Imperceptible adversarial attacks against traffic scene recognition

Yinghui Zhu · Yuzhen Jiang

Accepted: 13 August 2021 / Published online: 30 August 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
Adversarial examples have begun to receive widespread attention owing 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 detection algorithms can notice them, and this imperceptibility makes them more covert and dangerous. 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 a 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 C&W optimization and achieve 100% attack rate with very low distortion, among which, our experimental results indicate that the 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 1.503 and 1.574. We believe that our research on imperceptible adversarial attacks has a certain reference value 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 (Szegedy et al. 2014; Goodfellow et al. 2015). With some special perturbation on raw images, the adversarial examples can easily fool the models and make wrong predictions. Adversarial examples can attack not only image classifiers (Szegedy et al. 2014; Kurakin et al. 2017; Carlini and Wagner 2017) but also semantic segmentation and object detection models (Arnab et al. 2018; Alassad et al. 2021; Merrer et al. 2020; Ma et al. 2020; Taheri et al. 2020). However, most existing attack algorithms are with visible distortion or too much additional noise (Shen et al. 2018; Naseer et al. 2018; Boloor et al. 2020; Failed 2017), 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 detection algorithms but also human eyes. In this paper, we focus more on the imperceptible adversarial examples, study imperceptible adversarial attacks against different traffic scene recognition, including traffic sign classification, lane detection and street scene segmentation. Figure 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 detection algorithm. Our attacks are all white-box attacks (Ma et al. 2020; Goodfellow et al. 2015), the three networks involved are the famous MobileNetV2 (Sandler et al. 2018), U-net (Ronneberger et al. 2015) and DeepLabV3+ (Chen et al. 2018), which have been trained on well-known datasets: BelgiumTS (Timofte et al. 2014), Pascal VOC (Everingham et al. 2010) and Kitti (Geiger et al. 2012), respectively.

The contributions of this paper include:

- We study the adversarial attacks of automatic traffic identification system and reveal the potential security

risks of DNNS system from different application perspectives.

Aiming at the semantic image segmentation systems which are more difficult to attack, we propose a universal logits map-based attack architecture and on the basis of which we develop two effective and purposeful attack algorithms against traffic scene recognition.

We have proved it by experiments that, combined with well-designed targeted Logits map, the highly destructive adversarial samples of image semantic segmentation can also be obtained by applying C&W optimization strategies.

2 Adversarial attacks

The concept of adversarial example is first mentioned in Szegedy et al. (2014). Szegedy et al. revealed that minor perturbations in an image can easily confuse the classifier, yet people can hardly notice them. The existence of adversarial examples is 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) (Goodfellow et al. 2015) is regarded as a benchmark method owning to its simple and efficient operation. And later, Kurakin et al. upgraded FGSM to an iterative extension version, the basic iterative method (BIM) (Kurakin et al. 2017). They also proposed a targeted attack method of BIM: iterative least-likely class method (ILCM) (Kurakin et al. 2017, 2016). 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 (Carlini and Wagner 2017), the approach is regarded as more robust while with less distortion. Noting that, at that stage, almost all the approaches were designed for attacking the image classifiers at the beginning stage.

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 could not be directly introduced into semantic segmentation models. It is commonly recognized that it was much harder to generate the adversarial examples on semantic segmentation models than on classifiers (Arnab et al. 2018; Alassad et al. 2021; Osahor and Nasrabadi 2019; Klingner et al. 2020), 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 investigate 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 are of good quality, 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) (Timofte et al. 2014), which contains 62 types of traffic signs and has 7125 images in total. After 50 epochs, the model achieved 0.974 training accuracy and 0.887 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, the C&W strategy is adopted in our program which applies gradient descent optimization to solve the following problem:

$$\min_{\|I' - I\|_2 + a \cdot f(I')}$$

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

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. The parameter $a$ is the super-parameter that is used to adjust the proportion between $L_2$-norm restraint and attack strength. The parameter $\kappa$ is used to ensure the high confidence of the misled class. For searching the best combination of coefficients, let $x$ start with 0 and add 5 each time to search the best combination of coefficients, and $x = 100, \kappa = 20$ is the final coefficient values which can gain the superior attack performance and ensure the 100% attacking rate (Carlini and Wagner 2017). In addition, to reduce the iterative $[0, 1]$ clips in optimization, we also apply a variable $w$ to substitute $I'$:

