Ning Ding*, Knut Möller

Using adaptive learning rate to generate adversarial images

https://doi.org/10.1515/cdbme-2023-1090

Abstract: Convolutional neural networks (CNNs) have proved their efficiency in performing image classification tasks, as they can automatically extract the image features and make the corresponding prediction. Meanwhile, the CNNs application is highly challenged by their vulnerability to adversarial samples. These samples are slightly different from the legitimate samples, but the CNN gives wrong classification. There are various ways to find the adversarial samples. The most common method is using backpropagation to generate gradients as the directed perturbation. Contrarily to set a constrained limitation, in this paper, we use iterative fast gradient sign method to generate adversarial images with the minimum perturbation. The CNNs were trained to perform surgical tool recognition as a configuration for the modern operation room. The coefficient or the learning rate which influenced the modification per iteration, was set to be adaptive instead of a fixed number. A few functions were utilized to perform the learning rate decay to compare the performance. Especially, we propose a new adaptive learning rate algorithm that consider the loss as a part of influence factor constitute the learning rate for the rest iterations. According to the experiments, our loss adaptive learning rate method was proved to be efficient to get the minimal perturbations for adversarial attack.

Keywords: Convolutional neural network, adversarial attack, surgical tool recognition.

1 Introduction

Convolutional neural networks have been broadly applied in the image classification area because of their efficiency. Nevertheless, CNNs are sensitive to the adversarial attacks, even though the modifications are invisible. For safety concern in the medical application area, the robustness of the model should be ensured. To evaluate the robustness of the model, we use the adversarial attack technique to identify the vulnerability of the model. Adversarial attack is trying to slightly modify the legitimate samples in order to fool the model to wrong predictions. The most common adversarial attack algorithm is using the backpropagation method to craft the gradient as the perturbations. These methods were proposed such as the fast gradient sign method (FGSM) [1], the basic iterative fast gradient sign method (BIM) [2], the momentum iterative fast gradient sign method (MI-FGSM) [3], etc. So far, in the iterative form of fast gradient sign method, a fixed learning rate per iteration was used. In our experiments, we used a few decay functions to craft adaptive learning rate, in order to get the adversarial images with minimum perturbations representing the borderline samples. In addition to measuring the classification robustness, these borderline samples would help us to understand the decision boundary of the classification space.

As the black-box property of the CNNs training procedure, and the higher dimensions of parameter space and the learned feature space, understanding the CNNs learning procedure is still a difficult task. However, we could use some techniques to assess the learning performance. Additional to the classic evaluation metrics (e.g. accuracy, f1-score and mean average precision), robustness assessment is another interesting criterion to measure the CNNs performance. In this work, the adversarial attack which performed by the adaptive learning rate fast gradient sign method was applied on a simple trained CNN model, such as the AlexNet [4], at different training states to measure its classification performance on both of clean data and adversarial data.

In order to craft a robustness index, the minimal perturbations where eligible to change the model classification will be quantified as a measurement metric. Contrary to assign a limitation parameter to define the adversarial area of a given sample [5], we explored the smallest perturbations that could change a given legitimate sample to a borderline adversarial sample with a specific target class. The distance between the original sample and the generated sample is utilized as the safeness index around this particular sample, which indicates the required difficultness to modify a sample from the original class to an adversarial class. To represent the robustness for a particular class, the average modification was calculated using a sufficient number of random samples from a particular class.

***Corresponding author: Ning Ding:** Institute of Technical Medicine(ITeM), Villingen-Schwenningen, Germany, e-mail: din@hs-furtwangen.de
Knut Möller: Institute of Technical Medicine(ITeM), Villingen-Schwenningen, Germany

By witnessing these minimum perturbations, we can measure the robustness change at different training states.

Surgical tool recognition is a popular application of CNNs in the medical area. It has many potential applications such as monitoring the surgical process, and segmenting the surgical workflow. The ultimate aim of such applications is to develop a context aware system provide technique support in modern operating rooms [6,7]. As the CNNs can automatically learn the visual features from surgical videos, it would be dramatically improving the efficiency of the surgical procedure. Nevertheless, the security concern is essential medical domain. In this paper, the CNN models were trained to perform surgical tool classification in cholecystectomy. Subsequently, the classification robustness was evaluated with adversarial attack technique.

