Proceedings on Automation in Medical Engineering
Proc AUTOMED, Vol 2, No 1 (2023): Article Ref 754
www.journals.infinite-science.de/automed/article/view/754

Infinite Science Publishing

Generate adversarial images with gradient search

N. Ding[1*], K. Möller[1]

[1] *Institute of Technical Medicine, Furtwangen University, Villingen-schwenningen, Germany*
[*] *Corresponding author, email: din@hs-furtwangen.de*

Abstract: Convolutional neural networks have been successfully applied in many areas, but the security concern was challenged by the vulnerability to adversarial samples, which were crafted by minor modification on the legitimate samples. These adversarial samples can easily fool the neural network model with high success rate, therefore, to analyze the models' classification robustness, the adversarial samples can be developed into an index of model robustness. In this paper, we use a gradient search method to generate adversarial samples from the real samples. The model was trained to perform the surgical tool recognition task from cholecystectomy videos. Instead of setting a target misclassified class, we use non-target gradient space search to determine the nearest adversarial class region.

I. Introduction

Convolutional neural networks have drawn increasing attention as they have shown successful in image classification tasks. However, a drawback cannot be ignored is their vulnerability to adversarial samples, which are of concern especially in safety-critical application areas. Adversarial samples are generated by slightly modifying legitimate samples, so that they show invisible changes to the human observer but lead to misclassification by the model. To analyze the safe area around a benign sample, we will use a gradient space search to generate adversarial samples right alongside the borderline of original class and a misclassification. As expectation, the distance between the benign sample and the final generated image will be quantified as an index of the model robustness. Nevertheless, first of all, the adversarial images need to be successfully generated

I.I. Related work

Many approaches have been proposed to generate adversarial samples. The most popular gradient-based method is fast gradient sign method (FGSM)[1], or gradient method which use real gradient instead of the sign to modify the input. By the different cross-entropy loss function setting, we can choose the non-targeted gradient search or the targeted gradient search. As a comparison of our previous work which used the target loss function to seek for a targeted adversarial classification, in this paper, we use non-targeted gradient search to find out the nearest adversarial classification space.

I.II. Previous work

In the previous work, we tried to set a target classification to get the target adversarial image, at different training states. The result could be interpreted as different robustness according to trainings success.[2]

II. Material and methods

We fine-tuned the convolutional neural network model AlexNet [3] to perform surgical tool classification task in the cholecystectomy videos. The original dataset cholec80 [4] contains 80 cholecystectomy videos, including 7 different surgical tools, we extract 80,190 1-class images from the original dataset, from these images 25,000 were used to train the model, the training states were stopped at training accuracy 75%,85%,95%,99%, these snapshots were named as model 75, model 85, model 95 and model 99, respectively. Few samples were selected for the experiment, each class (or surgical tool) has maximum 20 images, but due to the low accuracy of class 5 and class 7, there were only 3 images for this test.

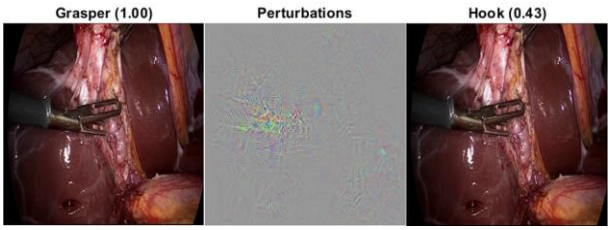

Figure 1: An example of using gradient method to generate an adversarial image.

The adversarial images were generated by iteratively adding the gradient of the loss to get an image that is far enough from the original for misclassification.

$$x_0^* = x;$$
$$x_n^* = x_{n-1}^* + \alpha \nabla_x J\left(x_{n-1}^*, y_{origin}\right); \quad (1)$$

In this function, x_n^* is the generated adversarial image, x_{n-1}^* is the generated adversarial image from the last iteration. y_{origin} is the original class where the image is correctly assigned. The learning rate α was set to 3 different values 1000, 10,000, and 100,000.

III. Results and discussion

Figure 2 shows an example of using gradient search to find an adversarial classification. At the beginning, the gradients generated were quite small, because the loss was nearly 0, after few iterations, the gradients were increased with larger loss value. The probability of an adversarial class, the pixel-wise modification of each iteration and the current loss all showed an exponential increment.

