to the optimal direction. Specifically, we randomly pick the noise tensor from specified sets and then attack the data by adding or subtracting the direction tensor into the singular value tensor. We utilize the confidence scores to check if the result is away from the decision boundary. The contributions of this paper are summarized as follows:

1. In this paper, we first try the tensor method in adversarial attack technology. The attacked image is processed by tensor singular value decomposition, and we add the noise tensor in singular value diagonal tensor to create perturbation instead of changing the pixel of original image with the same size. We also impose restrictions on noise tensor to generate less L_2 norm of the image.

2. We design a simple and fast algorithm to attack the targeted ML model by adding perturbation to images effectively. The noise tensor is randomly picked from the pre-specified set and then add or subtract it to the pre-acquired diagonal tensor.

3. The result shows that without adding the perturbation to the original image, our method achieves better query efficiency compared with the state-of-the-art method. We also attack different ML models to demonstrate the robustness of our method.

II. BACKGROUND

When constructing adversarial perturbation in image classification, the purpose is to change the output of the model predictions by adding imperceptible perturbation to original images. The perturbation should be restricted and they are imperceptible to humans. Generally, the same images should be classified into the same label and prediction, but the same images may have different outputs for machine learning classifiers. In this paper, we define the classifier model as h, and the image data as $\mathcal{X}$ with the model correctly predicts $y = \mathrm{h}(\mathcal{X})$, the purpose of the adversary attack is going to find a perturbed image $\mathcal{X}'$ to change the output:

$$\mathrm{h}(\mathcal{X}') = \mathcal{X}' \quad \boldsymbol{subject\ to} \quad \forall \mathcal{X}' \in \{\delta(\mathcal{X}, \mathcal{X}')\} \leq \rho \tag{1}$$

the $\delta(\mathcal{X}, \mathcal{X}')$ is the perceptual difference between the original and perturbed images, and it can be defined by the L_0, L_1 and L_∞. Following [14] [15], we choose $\delta(\mathcal{X}, \mathcal{X}') = \|\mathcal{X} - \mathcal{X}'\|_2$ as perceptual difference. For a successful adversarial attack algorithm, the perceptual difference should be as small as possible to the extent that the perturbed image is imperceptibly different.

- **Untargeted and targeted attack**

There are two different kinds of successful attack conditions. The simple one is the untargeted attack and it is defined as $\mathrm{h}(\mathcal{X}') \neq \mathrm{y}$, the objective of this attack is to change the output of original prediction. Another kind attack is targeted attack and it is represented as $\mathrm{h}(\mathcal{X}') = y'$, y' is an incorrect pre-chosen prediction of the model.

Adding adversarial perturbation to original data to change the output is a discrete optimization problem. Therefore it is necessary to define a surrogate loss $l_y(\cdot)$ to measure the degree between model $\mathbf{h}$ and output $\mathbf{y}$. The problem can be described as:

$$\min_{\delta} l_y(\mathcal{X} + \delta) \qquad \textbf{\textit{subject to}} \qquad \|\delta\|_2 < \rho \tag{2}$$

- **Attack models**

There are two kinds of attack models, they are white-box attacks and black-box attacks. If attackers know the information about classifier model h, back-propagation can be utilized on the target model because the model structure and parameter settings are exposed to the attacker. Gradient descent can be performed on the loss function $l_y(\mathcal{X}')$, $\mathbf{y}$ represents correct class.

In fact, for most real-world scenarios, attackers do not have information about the target model, white-box attacks are restricted to be applied. For black-box attacks, the most valid operation is to input the data to the model and get the corresponding output. The black-box attack method is much more practical for the adversary. For example, when we choose to attack Google Cloud Vision, it will cost time and money in each query, therefore in addition to remaining the perturbed image is imperceptible, minimize the number of queries should also be considered. The new optimization problem can be represented as:

$$\min_{\delta} l_y(\mathcal{X} \mid \delta) \qquad \textbf{\textit{subject to}} \qquad \|\delta\|_2 < \rho, \mathbf{queries} \leq \mathbf{B} \tag{3}$$

where $\mathbf{B}$ is the maximum of the queries we fix in the algorithm.