2 Method

2.1 Material

In this study the convolutional neural network model AlexNet [4] was trained for surgical tool classification using laparoscopic images. The Cholec80 dataset is a large dataset containing 80 cholecystectomy videos, including seven different surgical tools [7]. 80,190 images were extracted with at most one kind of tool present (single-class images). Single-class images of the first 40 videos (31477 images) were used to train the model the remaining single-class images (48713 images) were used for testing. To evaluate the robustness, 50 correctly classified images from each kind of surgical tool were selected to perform the adversarial attack [8,9].

2.2 Adversarial Attack

In this experiment, we only considered the correctly classified images from each class. For instance, we selected a correctly classified image x from class A, and set another incorrect class B as the target class. The adversarial attack was implemented by minimizing the cross entropy loss between the prediction of the image and the incorrect target class. The perturbations were generated by iteratively using the loss gradient search in the input space to minimize the distance of the current input sample to the wrong classification region. The iterations were stopped exactly when the classification was changed into the target class. Thus, these generated adversarial samples can be considered as the borderline samples right alongside the original class and the target class (only if there is no interfere class showing up during this procedure). The basic iterative fast gradient sign method function is below:

$$x_0^* = x;$$

$$x_n^* = x_{n-1}^* - \alpha \nabla_x J\left(x_{n-1}^*, y_{target}\right) \quad (1)$$

Where x_n^* is the generated adversarial image at nth iteration, x_{n-1}^* is the generated adversarial image from the last iteration. y_{target} is the target class (e.g. class B) [8,9]. Learning rate α was set to be adaptive to the iterations.

2.3 Learning Rate Function

Inspired by the adaptive learning rate schedule implemented to train the CNN model [10], the same learning rate scheduled was applied to generate the adversarial images. In addition to the default constant learning rate, the common learning rate schedules included many decay functions, such as iteration-based decay, step decay, exponential decay, etc.

Besides, in our experiment, by observing the loss decrement in the iterations, the loss influences the gradient (or the sign of gradient) further influence on the perturbation of each iteration, as the generated samples closer to the target class region, the loss can be rescaled as a decreasingly factor to define the current learning rate state. The specific method for our loss adaptive learning rate is also listed in the algorithm 1. The adversarial images generating procedure described below:

Algorithm 1: Generate adversarial images with adaptive learning rate

Input: Trained model at different epochs, test sample set{x,y}, original class Y_{orig}, target class Y_{target}, the generated image and its classification at current iteration $\{X_n, Y_n\}$, the cross-entropy loss at current iteration L_n, gradient sign map S_g, iterations Itr, learning rate lr, the initial learning rate lr_0, stopping criterion with maximum iteration limitation 100.

Output: The adversarial images misclassified as target class.

For Itr < 100:

If Y_n ~= Y_{target}, Calculate the current cross-entropy loss L_n between the prediction Y_n and the Y_{target}. Backpropagation to get the gradient sign map S_g for updating the input X_n.

Choose the different learning rate algorithms:

1. Constant learning rate: $lr = lr_0$;
2. Iteration based decay learning rate: $lr = lr_0 / (1 + 0.5 * (Itr - 1))$;
3. Exponential decay: $lr = lr_0 * \exp(-0.5 * (Itr - 1))$;
4. Step decay: $lr = lr_0 * 0.5 * \text{floor}((Itr - 1)/1)$;
5. Loss adaptive decay(ours): Concatenate with the previous loss $\{L_1, L_2, \ldots L_n, 0\}$, rescale as the learning rate factor between [0,1]: $\{F_1, F_2, \ldots F_n, 0\}$, F_n is the learning rate for the next iteration.

Update X_n: $X_n = X_{n-1} - lr * S_{gn}$;

$Itr = Itr + 1$;

Elseif Y_n == Y_{target}, break;

Return: The generated image with perturbations.

