Demonstrating Transferability of Adversarial Attacks via Fast Gradient Sign Method and Projected Gradient Descent on CIFAR-100

b09605039 JYUN SYUN, SU

Abstract

In this experiment, we tested the accuracy generated by ADX produced by many models using FGSM on these models, and plotted the heatmap of accuracy and gradient alignment. Meanwhile, we also ensembled the models that can cause the lowest average accuracy of the top 10, and found that lower accuracy can be obtained. Because it was found that several models still maintained prominent accuracy after Ensemble, we sequentially attempted to Ensemble these models and attacked using the PGD method. We successfully lowered the accuracy of the three models, and fortunately obtained an overall lower accuracy, unlike the case of several particularly high accuracies with FGSM.

1 Introduction

In the field of machine learning security, Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) are two commonly used and simple yet effective attack methods. The adversarial examples (ADX) generated by attacking methods, e.g. FGSM and PGD, exhibit transferability, and FGSM and PGD have different characteristics in this regard. FGSM achieves relative lower attack effectiveness but with higher transferability, while PGD exhibits the opposite. In this study we evaluated 25 pre-trained models for CIFAR-100 dataset. Firstly, we tested the effectiveness of FGSM by generating ADX from each model and evaluated their accuracy and gradient alignment across all models by plotting heatmap, observing the transferability of each model as a surrogate model, as well as the phenomena mentioned by Ambra Demontis et al.. Next, we utilized ensemble technique on surrogate models that produced adversarial examples causing the top 10 lowest average accuracy, and reapplied FGSM, achieving even lower average accuracy. However, as three models still maintained moderately higher accuracy after FGSM attack, we further attempted PGD attack and applied ensemble technique, resulting in not only lowering the accuracy of the model which are ensembled, but also fortunately achieved overall lower accuracy without any exceptionally high accuracy instances.

Through these experiments, we demonstrated the different characteristics of FGSM and PGD in terms of transferability and experienced the enhanced attack capability with ensemble technique. Furthermore, the heatmap of gradient alignment helped identify models less susceptible to transfer attacks, as they exhibited smaller cosine similarity in gradient alignment.

2 Threat Model

In this experiment, we assume that the attacker can only generate ADX using their own models to conduct transfer attacks on other unknown models, with the attacker only being aware that the target models are also image classification models. We assume that the defense side has not implemented any defense measures.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Parameters Attacker are allowed to change each pixel of the input image up to $\epsilon = 8$ on the $0 - 255$ pixel scale, and We set the update steps of PGD to 10.

3 Methods

We selected 500 images from CIFAR-100, with 5 images from each class, to be used for generating ADX.

We selected 25 models directly applicable to CIFAR-100 from the following two websites, including well-known models such as ResNet (Kaiming He et al.) and DenseNet (Gao Huang et al.).

`https://github.com/osmr/imgclsmob`
`https://github.com/chenyaofo/pytorch-cifar-models`

Next, we used these 25 models as surrogate models and generated ADX using FGSM, and recorded the gradients of the loss function in the meanwhile. Subsequently, we tested the accuracy on each model. Finally, we created heatmaps for accuracy and the cosine similarity of gradient.

3.1 Ensemble

We took the surrogate models of the top 10 ADX with the lowest average accuracy among all models, and used FGSM attack with ensemble technique on these 10 models.

3.2 PGD and Ensemble

Since three models maintained relatively higher accuracy compared to other models after being attacking by ADX generated by ensemble models, we sequentially ensembled these three models in an order, and applied PGD attacks to generate ADX. We followed the order that the first time, only one model was used for PGD; the second time, two models were ensembled and then were used to PGD; the third time, three models were ensembled and then were used to PGD.

4 Experiment

FGSM We selected the following 25 models from the two sites mentioned above, which have been pre-trained and can be directly used for CIFAR-100.

