Boosting the transferability of adversarial attacks with Expectation Over Transformation

Yuantao Zhang

October 31, 2022

1 Introduction

Deep neural network (DNN) has been shown as a successful tool in machine learning and has been widely applied in many fields. However, recent research found that DNN models are vulnerable to adversarial examples (Goodfellow et al., 2014; Szegedy et al., 2013), which are very similar to natural examples but let a model make wrong predictions.

Under a white-box setting, where attackers can get information about the target model, effective attack algorithms mainly generate adversarial examples by calculating the gradient of the objective function or directly solving an optimization problem. Gradient-based methods include FGSM (Goodfellow et al., 2014) and I-FGSM (Kurakin et al., 2018). Optimization-based methods like L-BFGS (Szegedy et al., 2013) also demonstrated great attack performance. Furthermore, to enhance the robustness of adversarial examples, methods like Expectation over Transformation (Athalye et al., 2018) are also proposed.

However, in pratice, model parameters are often invisible to users. Thus, attackers need to conduct attacks under a black-box setting, where they have limited information about the target model. There are two main categories among the existing black-box attack algorithms, query-based attack and transfer attack. Query-based attacks use consecutive queries to get the attack direction based on the feedback of the target model, while transfer

attacks generate adversarial examples on a white-box surrogate model.

In this work, we apply an existed white-box attack method called Expectation over Transformation (EOT) in transfer attacks. The key idea is to introduce a distribution of transformation to the objective function of the surrogate model. We provide a detailed theoretical analysis about the rationality to apply the EOT method in transfer attacks. Besides, we also find that the EOT method can be combined with other black-box attack algorithms like MI-FGSM (Dong et al., 2018) to achieve better transferability improvements.

2 Related Work

The Expectation Over Transformation method is first proposed to synthesize more robust adversarial examples and exapnd their real-world applications (Athalye et al., 2018). Under the setting that model parameters are known, a distribution of transfomration can be added to input images and the actual attacking process is to attack transformed images. This idea can also be extended to the 3D space and combined with 3D printing technology to manufacture physical adversarial objects.

After the existence of the EOT method, researchers found its value in black-box attacks. There are some previous works that apply the EOT method in query-based attacks. For example, Qin et al. (2021) studied the combination of ZO-optimization and EOT as adaptive attacks. They used theoretical analysis and experiments to verify that the EOT method could help to obtain a more accurate attack direction. However, few works apply the EOT method in transfer attacks.

Furthermore, for transfer attacks, previous works found that the transferability of adversarial examples is a significant influence factor to the success rate (Qin et al., 2022). However, white-box attack algorithms like I-FGSM increase success rates by sacrificing the transferability of adversarial examples (Dong et al., 2018). Therefore, many algorithms are proposed to enhance the transferability of adversarial examples, such as MI-FGSM (Dong et al., 2018), DI-FGSM (Xie et al., 2019), TI-FGSM (Dong et al., 2019). MI-FGSM mainly utilize the accumulated momentum to modify traditional gradeint descent/ascent steps when generating adversarial examples. Xie et al. (2019) found that introducing transformation factors may help to improve the transferability and apply random tranformations to get adversarial ex-

amples. Motivated by the fact that the core idea of DI-FGSM and EOT is very similar, we aim to apply the EOT method in transfer attacks and try to combine it with MI-FGSM, which improves the transferability from an aspect other than introducing transformations.

3 Methodology

3.1 Transfer adversarial attack

Given an input image x and its label y, transfer adversarial attack aims to generate an adversarial example $x^{'}$ which satisfies $\left\|x^{'}-x\right\|_p \leq \epsilon$ by attacking a white-box surrogate model $M^s(x,\theta)$ (θ represents model parameters). Then the adversarial example $x^{'}$ will be directly utilized to attack the black-box target model $M^t(x,\phi)$. The attack goal is to let the target model make wrong predictions, i.e., $M^t(x^{'},\phi) \neq y$(untargeted attack), or $M^t(x^{'},\phi) = y_t$(targeted attack, y_t represents the target class). Taking the targeted attack as example, the above problem can be formulated as the following optimization problem:

$$\min_{x^{'}} L(M^s(x^{'},\theta),y_t) \qquad s.t. \quad \left\|x^{'}-x\right\|_p \leq \epsilon$$