- **Tensor decomposition**

The color image is a 3-dimensional tensor, in order to keep its adjacent structure information for data, we introduce the tensor method to process the image data [16] [17]. Tensor methods have been applied more and more widely in the field of image processing. In the paper, the t-product $*$ is introduced to tensor calculation. The t-product of $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, and $\mathcal{B} \in \mathbb{R}^{n_2 \times n_4 \times n_3}$ is a tensor $\mathcal{C} \in \mathbb{R}^{n_1 \times n_4 \times n_3}$ is given by:

$$\mathcal{C} = \mathcal{A} * \mathcal{B} = \mathbf{Fold}(\mathbf{Circ}(\mathcal{A}) \times \mathbf{Vec}(\mathcal{B})) \tag{4}$$

where $\mathbf{Fold}()$ is an operation that takes $\mathbf{Vec}(\mathcal{B})$ into tensor $\mathcal{B}$ and it can be described as:

$$\mathbf{Vec}(\mathcal{B}) = \begin{bmatrix} \mathcal{B}^{(1)} \\ \mathcal{B}^{(2)} \\ \cdots \\ \mathcal{B}^{n_3} \end{bmatrix} \tag{5}$$

and $\mathbf{Circ}()$ is described as:

$$\mathbf{Circ}(\mathcal{A}) = \begin{bmatrix} \mathcal{A}^{(1)} & \mathcal{A}^{(n_3)} & \cdots & \mathcal{A}^{(1)} & \mathcal{A}^{(1)} \\ \mathcal{A}^{(2)} & \mathcal{A}^{(1)} & \mathcal{A}^{(n_3)} & \cdots & \mathcal{A}^{(3)} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathcal{A}^{(n_3)} & \mathcal{A}^{(n_3-1)} & \cdots & \mathcal{A}^{(2)} & \mathcal{A}^{(1)} \end{bmatrix} \tag{6}$$

Theorem 1 There is a tensor $\mathcal{A}$ with size $\mathbb{R}^{n_1 \times n_2 \times n_3}$, a tensor $\mathcal{B}$ with size $\mathbb{R}^{n_2 \times n_4 \times n_3}$, and a tensor $\mathcal{C}$ with same size with tensor $\mathcal{B}$, and they satisfy the commutative law:

$$\mathcal{A} * (\mathcal{B} + \mathcal{C}) = \mathcal{A} * \mathcal{B} + \mathcal{A} * \mathcal{C} \tag{7}$$

Theorem 2 If a tensor $\mathcal{A}$ with size $\mathbb{R}^{n_1 \times n_2 \times n_3}$, then we define the $\mathcal{A}^T$ by conjugate transposing each of the frontal slice of $\mathcal{A}$ and then reversing the order of transposed frontal slices 2 through n_3.

Theorem 3 A tensor $\mathcal{A}$ with size$\mathbb{R}^{n_1 \times n_2 \times n_3}$ is orthogonal, if it satisfies:

$$\mathcal{A}^T * \mathcal{A} = \mathcal{A} * \mathcal{A}^T = \mathcal{J} \tag{8}$$

Where $\mathcal{J}$ is identity tensor with size $\mathbb{R}^{n_1 \times 1 \times n_3}$ whose first frontal slice is identity matrix and other frontal slices are zero matrix.

Theorem 4 If $\mathcal{A}$ is an orthogonal tensor, the L_2 norm of $\mathcal{A} * \mathcal{B}$ can be denoted as:

$$\langle \mathcal{A} * \mathcal{B}, \mathcal{A} * \mathcal{B} \rangle = \langle \mathcal{B}, \mathcal{B} \rangle \tag{9}$$

For a color image data $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, , the t-SVD of $\mathcal{X}$ can be represented as:

$$\mathcal{X} = \mathcal{U} * \mathcal{S} * \mathcal{V}^T \tag{10}$$

where $\mathcal{U}$ and $\mathcal{V}$ are orthogonal tensors with size $\mathbb{R}^{n_1 \times n_1 \times n_3}$ and $\mathbb{R}^{n_2 \times n_2 \times n_3}$. $\mathcal{S}$ is the rectangular diagonal tensor with size$\mathbb{R}^{n_1 \times n_2 \times n_3}$. Although tensor $\mathcal{X}$ and tensor $\mathcal{S}$ have same size, $\mathcal{S}$ is a diagonal tensor and $\mathcal{X}$ is a tensor with full data, hence adding perturbation on tensor $\mathcal{X}$ is more efficient. The perturbed image can be formulated as $\mathcal{X}' = \mathcal{U} * (\mathcal{S}') * \mathcal{V}^T$, and the equation(3) can be rewritten as:

$$\min_{\delta} l_y(\mathcal{X}, \mathcal{X}') \qquad \boldsymbol{subject\ to} \qquad \|\delta\|_2 < \rho, \mathbf{queries} \leq \mathbf{B} \tag{11}$$

