Delving into Transferable Adversarial Examples and Black-box Attacks

Yanpei Liu , Xinyun Chen [1] Chang Liu, Dawn Song [2]

[1]Shanghai JiaoTong University

[2]University of the California, Berkeley

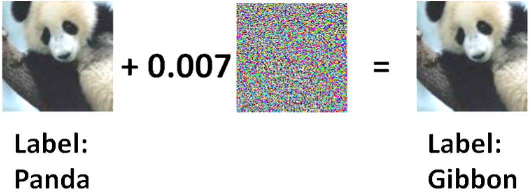
ICLR 2017/ Presenter: Ji Gao

Outline

1. Motivation
   - Transferability
2. Previous methods
3. Experiment I
4. Method
5. Experiment II
6. Experiment III: On realistic data

Motivation

- Adversarial examples: Samples are close to the normal seeds but are misclassified by the deep model



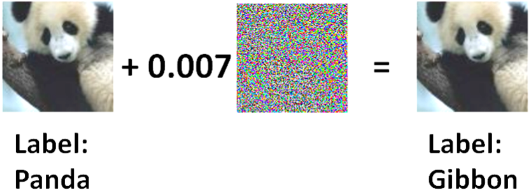
- Adversarial samples can transfer between models: Adversarial samples generated to fool one model can be also applied to another models.

Motivation: Transferability

- Transferability of the adversarial samples: The ability for adversarial samples generated can be transferred to another model.
- Research problem: How to represent transferability? How can we generate samples with more?
- If we can generate transferable sample, it means that we can attack a black-box model easily by generating adversarial sample on any model.

Targeted and non-targeted attack

- Targeted: Have an objective label.
- Non-targeted: Don't have an objective label, just want to mislead the model to a different label.



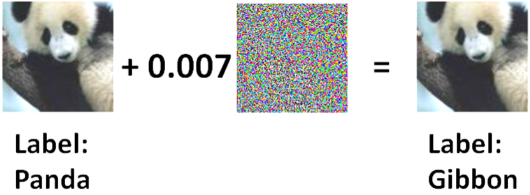
- Non-targeted samples have been shown easier to transfer.
- This paper focus on targeted attack.
- What if for the two models, the labels are different?

Adversarial samples

- Formally, adversarial sample can be defined as:

Adversarial sample

$$\begin{aligned} & d(x, x^*) \leq B, x^* \in \mathbb{X} \\ & F(x) \neq F(x^*) \end{aligned} \tag{1}$$



Previous solutions: Fast Gradient Sign

- The Fast Gradient Sign algorithm is one efficient algorithm for creating adversarial samples. Equation:

$$\Delta x = \epsilon \mathrm{sign}(\nabla_x \mathrm{Loss}(F(x^*;\theta), F(x;\theta))) \tag{2}$$

- This is motivated by controlling the L_∞ norm of Δx, i.e. perturbation in each feature dimension to be the same.
- If the sign function is not used, it become another method on L2 norm. It is called *fast gradient method* in this paper.

Previous solutions: Optimization-based approach

L2 attack (Carlini & Wagner (2016))

$$\arg\min_{x^*} \lambda d(x, x^*) + \text{Loss}(F(x;\theta)) \tag{3}$$

- This approach is more intuitive and accurate, but costly.

Targeted attack

- Previous method is for non-targeted attack, but it's easy to change to targeted attack.
- Simply use a different loss function:

Targeted L2 attack (Carlini & Wagner (2016))

$$\arg\min_{x^*} \lambda d(x, x^*) + \text{Loss}_{y^*}(F(x;\theta)) \tag{4}$$

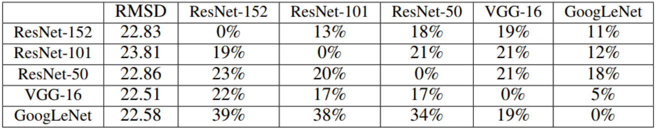
Experiment I

Experiment design:

