Imperceptible Adversarial Attacks against Traffic Scene Recognition

Yinghui Zhu
Hanshan Normal University
Yuzhen Jiang (jiangyz@hstc.edu.cn)
Hanshan Normal University

Research Article

Keywords: Adversarial Example, Scene Recognition, Image Classifier, Semantic Segmentation

Posted Date: July 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-652216/v1

Imperceptible Adversarial Attacks against Traffic Scene Recognition

Yinghui Zhu
School of Computer and Information Engineering, Hanshan Normal University, Chaozhou, Guangdong, China,
zyh366@163.com
Yuzhen Jiang (Corresponding Author)
School of Computer and Information Engineering, Hanshan Normal University, Chaozhou, Guangdong, China,
jiangyz@hstc.edu.cn

ABSTRACT:

Adversarial examples have begun to receive widespread attention owning to their potential destructions to the most popular DNNs. They are crafted from original images by embedding well calculated perturbations. In some cases the perturbations are so slight that neither human eyes nor monitoring systems can notice easily and such imperceptibility makes them have greater concealment and damage. For the sake of investigating the invisible dangers in the applications of traffic DNNs, we focus on imperceptible adversarial attacks on different traffic vision tasks, including traffic sign classification, lane detection and street scene recognition. We propose an universal logits map-based attack architecture against image semantic segmentation, and design two targeted attack approaches on it. All the attack algorithms generate the micro-noise adversarial examples by the iterative method of gradient descent optimization. All of them can achieve 100% attack rate but with very low distortion, among which, the mean MAE (Mean Absolute Error) of perturbation noise based on traffic sign classifier attack is as low as 0.562, and the other two algorithms based on semantic segmentation are only 1.574 and 1.503. We believe that our research on imperceptible adversarial attacks can be of substantial reference to the security of DNNs applications.

Keywords: Adversarial Example; Scene Recognition; Image Classifier; Semantic Segmentation

1. INTRODUCTION

The DNNs applications have shown significant potential in computer vision. However some research works in recent years disclose that DNNs are vulnerable to adversarial attacks[6,10]. With some special perturbation on raw images, the adversarial examples can easily fool the models and make wrong outputs. Adversarial examples can attack not only image classifiers[6,12,13] but also semantic segmentation and object detection models[4,5,8,9,11]. However, most existing attack algorithms are with visible distortion or too much additional noise[7,16,18,19], which can easily be detected and thus lose their attacks. It is still a large challenge to design adversarial examples which is able to fool not only computer algorithms but also human eyes. In this paper, we focus more on the imperceptible adversarial examples, and propose three imperceptible adversarial attacks against different traffic scene recognition, including traffic sign classification, lane detection and street scene segmentation. Fig.1 shows the three adversarial attacks. Our experiment results reveal that adversarial attacks can be implemented in various network models, and the attack effects can be designed arbitrarily. Moreover, the perturbation noise can be very small and imperceptible to both human and computer. Our attacks are all white-box

attacks[9,10], the three networks involved are the famous MobileNetV2[1], U-net[2] and DeepLabV3+[3], which have been trained on well-known datasets: BelgiumTS[21], Pascal VOC[22] and Kitti[23], respectively.