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

Figure 2 shows the diagram of Formula (3), it indicates that the upper function can well protect the white or bright color areas from severe modification, thus keeping 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. Figure 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 evaluating the perturbations of each algorithm more comprehensively, we randomly select 100 images from the validation set of BelgiumTs to generate adversarial examples by upper methods. The images are considered in $[0, 255]$ range, their target class is set by $C_{\text{target}} = \text{mod}(C_{\text{correct}} + 30, 62)$, where $C_{\text{correct}}$ is the correct class of original image. Three criteria, mean squared error(MSE), mean absolute error(MAE) and $L$-infinity norm ($L_{\infty}$), are introduced to detect the perturbation strength. MSE and MAE represent two different distances between the original image and the adversarial image, while $L_{\infty}$ indicates the maximum pixel modification. The smaller the values of these three criteria, the smaller the amount of disturbance, the better the algorithm effect. 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($\varepsilon$, for ILCM) be 0.01, the adversarial sample is iteratively optimized until being identified as the target class with a confidence of not less than 0.9. The least threshold means the least perturbations to mislead model and we search them by doubling the $lr$ (or the $\varepsilon$) from 0.0001. Table 1 shows the results. On the Lowest Threshold, surprisingly, the MAEs of all methods are less than 1. Compared to the [0 255] range of image value, 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 cannot get the minimum $L_{\infty}$, the adoption of Formula (3) can effectively avoid creating obvious noise in bright areas (i.e., eye sensitive areas) and protect its image quality.

Inspired by the superiority of C&W, we try to develop our semantic segmentation attacks by using the same optimization strategy. 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, a 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. Figure 4 shows four attack patterns, the first two rows show two non-targeted methods which output chaotic and striped maps (Poursaeed et al. 2018). The third row shows a category-hidden attack which outputs a "normal" map with little pedestrian information (Failed 2017). The fourth row shows an explicit attack which can indicate wrong driving direction by painting black lines on road image (Boloor et al. 2020). By contrast, the third
method has less perturbation noise and the prediction map looks more normal, so it is the most difficult to detect.

4.1 Universal attack architecture

To support the adversarial attacks on different models, we propose a 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

![Fig. 2 The Tanh function in C&W](image)

![Fig. 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](image)

| 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 |

Bold values indicate the best results. The criteria are evaluated from two perspectives: the 0.9 Confidence Satisfied and the Lowest Threshold.
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). The adversarial example generation frame is in the top dotted rounded 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.

The key step of this attack architecture is to introduce the targeted logits map. According to the purpose of the attack, the attacker may tamper the normal prediction map into the desired target effect, and through iterative

Fig. 4 Various adversarial attack patterns

Fig. 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
optimization, the adversarial samples that can predict the desired output map will be obtained. In the following sections, we will propose two attack methods against traffic scene segmentation according different tasks. The C&W optimization strategy is adopted in both methods to generate adversarial examples, and Tanh function described in Formula (3) is also used 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 being trained on the KITTI dataset (a famous road scenes dataset) (Geiger et al. 2012), the model achieves 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:

\[
x' = \text{mod}(x - \tan(\theta) \cdot y, W) \tag{5}
\]

The \(\theta\) in 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 C&W optimization problem:

\[
\text{minimize} \| I' - I \|_2 + a \cdot \| Z' - Z \|_2 \tag{6}
\]

where \(a = 100\), which is also obtained from pretest and can implement most attack within a configured iterations number.

Figure 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 \(\pm 45\) degrees yet the adversarial examples seem the same as the raw image. Figure 7 shows another instance with \(\pm 60\) degrees deflection attacks.

Hundred random images are taken from the KITTI validation dataset to be tested the same \(+45\) degree attack. And Table 2 shows the mean MSE, MAE and L_\(\infty\) which 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.

4.3 Street scene recognition attack

In this section, we demonstrate another semantic segmentation attack, people-hidden attack. We will fool the model to see none of peoples. The model adopted in this task is the U-net networks (Ronneberger et al. 2015), and the training and testing datasets are created from the Pascal VOC2012 datasets (Everingham et al. 2010). The model was trained to identify five categories: people, cars, motorcycles, bicycles and background. The category of background represents all other objects. The trained model achieves 0.967 training accuracy and 0.863 validation accuracy. The attack purpose of this task is to hide all the people in the scenes without affecting the identifications of other categories. So the targeted logits maps, e.g., \(Z'\), must be crafted in advance.

To hide the “people,” we can set all \(C_{\text{people}}\) in \(Z'\) be 0, but that would make it hard for algorithm to convergence and generate an adversarial example. Here, \(C_{\text{people}}\) 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:

\[
Z'_{x,y,C_{\text{others}}} = Z_{x,y,C_{\text{others}}} \tag{7}
\]

where \(C_{\text{others}}\) represents all categories except people.

\[
Z'_{x,y,C_{\text{max}}} = Z'_{x,y,C_{\text{max}}} + Z_{x,y,C_{\text{people}}} \tag{8}
\]

where \(C_{\text{max}}\) is the selected channel to be added the “people” data, it is calculated from following:

\[
C_{\text{max}} = \text{argmax}Z'_{x,y,C_{\text{others}}} \tag{9}
\]

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 that none of pixels in final prediction map belongs to “people” category. Figure 8 shows two adversarial instances. Their final output maps demonstrate that all the segmentation areas of “people” completely disappear, and they have been replaced by other category pixels. Hundred images with people, which are selected from validation datasets, were
tested with two intensities: ordinary attack ($lr = 0.001$) and low-distortion attack ($lr = 0.0001$). Table 3 shows the mean MSE, MAE and $L_\infty$. 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. All the experiments above are run on a workstation with NVIDIA RTX3000.

