Hindawi
Journal of Sensors
Volume 2022, Article ID 4390413, 13 pages
https://doi.org/10.1155/2022/4390413

Research Article

AdvU-Net: Generating Adversarial Example Based on Medical Image and Targeting U-Net Model

Hyun Kwon[1] **and Jongwook Jeong**[2]

[1]*Department of Artificial Intelligence and Data Science, Korea Military Academy, 574 Hwarang-ro, Nowon-gu, Seoul 01819, Republic of Korea*
[2]*Department of Computer Science, Korea Military Academy, 574 Hwarang-ro, Nowon-gu, Seoul 01819, Republic of Korea*

Correspondence should be addressed to Jongwook Jeong; jwjeong55@gmail.com

Received 25 August 2021; Revised 12 March 2022; Accepted 22 March 2022; Published 8 June 2022

Academic Editor: Carlos Marques

Deep neural networks (DNNs) provide excellent performance in image recognition, speech recognition, video recognition, and pattern analysis. These neural networks are often applied in the medical field to predict or classify patients' illnesses. One such network, the U-Net model, has shown good performance in data segmentation and is an important technology in medical imaging. However, deep neural networks such as those applied in medicine are vulnerable to attack by adversarial examples. Adversarial examples are samples created by adding a small amount of noise to an original data sample that is difficult for a human to see but that induces misclassification by the targeted model. In this paper, we propose AdvU-Net, a method for generating an adversarial example targeting the U-Net model used in segmentation. Performance was analyzed according to epsilon, using the fast gradient sign method (FGSM) for generating adversarial examples. We used ISBI 2012 as the dataset and TensorFlow as the machine learning library. In the experiment, when an adversarial example was generated using an epsilon value of 0.3, the pixel error was 3.54 or greater while the pixel error of the original sample was maintained at 0.15 or less.

1. Introduction

Deep learning algorithms have revolutionized the field of computer image processing. In particular, the deep neural network [1] has begun to attract attention, showing results that exceed the performance of existing machine learning models on classification tasks.

In the medical field, studies in radiology, pathology, and ophthalmology have recently been published in which deep learning technology [2] provided performance equivalent to that of physicians. For example, in research on diabetic retinopathy diagnosis [3] using fundus images, results have shown that Google's machine learning technology delivers the same or better performance than that provided by ophthalmologists.

The purpose of a computer-assisted diagnosis system based on medical image processing is to assist physicians in reading and diagnosing images by detecting lesions or presenting classification results for a given medical image. In such systems, an important role is played by segmentation technology. Deep learning is applied mainly in detection tasks that require cutting out a specific area from a medical image. However, these deep neural networks have weaknesses. For example, if an attacker intentionally modifies an image by adding a small amount of noise that is optimized for medical data, although the physician will not notice any problem, the image could be segmented incorrectly by the image segmentation model. An image modified in such a way is called an adversarial example. Adversarial examples [4–6] are original samples that have been modified by the addition of a small amount of noise that is difficult for humans to see but that induces misclassification by a target model. As mis-segmentation by an image segmentation model being used in a medical application can lead directly to problems for a patient's health, the adversarial example can be a serious threat for deep learning models used in the medical field. Adversarial examples have been studied mainly in the context of image classification models; this study is one of the few that relates to image segmentation models.

In this paper, we propose AdvU-Net, a method for generating adversarial examples targeting a U-Net model [7] used for image segmentation. To evaluate the proposed method, the attack effectiveness and segmentation results for the adversarial example were analyzed for various values of epsilon in the fast gradient sign method (FGSM) [8], which is a method for generating adversarial examples. The present study is the first to analyze adversarial examples based on the ISBI 2012 dataset and generated for a U-Net model used for image segmentation. In addition, unlike previous studies, this study analyzed the adversarial example performance in terms of pixel error according to the value of epsilon, and it was shown that unlike existing methods, the proposed method allows the degree of misegmentation to be controlled. The contributions of this paper are as follows. First, we propose the generation of adversarial examples targeting the U-Net model used in the medical field. We describe the principle of the proposed method and its system for generating adversarial examples. Second, we report the results of the experiment we performed to verify the proposed method using the ISBI 2012 dataset and the U-Net model. Third, we analyze the image distortion as well as the pixel error and adversarial noise for the generated adversarial examples systematically according to the value of epsilon. This study is the first to use the ISBI 2012 dataset for this analysis. In addition, we mention possible applications in which the proposed method might be used as a method of attack.

The rest of this paper is organized as follows. Section 2 reviews relevant concepts. In Section 3, the generation of adversarial examples is explained. Section 4 provides information on the experimental environment and reports the experimental results. Discussion is given in Section 5, and in Section 6, the conclusions are presented.

2. Relevant Concepts

This section provides descriptions of the target model U-Net and of adversarial examples.

2.1. U-Net Model. U-Net [7] is an end-to-end fully convolutional network [9] used for image segmentation in the biomedical field. U-Net consists of a symmetrical pair of networks, one for obtaining the overall information content of an image and one for performing localization. In particular, U-Net trains efficiently through data augmentation and displays state-of-the-art performance with a variety of medical datasets. The structure of the U-Net model consists of a contracting path and an expansive path. The contracting path serves to capture the context of the image, and in the expansive path, the feature map is upsampled, and the context of the feature map captured in the contracting path is combined with the feature map to perform accurate localization. The U-Net model has a large number of feature channels during the upsampling process; this means that context can be propagated to successive layers with their corresponding resolutions. In this model, only the valid part of each convolution is used, where a "valid part" is a segmentation map that contains full context. This enables smooth segmentation by the U-Net model, using the overlap-tile technique.