The formulation of untargeted attack can be easily derived by replacing the loss function L and y_t by $-L$ and y, respectively. Besides, the loss function L is often set as the cross entrophy loss:

$$L(M^s(x^{'},\theta),y_t) = -\log P(M^s(x^{'},\theta) = y_t)$$

Given that M^s is a white-box model, we can directly apply some white-box attack algorithm to solve the optimization problem, such as FGSM and I-FGSM. Taking the projected gradient descent(PGD, very similar to I-FGSM) as an example to solve the above problem, we intially set $x^{'}$ equal to x. Then we can update $x^{'}$ iteratively:

$$x^{'}_{t+1} = x^{'}_t - \alpha \cdot sign(\nabla_{x^{'}_t} L(M^s(x^{'}_t,\theta),y_t))$$

$$x^{'}_{t+1} = Clip_{\{\hat{x}:\|\hat{x}-x\|_p \leq \epsilon\}} x^{'}_{t+1}$$

$x^{'}_i$ denotes the adversarial example in ith iteration, α denotes the learning rate, and the second step is required to ensure that the pertubation of the original image is invisible.

3.2 Transfer adversarial attacks with momentum

Instead of directly applying white-box attack algorithms to solve the optimization problem, a momentum-based method called MI-FGSM is proposed to boost the transferability of adversarial examples. In order to stablize update directions and escape from poor local minima, MI-FGSM accumulates a velocity vector in the gradient direction of the loss function during iterations (Dong et al., 2018). In this case, the update of $x^{'}$ changes to:

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_{x_t^{'}} L(M^s(x_t^{'}, \theta), y_t)}{\left\| \nabla_{x_t^{'}} L(M^s(x_t^{'}, \theta), y_t) \right\|_1}$$

$$x_{t+1}^{'} = x_t^{'} - \alpha \cdot sign(g_{t+1})$$

$$x_{t+1}^{'} = Clip_{\{\hat{x}: \|\hat{x}-x\|_p \leq \epsilon\}} x_{t+1}^{'}$$

Here g_i is the accumulated velocity vector across i iterations, the value of g_0 is set as 0, and μ is the decay factor which is usually set as 1($\mu = 1$ contributes to the highest success rate in the paper proposing this method).

3.3 EOT in black-box attacks

Rather than solve the optimization problem in 3.1, the Expectation Over Transformation method uses a chosen distribution F of transformation f taking x as the input and make the transformed original image $f(x)$ perceived by the classifier. Rather than set the constraint on $\left\| x^{'} - x \right\|_p$, EOT aims to constrain the expected distance of the original input and the adversarial example after transformation. So the optimization problem can be reformulated as:

$$\min_{x^{'}} L(E_{f\sim F}[M^s(f(x^{'}), \theta)], y_t) \qquad s.t. \quad E_{f\sim F}[\left\| f(x^{'}) - f(x) \right\|_p] \leq \epsilon$$

The distribution of transformation here imports randomness to the optimization problem and prevents the final adversarial example from overfitting certain transformation. To simplify the analysis afterwards, we first make the following assumptions:

1. $E_{f\sim F}[M^s(f(x^{'}), \theta)] = M^s(f_0(x^{'}), \theta)$, which means that the expectation of classification results is equivalent to the classification result after applying a single transformation f_0.

2. Utilizing the f_0 in 1, $E_{f\sim F}[\left\| f(x^{'}) - f(x) \right\|_p] \leq \left\| f_0(x^{'}) - f_0(x) \right\|_p$

Using these two assumptions, the objective changes to $L(M^s(f(x^{'}),\theta),y_t)$ and the optimization contraint reduces to $\left\|f_0(x^{'})-f_0(x)\right\|_p \leq \epsilon$. Thus we can still use the PGD algorithm. After initializing $x^{'}$ as x, let $z=f_0(x)$ and $z^{'}=f_0(x^{'})$, we can update $z^{'}$ iteratively:

$$z^{'}_{t+1}=z^{'}_t-\alpha\cdot sign(\nabla_{z^{'}_t}L(M^s(z^{'}_t,\theta),y_t))$$

$$z^{'}_{t+1}=Clip_{\{\hat{z}:\|\hat{z}-z\|_p\leq\epsilon\}}z^{'}_{t+1}$$

Noted that do the optimization w.r.t z is equivalent to do the optimization w.r.t x. After achieving the final $z^{'}$, we can get $x^{'}=f_0^{-1}(z^{'})$.