Theorem 5 For a tensor $\mathcal{A}$ with size$\mathbb{R}^{n_1 \times n_2 \times n_3}$, and the t-SVD of $\mathcal{X}$ is decomposed as $\mathcal{X} = \mathcal{U} * \mathcal{S} * \mathcal{V}^T$. The L_2 norm of $\mathcal{X}$ can be written as:

$$\langle \mathcal{X}, \mathcal{X} \rangle = \langle \mathcal{S}, \mathcal{S} \rangle \tag{12}$$

III. PROPOSED METHOD

- **Algorithm**

In this section, we are going to introduce our method. There are some original images, and we define them as $\mathcal{X}$. Through a neural network classifier model $\mathbf{h}$, the output of label $\mathbf{y}$ is classified with predicted confidence or probability $p_h(y|\mathcal{X})$. The proposed algorithm is to add perturbation δ to change the output $\mathrm{h}(\mathcal{X} + \delta) \neq \mathrm{y}$. Because we are blind to the model h, the output of each query $\mathrm{h}(\mathcal{X} + \delta)$ is valuable and exclusive information for us.

The algorithm is proposed in this section. Firstly, we decompose the original image $\mathcal{X}$ by t-SVD and the diagonal tensor $\mathcal{S}$ can be calculated, which is the objective to be attacked. In our method, we represent the noise tensor as $\mathcal{Q}$ and step size as ϵ, and the perturbation can be written as $\mathcal{U} * \alpha\mathcal{Q} * \mathcal{V}^T$. The perturbation will be added to the original image, if the output probabilities of image $p_h(y|\mathcal{X} + \delta)$ is decreasing, we consider the step of attack can be kept to the data $\mathcal{X}$ and next attack perturbation can be written as $\delta + \mathcal{U} * \alpha\mathcal{Q} * \mathcal{V}^T$, otherwise we subtract perturbation. If neither adding nor subtracting perturbation can reduce the probability of the result, we consider the step as an invalid attack and the perturbation will be discarded. The noise tensor $\mathcal{Q}$ is randomly picked from the set W.

The candidate diagonal tensor $\mathbb{W}$ can be comprised of some different kinds of basis tensor, they are the standard basis, random orthogonal diagonal basis and some specified diagonal basis. The first choice for the attack direction is the standard basis $\mathcal{I}$. Recent work has discovered that orthogonal noise is more likely to be adversarial [18]. The random diagonal basis attack is effective, but we found that compared with standard basis and random orthogonal diagonal basis, adding specific orthogonal diagonal basis noise into W will increase the efficiency of the attack and natural suitability to images [18]. In this paper, we prescribe each direction Q_i have two characteristics, the one is $\langle Q, Q\rangle = 1$ and another is $\langle Q_i, Q_{\neq i}\rangle = 0$.

Table 1.Algorithm

Simple black-box adversarial attacks by t-SVD

```
Input: Original image $\mathcal{X}$, query direction $Q$ that belong to sets $\mathbf{W}$, step size $\epsilon$.
$\mathcal{X} = \mathcal{U} * \mathcal{S} * \mathcal{V}^T$, $\delta = 0$
$P = P_y(y|\mathcal{X})$
if $P_y = max_{y'} P_{y'}$ do
    for $\alpha \in (0, \epsilon)$ do
        $P' = P_h(y|\mathcal{X} + \delta + \mathcal{U} * \alpha Q * \mathcal{V}^T)$
        if $P' < P_y$ then
            $\delta = \delta + \alpha Q$
            $P = P'$
            break
return $\delta$
```

