Transferable Adversarial Attacks for Deep Scene Text Detection

Shudeng Wu*, Tao Dai*†§, Guanghao Meng*, Bin Chen*†, Jian Lu‡, Shu-Tao Xia*†
*Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
†PCL Research Center of Networks and Communications, Peng Cheng Laboratory, Shenzhen, China
‡ Shenzhen Key Laboratory of Advanced Machine Learning and Applications, Shenzhen University, China
{wsd19,mgh19, cb17}@mails.tsinghua.edu.cn,
daitao.edu@gmail.com, jianlu@szu.edu.cn, xiast@sz.tsinghua.edu.cn

Introduction

- **Scene Text Detection** aim to locate text from images in natural scene and have been widely used in license plate recognition, road sign recognition, image retrieval, commercial recommendation, etc.
- **Traditional STD methods** (using hand-crafted features): 1) sliding window; 2) connected component.
- **DNN-based models:**
  - 1) regression-based; directly regress bounding box of text instances.
  - 2) segmentation-based; segment text instances out of a scene image and then post-process the obtained segmentation map.
- **Challenges:**
  - Various architectures of models (regression-based and segmentation-based); **We define a probability map to formulate them.**
  - Objects are smaller and of great amount; **We design a threshold loss to avoid attacking objects from missing.**

Introduction

- **Contributions:**
  - Our attack is capable to craft adversarial examples for various DNN-based STD methods (regression-based and segmentation-based). We adopt a threshold loss that strictly prevents attacking targets from missing.
  - We design specific and universal attacks and both of them have good attack performance.
  - We make black-box evaluation by attacking a real-world STD engine of Google OCR, which verifies the potential applications of our attacks in practice.

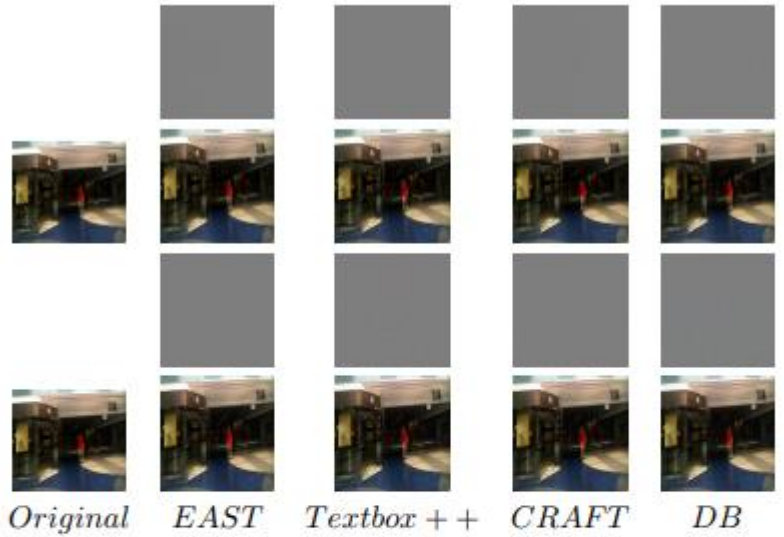


Fig. 1: Perturbations ($\epsilon = 11/255$) and adversarial examples crafted by the specific attack and universal attack. The first and second rows show perturbations and adversarial examples crafted by the specific attack, and the third and fourth rows show perturbations and adversarial examples crafted by the universal attack. Note that RGB values (127, 127, 127) are already added to the perturbations, resulting in the grey background. This is required for visualization as the perturbation can be negative.

Method

- **Attacked models:**
  - Regression-based:
    1. EAST: an anchor-free method, regress relative positions of corner points based on a single feature vector in the feature map.
    2. Textbox++: is a variant of SSD, applies quadrilateral regression to detect multi-oriented text.
  - Segmentation-based:
    1. CRAFT: is a character-level detector that treats the segmentation of a text instance as a heatmap, whose central points have higher values.
    2. DB: adaptive thresholds to distinguish which pixels are text or not by comparing its probability with the corresponding threshold.
- **Problem Formulation:**
  - Both regression-based and segmentation-based methods make predictions based on the final feature maps. 1) For regression-based, every position in the feature map correspond with one or more proposals; 2) for segmentation-based, every position in the feature map correspond with a probability indicating it's text or not, we dub it proposal as well.
  - Probability map: $M \in R^{h \times w \times K}$
    - Anchor-free or segmentation-based: $K = 1$;
    - h, w are the height and width of final feature map.