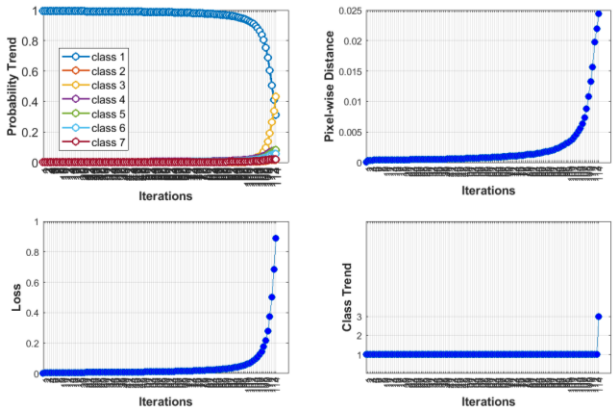

Figure 2: The classification probabilities, per-pixel modification at each iteration, the current loss curve and classification in an adversarial image generation process.

To reduce the calculation time, the maximum iterations was limited to 1000, when the classification cannot be changed within 1000 iterations, the case will be considered as fail.

Different from the result in the previous experiment which generate adversarial images with a set target adversarial class, which is 100% success rate for all the models to get the adversarial images within 1000 iterations, in this experiment, the success rates were lower.

Figure 2 shows the success rate with different learning rate at different training states, the success rate was consistently decreasing while the training state improved, only the model 75 can approximately 100% get adversarial images at 3 different learning rates.

However, a larger learning rate can slightly improve the success rate at better training states. Therefore, to successfully generate an adversarial image, the learning rate need to be increased at a better training state.

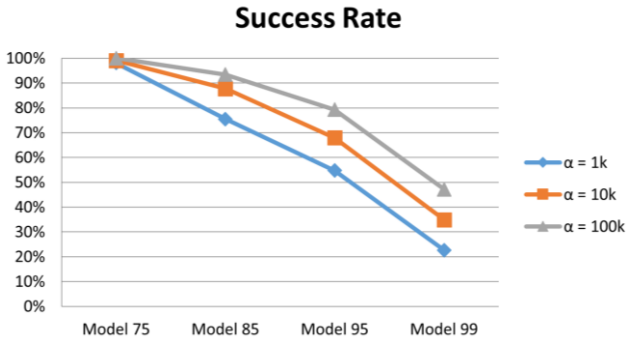

Figure 3: The success rate to generate adversarial images decreased at better training states. A larger learning rate can slightly improve the success rate.

The further evaluation focus on the nearest wrong classification space by gradient search method, we noticed some sample from a class goes more to a specific wrong class, then this class could be the nearest adversarial class that the sample can be modified to.

Table 1 shows the most frequently shown adversarial class, however, when the success rate is too low, most samples will remain as the original class.

Table 1: Frequently showing final class within 1000 iterations and the number of samples when the learning rate α=1000. (The cells marked as red indicate that most samples were not successfully changed their label.)

	State	Class	State	Class	State	Class	State	Class
1	75%	3 13/20	85%	3 7/20	95%	3 8/20	99%	1 10/20
2	75%	1 13/20	85%	2 14/20	95%	2 15/20	99%	2 20/20
3	75%	1 15/20	85%	3 10/20	95%	3 17/20	99%	3 18/20
4	75%	2 8/20	85%	1,5 10/20	95%	4 10/20	99%	4 19/20
5	75%	1 2/3	85%	6 2/3	95%	4 3/3	99%	5 2/3
6	75%	1,3 10/20	85%	1,3,5 6/20	95%	1 8/20	99%	6 10/20
7	75%	1 3/3	85%	1 3/3	95%	1 2/3	99%	7 3/3

Table 1 indicates the most frequently showing adversarial class for each original class. In a summary, most class has only 1 or 2 adversarial class tendency, for instance, in the successful generated images, class 1 tend to misclassify as class 3, but class 2,3,7 tend to misclassify as class 1. However, for class 4,5,6, they have relatively random tendency to an adversarial misclassification.

IV. Conclusions

Gradient search is an efficient method to change an input to a different classification label. When the loss was calculated between prediction and the original classification, as the loss decrease at a better training states, makes it more difficult to get an adversarial sample. Therefore, the learning rate need to be increased at well-trained states to improve the success rate. Similarly, to generate adversarial images for additional adversarial training, the learning rate need to be adaptive with the training states to prevent overfitting with the data points.

Author's Statement

Research funding: This work was supported by the German Federal Ministry of Research and Education (BMBF under grant CoHMed/PersonaMed KFZ 13FH5I06IA). Conflict of interest: Authors state no conflict of interest. Informed consent: Informed consent has been obtained from all individuals included in this study. Ethical approval: The research related to human use complies with all the relevant national regulations, institutional policies and was performed in accordance with the tenets of the Helsinki Declaration, and has been approved by the authors' institutional review board or equivalent committee.

References

[1] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).

[2] Ding, Ning, and Knut Möller. "Robustness evaluation on different training state of a CNN model." Current Directions in Biomedical Engineering 8.2 (2022): 497-500.

[3] Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[4] Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M. and Padoy, N., 2016. Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging, 36(1), pp.86-97.