2.4 Learning Rate Function

There are seven surgical tools used in the Cholec80 dataset, each surgical tool is considered as a specific class. To perform the adversarial attack, the sample from a specific class was

modified to all 6 other classes, and the modifications were quantified as a measurement metric. For instance, we have 50 images from class A, all the images are modified to another class B, and the mean modification from these samples would approximately represent the distance of classification region from A to B. However, according to some observations, due to the limitations of experiment setting and the property of some specific samples, some images cannot be successfully modified to the target class. Hence, we consider the success rate as one of evaluation metrics [9].

– The classic model performance measurement: accuracy and F1-score. The function of F1-score is list below:

$$F-score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (2)$$

– The maximum iteration of the gradient search procedure was limited to 100. Normally, the iterations stopped exactly when the generated image successfully misclassified as the target class. However, when the image cannot be changed to the target adversarial class within 100 iterations was considered as a failed case.

– The difference between the original image and the generated adversarial image was summed as pixel-wise L1-norm distance:

$$D(x, x^*) = \frac{1}{n} \|x^* - x\|_1 \quad (3)$$

n is the number of pixels, x is the original image and $x^*$is the generated adversarial image.

3 Results

In this experiment, we trained a shallow convolution neural network AlexNet [4] to perform the surgical tool classification task. We trained the model with 10 epochs in total. Trained model at every epoch was evaluated for both the classification performance and the classification robustness. Initially, the classification performance on the test dataset was evaluated (see figure 1).

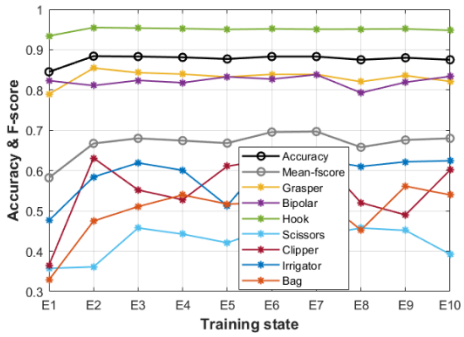

Figure 1: The classification performance measurement by accuracy and f1-score at different training epochs.

The robustness of classification was also evaluated. As mentioned before, correctly classified samples were minimally Even though there are some fluctuations of some surgical tools, the classification performance on the original test set has not much difference at different training epochs.

modified to create patterns not noticeable different for human observers with incorrect classifications. The hypothesis is that a larger value of the L1-norm distance of the generated to the original input sample indicates a higher level of robustness of the current trainings state.

Unfortunately, this distance is affected by the learning rate. Thus, different algorithms were explored to identify the adversarial attack success rate within the limitation of 100 iterations and to compare their average distances. Figure 2 shows the success rate of different learning rate algorithms. The results show that the constant learning rate, iteration decay

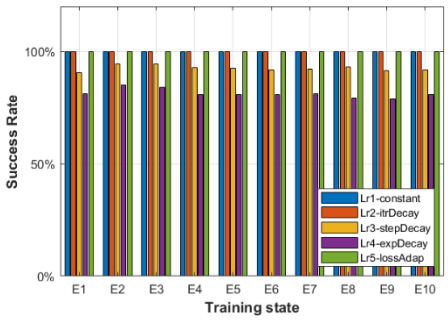

Figure 2: The success rate to generate adversarial images within 100 iterations. The adaptive learning rate were listed accordingly: the constant learning rate, iteration decay, step decay, exponential decay, and loss adaptive decay(ours).

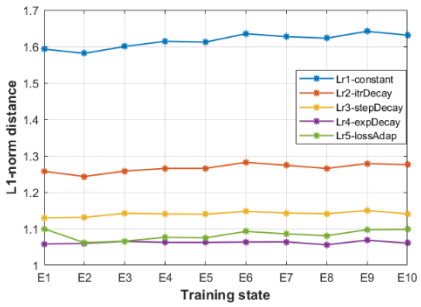

Figure 3: The L1-norm distance of original image and generate adversarial images as the reference to measure the robustness. The adaptive learning rate were listed accordingly: the constant learning rate, iteration decay, step decay, exponential decay, and loss adaptive decay (ours).

and loss adaptive decay(ours) can generate adversarial images successfully in each case(100%) within 100 iterations. This demonstrates the efficiency and effectiveness of these decay algorithms in gradient space search. However, step decay and exponential decay were not able to 100% successfully produce the targeted adversarial images. Similar to the classification

performance on the clean data, the success rate of adversarial attack is approximately same at different training epochs.