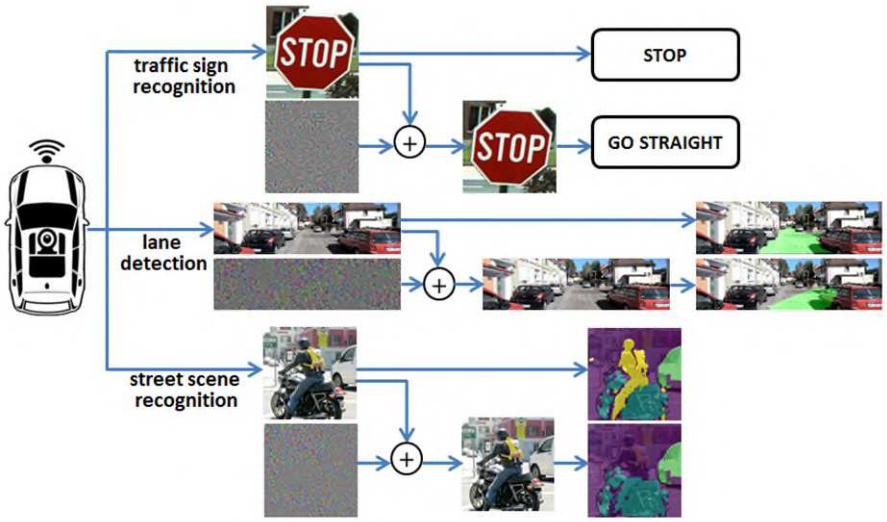


Figure 1 Adversarial attacks against traffic scene recognition

2. ADVERSARIAL ATTACKS

The concept of adversarial example is first mentioned in [6], Szegedy et al. revealed that, minor perturbations carefully crafted in an image can easily confuse the classifier, yet people can hardly notice them. The existence of adversarial examples are considered as a serious security threat to DNNs applications, therefore many scholars begin to develop the studies around adversarial attacks and adversarial defense. The fast gradient sign method(FGSM)[10] is regarded as a benchmark method owning to its simple and efficient operation. And later Kurakin et al. upgraded FGSM to the Basic Iterative Method(BIM)[12], an iterative extension version of FGSM. They also proposed a targeted attack method of BIM: Iterative Least-likely Class Method(ILCM)[12,20]. ILCM can fool the models to achieve any pre-set class, even the least-likely class. Besides that, Carlini and Wagner presented a powerful attack approach(C&W) by using the special objective function defined to suit for optimization[13], the approach is regarded as more robust while with less distortion. For seeking less distortion, Papernot et al. put forward the Jacobian-based Saliency Map Attack (JSMA) that can cause misclassification by explicitly modifying a few pixels[14]. Su et al. even introduced an extreme attack that can fool the DNNs by modifying only one pixel[15]. Noting that, at that stage, almost all the approaches were designed for attacking the image classifiers. Although validated on small datasets (like MNIST, etc.), most of them can be transfered to large datasets (like ImageNet, etc.). However some methods are found limitations in application such as JSMA and One-pixel attack, the former can not be used in large datasets because of the computational cost, and the latter has few actually affect due to its low success rate.

As the attack research progresses, many studies started trying to transfer the classifier attacks to other DNNs models such as semantic segmentation. But due to the architecture differences, most classifier attacking algorithms can not be directed used in semantic segmentation models. Anurag Arnab believed that it was much harder to generate the adversarial examples on semantic segmentation models than on classifiers [4,5], because: 1) The structures of semantic segmentation models are usually more complicated than classifier. 2) The attack purpose on semantic segmentation is flexible and need more

manipulations. 3) It is hard to set the success standard for an attack.

To investigating the feasibility and imperceptibility of adversarial attacks on various networks, we build three models regarding different application of traffic scene recognition, and try to attack them with the lowest distortion cost. Our experimental results show that, all the adversarial examples have nice qualities, and they are hard to be detected by both human or detection algorithm.

3. ADVERSARIAL ATTACK AGAINST TRAFFIC SIGN CLASSIFIER

In this section, we investigate several adversarial attacks against a traffic sign classifier. We adopt the MobileNetV2 pre-trained on ImageNet dataset as our target network. To train the model to classify various traffic signs, we fine-tune the MobileNerV2 by unlocking the untrainable properties of the last five parameter layers. The training dataset is Belgium traffic sign (BelgiumTS), which contains 62 types of traffic signs and have total 7125 images. After 50 epochs, the model achieved 0.971 training accuracy and 0.863 validation accuracy, and we turn to conduct various adversarial experiments on it to defeat the classification strategies.

3.1 Method

To generate the adversarial traffic sign images, we adopt C&W strategies into our program, which applies gradient descent optimization to solve the follow problem:

$$\text{minimize } \|I' - I\|_2 + a \cdot f(I') \tag{1}$$

$$f(I') = \max(\max\{Z(I')_i : i \neq t\} - Z(I')_i, -\kappa) \tag{2}$$

Where *I'* is the adversarial example of *I*, *f*(*I'*) represents the objective function of adversarial attack, $Z(I')_i$ denotes the *i-th* logits value. Logits is the prediction output before the final Softmax activation. The parameter *a* is the super-parameters used to adjust the proportion between L2_norm restraint and attack strength. The parameterκused to ensure the high confidence of the misled class. Combined with the empirical value in [13] and practically tested on the BelgiumTS datatset[28], we set α=100 andκ=20 (which can gain the superior attack performance and ensure the 100% success rate[13]). The method utilizes Adam optimizer to generate adversarial examples, and the learning rate is set with 0.01. In addition, to reduce the iterative [0,1] clips in optimization, we also apply a variable *w* to substitute *I'*:

$$I'_i = \frac{1}{2}(\tanh(w_i + 1)) \tag{3}$$

Fig. 2 show the diagram of Formula (3), no mater how *w* is modified, *I'* is always in [0,1] range. In addition, we also find that, the value of *I'* changes slightly when it near the 0 or 1, but when near the 0.5, it changes dramatically. Human eyes are more sensitive to white or bright color(pixel value is near 1), minor modifications in there are easily perceived. Therefore, the upper function can well protect the white or bright-color areas from severe modification thus keep the visual quality of whole image.