Method

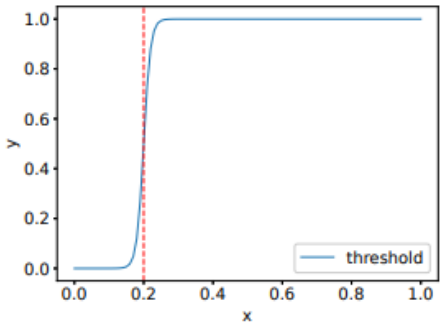
- **Specific Attack:**

$$L(\delta, x, D, \theta) = \sum_{k=1}^{K} \sum_{(i,j)\in D} \frac{1}{1 + e^{-c(f_{i,j}^{(k)}(x+\delta)-t)}} \quad (4)$$

$$s.t. \quad \delta \in (-\epsilon, \epsilon).$$

$$L(w, x, D, \theta) = \sum_{k=1}^{K} \sum_{(i,j)\in D} \frac{1}{1 + e^{-c(f_{i,j}^{(k)}(x+\epsilon(\frac{1-e^w}{1+e^w}))-t)}} \quad (5)$$

$$\delta = \epsilon(\frac{1 - e^w}{1 + e^w}).$$



(a)

Fig. 2: Illustration of threshold loss function $y = \frac{1}{1+e^{-c(x-t)}}$ $(c = 100, t = 0.2)$.

Algorithm 1 Specific Attack

Input: model function $f(\cdot)$; input image x; point set of text instances D; maximum iterations T; text threshold t; stop loss threshold l_{stop}; perturbation size ϵ; step α.

Output: the adversarial example x'

$w \leftarrow \mathbf{0}, i \leftarrow 0, l \leftarrow 0$
while $i < T$ **and** $l > l_{stop}$ **do**
  $l \leftarrow L(w, x, D, \theta)$
  $w \leftarrow w - \alpha \cdot sign(\nabla_w L(w, x, D, \theta))$
  $i \leftarrow i + 1$
end while
$\delta \leftarrow \epsilon(\frac{1-e^w}{1+e^w})$
return $clip_0^1(x + \delta)$

Method

- **Universal Attack:**

$$L(\delta, x, D, \theta) = \sum_{k=1}^{K} \sum_{(i,j)\in D} \frac{1}{1 + e^{-c(f_{i,j}^{(k)}(x+\delta)-t)}} \quad (4)$$

$$s.t. \quad \delta \in (-\epsilon, \epsilon).$$

$$L(w, x, D, \theta) = \sum_{k=1}^{K} \sum_{(i,j)\in D} \frac{1}{1 + e^{-c(f_{i,j}^{(k)}(x+\epsilon(\frac{1-e^w}{1+e^w}))-t)}} \quad (5)$$

$$\delta = \epsilon(\frac{1-e^w}{1+e^w}).$$

Algorithm 2 Universal Attack

Input: model function $f(\cdot)$; input images $\{x_1, x_2, ..., x_n\}$; point sets of text instances $\{D_1, D_2, ..., D_n\}$; epoch number e; text threshold t; perturbation size ϵ; step size α.

Output: an universal perturbation δ ; the adversarial example $\{x_1^{'}, x_2^{'}, ..., x_n^{'}\}$

$w \leftarrow \mathbf{0}, i \leftarrow 0$
while $i < e$ **do**
  $j \leftarrow 0$
  while $j < n$ **do**
    $w \leftarrow w - \alpha \cdot sign(\nabla_w L(w, x_j, D_j, \theta))$
    $j \leftarrow j + 1$
  end while
  $i \leftarrow i + 1$
end while
$\delta \leftarrow \epsilon(\frac{1-e^w}{1+e^w})$
$\{x_1^{'}, x_2^{'}, ..., x_n^{'}\} \leftarrow \{clip_0^1(x_1 + \delta), ..., clip_0^1(x_n + \delta)\}$
return δ, $\{x_1^{'}, x_2^{'}, ..., x_n^{'}\}$

Experiments Setting

- **Attacked models**: EAST (VGG16), Textbox++ (ResNet-50) , CRAFT (VGG16), DB (ResNet-50).
- **Dataset:** ICDAR 2015, multi-oriented; Total-Text, curve and multi-oriented.
- **Evaluation metrics:** precision, recall, mAP.