2.2. Adversarial Examples. Adversarial examples were first proposed in a study by Szegedy et al. [4]. Their study showed that the deep learning model has a weakness with regard to the image classification problem in that the addition of a small amount of noise that cannot be perceived by the human eye can cause an image to be classified incorrectly by the model. In an attack using an adversarial example, the prediction result given by the deep learning classification model for the image can be changed, and false predictions are reported with high reliability.

Such attack methods [10, 11] can be classified according to the information known about the target model, the specificity of the intended misclassification of the adversarial example, and the distortion metric used. First, the information known about the target model determines whether the attack is a white-box attack [12–16] or a black-box attack [17–21]. In a white-box attack, all information about the target model is known by the attacker, including the structure of the model, its parameters, and the output probability values corresponding to the possible result values. In a black-box attack, the attacker does not have information about the target model.

Second, the specificity for the intended misclassification of the adversarial example determines whether the attack is targeted or untargeted. In a targeted attack [5, 22, 23], the adversarial example is designed to be misclassified by the model as a specific target class chosen by the attacker. In an untargeted attack [24–26], the adversarial example is designed to be misclassified as any class other than the original class (i.e., any invalid class). An untargeted-attack method has the advantage of generating adversarial examples with fewer iterations and less distortion than a targeted-attack method.

The distortion metric [5, 27–30] is a third way of characterizing adversarial example generation methods. The possibilities include L_1, L_2, and L_∞, defined as follows:

$$L_p = \left(\sum_{i=1}^{n} |x - x^*|^p \right)^{1/p}, \tag{1}$$

where x is the original sample, and x^* is the adversarial example. With each of these, the smaller the distortion value, the less distortion there is between the original sample x and the adversarial example x^*.

Existing adversarial example generation methods [31–35] have been studied in the context of image recognition; research on image segmentation models has been lacking. In this paper, we propose a method for generating an adversarial example targeting the U-Net model, which is an image segmentation model. We analyze the image segmentation, pixel error, and adversarial noise for the adversarial example according to the value of epsilon.

3. Methodology

Figure 1 shows the architecture for the proposed method, AdvU-Net. FGSM was used as an adversarial example generation method. The transformer modulates the original

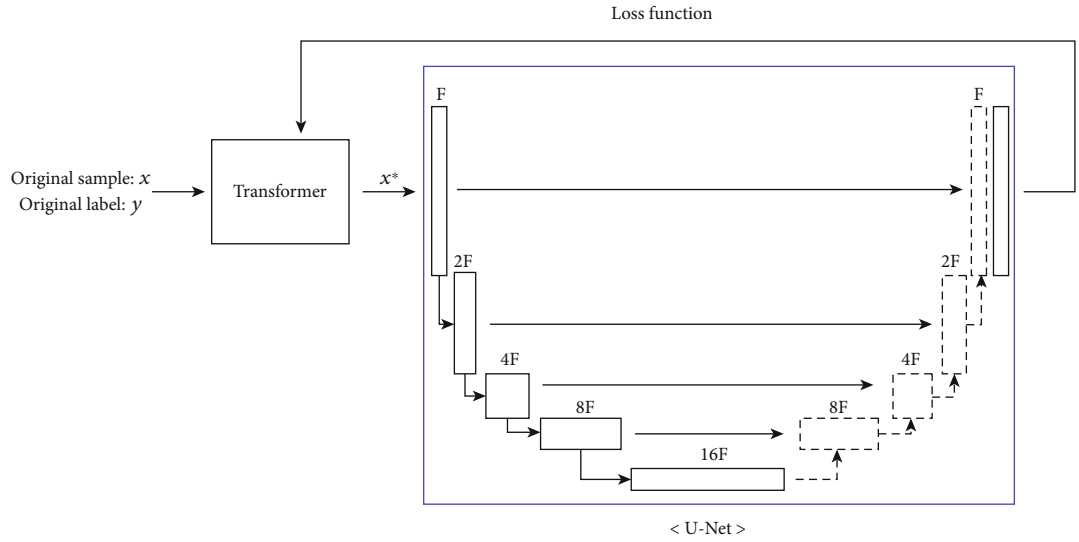


Figure 1: Overview of the methodology for the proposed method, AdvU-Net.

sample and provides it to the target model. The proposed method adds some noise to the original sample through the transformer and using feedback from the model. To the human eye, the noise is small, but it causes the image to be mis-segmented by the model.

In mathematical terms, the fast gradient sign method (FGSM) [8] finds x^* through L_∞:

$$x^* = x + \varepsilon \cdot \text{sign}\left(\nabla \text{loss}_{F,t}(x)\right), \tag{2}$$

where F is an objective function, t is the target class, and x^* is an adversarial example. In FGSM, the adversarial example is generated according to the value of ε from the input image x through gradient ascent. The gradient ascent method is simple but has excellent performance.

4. Experimental Setup and Evaluation

This section describes the experimental environment and presents the results for the adversarial examples generated for U-Net. We used the TensorFlow [36] machine learning library and a Xeon E5-2609 1.7-GHz server.