- **Budget considerations**

Considering the sets of noise tensor W, we find that the L_2 norm of perturbation $\|\delta\|_2$ can be restricted. For each attack iteration, the noise tensor is either added or subtracted to the tensor $\mathcal{S}$. If neither adding nor subtracting can change the output probability, we discard the picked noise tensor in this iteration. In this paper, we define$\alpha \in (0, \epsilon)$ as the step size and after T iteration, the perturbation can be represented as:

$$\delta_T = \delta_t + \mathcal{U} * \alpha\mathcal{Q} * \mathcal{V}^T \tag{13}$$

the perturbation can also be rewritten as the sum of these each search directions:

$$\delta_T = \mathcal{U} * \textstyle\sum_{t=1}^{T} \alpha\mathcal{Q}_t * \mathcal{V}^T \tag{14}$$

and the L_2 norm of the adversarial perturbation can be written as:

$$\|\delta_T\|_2^2 = \langle \mathcal{U} * \textstyle\sum_{t=1}^{T} \alpha\mathcal{Q}_t * \mathcal{V}^T, \mathcal{U} * \sum_{t=1}^{T} \alpha\mathcal{Q}_t * \mathcal{V}^T \rangle$$

$$= \alpha_t^2 \langle \mathcal{U} * \textstyle\sum_{t=1}^{T} \mathcal{Q}_t * \mathcal{V}^T, \mathcal{U} * \sum_{t=1}^{T} \mathcal{Q}_t * \mathcal{V}^T \rangle \tag{15}$$

since t-product satisfy the **Theorem 1**, the right part $\langle , \rangle$ can be unfolded as:

$$\langle \mathcal{U} * \textstyle\sum_{t=1}^{T} \mathcal{Q}_t * \mathcal{V}^T, \mathcal{U} * \sum_{t=1}^{T} \mathcal{Q}_t * \mathcal{V}^T \rangle \tag{16}$$

$$= \langle \mathcal{U} * \mathcal{Q}_1 * \mathcal{V}^T + \mathcal{U} * \mathcal{Q}_2 * \mathcal{V}^T + \cdots \mathcal{U} * \mathcal{Q}_T * \mathcal{V}^T, \mathcal{U} * \mathcal{Q}_1 * \mathcal{V}^T + \mathcal{U} * \mathcal{Q}_2 * \mathcal{V}^T + \cdots \mathcal{U} * \mathcal{Q}_T * \mathcal{V}^T \rangle$$

we assume the formula $\mathcal{U} * \mathcal{Q}_1 * \mathcal{V}^T$ as a_1, $\cdots$ and $\mathcal{U} * \mathcal{Q}_T * \mathcal{V}^T$ as a_T , for any $i_1, i_2 \in [0, T]$, according to the matrix triple product operational rule, the equation can be transformed into:

$$\langle a_1, a_1 \rangle + \langle a_1, a_2 \rangle + \cdots + \langle a_{i_1}, a_{i_2} \rangle + \cdots + \langle a_T, a_T \rangle \tag{17}$$

according to theorem 5 and $\langle \mathcal{Q}_i, \mathcal{Q}_{\neq i} \rangle = 0$., for any $i_1 \neq i_2$ we have:

$$\langle a_{i_1}, a_{i_2} \rangle = \langle \mathcal{U} * \mathcal{Q}_{i_1} * \mathcal{V}^T, \mathcal{U} * \mathcal{Q}_{i_2} * \mathcal{V}^T \rangle = 0 \tag{18}$$

hence the equation(15) can be rewritten as:

$$\|\delta_T\|_2^2 = \alpha_t^2 \langle \mathcal{U} * \sum_{t=1}^{T} \mathcal{Q}_t * \mathcal{V}^T, \mathcal{U} * \sum_{t=1}^{T} \mathcal{Q}_t * \mathcal{V}^T \rangle$$

$$= \alpha_t^2 \textstyle\sum_{t=1}^{T} \langle \mathcal{Q}_t, \mathcal{Q}_t \rangle \leq T\epsilon^2 \tag{19}$$

Since $\mathcal{U}$ and $\mathcal{V}$ are constant tensors. From equation (15), we can find that ϵ is a vital parameter to restrict the perturbation. Meanwhile, we found that if the query is restricted, we can set ϵ higher to reduce the number of iterations, thereby obtaining a higher disturbance L_2norm. Otherwise, if small-norm solutions are proposed, restrict ϵ will require more queries in the same L_2 norm.

IV. EXPERIMENT AND RESULTS

In this section, we are going to demonstrate the efficiency of the method by fooling the convolutional neural network (CNN) models with three types of performance evaluation: the cost of queries (***B***), the L_2 norm of perturbation (***P***), and the rate of the optimization problem to find a feasible point (*success rate*). Meanwhile, we compare the proposed method with other black-box algorithms: the QL attack [19], the SimBA and the SimBA-DCT [18]. We use standard dataset: ImageNet [20]. Firstly, we randomly choose 1000 images from the ImageNet and then classify them with the correct label. In the experiment, we try to minimize the probability of the correct label in untargeted attacks and maximize the probability of the target label in targeted attacks, we limit the maximal $\mathrm{T} = 10000$.