Name		
`resnet20`	`resnet56`	`diaresnet20`
`seresnet20`	`sepreresnet20`	`diapreresnet20`
`vgg11_bn`	`vgg19_bn`	`mobilenetv2_x0_5`
`repvgg_a0`	`repvgg_a2`	`mobilenetv2_x1_4`
`shufflenetv2_x0_5`	`shufflenetv2_x2_0`	`shake-shake-resnet20_2x16d`
`wrn16_10`	`wrn40_8`	`wrn20_10_32bit`
`densenet40_k12`	`densenet40_k12_bc`	`densenet250_k24_bc`
`pyramidnet110_a48`	`pyramidnet272_a200_bn`	
`ror3_56`	`ror3_164`	

We made some modifications to the two classes, FGSM and PGD, provided by the `torchattacks` module, to extend them into our own versions for generating ADX, and we recorded the gradients of 500 images on various models in the meanwhile.

We used the ADX generated by the 25 models to calculate their accuracy on all models, and created a heatmap (Fig 1).

The alignment of gradients is one of the metrics used to measure transferability (Ambra Demontis et al.). Therefore, we recorded the gradient of each of the 500 images when generating ADX. Subsequently, we calculated the average cosine similarity between the signed gradient vector of the surrogate model and the gradient vector of the target model across the 500 images.