- Models: 5 networks – ResNet-50, ResNet-101, ResNet-152, GoogLeNet and VGG-16
- Data: Imagenet 2012.
- Metric of transferability:
  - Non-targeted: Accuracy
  - Targeted: The number of samples predicted exactly the target class (on the transfered model), they called matching rate.
- Bound of distortion: RMSD

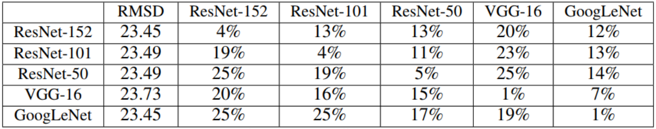
$$\sqrt{\sum_i (x_i^* - x_i)^2/N}$$

Experiment of non-targeted adversarial samples



	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
ResNet-152	22.83	0%	13%	18%	19%	11%
ResNet-101	23.81	19%	0%	21%	21%	12%
ResNet-50	22.86	23%	20%	0%	21%	18%
VGG-16	22.51	22%	17%	17%	0%	5%
GoogLeNet	22.58	39%	38%	34%	19%	0%

Panel A: Optimization-based approach

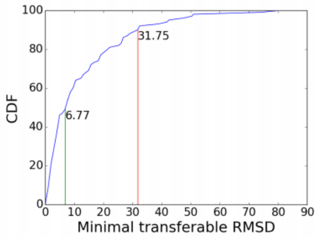


	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
ResNet-152	23.45	4%	13%	13%	20%	12%
ResNet-101	23.49	19%	4%	11%	23%	13%
ResNet-50	23.49	25%	19%	5%	25%	14%
VGG-16	23.73	20%	16%	15%	1%	7%
GoogLeNet	23.45	25%	25%	17%	19%	1%

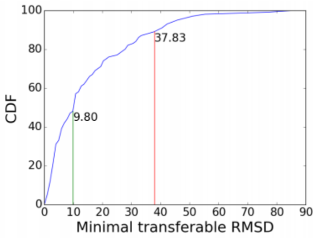
Panel B: Fast gradient approach

Transfers well with a large distortion.
FGS result is worse than FG and optimization-based approaches.

Minimal transferable RMSD

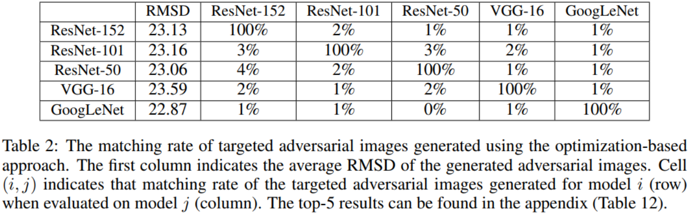


(a) Fast Gradient



(b) Fast Gradient Sign

Experiment of targeted adversarial samples



	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
ResNet-152	23.13	100%	2%	1%	1%	1%
ResNet-101	23.16	3%	100%	3%	2%	1%
ResNet-50	23.06	4%	2%	100%	1%	1%
VGG-16	23.59	2%	1%	2%	100%	1%
GoogLeNet	22.87	1%	1%	0%	1%	100%

Table 2: The matching rate of targeted adversarial images generated using the optimization-based approach. The first column indicates the average RMSD of the generated adversarial images. Cell (i, j) indicates that matching rate of the targeted adversarial images generated for model i (row) when evaluated on model j (column). The top-5 results can be found in the appendix (Table 12).

Doesn't transfer well.

Attack ensembles

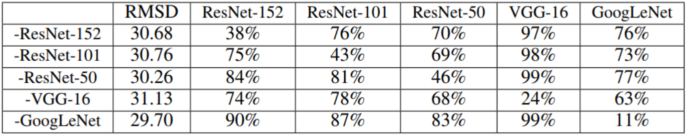
- Generate adversarial images for the ensemble of the (five) models.