For investigating the performance of our method above, the ILCM、L-BFGS algorithms are also tested on the same model. Fig.3 shows the images of adversarial instances generated by different attack algorithms. All the three methods are targeted attacks, we appoint the target class is “Straight”, and stop the iterative optimization when it achieves the target class confidence not less than 0.9. Under the same attack conditions, the C&W has the smoothest noise fluctuations and least distortions, so it shows the best quality.

For further evaluating the perturbations of each algorithm more comprehensively, from the traffic

scene images correctly classified in the validation set of BelgiumTs, we randomly select 100 ones to generate adversarial examples by upper methods. The images are considered in [0, 255] range, their target class is set by: C_{target}=mod($C_{correct}$+30,62), where $C_{correct}$ is the correct class of original image. Three criteria: Mean Squared Error(MSE), Mean Absolute Error(MAE) and L-infinity norm(L_∞), are introduced to indicate the perturbation strength. The criteria are evaluated from two perspectives: the 0.9 Confidence Satisfied and the Lowest Threshold. For the 0.9 Confidence Satisfied, setting the learning rate(*lr*, for L-BFGS and C&W) or the factor of attack strength(ε, for ILCM) be 0.01, the adversarial sample is iteratively optimized until being identified as the target class with more than 0.9 confidence. The least threshold means the least perturbations to mislead model and we search them by doubling the *lr* (or theε)from 0.0001. Table 1 shows the results. MSE and MAE indicate the mean distortion of an image, L_∞ indicates the maximum pixel modification. On the Lowest Threshold, surprisingly, the MAEs of all methods are less than 1. Compared to the image's 0~255 value range, our modification scale is very slight and hard to be found. C&W achieves both the lowest MSE and the lowest MAE, which indicated its minimum amount of perturbation noise. Although it can't get the minimum L_∞ the adoption of Formula(3) can effectively avoid creating obvious noise in bright areas (i.e. eye sensitive areas) and protect its image quality .

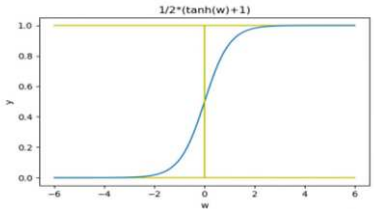


Figure 2 The Tanh function in C&W

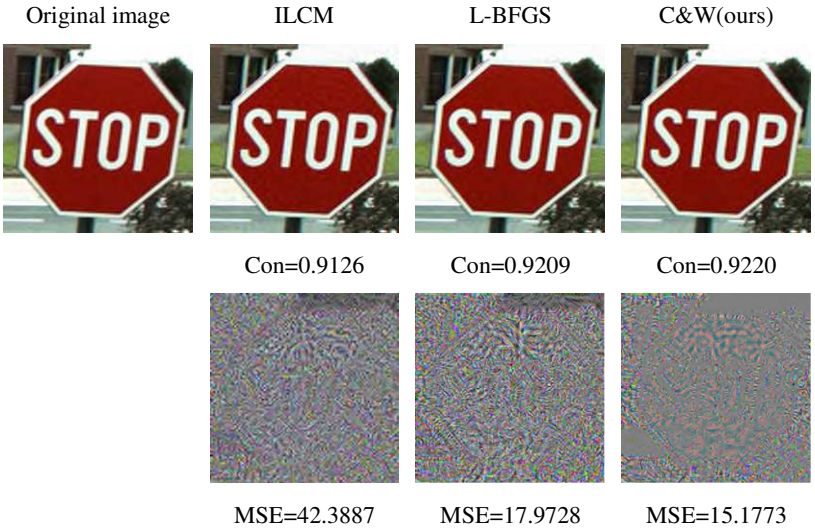


Figure 3: Adversarial examples and perturbations. The left image in first row is the original traffic sign "Stop", the other three images are all the adversarial examples that can fool the classifier to misclassify as sign "Go straight". The second row images are the respective perturbations which are all magnified and mapped to [0,1] range.