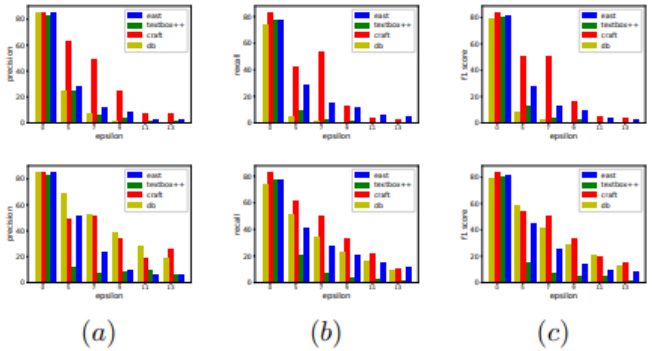
Overall Results

TABLE I: Results of original images and adversarial examples crafted by specific attack ($\epsilon = 11/255$).The values in each table cell "a/b" means metric value on original images (a) and on adversarial examples (b). "P", "R", "F" indicate precision, recall and f1 score, respectively.

	ICDAR 2015			Total-Text		
	P	R	F	P	R	F
EAST	85.9/2.5	77.9/5.3	81.7/3.4	-	-	-
Textbox++	83.7/2.0	77.9/0.6	80.7/1.0	-	-	-
CRAFT	85.4/7.2	83.7/3.3	84.6/4.5	78.6/8.6	80.0/9.9	79.3/9.2
DB	85.0/0.6	74.6/0.2	79.4/0.3	79.5/0.9	79.4/1.0	79.4/0.9

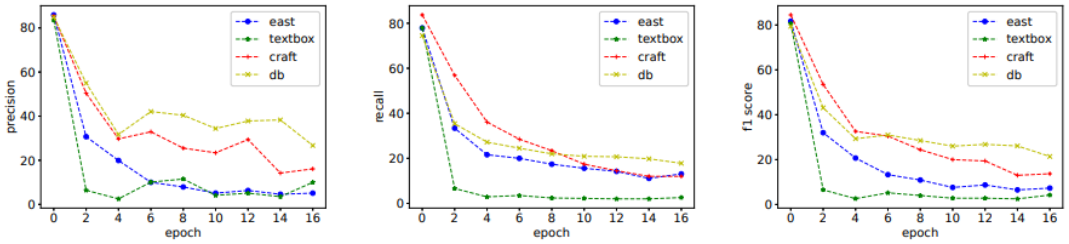
TABLE II: Results of original images and adversarial examples crafted by universal attack ($\epsilon = 11/255$). The values in each table cell "a/b" means metric value on original images (a) and on adversarial examples (b). "P", "R", "F" indicate precision, recall and f1 score, respectively.

	ICDAR 2015			Total-Text		
	P	R	F	P	R	F
EAST	85.9/6.2	77.9/12.0	81.7/8.2	-	-	-
Textbox++	83.7/5.9	77.9/1.3	80.7/2.1	-	-	-
CRAFT	85.4/26.2	83.7/11.3	84.6/15.8	78.6/17.3	80.0/12.6	79.3/14.6
DB	85.0/19.0	74.6/10.0	79.4/13.1	79.5/21.8	79.4/24.3	79.4/22.9



(a) (b) (c)

Fig. 3: Comparison of the precision (a), recall (b) and f1 score (c) with respect to different ϵ values for the specific attack and the universal attack. The subfigures in the first row show the metrics of the specific attack and the subfigures in the second row show the metrics of the universal attack.