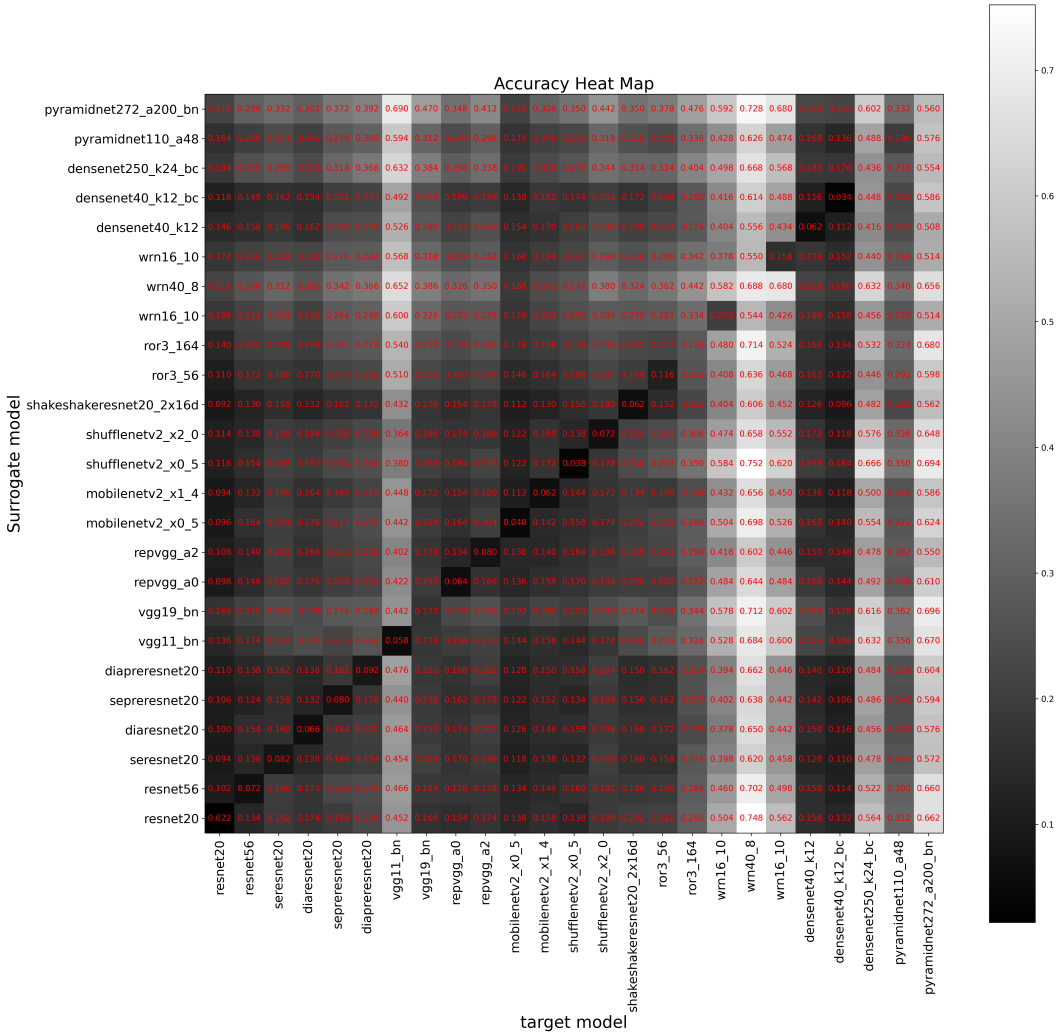

Figure 1: The heatmap of accuracy. The models of the y axis are surrogate model, which generate ADX to attck the models of the x axis

$$l(y, x + \hat{\delta}, w) \approx l(y, x, w) + \hat{\delta} \nabla_x l(y, x, w) \tag{1}$$

The Equation (1) is mentioned in Ambra Demontis et al.. In this study we use l_p norm as our distance measurement, so $\hat{\delta} = \epsilon \cdot \text{sign} \nabla_x l(y, x, w)$ which is different from the paper.

$$l(y, x, w) + \epsilon \cdot \text{sign}(\nabla_x l(y, x, \hat{w}))^T \nabla_x l(y, x, w) \tag{2}$$

Hence we define the cosine similarity in Equation 3., and make the heatmap (Fig 2.).

$$cos(\theta) = \frac{\text{sign}(\nabla_x l(y, x, \hat{w}))^T \nabla_x l(y, x, w)}{||\text{sign}(\nabla_x l(y, x, \hat{w}))||_2 \cdot ||\nabla_x l(y, x, w)||_2} \tag{3}$$

Ensemble Observing Figure 3, it becomes apparent that the ADX generated by certain models lack significant transferability. Consequently, we utilized Ensemble technique on the top 10 models with the lowest average accuracy and conducted a FGSM attack.