	Targeted confidence≥0.9 (lr=0.01)			Lowest threshold (lr≥0.0001)		
	MSE	MAE	L_∞	MSE	MAE	L_∞
ILCM	47.583	5.983	25.450	0.589	0.630	**1.296**

L-BFGS	23.796	4.466	**16.302**	0.698	0.691	1.569
C&W(ours)	**16.908**	**3.032**	19.632	**0.572**	**0.562**	2.242

Table 1 The MSE, MAE and L_∞ of different adversarial examples. The criteria are evaluated from two perspectives: the 0.9 Confidence Satisfied and the Lowest Threshold.

Inspired by the C&W superiority, we develop our semantic segmentation attacks refer to C&W optimization idea. However, the C&W objective function is not applicable to segmentation model directly, so we have to recraft proper optimization ways considering both the attack intent and the segmentation network feature. In the following, an universal attack architecture against semantic segmentation will be put forward.

4. SEMANTIC SEGMENTATION ATTACK

On semantic segmentation attacks, Non-targeted methods simply fool the models to get unrecognizable maps which are useless and meaningless while targeted attacks achieve the available maps that look "normal" but with covert destruction intent.Fig.4 shows four attack patterns, the first two rows show two non-targeted methods which output chaotic and striped maps [17]. The third row shows a category-hidden attack which outputs a "normal" map with little pedestrian information [19]. The fourth row shows an explicit attack which can indicate wrong driving direction by painting black lines on road image[18]. By contrast, the non-targeted methods with chaos or meaningless output map would be found out more easily. However, method in third row seems hard to remove all pedestrian areas(the red areas), while the clear addiction in the fourth method is also easy to notice.

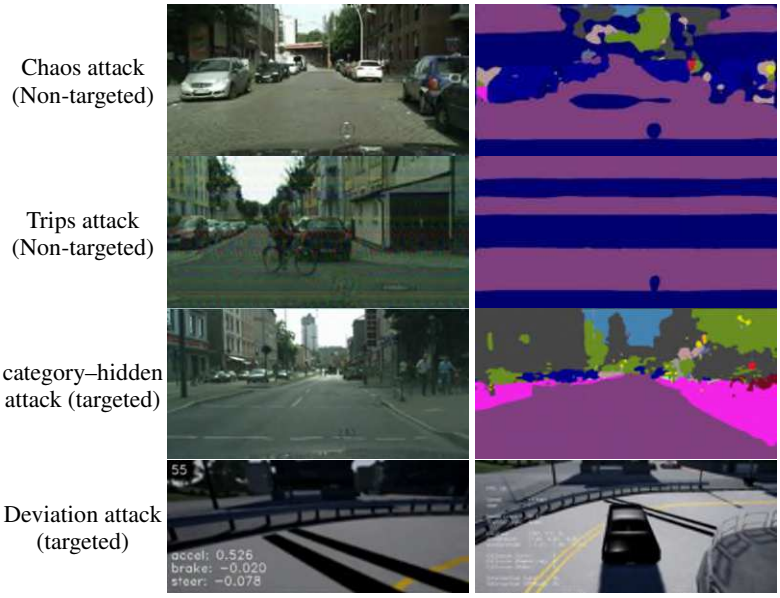


Figure 4: Various adversarial attack patterns

4.1 Universal attack architecture

To support the adversarial attacks on different models, we propose an universal adversarial architecture of semantic segmentation based on the targeted logits map in Fig. 5. In the architecture, the bottom black flow line represents the regular semantic segmentation prediction. The pixel-case targeted logits map is extracted from it and modified to be the targeted logits map(the yellow cuboid). Our adversarial example generation frame is in the top dotted rectangle, the red flow lines form a cycle

structure, where the adversarial example is iteratively optimized by constantly reducing the loss between two logits maps, till the attack purpose is satisfied.