4.1. Dataset. The dataset used in the experiment was the ISBI 2012 dataset [37], which is used in segmentation of medical images. It consists of 30 data items and their corresponding labels, from a serial section transmission electron microscopy (ssTEM) dataset of the Drosophila first instar larva ventral nerve cord (VNC). The microcube measures approximately $2 \times 2 \times 1.5$ microns with a resolution of $4 \times 4 \times 50$ nm/pixel. The label is binary and is represented by a black and white picture, with the segmented objects in white and the rest in black. Testing was conducted by the k-fold cross-validation method using six of the data items. Although the number of items in the ISBI 2012 dataset is small, the data augmentation by the U-Net model provides very high performance.

4.2. Target Model. The target model was a U-Net model used for data segmentation. Its structure is shown in Table 1. The Adam algorithm [38] was used as the model's optimization algorithm, and ReLU [39] was used as the activation function. The parameter values are shown in Table 2.

4.3. Generation of the Adversarial Examples. FGSM was used as the adversarial example generation method. Adversarial examples were produced for a range of epsilon values from 0.1 to 0.9, 30 adversarial examples for each epsilon value. This made it possible to analyze the pixel error given by the target model and the adversarial noise for the adversarial examples generated for each epsilon value.

4.4. Experimental Results. In this section, the adversarial example images, adversarial noise, the output result according to epsilon, and the pixel error between the original label and the output result are analyzed for the adversarial examples generated by FGSM.

4.4.1. Visual Comparison between Original Sample Image and Adversarial Example Image for Three Example Cases. Table 3 shows images of original samples, adversarial noise, and adversarial examples. The epsilon value was set to 0.3 to generate adversarial noise for each data sample. It can be seen in the table that in terms of human perception, there is little difference between the original sample and its corresponding adversarial example. Although it is difficult to perceive the difference between the original sample and the adversarial example, the segmentation of each adversarial example by the target model is incorrect.

4.4.2. Comparison of Image Segmentation Performance by Target Model on Original Sample and Adversarial Example. Table 4 compares the original label with the output results for the corresponding original sample and adversarial example. The adversarial examples were generated using an epsilon value of 0.3. As can be seen in the table, the output results for the original samples are similar to the

Table 1: Architecture of the target model.

	Level	Conv layer	Filter	Stride	Output size
Input					244, 244, 3
Contracting path	Level 1	Conv 1	3, 3, 64	1	224, 244, 64
		Conv 2	3, 3, 64	1	224, 244, 64
	Level 2	Conv 3	3, 3, 128	2	112, 112, 128
		Conv 4	3, 3, 128	1	112, 112, 128
	Level 3	Conv 5	3, 3, 256	2	56, 56, 256
		Conv 6	3, 3, 256	1	56, 56, 256
	Level 4	Conv 7	3, 3, 512	2	28, 28, 512
		Conv 8	3, 3, 512	1	28, 28, 512
Bridge	Level 5	Conv 9	3, 3, 1024	2	14, 14, 1024
		Conv 10	3, 3, 1024	1	14, 14, 1024
Expansive path	Level 6	Conv 11	3, 3, 512	1	28, 28, 512
		Conv 12	3, 3, 512	1	28, 28, 512
	Level 7	Conv 13	3, 3, 256	1	56, 56, 256
		Conv 14	3, 3, 256	1	56, 56, 256
	Level 8	Conv 15	3, 3, 128	1	112, 112, 128
		Conv 16	3, 3, 128	1	112, 112, 128
	Level 9	Conv 17	3, 3, 64	1	224, 244, 64
		Conv 18	3, 3, 64	1	224, 244, 64
Output		Conv 19	1, 1	1	224, 244, 1

Table 2: Parameters of the target model.

Parameter	Values
Learning rate	0.001
Momentum	0.9
Dropout	0.5
Batch size	4
Epochs	100

corresponding original labels, and thus, the samples appear to be properly segmented. On the other hand, it can be seen by a comparison with the original labels that the adversarial examples are not properly segmented. Thus, the adversarial examples (with adversarial noise) have been mis-segmented by the model.

4.4.3. Analysis of Image Segmentation Performance by Target Model on Adversarial Examples according to Epsilon Value. Table 5 shows images and output results for adversarial examples according to epsilon value. It can be seen in the table that as the epsilon value increases, the adversarial examples become less well segmented by the model. On the other hand, it can also be seen that as the epsilon value increases, the distortion in the adversarial example image increases. Therefore, when generating adversarial examples, it is necessary to select an appropriate epsilon value to find a point at which segmentation performance is degraded without the image distortion being discernible to human perception. It can be seen that when the epsilon value used to generate the adversarial examples is 0.3, the resulting noise is difficult for humans to discern, and the adversarial examples are not well segmented by the model.

4.4.4. Analysis of Image Segmentation Performance by Target Model on Adversarial Examples according to Epsilon Value in terms of Pixel Error. Figure 2 shows the pixel error between the original label and the output results for the adversarial examples according to epsilon value. The pixel error is the difference between the pixels of the original label and the segmented output, applying the L_2 metric. From the pixel error chart, it can be seen that the output results for the adversarial examples have a greater pixel error than the original label. In addition, it can be seen that as epsilon increases, the pixel error increases. It can be seen that the favorable point at which the pixel error for the segmentation by the model is substantially increased while human perception maintained is at an epsilon value of 0.3. Based on these results, it is found that the adversarial examples degrade the segmentation performance of the model.