Furthermore, we can combine the momentum-based method with the Expectation Over Transformation method. In this case, the optimization problem is unchanged, the optimization steps change to:

$$g_{t+1}=\mu\cdot g_t+\frac{\nabla_{z^{'}_t}L(M^s(z^{'}_t,\theta),y_t)}{\left\|\nabla_{z^{'}_t}L(M^s(z^{'}_t,\theta),y_t)\right\|_1}$$

$$z^{'}_{t+1}=z^{'}_t-\alpha\cdot sign(g_{t+1})$$
$$z^{'}_{t+1}=Clip_{\{\hat{z}:\|\hat{z}-z\|_p\leq\epsilon\}}z^{'}_{t+1}$$

3.4 Rationality analysis of EOT's application in transfer attack

This part will analysize the rationality to apply EOT in transfer attack, intuition will be combined with mathematical analysis to strengthen the analysis. First we introduce a concept called adversarial decison boundary. The adversarial decision boundary of a benign example x on model $M(x,\theta)$ is a collection of points p_b satisfying the following conditions:

$$\lim_{\delta\to 0}M(p_b-\delta\frac{(p_b-x)}{\|p_b-x\|_2},\theta)=M(x,\theta)$$

$$\lim_{\delta\to 0}M(p_b+\delta\frac{(p_b-x)}{\|p_b-x\|_2},\theta)\neq M(x,\theta)$$

The definition of the adversarial decision boundary means that if the point on the boundary moves a little bit closer to x, the point has the same classification as x; if the point on the boundary moves a little bit further from x, it will be regarded as an adversarial example. Suppose the adversarial

examples are bounded by the 2-norm, for any adversarial example x' w.r.t model $M(x, \theta)$, we have:

$$\mathbb{I}(\frac{x^{'} - x}{\|x' - x\|_2} = \frac{p_b - x}{\|p_b - x\|_2}) \|p_b - x\|_2 \leq \mathbb{I}(\frac{x^{'} - x}{\|x' - x\|_2} = \frac{p_b - x}{\|p_b - x\|_2}) \left\|x^{'} - x\right\|_2 \leq \epsilon$$

Here $\mathbb{I}(\cdot)$ returns 1 if $\cdot$ is correct and returns 0 otherwise. In this case, we can roughly define the transferability of an adversarial example. Because the surrogate model $M^s(x, \theta)$ needs to be similar to the target model $M^t(x, \phi)$, we can use the adversarial decision boundary to define the similarity:

$$\mathbb{I}(\frac{p_b^s - x}{\|p_b^s - x\|_2} = \frac{p_b^t - x}{\|p_b^t - x\|_2}) \left\|p_b^s - p_b^t\right\|_2 \leq c \qquad , \forall x$$

Here p_b^s denotes the adversarial decision boundary for x w.r.t. the surrogate model M^s, p_b^t denotes the adversarial decision boundary for x w.r.t the target model M^t, and c is a constant which specifies the extent of similarity. If we fix the target model, the above inequality needs to be satisfied by all surrogate model. Alternatively, if we fix the surrogate model, all possible target model also needs to satisfy the inequality. After that, the transferability of the an adversarial example $x^{'}$ w.r.t x can be defined by:

$$transferability = \mathbb{I}(\frac{x^{'} - x}{\|x' - x\|_2} = \frac{p_b^s - x}{\|p_b^s - x\|_2}) \left\|x^{'} - p_b^s\right\|_2$$

Noted that if transferability of an adversarial example is larget than c, it can attack all target models corresponding to a fix surrogate model successfully.

