Improved Forward-backward Propagation to Generate Adversarial Examples

Yuying Hao[1], Tuanhui Li[2], Yang Bai[1], Li Li[2], Yong Jiang[2], and Xuanye Cheng[3]
[1]Tsinghua-Berkely Shenzhen Institute, Tsinghua University, Shenzhen, China
[2]Graduate School at Shenzhen, Tsinghua University, Shenzhen, China
[3]SenseTime Research, SenseTime, Shenzhen, China
Contact Author: (haoyy17@mails.tsinghua.edu.cn)

ICANN19
28th International Conference on Artificial Neural Networks

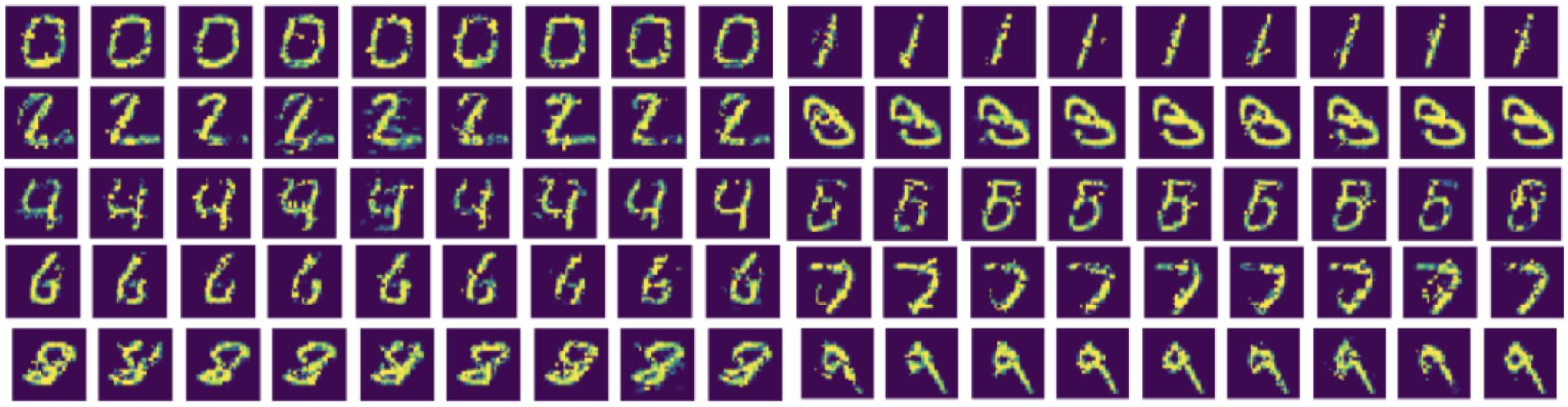


Fig.1 Our method applied on MNIST performs targeted attacks. The generated images in each row have all labels in order except its original label.(For example, in the first row for 0, adversarial images are listed with targeted label 1-2-3-4-5-6-7-8-9.)

Contributions

(1)We combine forward and backward propagation to add sparse perturbations and introduce an excellent approach to select sensitive pixels for misclassification.

(2)We introduce a novel loss function, which can smooth and reduce the perturbations and achieve the goals of targeted attack. More specifically, the l_0 norm can be converted into a derivable function.

The proposed method

1. Forward Derivative Local Attack

$$\nabla \mathbf{Z}_t(\mathbf{X}') = \frac{\partial \mathbf{Z}_t(\mathbf{X}')}{\partial \mathbf{X}'}$$

$$\mathbf{Pert}[p,q,n] = d_i / (\sum_{i=1} |d_i|)$$

where $\mathbf{Z}_t(\mathbf{X}')$ is the output of logit layer with $\mathbf{X}'$ and $[p,q,n]$ to represent the location of selected pixel.

2. Modeling for Loss Function

$$\begin{aligned} \min \quad & c_1 \|\mathbf{X} - \mathbf{X}'\|_2^2 \\ & + c_2 \sum_{x \in \mathbf{X}} clip\{255 * (|x - x'| - 0.0039), 0, 1\} \\ & + c_3 \max\{0, \mathbf{Z}_{max} - \mathbf{Z}_t\} \\ & + c_4 \max\{0, \mathbf{Z}_0 - \mathbf{Z}_t\} \\ \text{s.t.} \quad & \mathbf{X}' \in [0,1]^n . \end{aligned}$$