5. Discussion

5.1. Assumption. This method assumes a white-box attack, with information on the target classifier known by the attacker. This is because FGSM needs to have access to the feedback from the target model. In a white-box attack, all information is known about the parameters of the model, the structure of the model, the result for a given input value, and the associated probability values. Thus, the attacker

Table 3: Original sample image, adversarial noise, and adversarial example image for three example cases.

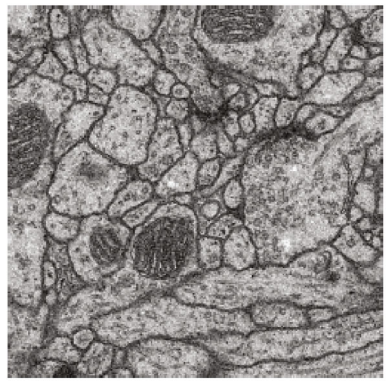
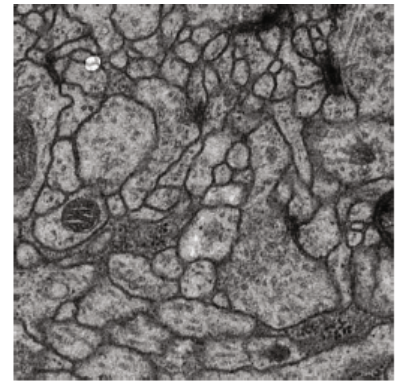
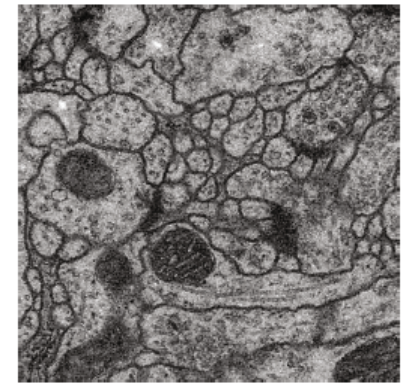
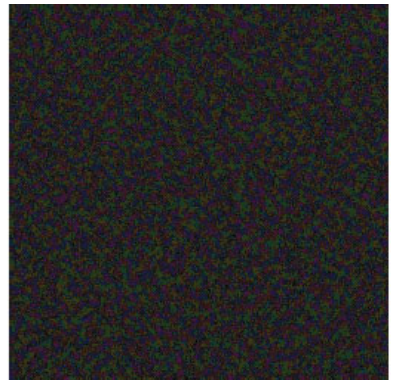
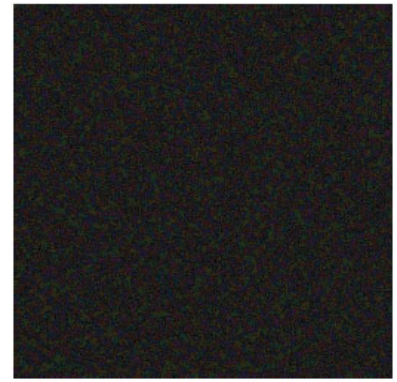
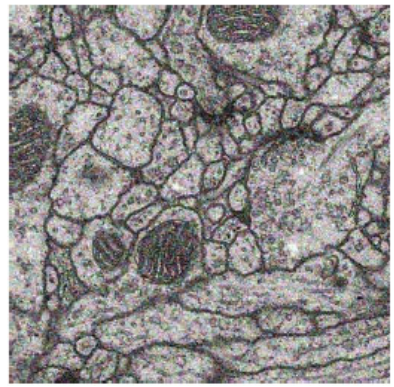
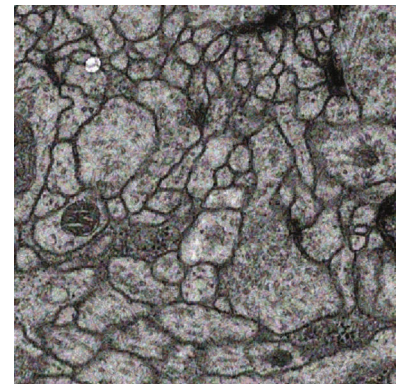
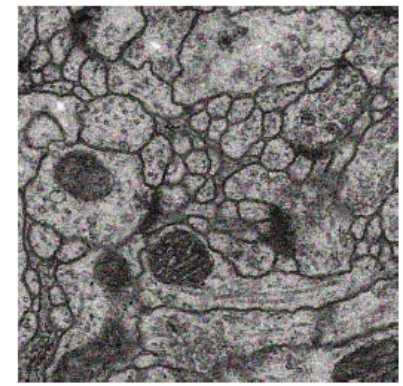
Description	Case1	Case2	Case3
Original sample			
Adversarial noise			
Adversarial example			

Table 4: Comparison between the original label and the output results for original samples and adversarial examples.