For the EOT method, it has already been verified that the EOT method performs very well in the white-box setting (Athalye et al., 2018). Previous motivation to design EOT comes from the fact that the traditional adversarial examples will not be adversarial if we add transformations to them. To analysize the EOT method in transfer attack, we first consider the following assumptions:

1. Here we consider the worst case of $f_0(x)$ for the adversarial example generated by traditional attack methods, $f_0(x)$ satisfies the condition:

$$\frac{x^{'} - x}{\|x' - x\|_2} = -\frac{f_0(x) - x}{\|f_0(x) - x\|_2}$$

   The condition means that the traditional adversarial exmaple $x^{'}$ will go back vertically cross the adversarial decision boundary towards x.

2. The adversarial example x^{adv} generated by the EOT method is different from the adversarial example $x^{'}$ generated by tranditional attack methods, but satisfies the following condition:

$$\frac{x^{adv} - x}{\|x^{adv} - x\|_2} = \frac{x^{'} - x}{\|x^{'} - x\|_2}$$

Although x^{adv} can locate at anywhere, without loss of generality, we let x^{adv},$x^{'}$,x on the same line.

3. for the transformation $f_0(x)$, it satisfies the condition that $f_0(x) - x$ is the same for all x.

Given the fact that traditional adversarial examples will not be adversarial if transformations are added. We have:

$$\mathbb{I}(\frac{x^{'} - x}{\|x^{'} - x\|_2} = \frac{p_b^s - x}{\|p_b^s - x\|_2})(\left\|f_0(x^{'}) - x\right\|_2 < \|p_b^s - x\|_2)$$

Given $f_0(x^{adv})$ is adversarial, we have:

$$\mathbb{I}(\frac{x^{'} - x}{\|x^{'} - x\|_2} = \frac{p_b^s - x}{\|p_b^s - x\|_2})(\|p_b^s - f_0(x)\|_2 < \left\|f_0(x^{adv}) - f_0(x)\right\|_2)$$

Noted that the intersection of the adversarial decision boundary and the line composed of x^{adv},$x^{'}$,x is p_b^s, so p_b^s is also the on the adversarial decision boundary of $f_0(x)$ w.r.t to the surrogate model.

Because the transformation is identical, by assumption 3, we also have:

$$x - f_0(x) = x^{'} - f_0(x^{'}) = x^{adv} - f_0(x^{adv})$$

Thus we can calculate the transferability of x^{adv} and $x^{'}$. Denote the transferability for $\cdot$ as $trans(\cdot)$, we have:

$$trans(x^{'}) = \mathbb{I}(\frac{x^{'} - x}{\|x^{'} - x\|_2} = \frac{p_b^s - x}{\|p_b^s - x\|_2}) \left\|x^{'} - p_b^s\right\|_2$$

$$= \mathbb{I}(\frac{x^{'} - x}{\|x^{'} - x\|_2} = \frac{p_b^s - x}{\|p_b^s - x\|_2})(\left\|x^{'} - f_0(x^{'})\right\|_2 + \left\|f_0(x^{'}) - p_b^s\right\|_2)$$

$$= \mathbb{I}(\frac{x^{'} - x}{\|x^{'} - x\|_2} = \frac{p_b^s - x}{\|p_b^s - x\|_2})(\left\|x^{'} - f_0(x^{'})\right\|_2 + \left\|f_0(x^{'}) - f_0(x)\right\|_2 - \|p_b^s - f_0(x)\|_2)$$

$$trans(x^{adv}) = \mathbb{I}(\frac{x^{adv}-x}{\|x^{adv}-x\|_2} = \frac{p_b^s - x}{\|p_b^s - x\|_2}) \left\|x^{adv} - p_b^s\right\|_2$$

$$= \mathbb{I}(\frac{x^{'}-x}{\|x^{'}-x\|_2} = \frac{p_b^s - x}{\|p_b^s - x\|_2}) \left\|x^{adv} - p_b^s\right\|_2$$

$$= \mathbb{I}(\frac{x^{'}-x}{\|x^{'}-x\|_2} = \frac{p_b^s - x}{\|p_b^s - x\|_2})(\left\|x^{adv} - f_0(x^{adv})\right\|_2 + \left\|f_0(x^{adv}) - p_b^s\right\|_2)$$

$$= \mathbb{I}(\frac{x^{'}-x}{\|x^{'}-x\|_2} = \frac{p_b^s - x}{\|p_b^s - x\|_2})(\left\|x^{'} - f_0(x^{'})\right\|_2 + \left\|f_0(x^{adv}) - f_0(x)\right\|_2 - \|p_b^s - f_0(x)\|_2)$$