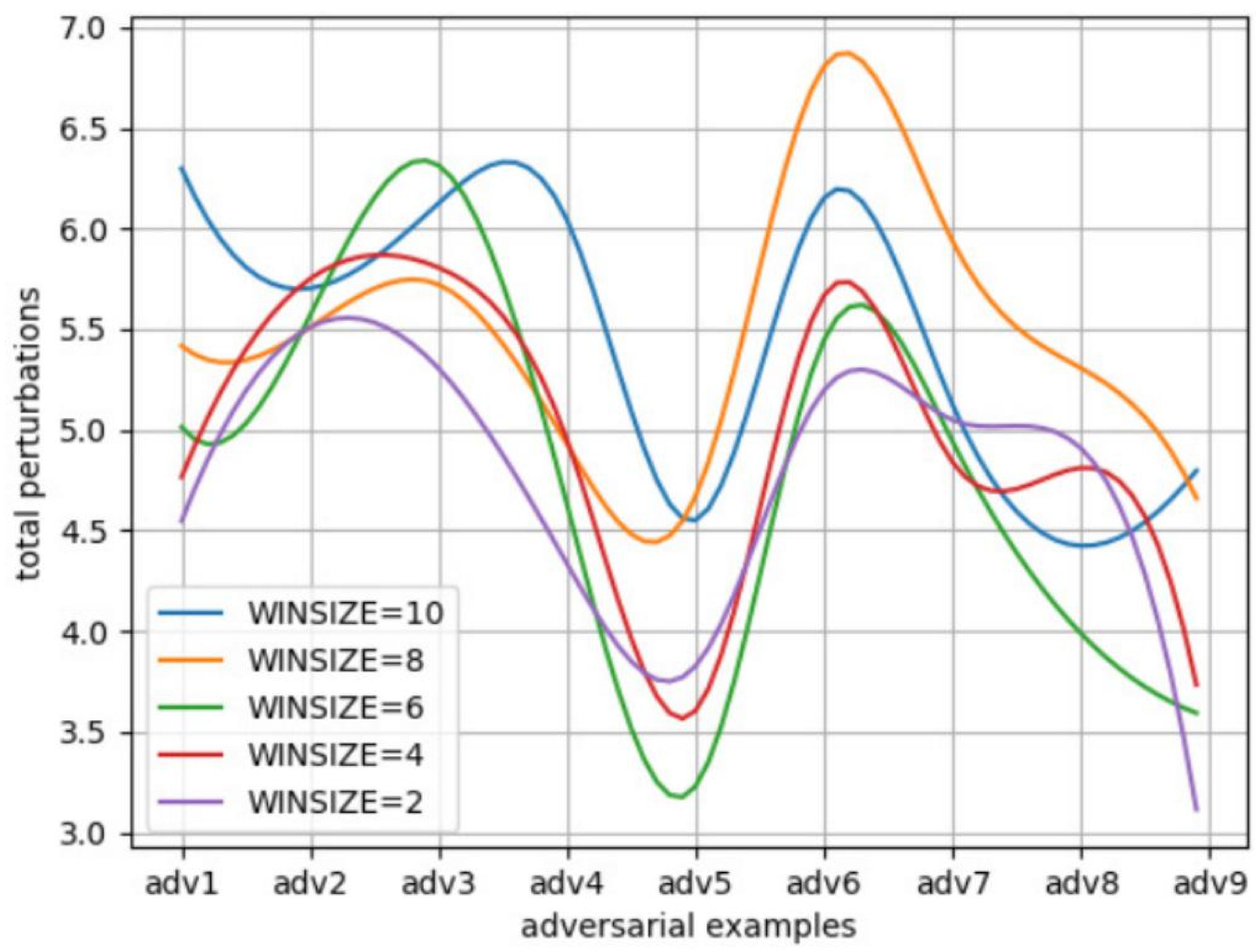
3. Algorithm

Algorithm 1 The propose algorithm for adversarial attack

Input: benign image $\mathbf{X}$; ground-truth y_0; target label y_t; local attack pixel number k; maximum iterations *maxiter*;
Output: $\mathbf{X}'$
while $iter < maxiter$ **do**
  Compute $\nabla \mathbf{Z}_t(\mathbf{X}')$ according to (5)
  Select the top-k magnitudes in $\nabla \mathbf{Z}_t(\mathbf{X}')$ and record the correspond position $[p,q,n]$.
  Calculate $\mathbf{Pert}[p,q,n]$ according to (6)
  if $m \leq \mathbf{Round}(k/2)$ **then**
    $\mathbf{Pert}[p,q,n] = \text{Floor}\ (\mathbf{Pert}[p,q,n])$
  else
    $\mathbf{Pert}[p,q,n] = \text{Ceil}\ (\mathbf{Pert}[p,q,n])$
  end if
  Compute $\mathbf{Mod} = \nabla F_{loss}(\mathbf{X}')$ according to (10)
  $\mathbf{X}' = \mathbf{X}' - r_1 * \mathbf{Mod} + r_2 * \mathbf{Pert}$
  if $\arg\max\ (\mathbf{Z}(\mathbf{X}')) = y_t$ **then**
    Return $\mathbf{X}'$
  end if
end while

Experiments

➢ The effect of WINSIZE on total perturbations on MNIST.



➢ White-box attack on MNIST, T and K are different CNN structures.

Model	Method			
	FGSM	Deepfool	C&W-ℓ_2	Ours
K	43.55	15.58	0.75	0
T	1.32	1.885	1.5	0

➢ Transferability of different methods on MNIST dataset.

method	type	acc(%)
FGSM	Adversarial training	30.93
	Black-box attack	32.73
Virtual	Adversarial training	33.53
	Black-box attack	36.04
C&W-ℓ_2	Adversarial training	34.83
	Black-box attack	40.24
Deepfool	Adversarial training	6.61
	Black-box attack	11.41
Ours	Adversarial training	6.06
	Black-box attack	6.17

Contact us by Wechat