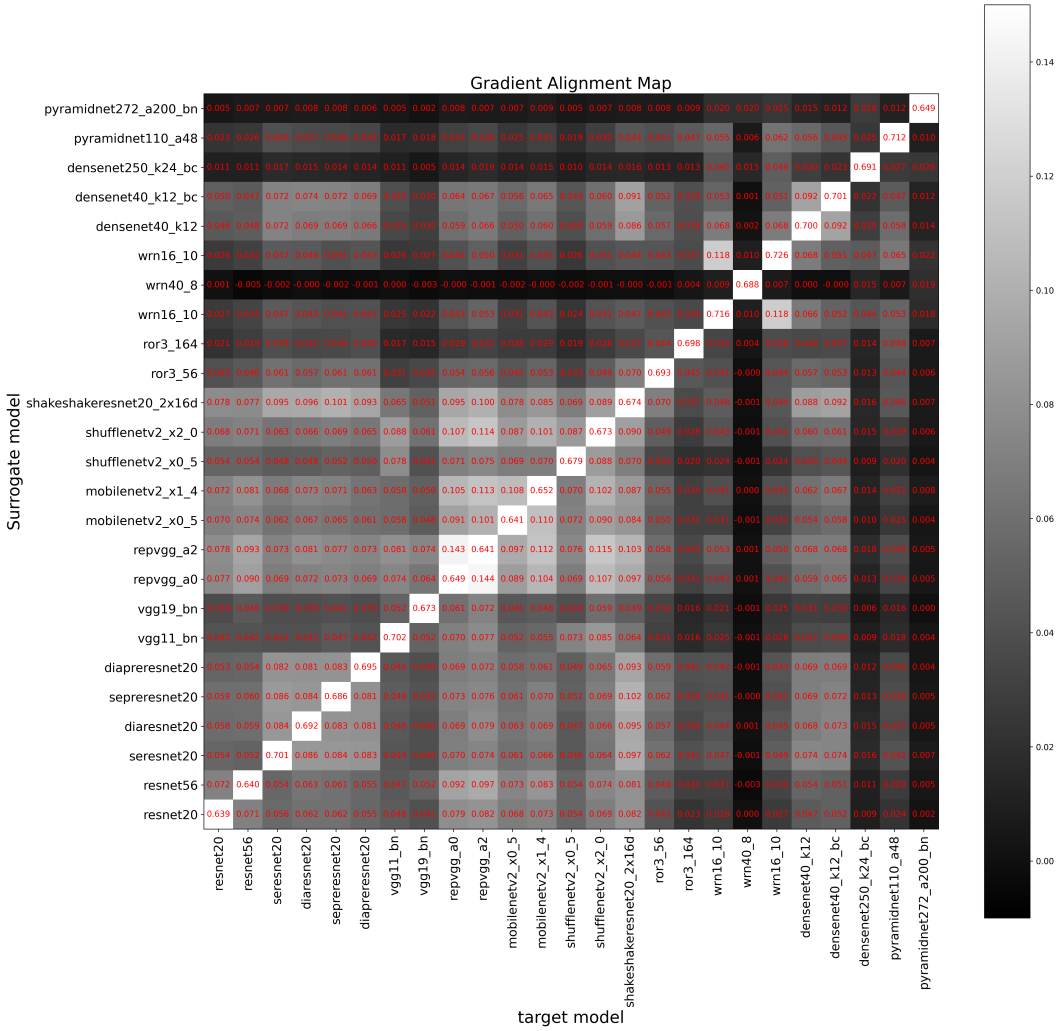

Figure 2: The heatmap represents the cosine similarity. The models listed on the y-axis are surrogate models, responsible for generating ADX to attack the models listed on the x-axis. We set the upper limit to 0.15 to ensure that the heatmap's effectiveness is not compromised by larger values among models of the same type.

PGD and Ensemble In Fig 4, we can observe that three models exhibit consistently high accuracy even under ensemble FGSM attack. Therefore, we first selected `wrn40_8` as the surrogate model and conducted a PGD attack. Subsequently, we ensembled `wrn40_8` and `pyramidnet272_a200_bn` as surrogate models and performed a PGD attack. Finally, we ensembled `wrn40_8`, `pyramidnet272_a200_bn`, and `vgg11_bn` as surrogate models and conducted a PGD attack. The results of these three attacks, including the accuracy on the original images and the accuracy of ensemble FGSM ADX, are depicted in Fig 5.

5 Results

Gradient Alignment From the heatmap of cosine similarity (Fig 2.), we can find that target models which are less susceptible to attacks exhibit cosine similarity values close to 0 with all surrogate models, while target models that are more susceptible to attacks have cosine similarity values that are not as close to 0.

