Homework 1 – Graybox Attack

Walker
Department of Electrical Engineering
B10901036

Abstract

This homework is about implementing graybox attack. By using different kinds of attacks, such as FGSM, MI-FGSM and PGD, we can generate adversarial examples to attack the model. Also, I would like to test the transferbility of the adversarial attacks on each attack. Moreover, different hyperparameters will affect the robustness and transferbility of the adversarial examples. In the end, I will try adversarial training to see the effect of adversarial training on the model's robustness.

1 Attack

In this homework, I use the CIFAR-100 dataset, which consists of 500 inference images in 100 classes, with 5 images per class. I have tried various attacks on the CIFAR-100 dataset, including FGSM, MI-FGSM and PGD attack. For each attack, I want to check the robustness and the transferbility of these adversarial examples. To do so, I generate te adversarial examples in two ways:

1. Single model attack
2. Ensemble attack

1.1 Three Attack Methods

I implement three different attacks on the CIFAR-100 dataset to generate adversarial examples.:

1. Fast Gradient Sign Method (FGSM)
2. Momentum Iterative Fast Gradient Sign Method (MI-FGSM)
3. Projected Gradient Descent (PGD)

For all of the attacks, the perturbation is limited by ℓ_∞-norm, with ϵ set to $\frac{8}{255}$. For both MI-FGSM and PGD attacks, the iteration number is set to 4. The results of the attacks are shown in Table 1.1.

There are plenty of models we can choose to attack from the model lists[1] After trials and error, I use $resnet164bn\_cifar100$ to generate the adversarial examples for single model attack, and use three different models below to generate the adversarial examples for ensemble attack.

1. $resnet164bn\_cifar100$
2. $preresnet110\_cifar100$
3. $wrn16\_10\_cifar100$

[1] The available models are listed in the github `https://github.com/osmr/imgclsmob/blob/master/pytorch/pytorchcv/model_provider.py`

Attack type	Attack method	Tested Model	Acc.
No Attack	Original	$resnet164bn\_cifar100$	0.808
		$wrn16\_10\_cifar100$	0.978
		$ror3\_164\_cifar100$	0.862
		$rir\_cifar100$	0.952
		$pyramidnet110\_a270\_cifar100$	0.986
Single Model Attack	FGSM	$resnet164bn\_cifar100$	0.140
		$wrn16\_10\_cifar100$	0.346
		$ror3\_164\_cifar100$	0.208
		$rir\_cifar100$	0.318
		$pyramidnet110\_a270\_cifar100$	0.474
	MIFGSM	$resnet164bn\_cifar100$	**0.006**
		$wrn16\_10\_cifar100$	0.188
		$ror3\_164\_cifar100$	**0.104**
		$rir\_cifar100$	**0.16**
		$pyramidnet110\_a270\_cifar100$	**0.352**
	PGD	$resnet164bn\_cifar100$	0.008
		$wrn16\_10\_cifar100$	0.486
		$ror3\_164\_cifar100$	0.248
		$rir\_cifar100$	0.402
		$pyramidnet110\_a270\_cifar100$	0.648
Ensemble Attack	FGSM	$resnet164bn\_cifar100$	0.214
		$wrn16\_10\_cifar100$	0.214
		$ror3\_164\_cifar100$	0.310
		$rir\_cifar100$	0.438
		$pyramidnet110\_a270\_cifar100$	0.526
	MIFGSM	$resnet164bn\_cifar100$	0.042
		$wrn16\_10\_cifar100$	**0.018**
		$ror3\_164\_cifar100$	0.242
		$rir\_cifar100$	0.314
		$pyramidnet110\_a270\_cifar100$	0.438
	PGD	$resnet164bn\_cifar100$	0.126
		$wrn16\_10\_cifar100$	0.158
		$ror3\_164\_cifar100$	0.456
		$rir\_cifar100$	0.616
		$pyramidnet110\_a270\_cifar100$	0.770

Table 1: The table shows the accuracy of the adversarial examples on the tested models with different attacks. The **bold black number** represents the best attack (lowest accuracy) on the same model for all kinds of attack.

For ensemble attack, three of the models above will generate the adversarial examples respectively and then return the average image of the three adveresarial examples.

To check the relationship between the effect of adversarial attack and the transferbility of the adersarial examples, I use the generated adversarial examples above on 5 different models:

1. $resnet164bn\_cifar100$
2. $wrn16\_10\_cifar100$
3. $ror3\_164\_cifar100$
4. $rir\_cifar100$
5. $pyramidnet110\_a270\_cifar100$

The first model is the same as the model used to generate the adversarial examples (both single model attack and ensemble attack), and the second model is a model used to generate the adversarial examples on ensemble attack. The other three models are used to test the transferbility of the adversarial examples.