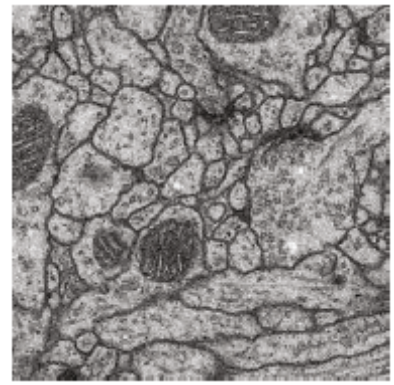
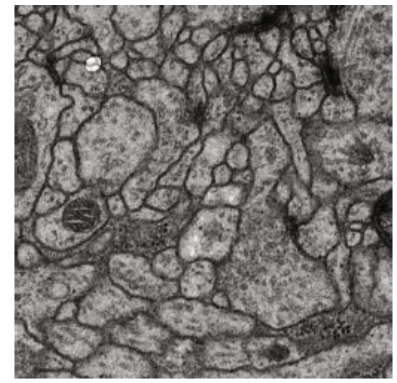
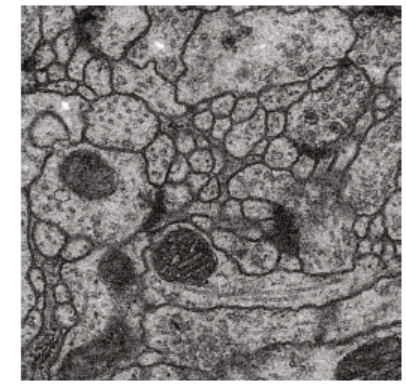
Description	Case1	Case2	Case3
Original labels			
Original sample			
Output of original sample			

Table 4: Continued.

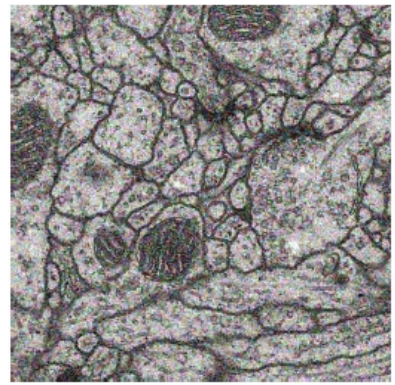
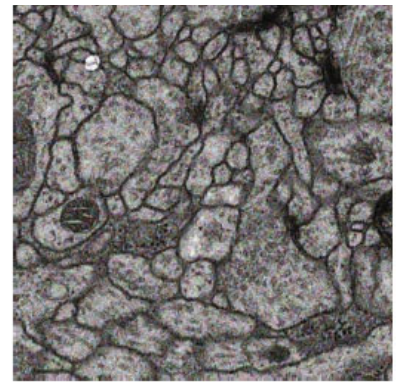
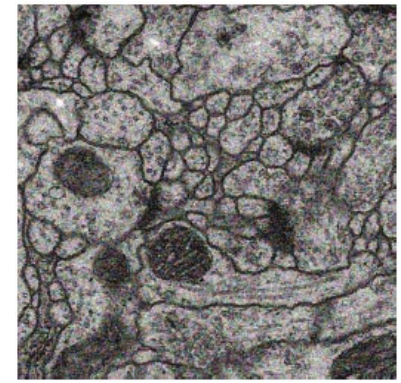
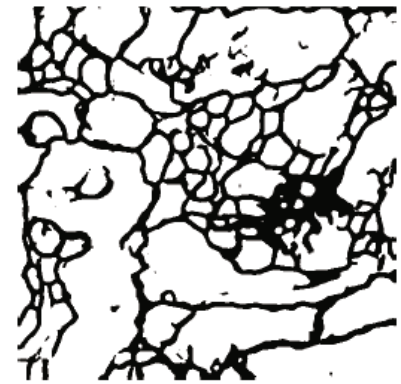
Description	Case1	Case2	Case3
Adversarial examples			
Output of adversarial example			

Table 5: Image and output result for adversarial examples according to epsilon value.

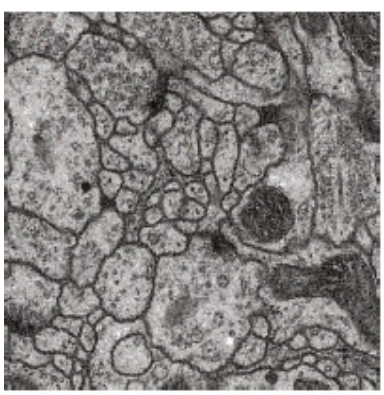
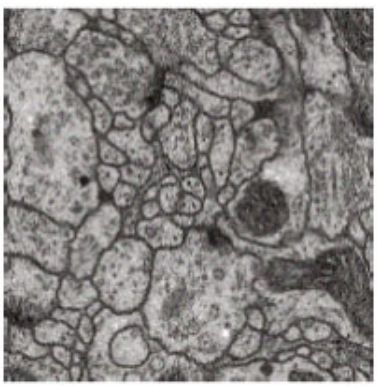
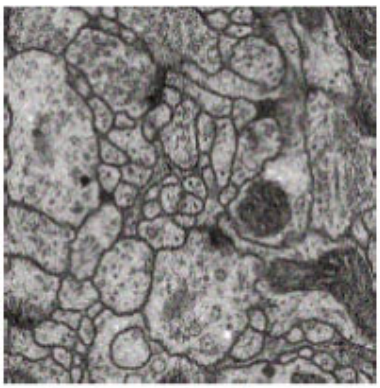
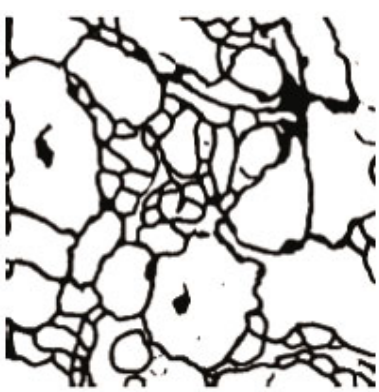
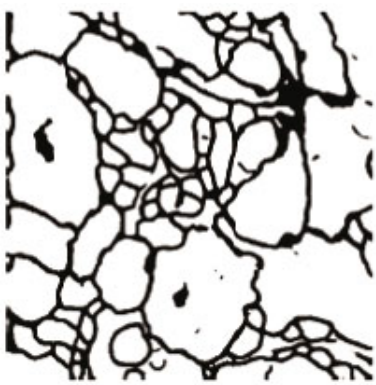
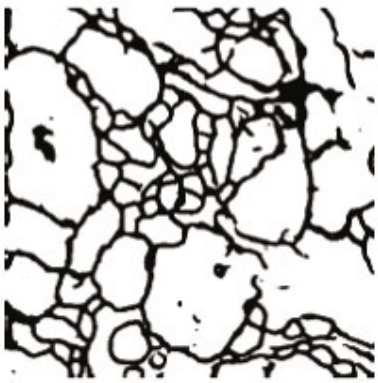
0.1	0.2	0.3
		