By the inequality above:

$$\mathbb{I}(\frac{x^{'}-x}{\|x^{'}-x\|_2} = \frac{p_b^s - x}{\|p_b^s - x\|_2})(\left\|f_0(x^{adv}) - f_0(x)\right\|_2 - \|p_b^s - f_0(x)\|_2) > 0$$

$$> \mathbb{I}(\frac{x^{'}-x}{\|x^{'}-x\|_2} = \frac{p_b^s - x}{\|p_b^s - x\|_2})(\left\|f_0(x^{'}) - f_0(x)\right\|_2 - \|p_b^s - f_0(x)\|_2)$$

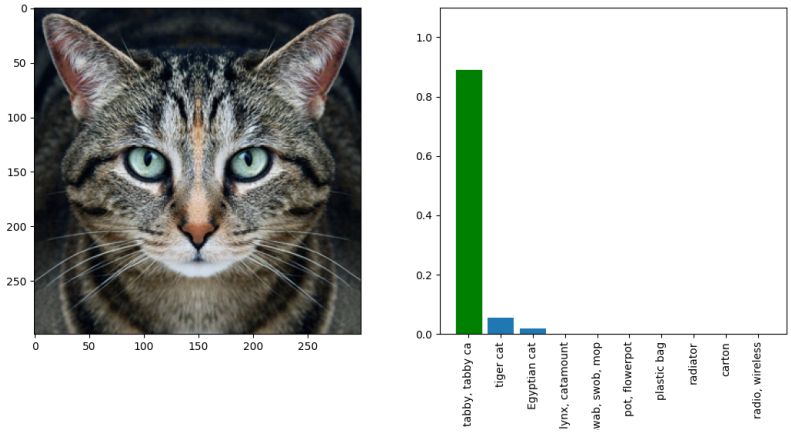
Finally, we derive that the transferability of x^{adv} is larger than the transferability of $x^{'}$, which means that compared with traditional transfer attacks, the EOT method can increase the transferability of the generated adversarial example. The transferability can be much closer to the similarity extent c or exceed c. The application of the EOT method in transfer attacks is reasonable to certain extent.

4 Experiments

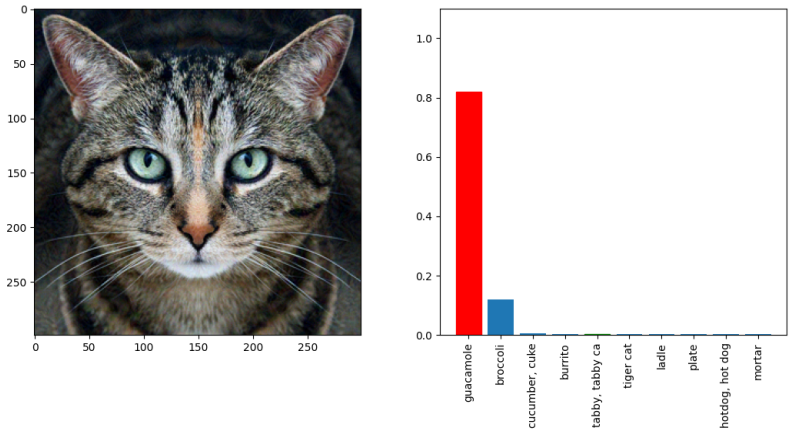
In out experiments, we first test the EOT method based on the pretrained Inception v3 model (Szegedy et al., 2016). Then we apply the EOT method in transfer attacks. In transfer attacks, We adopt the VGG-16 (Simonyan & Zisserman, 2014) as the target model and ResNet-50 (He et al., 2016) as the surrogate model. The data we use comes from the most widely used benchmerk dataset in adversarial machine learning: Imagenet.