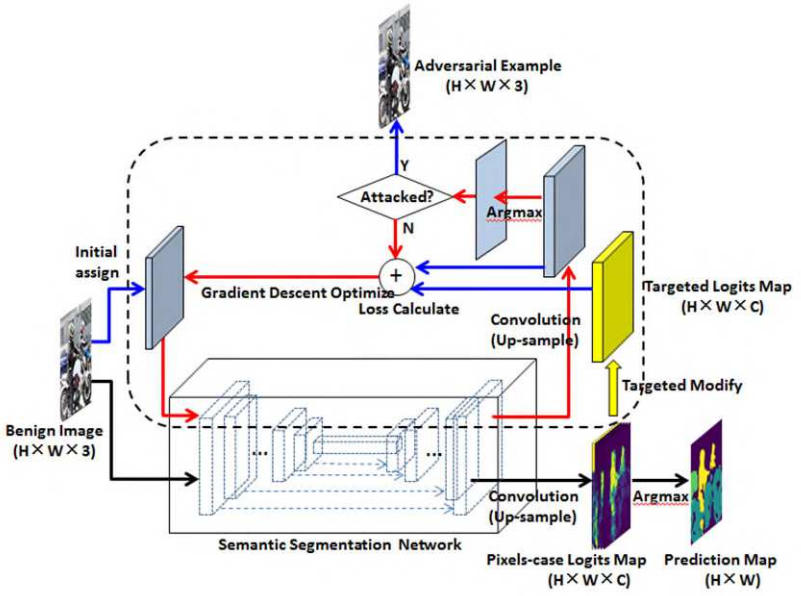


Figure 5 Universal adversarial architecture of semantic segmentation. The yellow cuboid represents the Targeted logits maps that are targeted modified by the pixels-case logits maps, which are the output layers before the final Argmax operation.

Based on this attack architecture, we will propose two attack methods against traffic scene segmentation according different tasks. Both methods also use Tanh function described in formula (3) to improve their imperceptibility.

4.2 Lane attack

In this section, we provide an adversarial attack method with the purpose to deflect the detected lane. Both the target direction and deflection angle can be set arbitrarily. The state-of-the-art DeepLabV3+ network is adopted as our attack model. After training on the KITTI dataset(a famous road scenes dataset)[23], the model achieve the 0.974 training accuracy and 0.921 validation accuracy. Then we begin to perform the adversarial attack against it. We construct the targeted logits maps by shearing the original logits maps: define $Z_{x,y,c}$ to be the logits value at (x,y,c). Here x,y represent the pixel locations and c represents current class channel. $Z'_{x,y,c}$ represents the targeted logits at (x,y,c) and it is set as:

$$Z'_{x,y,c} = Z_{x',y,c} \tag{4}$$

Where x' is obtained with：

$$x' = mod(x + \lfloor -tan(\theta) \cdot y \rfloor, W) \tag{5}$$

Theθin the formula is the attack angle, positive is to left and negative is to right. W stands for the width of original image. After constructing Z', the adversarial example can be generated by solving the following optimization problem:

$$minimize \ \ \|I' - I\|_2 + a \cdot \|Z' - Z\|_2 \tag{6}$$

The parameter a is the balance factor between the two terms, perturbation quantity and attack magnitude. We empirically set a=100, which can enable most images to implement attack within a configured iterations number.

Fig. 6 demonstrates the attack against lane scene detection. We extract its pixels-case logits map

while predicting and shear it by left and right 45 degrees. Here lr = 0.001, and the adversarial examples are generated after 300 iterations of optimization. Attacked results show obvious deflections with about ±45 degrees yet the adversarial examples seem the same as the raw image. Fig. 7 shows an anther instance with ±60 degrees deflection attacks.

100 random images are taken from the KITTI validation dataset to be tested the same +45 degree attack. And the mean MSE, MAE and L_∞ are evaluated from two attacking intensities: Ordinary attack(lr=0.001 with 300 iterations) and Low-distortion attack (lr=0.0001 with 1000 iterations). The two settings can obtain similar attacking results, but bring different perturbations. Actually the perturbations under the two settings are both imperceptible, the less learning rates the less perturbations, but with more iterations. By setting lr=0.0001, our methods can significantly mislead the lane detection model by a very slight perturbation whose average MAE is only 1.503.