		

Table 5: Continued.

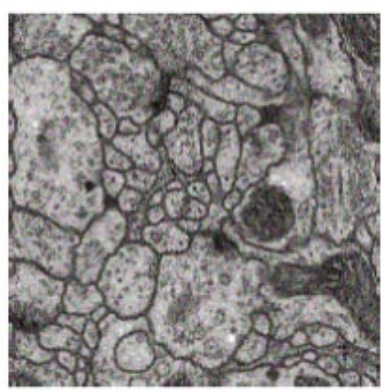
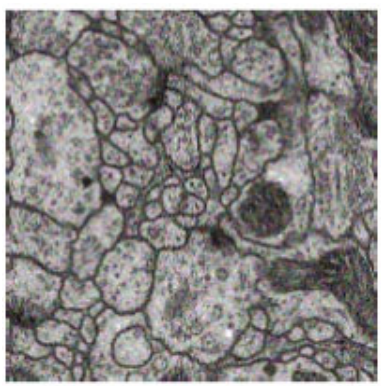
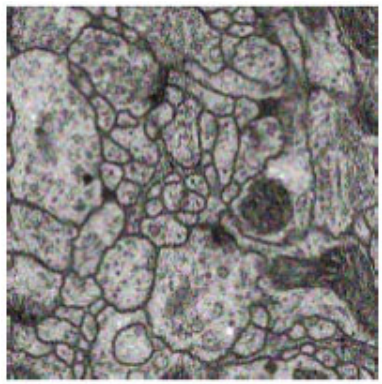
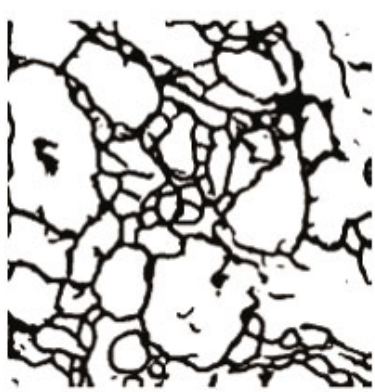
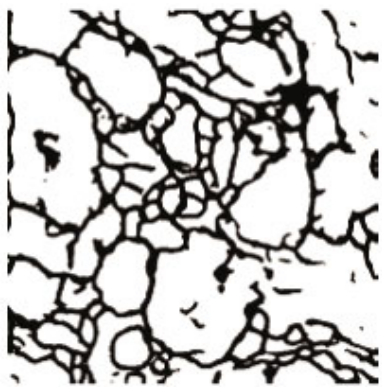
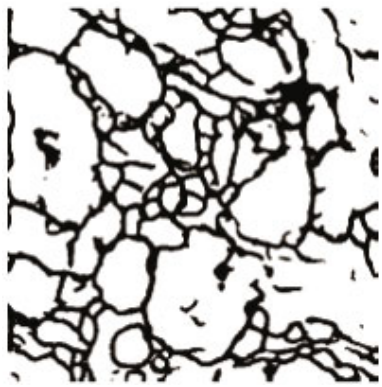
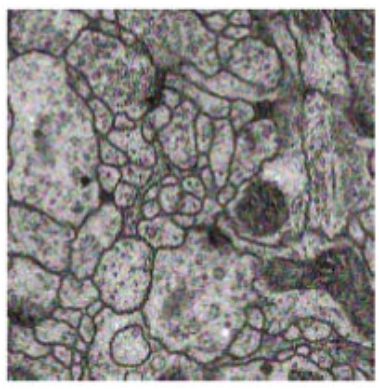
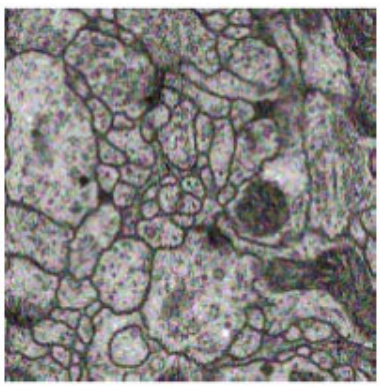
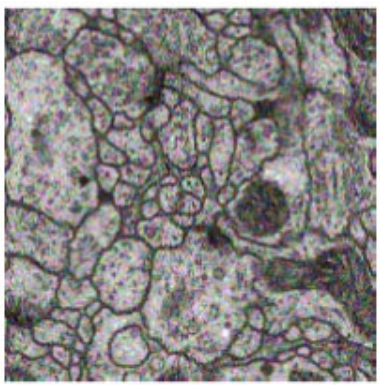
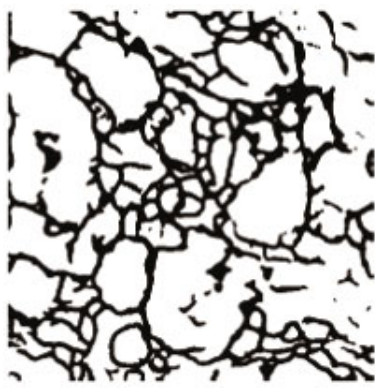
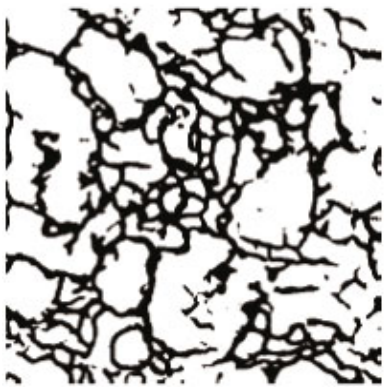
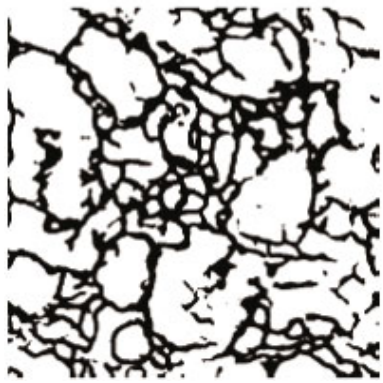
0.4	0.5	0.6
		