4.1 Implementation of EOT in white-box setting

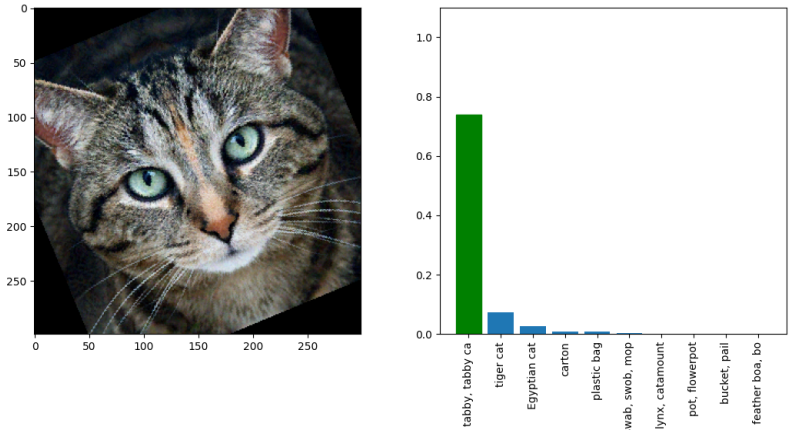
In our implementation of the EOT, we select a cat image as an example. To guarantee the effectiveness of the EOT method, we first use the inception v3 to classify the image and see whether it can be correctly classfied (Here we demonstrate the classification probability of top10 class).



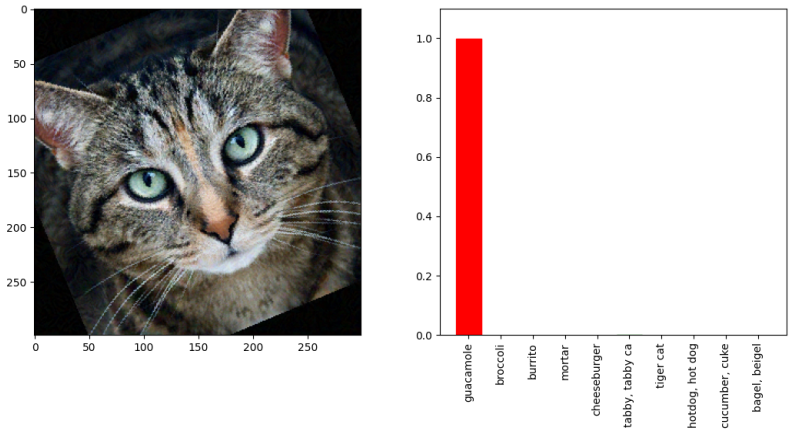
As we can see, the cat image can be correctly classfied to the "tabby, tabby cat" class. Then we first use traditional white-box attack algorithm(the PGD algorithm is used here) to attack the inception v3 model. In our experiment, we set the step size of the gradient descent to $\frac{2}{255}$, the maximum allowable pertubation ϵ to $\frac{8}{255}$, and the target class is set as the "guacamole" class:



But if we simply add a transformation to the generated adversarial example, the adversarial example is no longer an adversarial example.

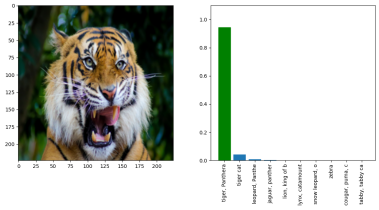


Therefore, we need to use the EOT method to synthesize a more robust adversarial example. Here, we specify the transfomration f_0 as rotating the image counterclockwise by 22.5^o, the step size and the maximum allowable pertubation are still $\frac{2}{255}$ and $\frac{8}{255}$, respectively.