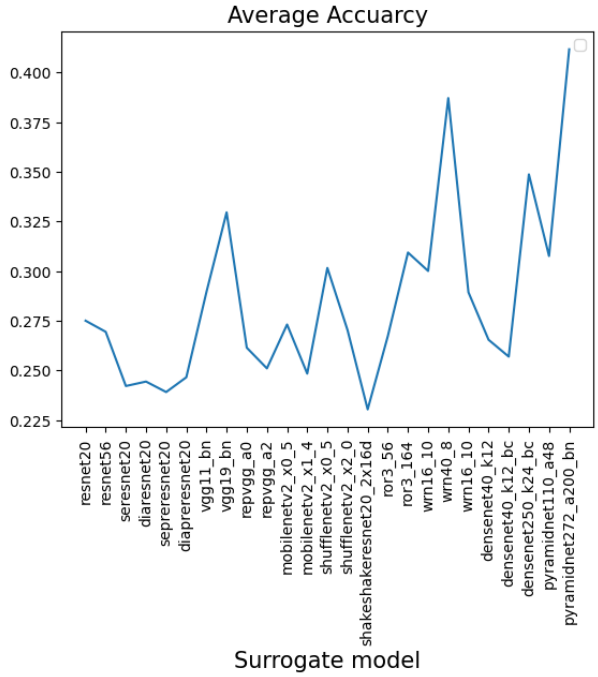

Figure 3: The y-axis represents the average accuracy among all models, while the x-axis represents the surrogate model responsible for generating the ADX

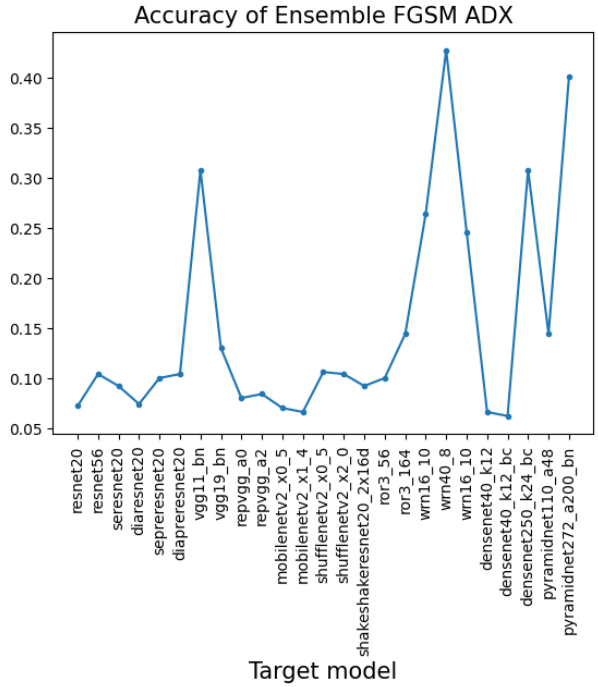

Figure 4: The graph shows the accuracy of each model when facing ADX generated by applying Ensemble techinque. The y-axis represents the accuracy of certain models, while the x-axis represents the target model