- **Untargeted attack on google Cloud Vision**

For the untargeted attack, the purpose is to change the correctly labeled image into the incorrect label. In this experiment, we test our proposed method by attacking the Google Cloud Vision API, and Fig 1 shows its efficiency, we also compare our method with SimBA and SimBA-DCT. The result shows that our method ultimately achieves a relatively high success rate and our method increases dramatically faster in success rate than SimBA and SimBA-DCT.

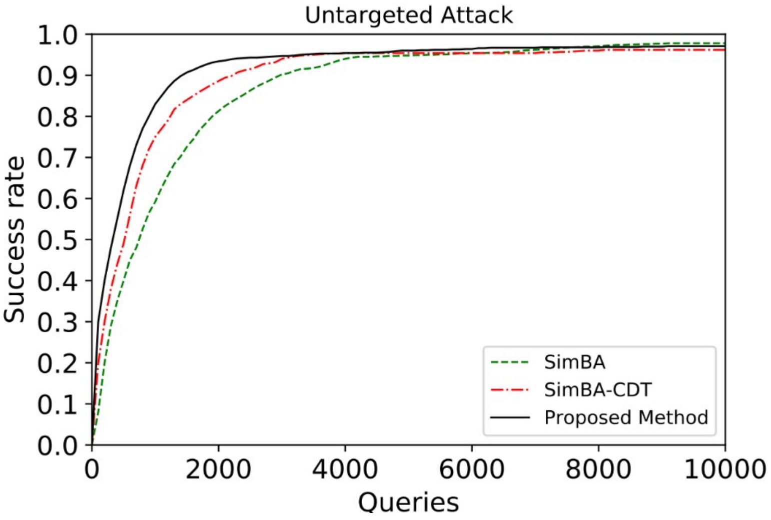

Fig. 1. *The success rate and the number of cost queries compared with SimBA, SimBA-DCT and our proposed method by untargeted attacks. The successrate of proposed method increases faster than SimBA and SimBA-DCT methods.*

- **Untargeted and targeted attack on ResNet-50**

Attack Method	Avg queries		Avg L_2 norm		Success rate	
	Untargeted	Targeted	Untargeted	Targeted	Untargeted	Targeted
QL-attack	28185	20857	8.54	11.48	85.7%	98.9%
SimBA	1957	7902	4.31	9.48	98.7%	100%
SimBA-DCT	1539	8759	3.89	7.08	97.4%	96.4%
Proposed	1207	5783	4.76	8.76	96.8%	97.8%

Table 2. Four attack methods are performed on ImageNet by the untargeted and targeted attack, and we choose three different metrics to evaluate the methods: the number of cost queries (lower is better), average L2-norm of average perturbation (lower is better), and success rate (higher is better). The proposed method achieves close to 98% success rate slightly lower than other methods but requires significantly fewer model queries.

In the second experiment, we test the performance of our method by attacking the ResNet-50 network [22] and compare it with QL-attack, SimBA and SimBA-DCT. Furthermore, untargeted attack and targeted attack are performed and the number of cost queries, success rate and average L_2 norm of perturbation is utilized to evaluate the performance of our method.

Ideally, we ensure that the success rate of each algorithm attack is as high as possible. We believe that the successful method constructs the perturbation with lower L_2 norm and the lower queries. From Table 2, we can find that our method has significantly lower queries than other methods. In the untargeted attack experiment, QL-attack only gets 85% but costs 28000 queries. Although compared to SimBA and SimBA-DCT, we do not achieve a higher success rate, but our method costs fewer queries. In the targeted attack experiment, the test methods are much more comparable, but our method still requires fewer queries than other methods.

- **The qualitative comparison of different methods**

In this part, we randomly selected several images to verify the qualitative results of different methods. In this experiment, we choose SimBA and SimBA-DCT for comparison. Figure 2 shows the original images and the attacked images, as well as the L_2 norm of adversarial perturbation of each image and the number of cost queries. All methods have successfully attacked the original image. Although our method cannot always achieve the smallest L_2 norm, the number of queries consumed by our method is significantly less than other methods.