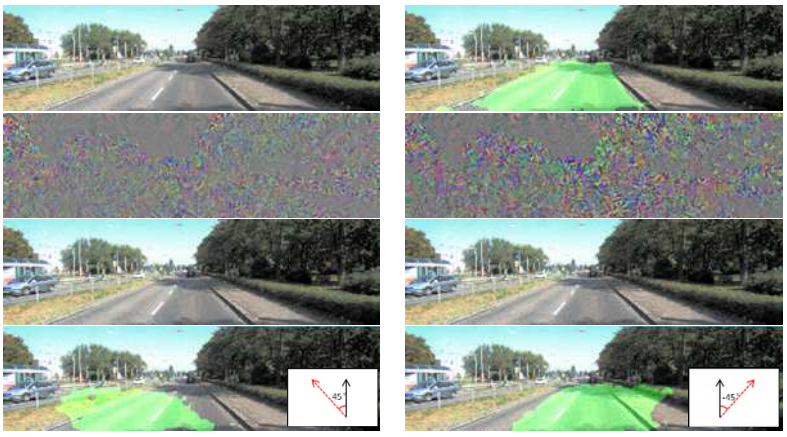


Figure 6 ±45 degrees deflection attacks on lane detection. The left image in the first row is the original image. The right one is its normal prediction output in DeepLabV3+ model. The second row shows the magnified perturbations (+45 degrees attack in the left image and -45 degrees attack in the right image) and the third row shows the adversarial examples. The images in last row are the final attacked prediction.

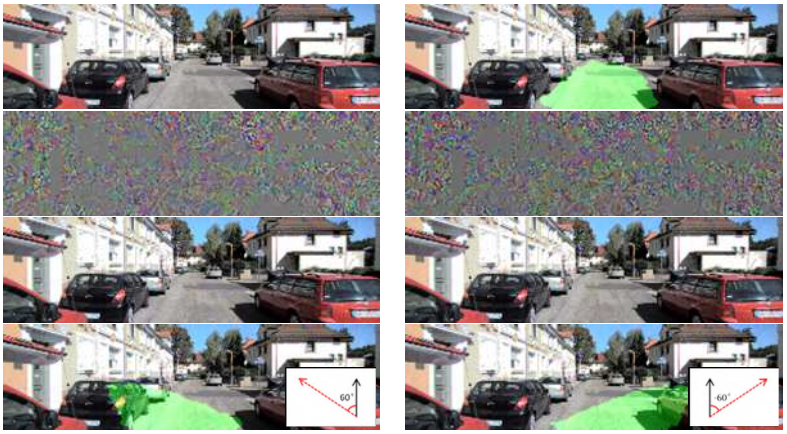


Figure 7: ±60 degrees deflection attacks against lane detection

Ordinary attack			Low-distortion attack		
MSE	MAE	L_∞	MSE	MAE	L_∞
7.364	2.055	16.494	4.014	1.503	8.828

Table 2 MSE,MAE and L∞ of adversarial examples against street scene recognition. The criteria are evaluated from two perspectives: Ordinary attack(lr=0.001, iter=300) and Low-distortion attack (lr=0.0001, iter=1000).

4.3 Street scene recognition attack

In this Section we demonstrate another semantic segmentation attack, category-hidden attack. We will fool the model to see none of peoples. The model adopted in this task is the U-net networks[2], and the training and testing datasets are created from the Pascal VOC2012 datasets[22]. We select all the images with people, cars, motorcycles and bicycles from Pascal VOC2012 datasets to form a 5–category subset. Besides the four categories above, the 5th category represents background. After training, the model achieves 0.971 training accuracy and 0.863 validation accuracy. The attack purpose of this task is to hide all the people in the scenes yet not affect the identifications of other categories. So the targeted logits maps must be crafted in advance.

To hide the "people", the first idea is to directly set all C_{people} in Z' be 0, but that would make it hard for algorithm to convergence and generate an adversarial example. Here C_{people} represents the "people" channel. Therefore, before setting it to 0, we distribute the "people" logits data to other channels. Each "people" data in logits map is added to another logits channel, e.g. another category, which is in the same pixel and has the maximum data among all categories except "people". Z' can be obtained as follows:

$$Z' = Z \tag{7}$$

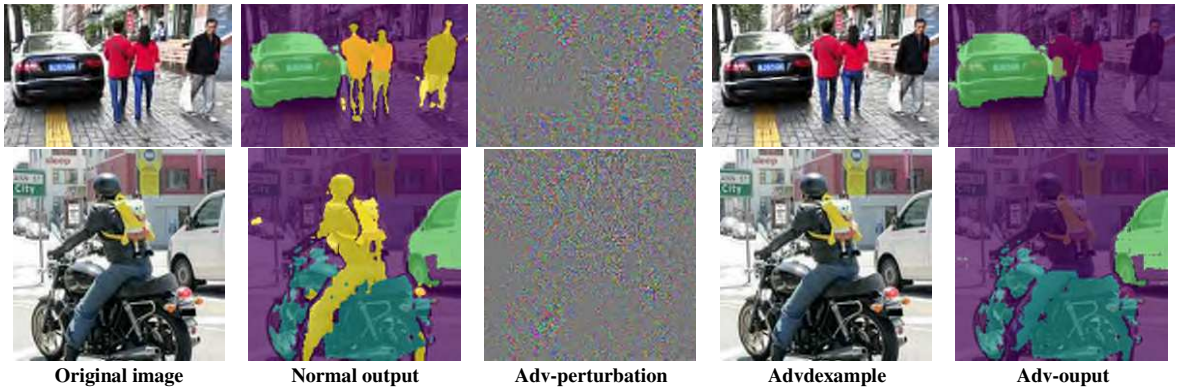
$$Z'_{i,j,C_{embed}} = Z_{i,j,C_{embed}} + Z_{i,j,C_{people}} \tag{8}$$