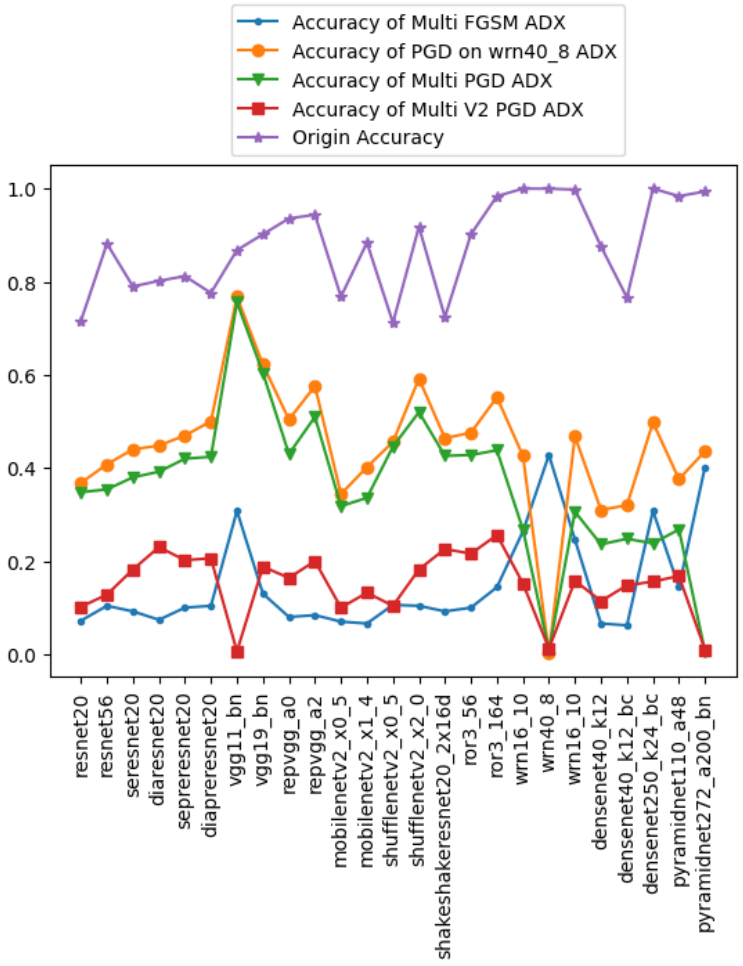

Figure 5: Accuracy of Multi FGSM ADX is the result of ensembling the top 10 models with the lowest average accuracy, i.e. Fig. 4 above. Accuracy of Multi PGD ADX is the result of ensembling `wrn40_8` and `pyramidnet272_a200_bn` as surrogate model. Accuracy of Multi V2 PGD ADX is the result of ensembling `wrn40_8`, `pyramidnet272_a200_bn` and `vgg11_bn` as surrogate model. Except for the models of Origin Accuracy whose results are based on the accuracy of original images as input, all other results are based on the respective ADX as input.

Ensemble From the experiments conducted, we observed that FGSM generally results in low accuracy for most models. However, models `wrn40_8`, `pyramidnet272_a200_bn`, and `vgg11_bn` still maintained higher accuracy compared to other models. Therefore, we proceeded with experimenting with PGD to see its effectiveness, and the results are presented in Fig. 5. Firstly, we used `wrn40_8` as a surrogate model, and after the attack, we observed the characteristics of PGD that the accuracy of `wrn40_8` was notably low, while the other models were slightly affected. Next, we ensembled `wrn40_8` and `pyramidnet272_a200_bn`, and performed PGD attacks. It resulted in decreased accuracy for both models. Finally, when we ensembled `wrn40_8`, `pyramidnet272_a200_bn` and `vgg11_bn`, the accuracy of all three models decreased.

It is important to note that although it may seem that the accuracy of other models gradually decreases as the number of models which are ensembled increases, the random start step in PGD provides an opportunity for other models to maintain higher accuracy. Using the Ensemble technique with PGD only ensures a powerful attack on these three models. The examples above just happen to coincide with decreased accuracy for other models.

6 Conclusion

From the experiments conducted, we have demonstrated the effectiveness of FGSM and PGD attacks, as well as the enhanced transferability achieved through the application of ensemble techniques. Additionally, we have illustrated that models less susceptible to transfer attacks tend to exhibit cosine similarity values closer to 0. We hope that this experiment serves as an example to highlight the significant impact simple attacks can have on model performance.

Overall, our findings underscore the importance of robustness testing and the need for effective defense strategies in safeguarding machine learning models against adversarial attacks.

References

[1] Alex Krizhevsky. (2009) Learning Multiple Layers of Features from Tiny Images. Available at : https://www.cs.toronto.edu/ kriz/cifar.html

[2] PyTorch. (n.d.) Docs. Available at : https://pytorch.org/docs/stable/index.html

[3] chenyaofo. (2021) PyTorch CIFAR Models. Available at : https://github.com/chenyaofo/pytorch-cifar-models

[4] osmr (Oleg Sémery). (2021) Deep learning networks (imgclsmob). Available at : https://github.com/osmr/imgclsmob/blob/master/README.md

[5] Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, & Fabio Roli. (2019) Why Do Adversarial Attacks Transfer? Explaining Transferability of Evasion and Poisoning Attacks.

[6] Kim & Hoki. (2020) Torchattacks: A pytorch repository for adversarial attacks. Available at : https://github.com/Harry24k/adversarial-attacks-pytorch?tab=readme-ov-file

[7] OpenAI. (2022). ChatGPT [Computer software]. Retrieved from https://openai.com/chatgpt