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. In our scheme, we can design arbitrary attack effects by using the proposed universal adversarial architecture of semantic segmentation. Our experimental results reveal DNNs’ vulnerability from the following two 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 detection algorithm, and such attacks are more covert and dangerous.

The ultimate goal of the research against attacks is to improve the robustness and security of the DNNs application system. Defense against imperceptible adversarial examples is a very challenging study in our future work, and we will focus the two directions: (1) Researching on the integration of signal acquisition and recognition, because reducing the transmission and preprocessing links can reduce the risk of attack. (2) Enhancing the DNNs defense performance by expanding the defense training, so as to improve the robustness and security of these application systems.

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

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

**Declarations**

**Conflict of interest** The author declares no conflict of interest.

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

**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**

Alassad M, Spann B, Agarwal N (2021) Combining advanced computational social science and graph theoretic techniques to reveal adversarial information operations. Inf Process Manag 58(1):102385

Arnab A, Miksik O, Torr PH (2018) On the robustness of semantic segmentation models to adversarial attacks. In: Proceedings of the IEEE international conference on computer vision and pattern recognition (CVPR), pp 888–897

Boloor A, Garamella K, He X, Gill C, Vorobeychik Y, Zhang X (2020) Attacking vision-based perception in end-to-end autonomous driving models. J Syst Archit 110:101766

Carlini N, Wagner D (2017) Towards evaluating the robustness of neural networks. In: 2017 IEEE symposium on security and privacy, pp 39–57

Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp 833–851

Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2010) The PASCAL Visual Object Classes (VOC) challenge. Int J Comput Vis 88(2): 303–338

Geiger A, Lenz P, Urtasun R (2012) Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proceedings of
the IEEE international conference on computer vision and pattern recognition (CVPR), pp 3354–3361
Goodfellow IJ, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples. In: International conference on learning representations (ICML), pp 1–10
Klingner M, Br A, Fingscheidt T (2020) Improved noise and attack robustness for semantic segmentation by using multi-task training with self-supervised depth estimation. In: The IEEE/CVF conference on computer vision and pattern recognition (CVPR) workshops. IEEE
Kurakin A, Goodfellow IJ, Bengio S (2016) Adversarial examples in the physical world. In: The international conference on learning representations (ICLR), 1607.02533.2016
Kurakin A, Goodfellow IJ, Bengio S (2016) Adversarial machine learning at scale. In: The international conference on learning representations (ICLR 2017), 1611.01236
Le Merrer E, Pérez P, Trédan G (2020) Adversarial frontier stitching for remote neural network watermarking. Neural Comput Appl 32: 9233–9244
Ma X, Niu Y, Gu L, Wang Y, Zhao Y, Bailey J, Lu F (2020) Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recognit 107332
Metzen JH, Kumar MC, Brox T, Fischer V (2017) Universal adversarial perturbations against semantic image segmentation. In: 2017 IEEE international conference on computer vision (ICCV), Venice, pp 2774–2783
Naseer M, Khan SH, Rahman S, et al (2018) Task-generalizable adversarial attack based on perceptual metric. Comput Vis Pattern Recognit, 1811.09020
Osahor U, Nasrabadi N (2019) Deep adversarial attack on target detection systems. In: Artificial intelligence and machine learning for multi-domain operations applications, International Society for Optics and Photonics, Baltimore, MD, USA
Poursaeed O, Katsman I, Gao B, Belongie S (2018) Generative adversarial perturbations. In: 2018 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 4422–4431
Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In: Proceedings of the international conference on medical image computing and computer assisted intervention (MICCAI). Springer, Berlin, pp 234–241
Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L (2018) MobileNetV2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE international conference on computer vision and pattern recognition (CVPR), pp 4510–4520
Shen G, Mao C, Yang J, Ray B (2018) AdvSPADE: realistic unrestricted attacks for semantic. In: Proceedings of the IEEE international conference on computer vision and pattern recognition (CVPR)
Szegedy C, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2014) Intriguing properties of neural networks. In: The International Conference on Learning Representations (ICLR), pp 1–10
Taheri R, Javidan R, Shojafar M et al (2020) On defending against label flipping attacks on malware detection systems. Neural Comput Appl 32:14781–14800
Timofte R, Zimmermann K, Gool LV (2014) Multi-view traffic sign detection, recognition and 3D localization. In: IEEE workshop on applications of computer vision (vol. 25), pp 633–647

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.