		

Table 5: Continued.

0.7	0.8	0.9
		
		

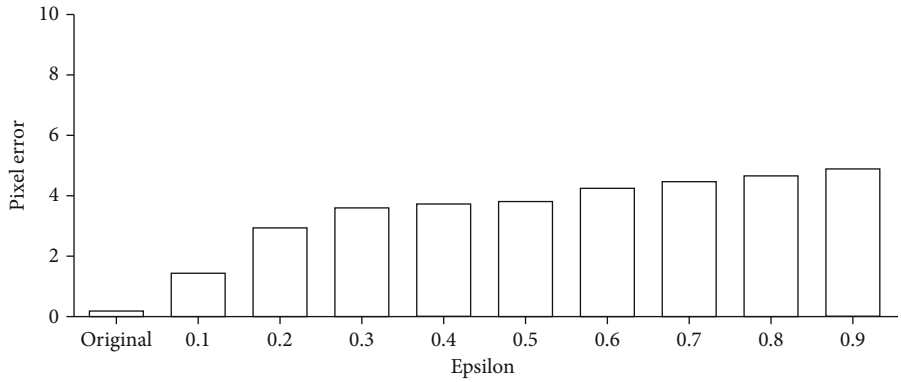


Figure 2: Pixel error between the original label and the output result for the adversarial example according to epsilon value.

creates an adversarial example only in a scenario in which all information about the target model is known.

5.2. Epsilon. In FGSM, epsilon is a parameter that controls the amount of adversarial noise. In the generation of adversarial examples, as epsilon increases, the pixel error of the segmentation result by the target model increases. On the other hand, the amount of adversarial noise added in the adversarial example also increases as epsilon increases. Therefore, it is important to select a value for epsilon that produces a sufficiently high pixel error of segmentation by the model but does not allow the adversarial noise to be identifiable by the human eye. It can be seen from the results of this study that an epsilon value of approximately 0.3 provides a favorable trade-off.

5.3. Pixel Error. The performance on the original samples and the results for the adversarial examples were analyzed in terms of the pixel error. From the results of this analysis, it can be seen that the pixel error for the original sample is very low, at 0.15. For the adversarial examples, on the other hand, it can be seen that the pixel error becomes greater than 3.54 when epsilon is greater than 0.3. Thus, it was demonstrated that the model does not provide a good segmentation result for adversarial examples.

5.4. Applications. The proposed adversarial example generation method can be used as a method of attack in medical fields where adversarial examples can cause misclassification. The experimental analyses in this study were performed for adversarial examples created in the context of segmentation, an important technology for MRI and tumor identification in the medical field. If a segmentation is improperly performed as a result of the application of adversarial examples in such medical projects, it can pose a serious threat to the patient receiving medical care. The risk is also present in laser-assisted in situ keratomileusis (LASIK) surgery and other laser treatments relying on image segmentation. In military contexts, the method could be applied in assassination operations to threaten a specific person.

5.5. Limitations. In order to generate adversarial examples by FGSM, feedback from the target model is required. In the case of a black-box attack, in which feedback is not received from the target model, the generation of adversarial examples by FGSM is limited. When generating adversarial examples, test data can be obtained, and additional time is required to add real-time adversarial noise.

6. Conclusion

In this paper, we have proposed AdvU-Net, a method for generating an adversarial example that targets the U-Net model used in segmentation. In this study, adversarial examples generated using the method were analyzed with respect to pixel error, image distortion, epsilon value, and adversarial noise. Using the proposed method, it was possible to generate an adversarial example for the image segmentation model used in the medical field, and the vulnerability of the U-Net model to adversarial examples was confirmed.

In future research, the investigation could be expanded to include experimentation with other datasets. In addition, the adversarial example could be applied to medical data using a generative adversarial net [40]. Finally, it would be useful to develop possible defenses against the proposed method for use in the medical field.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request after acceptance.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the AI R&D Center of Korea Military Academy, the Hwarang-Dae Research Institute of Korea Military Academy, and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1I1A1A01040308).