- **Evaluating different networks**

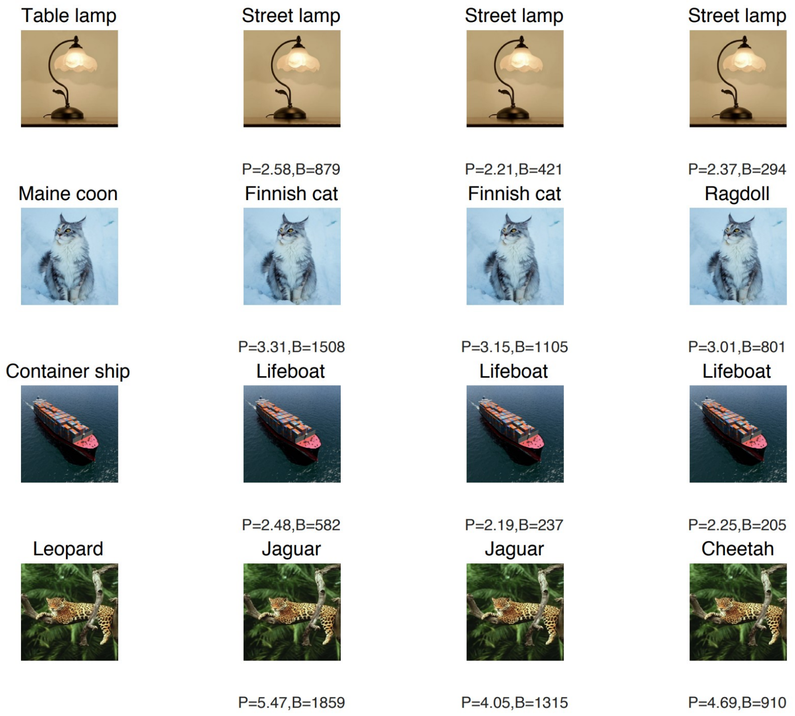

Fig. 2. *The first row of the figure is original image, the other rows are the result attacked by SimBA, SimBA-DCT, and the proposed method. P means the L2 norm of adversarial perturbation and B means the cost number of queries. Comparing SimBA and SimBA-DCT, our method cannot guarantee the lowest L2 norm of perturbation, but the number of queries is significantly less than the other two methods.*

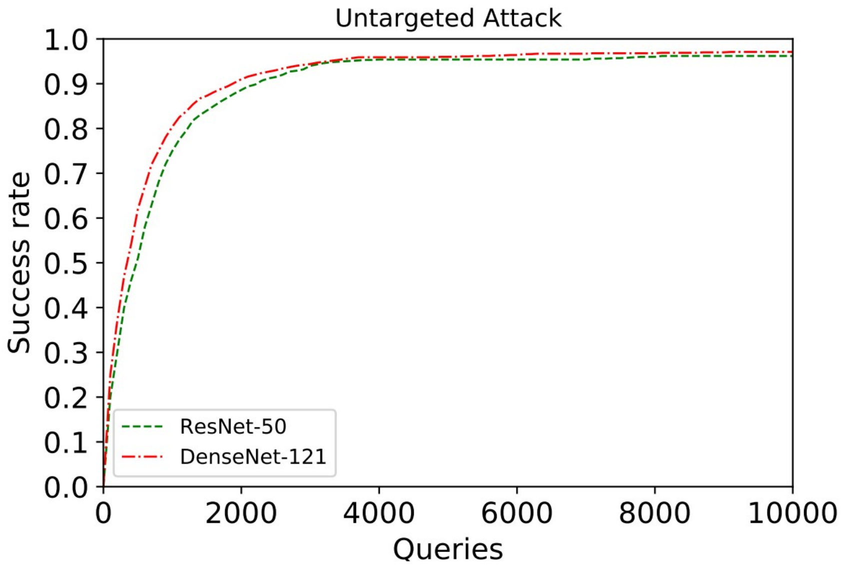

Fig. 3. *The success rate and the number of queries through ResNet-50 and DenseNet-121 models for untargeted attacks. Our method can fool both ResNet-50 and DenseNet-121 successfully within 10000 queries with high probability. Compared with ResNet-50 model, DenseNet is more vulnerable against untargeted attacks.*

In order to verify that our proposed method is also effective for other convolutional neural networks models, we choose DenseNet-121 [23] as our objective model for the untargeted attack. The result shows the success rate and the number of model queries with DenseNet-121 and ResNet-50 models. From Fig 3, we find that whether DenseNet-121 or ResNet-50 model are both vulnerable to our attack method, and DenseNet-121 model is trended to be fooled easier. From the experimental results, our method successfully attacks different CNN models with high probability.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we are the first to utilize the tensor method to construct adversarial perturbations. A simple and effective black-box algorithm is proposed. We use tensor singular value decomposition to process the image and add specific perturbation into the singular value tensor to create perturbation. Our attack method is not only effective for different CNN models, but also more efficient than other methods (our method has a higher success rate in the first 1000 queries).