$$Z'_{i,j,C_{people}} = 0 \tag{9}$$

Where C_{embed} is the special channel to receive the "people" data, it would be found as follow:

$$C_{embed} = \arg\max_{C} Z'_{i,j,C} \quad s.t. \quad C \neq C_{target} \tag{7}$$

With this modified Z', we conduct the street scene attack whose optimization mechanism is similar to the our Lane attack. The termination condition of this task is, none of pixels in final prediction map belongs to "people" category. Fig. 8 shows two adversarial instances. Their final output maps demonstrate, all the segmentation areas of "people" are completely disappears, and they have been replaced by other category pixels. 100 images with people are selected from validation datasets and attacks tests are conducted on them with two intensities: Ordinary attack(lr=0.001) and Low-distortion attack(lr=0.0001). Table 3 shows the mean MSE，MAE and L_∞. It tends to the same conclusion, that our people-hidden method can also imperceptibly fool the recognition of street scenes. With Low-distortion attack, the mean MAE is only 1.574.



Original image **Normal output** **Adv-perturbation** **Advdexample** **Adv-ouput**

Figure 8 People-hidden attacks on street scene recognition. The original images are given on the left column. The normal recognition results are shown on the second column. The attack perturbations and adversarial examples are shown respectively on the third and four columns. The final recognitions of each adversarial example are

shown on the right column.

Ordinary attack			Low-distortion attack		
MSE	MAE	L_∞	MSE	MAE	L_∞
13.665	2.580	14.358	4.683	1.574	11.430

Table3 MSE,MAE and L∞ of adversarial examples against street scene recognition. The criteria are evaluated from two perspectives: Ordinary attack and Low-distortion attack. Their termination condition is that none of the pixels belong to the "people".

5. CONCLUSION

As a substitute for the human eye, DNNs is widely used in many recognition tasks. For some highly critical applications, such as self-driving cars, safety is a major concern.However, recent research has shown that DNNs are vulnerable. Our experimental results confirm DNNs vulnerability.In addition, it also reveals the following points :1) There is great flexibility in the design and implementation of the purpose of adversarial attacks.Attacks can be customized arbitrarily, which is a great threat to the security of system applications. 2) The disturbing noise in the adversarial example may be very slight, which is difficult to detect by human eyes or monitoring system, and such attacks are more covert and dangerous.

Therefore, we also put forward some Suggestions for the safe use of DNNs. In addition to enhancing the robustness of the model through adversarial training, attention should also be paid to :1) In order to ensure the reality of the image input, it is better to integrate video/image acquisition and recognition processing to reduce the transit link in the transmission process, because this stage is the easiest to embed malicious disturbance. 2) In the recognition stage, it is better to carry out multi-path synchronous prediction, and the additional path will carry out some adversarial defense preprocessing before the prediction, such as color compression, local clipping, translation, etc.If the results of the multipath prediction are different, the original input may be attacked by adversarial interference. Defense against imperceptible adversarial examples is a very challenging study and will be our next research direction.

Ethics approval

This article does not contain any studies with human participants or animals performed by any of the author.

Funding details

The study was supported by "Teaching Quality and Teaching Reform Project of Guangdong Universities,China (Grant No.191171-DXSSJJXJD-32)"and"Research Project of Hanshan Normal University,China (Grant No. XS201908, XN202034)"and"Philosophy and Social Science '13th Five-Year plan' Project of Chaozhou,China(Grant No.2019-A-05, 2020-C-17)".

Conflict of interest

The author declares no conflict of interest.

Contributions

Yinghui Zhu is responsible for the collection of experimental data, and Yuzhen Jiang is responsible for the writing of the paper.