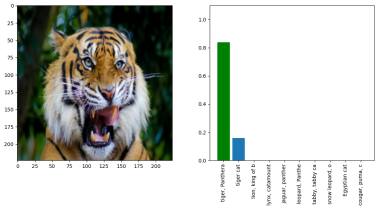


4.2 Implementation of EOT in transfer attacks

In the implementation of EOT in transfer attacks, we select a tiger image as the input. First we evaluate the tiger image in both the VGG-16 model and ResNet-50 model. For the following demonstrations, if not specially emphasize, the left side corresponds to the classification results of the surrogate model ResNet-50 and the right side corresponds to the classification reuslts of the target model VGG-16.

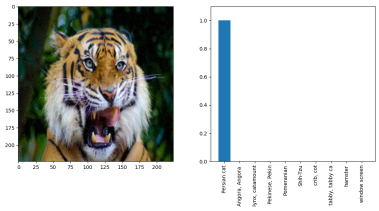


(a) ResNet

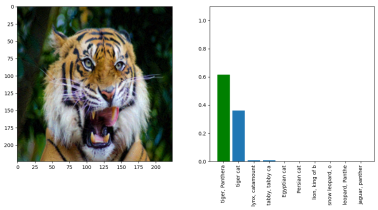


(b) VGG

Noted that both VGG and ResNet can correctly classify the tiger image to the correct class "tiger, Panthera". Then we first use traditional transfer attack to attack the model. Here we adopt the PGD algorithm and untargeted attack, the step size and maximum allowable pertubation are set as $\frac{2}{255}$ and $\frac{10}{255}$, respectively.

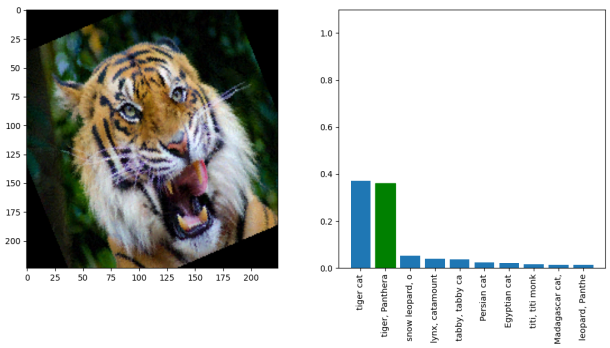


(c) ResNet

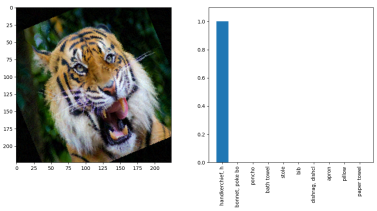


(d) VGG

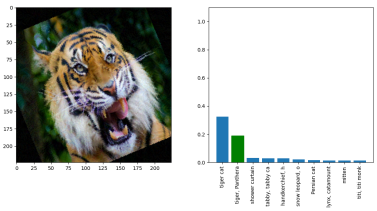
From the above result, we can easily observe that the adversarial example generated by traditional transfer attacks suffers from low transferability. The adversarial example let the surrogate model make wrong predictions but can still correctly classified by the target model. If we simply add a transformation to the adversarial example, the target model can still correctly classify the example(the example still has about 40 percentage to be classfied to the "tiger Panthera" class).



Thus we utilize the EOT method by applying the process similar to the description in our theoretical analysis. The step size and maximum allowable pertubation are still set as $\frac{2}{255}$ and $\frac{10}{255}$, respectively.



(e) ResNet

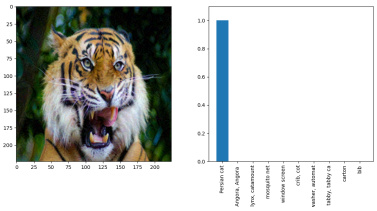


(f) VGG

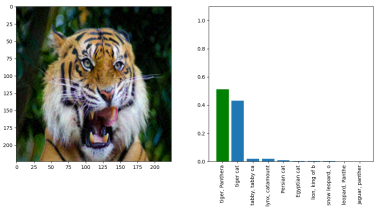
The result of the EOT method is quite satisfactory. The EOT method improves the transferability of the adversarial example to certain extent, which verifies the rationality to apply the EOT method in transfer attacks. However, the target class still has about 20 percentage to classify the adversarial example to the correct class.