(a) (b) (c)

Fig. 4: Comparison of precision (a), recall (b) and f1 score (c) of adversarial examples crafted by universal attack against different models with respect to different epoch number. The comparision is conducted on ICDAR 2015 ($\epsilon = 11/255$).

Cross-dataset transfer attack

TABLE III: Illustration of cross-dataset transfer attack. "P_i", "P_t" indicate universal perturbations ($\epsilon = 11/255$) crafted on ICDAR 2015 and Total-Text respectively. The gray rows are the results of cross-dataset transfer attack. "P", "R", "F" indicate precision, recall and f1 score, respectively.

	CRAFT			**DB**		
	P	R	F	P	R	F
ICDAR 2015	85.4	83.7	84.6	85.0	74.6	79.4
ICDAR 2015+P_i	26.2	11.3	15.8	17.3	12.6	14.6
ICDAR 2015+P_t	21.0	65.4	31.8	21.0	33.9	25.9
Total-Text	78.6	80.0	79.3	79.5	79.4	79.4
Total-Text+P_i	79.5	74.1	76.7	61.6	65.2	63.3
Total-Text+P_t	19.0	10.0	13.1	21.8	24.3	22.9

Cross-model transfer attack

TABLE IV: Results of cross-model transfer attack and black-box attack with adversarial examples crafted by specific attack ($\epsilon = 11/255$). Each table cell is in the "P/R/F" format, where "P","R","F" indicate precision, recall and f1 score respectively.

	EAST	Textbox++	CRAFT	DB	Google OCR
Original	85.9/77.9/81.7	83.6/77.9/80.6	85.4/83.7/84.5	84.9/74.5/79.4	65.8/88.8/75.6
EAST	2.5/5.3/3.4	83.9/71.4/77.2	83.9/80.5/82.2	83.7/71.4/77.1	68.3/79.7/71.1
Textbox++	43.4/69.4/53.4	2.0/0.6/1.0	60.2/80.3/68.8	**85.7/66.0/74.6**	66.0/80.1/72.4
CRAFT	85.1/63.8.72.9	83.8/71.8/77.4	7.2/3.3/4.5	84.8/70.9/77.2	66.2/76.8/71.1
DB	**41.6/25.4/31.5**	**82.2/38.3/52.3**	**53.3/52.6/53.0**	0.6/0.2/0.3	**63.7/63.4/63.5**

TABLE V: Results of cross-model transfer attack and black-box attack with adversarial examples crafted by universal attack ($\epsilon = 11/255$). Each table cell is in the "P/R/F" format, where "P","R","F" indicate precision, recall and f1 score respectively.

	EAST	Textbox++	CRAFT	DB	Google OCR
Original	85.9/77.9/81.7	83.6/77.9/80.6	85.4/83.7/84.5	84.9/74.5/79.4	65.8/88.8/75.6
EAST	6.2/12.0/8.2	85.5/67.1/75.2	83.2/70.9/76.6	85.8/60.7/71.1	67.2/73.7/70.3
Textbox++	17.7/40.6/24.7	5.9/1.3/2.1	**9.1/60.4/15.8**	**78.3/37.9/51.1**	**54.9/67.4/60.5**
CRAFT	83.5/37.2/51.5	84.4/61.6/71.2	26.2/11.3/15.8	84.7/56.0/67.4	67.6/72.8/70.1
DB	**14.7/33.9/20.5**	**89.9/0.5/64.2**	43.8/60.0/50.6	19.0/10.0/13.1	69.2/62.1/65.4

Conclusion

- **Some findings:**
  - Adversarial examples crafted by more complicated backbone (ResNet) are more transferable and have better attack performance.
  - Models with complicated backbone (ResNet) are more robust and harder to be transfer attacked.
  - Adversarial examples crafted by universal attacks are more transferable.

- **Conclusions:**
  - we are the first to propose a generic and efficient attack method against both regression-based and segmentation-based STD models.
  - We define a generic probability map for all STD models and optimize a threshold loss that can prevent attacking targets from missing.
  - We then conduct extensive experiments to evaluate our attack method on four state-of-the-art STD models and a real-world STD engine of Google OCR, in which our method consistently degrades the detection accuracy of these models.