Figure 3 shows the mean L1-norm distance calculated from the different learning rate algorithms. The results only consider the successful misclassified images. Clearly, the constant learning rate generated more perturbations than other learning rate algorithms. On the other hand, step decay, exponential decay and our loss adaptive learning rate can generate relatively smaller perturbations. However, the success rate using step decay and exponential decay (see figure 3) is lower than 100% when the iterations are limited to 100. Thus, our loss adaptive learning rate could maintain a higher success rate, in the same time, generating smaller perturbations for the adversarial images.

4 Discussion

The gradient based search is an efficient method to generate directed, small perturbations in the input space to provoke incorrect classifications. Because of the high-dimensional property of the input space and the black-box property of the CNN model, it is difficult to define an exactly safe area around a given sample. Therefore, we have to rely on experimental methods to approximately detect the classification boundaries near the input space samples. Besides the L1-norm distance which can directly represent the robustness of the CNN, the effect of the applied learning rate function is another reference to compare the learning algorithms. Figure 4 shows the learning rate functions calculated from all the samples at the 10th epoch state. Compared to the constant learning rate and the iteration decay in the figure 4 (left), our loss adaptive learning rate is smaller than others, the generated perturbation is more precise and targeted. Similar to the result of success rate, the step decay and exponential decay function in figure 4 (right) cannot 100% successfully get adversarial images for all the samples. And a higher mean iterations indicate they are more time-consuming than the algorithms in the figure 4 (left).

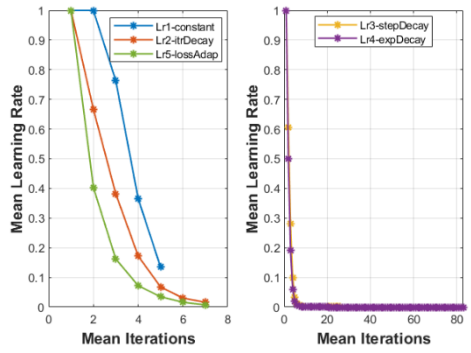

Figure 4: The learning rate summary based on the mean iterations and mean learning rate calculated from all the samples. (Mean iterations: step decay 50, exponential decay 83.

5 Conclusion

In this research, we used a gradient search method with adaptive learning rate to generate adversarial samples. Adversarial training is another interesting topic to improve model robustness, in the future work, we will use these adaptive learning rate gradient search method to generate adversarial images as additional training samples to enhance the training process, which might be helpful to improve CNN robustness.

Author Statement

Research funding: This work was supported by the German Federal Ministry of Research and Education (BMBF under grant CoHMed/PersonaMed KFZ 13FH5I06IA and DAAD Grant AIDE-ASD FKZ 57656657). Conflict of interest: Authors state no conflict of interest. Informed consent: Informed consent is not applicable. Ethical approval: The research is not related to either human or animals use.

References

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

[2] Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial machine learning at scale." arXiv preprint arXiv:1611.01236 (2016).

[3] Dong, Yinpeng, et al. "Boosting adversarial attacks with momentum." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[4] 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).

[5] Szegedy, Christian, et al. "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199 (2013).

[6] Alshirbaji, T. A., Ding, N., Jalal, N. A., & Möller, K.(2020). The effect of background pattern on training a deep convolutional neural network for surgical tool detection.Proceedings on Automation in Medical Engineering, 1(1), 024-024.

[7] 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.

[8] Ding, Ning, and Knut Möller. "Generating adversarial images to monitor the training state of a CNN model." Current Directions in Biomedical Engineering 7.2 (2021): 303-306.

[9] 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.

[10] https://towardsdatascience.com/learning-rate-schedules-and-adaptive-learning-rate-methods-for-deep-learning-2c8f433990d1.