Informed consent

All authors agree to submit this version and claim that no part of this manuscript has been

published or submitted elsewhere.

We appreciate your consideration of our manuscript, and we look forward to receive comments from the reviewers as soon as possible.

REFERENCES

[1] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks”, in Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition(CVPR), 2018, pp. 4510–4520.

[2] O. Ronneberger, P. Fischer , and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", in Proc. of the International Conference on Medical Image Computing and Computer Assisted Intervention(MICCAI): Springer, 2015, pp. 234-241.

[3] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation”. In Proc. of the European Conference on Computer Vision(ECCV), 2018, pp. 833–851.

[4] A. Arnab, O. Miksik, and P. H. Torr, "On the robustness of semantic segmentation models to adversarial attacks", in Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition(CVPR), 2018, pp. 888-897.

[5] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, "Adversarial examples for semantic segmentation and object detection", IEEE International Conference on Computer Vision(ICCV), 2017, pp. 1369-1378.

[6] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks”, The International Conference on Learning Representations(ICLR), 2014, pp.1-10.

[7] G. Shen, C. Mao, J. Yang, B. Ray, “AdvSPADE: Realistic Unrestricted Attacks for Semantic”, in Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition(CVPR), 2018.

[8] Y. Deng, X. Zheng, T. Zhang, C.Chen, G. Lou, and M. Kim, “An Analysis of Adversarial Attacks and Defenses on Autonomous Driving Models”, in The International Conference on Pervasive Computing and Communications(PerCom), arXiv:2002.02175,2020

[9] X. Ma , Y. Niu , L. Gu , Y. Wang, Y. Zhao, J. Bailey, and F. Lu, “Understanding Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems”, Pattern Recognition, 2020:107332.

[10] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples”, in International Conference on Learning Representations(ICML), 2015, pp. 1-10.

[11] L. Chen, W. Xu, “Attacking Optical Character Recognition(OCR) Systems with Adversarial Watermarks”, In 2020 Artificial Intelligence and Cognitive Science(AICS), arXiv:2002.03095, 2020

[12] A. Kurakin, I.J. Goodfellow, and S. Bengio, “Adversarial machine learning at scale”, The International Conference on Learning Representations(ICLR 2017), arXiv:1611.01236, 2016

[13] N. Carlini and D. Wagner,. “Towards Evaluating the Robustness of Neural Networks”, In 2017 IEEE Symposium on Security and Privacy, 2017, pp. 39–57.

[14] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A.Swami, “The Limitations of Deep Learning in Adversarial Settings”, In Proc.of 2016IEEE European Symposium on Security and Privacy (EuroSP), 2016, pp. 372–387.

[15] J. Su, D. V. Vargas and S. Kouichi, “One pixel attack for fooling deep neural networks”, IEEE Transactions on Evolutionary Computation, 2017, 23(5): 828-841.

[16] M. Naseer, S. H. Khan, S. Rahman, et al. Task-generalizable Adversarial Attack based on Perceptual Metric[J]. Computer Vision and Pattern Recognition (CVPR),2018, arXiv:1811.09020

[17] O. Poursaeed, I. Katsman, B. Gao, and S. Belongie, "Generative Adversarial Perturbations", 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4422-4431.

[18] A. Boloor, K. Garimella, X. He, C. Gill, Y. Vorobeychik, and X. Zhang, "Attacking vision-based perception in end-to-end autonomous driving models", 2020 Journal of Systems Architecture 110:101766

[19] J. H. Metzen, M. C. Kumar, T. Brox and V. Fischer, "Universal Adversarial Perturbations Against Semantic Image Segmentation," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 2774-2783.

[20] A. Kurakin, I.J. Goodfellow. and S.Bengio, “Adversarial examples in the physical world”, The International Conference on Learning Representations(ICLR), arXiv:1607.02533,2016

[21] R. Timofte, K. Zimmermann, and L. V. Gool, “Multi-view traffic sign detection, recognition and 3D localization”, IEEE Workshop on Applications of Computer Vision (Vol.25), 2014, pp. 633-647.

[22] M. Everingham, , Van Gool, L., Williams, C. K. I., Winn, J. and A. Zisserman, ”The PASCAL Visual Object Classes (VOC) Challenge”, International Journal of Computer Vision, 88(2), 2010, pp. 303-338.

[23] A. Geiger, P. Lenz, R. Urtasun, “Are we ready for autonomous driving? the KITTI vision benchmark suite”, In Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 3354–3361.