Ensemble attack

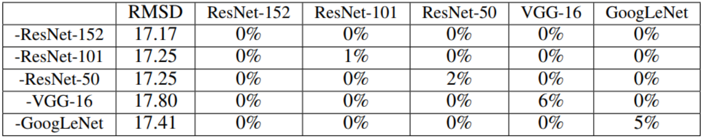
$$\arg\min_{x^*} \lambda d(x, x^*) + \text{Loss}(\sum_i \alpha_i F_i(x; \theta)) \tag{5}$$

Experiment result of attacking ensembles

- For each of the five models, we treat it as the black-box model to attack, and generate adversarial images for the ensemble of the rest four, which is considered as white-box.



	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
-ResNet-152	30.68	38%	76%	70%	97%	76%
-ResNet-101	30.76	75%	43%	69%	98%	73%
-ResNet-50	30.26	84%	81%	46%	99%	77%
-VGG-16	31.13	74%	78%	68%	24%	63%
-GoogLeNet	29.70	90%	87%	83%	99%	11%



	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
-ResNet-152	17.17	0%	0%	0%	0%	0%
-ResNet-101	17.25	0%	1%	0%	0%	0%
-ResNet-50	17.25	0%	0%	2%	0%	0%
-VGG-16	17.80	0%	0%	0%	6%	0%
-GoogLeNet	17.41	0%	0%	0%	0%	5%

Observations

- Several observations found:
- The gradient directions of different models in our evaluation are almost orthogonal to each other.
- Decision boundaries of the non-targeted approaches using different models align well.

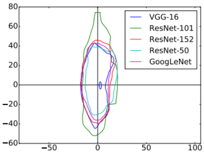


Figure 4: The decision boundary to separate the region within which all points are classified as the ground truth label (encircled by each closed curve) from others. The plane is the same one described in Figure 3. The origin of the coordinate plane corresponds to the original image. The units of both axises are 1 pixel values.

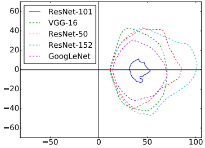
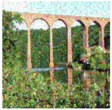
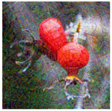
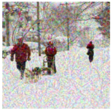
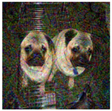


Figure 5: The decision boundary to separate the region within which all points are classified as the target label (encircled by each closed curve) from others. The plane is spanned by the targeted adversarial direction and a random orthogonal direction. The targeted adversarial direction is computed as the difference between the original image in Figure 2 and the adversarial image generated by the optimization-based approach for an ensemble. The ensemble contains all models except ResNet-101. The origin of the coordinate plane corresponds to the original image. The units of both axises are 1 pixel values.

On realistic data

- Clarifai.com: A commercial company providing state-of-the-art image classification services
- Submit adversarial samples to the website.
- Result:
  - Non-targeted: 57% of the targeted adversarial examples generated using VGG-16, and 76% of the ones generated using the ensemble can mislead Clarifai.com to predict labels irrelevent to the ground truth
  - Targeted: 18% of those generated using the ensemble model can be predicted as labels close to the target label by Clarifai.com. The corresponding number for the targeted adversarial examples generated using VGG-16 is 2%.

Some adversarial samples

original image	true label	Clarifai.com results of original image	target label	targeted adversarial example	Clarifai.com results of targeted adversarial example
	viaduct	bridge, sight, arch, river, sky	window screen		window, wall, old, decoration, design
	hip, rose hip, rosehip	fruit, fall, food, little, wildlife	stupa, tope		Buddha, gold, temple, celebration, artistic
	dogsled, dog sled, dog sleigh	group together, four, sledge, sled, enjoyment	hip, rose hip, rosehip		cherry, branch, fruit, food, season
	pug, pug-dog	pug, friendship, adorable, purebred, sit	sea lion		sea seal, ocean, head, sea, cute