Attack method	Tested Model	Small Iter Acc	Big Iter Acc
MI-FGSM Attack	$resnet164bn\_cifar100$	**0.000**	**0.000**
	$wrn16\_10\_cifar100$	**0.172**	0.184
	$ror3\_164\_cifar100$	**0.082**	0.104
	$rir\_cifar100$	**0.134**	0.146
	$pyramidnet110\_a270\_cifar100$	**0.334**	0.398
PGD Attack	$resnet164bn\_cifar100$	0.008	**0.000**
	$wrn16\_10\_cifar100$	**0.450**	0.502
	$ror3\_164\_cifar100$	**0.232**	0.260
	$rir\_cifar100$	**0.368**	0.436
	$pyramidnet110\_a270\_cifar100$	**0.608**	0.668

Table 2: The table shows the accuracy of the adversarial examples on the tested models with different iteration numbers for both MI-FGSM and PGD attack. For MI-FGSM attack, the iteration number is 16 and 64, and for PGD attack, the iteration number is 4 and 16. The **bold black number** represents the better attack (lower accuracy) on the same model for the same attack.

We can see that using MI-FGSM attack can not only generate the best adversarial examples for the same model attack, but the examples can also transfer to other models with low accuracy. Also, in my model selection, **single model attack can have better transferability than ensemble model attack.**

1.2 Hyperparameters

For MI-FGSM and PGD attack, I have tried different iteration number to see whether the model robustness and transferability will change. The result is shown in Table 1.2.

We can see that for both MI-FGSM and PGD attack, the iteration number acts as the trade-off factor between the robustness and transferability. The more iteration number, the better the model robustness on the trained model, but the worse the transferability on the other models that are not used for training.

2 Defense

To see the effect of adversarial training, I use the adversarial examples generated from the attack mentioned above to train the model.

There are a total of 5000 images, with 3 types of adversarial images, generated from

1. FGSM attack · 1000
2. MI-FGSM attack · 2000
3. PGD attack · 2000

These adversarial images are used to train the model with adversarial training respectively.

The model I choose for adversarial training is $resnet110\_cifar100$. To see whether adversarial training with one type of adveresarial images can improve the model's robustness on other adversarial examples, I train the model with three types of adversarial images respectively and test the model on the adversarial images. The result is shown in Table 2.

This result shows that for each training set of adversarial images, the accuracy tested on the same type of adversarial examples can have the best accuracy improvement. Also, FGSM-based adversarial exmaples are better to defend since for both MIFGSM-trained and PGD-trained models, the accuracy improvements on FGSM adversarial examples enhance quite a lot (about 28 percent) compared with the other two types of adversarial examples (about 14 percent). This can be an indirect sign that FGSM-based adversarial examples are weaker than the other two types of adversarial examples.

testing adversarial examples Training adversarial examples	FGSM	MI-FGSM	PGD
FGSM	**0.187** → **0.767**	0.014 → 0.058	0.019 → 0.049
MI-FGSM	0.016 → 0.286	**0.53** → **0.953**	0.006 → 0.145
PGD	0.027 → 0.294	0.011 → 0.147	**0.073** → **0.977**

Table 3: The table shows the accuracy before and after the adversarial training for different training and testing adversarial examples. The **bold black accuracy change** represents the best accuracy improvement in each row.

3 Conclusion

In this homework, I have implemented three different attacks on the CIFAR-100 dataset to generate adversarial examples. I have also tried different hyperparameters for each attack to see the effect of the robustness and transferability of the adversarial examples. The result shows that MI-FGSM attack can generate the best adversarial examples for the same model attack, and the examples can also transfer to the other model with the best accuracy. Also, in my model selection, single model attack can have better transferability than ensemble attack.

I have also tried adversarial training to see the effect of adversarial training on the model's robustness. The result shows that for each training set of adversarial images, the accuracy tested on the same type of adversarial examples can have the best accuracy improvement. Also, FGSM-based adversarial exmaples are better to defend since for both MIFGSM-trained and PGD-trained models,

Homework Submission

For submission, I use the adversarial examples generated from MI-FGSM model with single model attack. The iteration number is set to 16, and the model used to generate the adversarial examples is $resnet164bn\_cifar100$.

Github Code

The code for this homework is available on my github repository:
`https://github.com/walkerhsu/SPML/tree/main/greybox_attack`

References

[1] torchattack github link: `https://github.com/Harry24k/adversarial-attacks-pytorch`