In addition to the experiments introduced in this paper, we also performed some other experiments. In the experiment, we found that attacks on different positions of the singular value tensor, the perturbation had different characteristics. We believe that these characteristics have an impact on the output of the model. In the next research, we will conduct research on this characteristic to improve the efficiency of the algorithm.

Acknowledgments

The authors would like to thank Professor Qibin Zhao and Researcher Jianfu Zhang for their discussion and useful recommendation.

This work was supported by JST CREST Grant Number JPMJCR1784, JSPS KAKENHI (Grant No. 18K04178 and No.20H04249)

REFERENCES

[1] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks [J]. arXiv preprint arXiv:1312.6199, 2013.

[2] Biggio B, Corona I, Maiorca D, et al. Evasion attacks against machine learning at test time[C]//Joint European conference on machine learning and knowledge discovery in databases. Springer, Berlin, Heidelberg, 2013: 387-402.

[3] Cheng M, Le T, Chen P Y, et al. Query-efficient hard-label black-box attack: An optimization-based approach [J]. arXiv preprint arXiv:1807.04457, 2018.

[4] Athalye A, Carlini N. On the robustness of the cvpr 2018 white-box adversarial example defenses [J]. arXiv preprint arXiv:1804.03286, 2018.

[5] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks [J]. arXiv preprint arXiv:1706.06083, 2017.

[6] Guo C, Frank J S, Weinberger K Q. Low frequency adversarial perturbation [J]. arXiv preprint arXiv:1809.08758, 2018.

[7] Brendel W, Rauber J, Bethge M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models [J]. arXiv preprint arXiv:1712.04248, 2017.

[8] Chen Q P, Cao J T. Low tensor-train rank with total variation for magnetic resonance imaging reconstruction [J]. Science China Technological Sciences, 2021: 1-9.

[9] Zhang Z, Ely G, Aeron S, et al. Novel methods for multilinear data completion and de-noising based on tensor-SVD[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 3842-3849.

[10] Ely G, Aeron S, Miller E L. Exploiting structural complexity for robust and rapid hyperspectral imaging[C]//2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013: 2193-2197.

[11] Wright J, Ganesh A, Min K, et al. Compressive principal component pursuit[J]. Information and Inference: A Journal of the IMA, 2013, 2(1): 32-68.

[12] Ely G, Aeron S, Miller E L. Exploiting structural complexity for robust and rapid hyperspectral imaging[C]//2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013: 2193-2197.

[13] Entezari N, Papalexakis E E. TensorShield: Tensor-based defense against adversarial attacks on images [J]. arXiv preprint arXiv:2002.10252, 2020.

[14] Moosavi-Dezfooli S M, Fawzi A, Frossard P, et al. A simple and accurate method to fool deep neural networks[C]//Proceedings of the CVPR. 2574-2582.

[15] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Universal adversarial perturbations[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1765-1773.

[16] Cichocki A, Zdunek R, Phan A H, et al. Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation[M]. John Wiley Sons, 2009.

[17] Cichocki A, Mandic D, De Lathauwer L, et al. Tensor decompositions for signal processing applications: From two-way to multiway component analysis[J]. IEEE signal processing magazine, 2015, 32(2): 145-163.

[18] Guo C, Gardner J, You Y, et al. Simple black-box adversarial attacks[C]//International Conference on Machine Learning. PMLR, 2019: 2484-2493.

[19] Ilyas A, Engstrom L, Athalye A, et al. Black-box adversarial attacks with limited queries and information[C]//International Conference on Machine Learning. PMLR, 2018: 2137-2146.

[20] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]// 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.

[21] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[22] Huang S, Papernot N, Goodfellow I, et al. Adversarial attacks on neural network policies [J]. arXiv preprint arXiv:1702.02284, 2017.

[23] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708