4.3 Boosting the transferability with momentum

Still use the tiger image as the example, we also do the experiment using the MI-FGSM algorithm and try to boost the transferability of the adversarial example with momentum. Recall that the adversarial example generated by traditional transfer attacks suffer from low transferability, the MI-FGSM algorithm can also improve the transferability to certain extent.



(g) ResNet

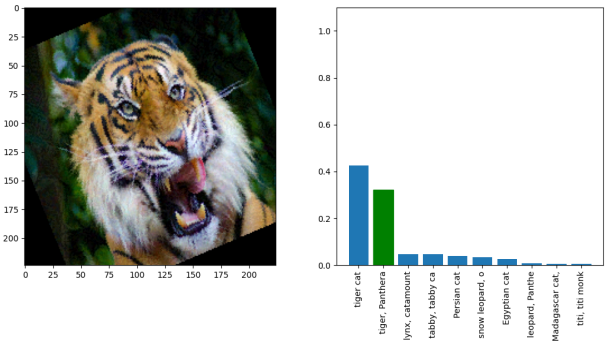


(h) VGG

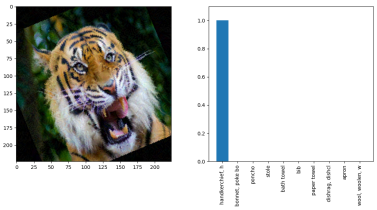
Noted that the probability that the adversarial example can be correctly classfied by the target model reduces from 60 percent to 50 percent. The transferability increases by a small amount.

4.4 Combination of the EOT method and MI-FGSM

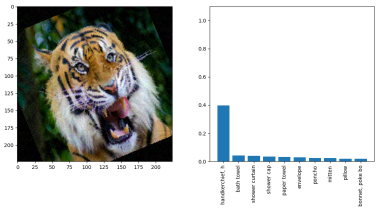
After doing the above experiments, we become interested in the combination of the EOT method and the MI-FGSM algorithm. First we test the classfication result of the target model after adding a transformation to the adversarial example generated by MI-FGSM.



By observation, the example is still not so adversarial. Thus we combine the EOT method and the MI-FGSM algorithm using the process similar to the description in our theoretical analysis.



(i) ResNet



(j) VGG

The result is very satisfactory and better than using the EOT method or MI-FGSM separately. The target model has no probability to classify the adversarial example to the correct class. According to the experiments, we can conclude that the combination of EOT mehtod and MI-FGSM can enhance the improvement of transferability performed by the two algorithm separately.

5 Conclusion

This work studies the application of a white-box attack method called Expectation Over Transformation. Motivated by the idea of adversarial decision

boundary, this work gives a detailed analysis of the rationality to apply the EOT method in transfer attacks and do the experiments to verify the theoretical anaysis. Furthermore, this work finds that both the EOT method and the MI-FGSM algorithm can improve the transferability of adversarial examples and the combination of them could achieve better improvements. However, this work still has some limitations. The experiment samples of this work may be not sufficient because of the time limits and the work may lack some novelty.

References

Athalye, A., Engstrom, L., Ilyas, A., & Kwok, K. (2018). Synthesizing robust adversarial examples. In *International conference on machine learning* (pp. 284–293).

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., & Li, J. (2018). Boosting adversarial attacks with momentum. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 9185–9193).

Dong, Y., Pang, T., Su, H., & Zhu, J. (2019). Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 4312–4321).

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Kurakin, A., Goodfellow, I. J., & Bengio, S. (2018). Adversarial examples in the physical world. In *Artificial intelligence safety and security* (pp. 99–112). Chapman and Hall/CRC.

Qin, Z., Fan, Y., Liu, Y., Shen, L., Zhang, Y., Wang, J., & Wu, B. (2022). Boosting the transferability of adversarial attacks with reverse adversarial perturbation. *arXiv preprint arXiv:2210.05968*.

Qin, Z., Fan, Y., Zha, H., & Wu, B. (2021). Random noise defense against query-based black-box attacks. *Advances in Neural Information Processing Systems*, *34*, 7650–7663.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2818–2826).

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., & Yuille, A. L. (2019). Improving transferability of adversarial examples with input diversity. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 2730–2739).