References

[1] J. Schmidhuber, "Deep learning in neural networks: an overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[2] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, no. 1, pp. 221–248, 2017.

[3] V. Gulshan, L. Peng, M. Coram et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.

[4] C. Szegedy, W. Zaremba, I. Sutskever et al., "Intriguing properties of neural networks," in *In International Conference on Learning Representations*, 2014.

[5] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *In Security and Privacy (SP), 2017 IEEE Symposium on*, pp. 39–57, San Jose, CA, USA, 2017.

[6] H. Kwon, "Medicalguard: U-net model robust against adversarially perturbed images," *Security and Communication Networks*, vol. 2021, 8 pages, 2021.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *In International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, Cham, 2015.

[8] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *In International Conference on Learning Representations*, 2015.

[9] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," 2016, http://arxiv.org/abs/1610.09585.

[10] M. Barreno, B. Nelson, A. D. Joseph, and J. Tygar, "The security of machine learning," *Machine Learning*, vol. 81, no. 2, pp. 121–148, 2010.

[11] H. Kwon and J. Lee, "AdvGuard: fortifying deep neural networks against optimized adversarial example attack," *IEEE Access*, 2020.

[12] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European Symposium on In Security and Privacy (EuroS&P)*, pp. 372–387, Saarbruecken, Germany, 2016.

[13] P. McDaniel, N. Papernot, and Z. B. Celik, "Machine learning in adversarial settings," *IEEE Security & Privacy*, vol. 14, no. 3, pp. 68–72, 2016.

[14] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2574–2582, 2016.

[15] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou, "Hotflip: white-box adversarial examples for text classification," 2017, http://arxiv.org/abs/1712.06751.

[16] Q. Huang, I. Katsman, H. He, Z. Gu, S. Belongie, and S.-N. Lim, "Enhancing adversarial example transferability with an intermediate level attack," in *In Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4733–4742, 2019.

[17] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," 2016, http://arxiv.org/abs/1611.02770.

[18] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pp. 506–519, New York, 2017.

[19] S. M. Moosavi Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *In Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[20] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial Intelligence Safety and Security*, ICLR Workshop, 2018.

[21] S. Cheng, Y. Dong, T. Pang, H. Su, and J. Zhu, "Improving black-box adversarial attacks with a transfer-based prior," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[22] D. Wang, L. Dong, R. Wang, D. Yan, and J. Wang, "Targeted speech adversarial example generation with generative adversarial network," *IEEE Access*, vol. 8, pp. 124503–124513, 2020.

[23] H. Kwon, Y. Kim, K.-W. Park, H. Yoon, and D. Choi, "Multi-targeted adversarial example in evasion attack on deep neural network," *IEEE Access*, vol. 6, pp. 46084–46096, 2018.

[24] N. Narodytska and S. Kasiviswanathan, "Simple black-box adversarial attacks on deep neural networks," in *In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1310–1318, 2017.

[25] Y. Dong, F. Liao, T. Pang et al., "Boosting adversarial attacks with momentum," in *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9185–9193, 2018.

[26] A. Wu, Y. Han, Q. Zhang, and X. Kuang, "Untargeted adversarial attack via expanding the semantic gap," in *In 2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 514–519, Shanghai, China, 2019.

[27] J. Rony, L. G. Hafemann, L. S. Oliveira, I. B. Ayed, R. Sabourin, and E. Granger, "Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses," in *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4322–4330, 2019.

[28] F. Zuo and Q. Zeng, "Exploiting the sensitivity of l2 adversarial examples to erase-and-restore," in *In Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, pp. 40–51, New York, 2021.

[29] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel, "Adversarial attacks on neural network policies," 2017, http://arxiv.org/abs/1702.02284.

[30] M. Andriushchenko and N. Flammarion, "Understanding and improving fast adversarial training," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16048–16059, 2020.

[31] B. Luo, Y. Liu, L. Wei, and Q. Xu, "Towards imperceptible and robust adversarial example attacks against neural networks," *In Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[32] L. Sun, M. Tan, and Z. Zhou, "A survey of practical adversarial example attacks," *Cybersecurity*, vol. 1, no. 1, pp. 1–9, 2018.

[33] L. Chen, G. Zhu, Q. Li, and H. Li, "Adversarial example in remote sensing image recognition," 2019, http://arxiv.org/abs/1910.13222.

[34] H. Kwon and S. Lee, "Ensemble transfer attack targeting text classification systems," *Computers & Security*, vol. 117, p. 102695, 2022.

[35] G. Du, X. Cao, J. Liang, X. Chen, and Y. Zhan, "Medical image segmentation based on U-Net: a review," *Journal of Imaging Science and Technology*, vol. 64, no. 2, 2020.

[36] M. Abadi, P. Barham, J. Chen et al., "Tensorflow: a system for large-scale machine learning," *In 12th Usenix Symposium On Operating Systems Design And Implementation (Osdi 16)*, vol. 16, pp. 265–283, 2016.

[37] A. Persekian, M. Jiao, and L. Tindall, "U-net on biomedical images".

[38] D. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *The International Conference on Learning Representations (ICLR)*, San Diego, 2015.

[39] H. Fallahgoul, V. Franstianto, and G. Loeper, "Towards explaining the ReLU feed-forward network," *Available at SSRN*, vol. 10, 2019.

[40] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial nets," *In Advances In Neural Information Processing Systems*, vol. 27, pp. 2672–2680, 2014.