ADVERSARIAL ATTACK DRIVEN DATA AUGMENTATION FOR ACCURATE AND ROBUST MEDICAL IMAGE SEGMENTATION

Mst. Tasnim Pervin[⋆] *Linmi Tao*[⋆] *Aminul Huq*[⋆] *Zuoxiang He*[†] *Li Huo*[‡]

[⋆]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[⋆]Beijing National Research Center for Information Science and Technology, China
[⋆]Key Laboratory of Pervasive Computing, Ministry of Education, Beijing, China
[†]Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua University
[‡]M.D, Nuclear Medicine Department, Peking Union Medical College Hospital

ABSTRACT

Segmentation is considered to be a very crucial task in medical image analysis. This task has been easier since deep learning models have taken over with its high performing behavior. However, deep learning model's dependency on large data proves it to be an obstacle in medical image analysis because of insufficient data samples. Several data augmentation techniques have been used to mitigate this problem. We propose a new augmentation method by introducing adversarial learning attack techniques, specifically Fast Gradient Sign Method (FGSM). Furthermore, We have also introduced the concept of Inverse FGSM (InvFGSM), which works in the opposite manner of FGSM for the data augmentation. This two approaches worked together to improve the segmentation accuracy, as well as helped the model to gain robustness against adversarial attacks. The overall analysis of experiments indicates a novel use of adversarial machine learning along with robustness enhancement.

Index Terms— Medical Image Segmentation, Data Augmentation, U-Net, Adversarial Machine Learning, Adversarial Examples

1. INTRODUCTION

The substantial structures of Deep Learning (DL) models have been proved to be very effective for several discriminating tasks related to Computer Vision applications such as prediction, image segmentation, or object recognition with the help of progressive and efficient computational power, and in-depth network architecture [1]. However, DL networks can rarely perform well without the availability of a massive dataset. This limitation proves to be even larger for the field of medical image processing as the availability of a big data is not an easy coin to toss. To mitigate these obstacles, several supervised biomedical segmentation methods focus on manually-engineered preprocessing steps and architectures [2], [3]. Manually-tuned data augmentation is also used to increase the number of training examples [4], [5].

Data augmentation is a good solution for this problem as it will minimize the gap between validation and training portions. Augmentation functions such as random rotations or random nonlinear deformations are easy to apply and are fruitful at improving segmentation accuracy in some settings [2], [3], [5]. However, these functions have a limited capability to imitate real variations and can be highly subtle to the choice of parameters [6]. This phenomena has been reported in the paper of skin lesion classification by [7] and so for liver lesion classification [8].

In this paper, we carried out the task of data augmentation by employing adversarial learning. Recent works show that adversarial attacks are equally capable to degrade segmentation models' performance [9]. The traditional augmentation techniques can not make these models' robust against adversarial attacks. Wetstein et al. proved that powerful DL models such as Inception-v3 and DenseNet-121 tend to loose strength when attacked by FGSM and Projected Gradient Descent (PGD) attacks [10]. Paschali et al. also investigated the effect of adversarial attack in degrading performance of classification (Inception V3, Inception V4, MobileNet) and segmentation (SegNet, U-Net, DenseNet) models [9]. In a recent work, the authors Chen et al. used adversarial data augmentation for improved generalization ability by generating natural signal corruptions rather than pixel attacks [11]. Though adversarial attacks are meant to degrade ML/ DL model performance, but here we used an adversarial attack to enlarge the dataset for training segmentation model. We have also introduced a new perspective to use FGSM for further improvement by modifying its working technique. The key contribution of this paper can be listed as below:

- We used adversarial learning in favor of deep learning models rather that traditional aim of making models' vulnerable. We used FGSM attack to create adversarial examples (tuned to have minimum noise). This helped

Correspondences should be addressed to : linmi@tsinghua.edu.cn. This work is supported by the National Science Foundation of China at the Project 61672017.

to double the number of data to be augmented to mitigate overfitting.

第3个 inverse FGSM

- We used inverse approach of FGSM (InvFGSM) which targets to adjust input data towards minimum loss. This helped model with improved performance by adding positive noises instead of adversarial noises. This approach has not been used till now in any other literature.
- Generally all DL models are vulnerable to adversarial attacks unless are adversarial trained or use any defense mechanism. For our work, we used adversarial training based augmentation, so the model became prepared and become robust against similar attacks in future.

2. RELATED WORKS

FCN

Fully Convolutional Network

Medical image segmentation models uses similar architecture with fewer convolutional blocks and fewer parameters. Ronnerberger et al. proposed the most well-known segmentation network, U-Net, where they expanded the FCN into upsampling and downsampling layers to work with few training images and to create high accurate output [12]. Since then, many modification of U-Net has been proposed to improve the segmentation in various applications, for instance [13]. Yuan created a network called Convolutional-Deconvolutional Neural Networks that uses Jaccard distance based loss function and performs under different image acquisition conditions [14]. For consistent use of contextual information, Hwang el al. proposed a model called Network-Wise Training of Convolutional Networks that employs a multi-stage training strategy for distilled segmentation with smooth boundary [15].

Recent advances in adversarial learning has broken the phantasm of high-performing DL models. An adversarial example, created using attacks of adversarial learning, is a sample of modified inputs (image, text, audio, etc.) that can force a model to misbehave. These attacks corrupt DL models by injecting tiny amount of imperceptible noise into the input and generates original input look-alike adversarial examples. These examples pose a potential security threats for practical applications of DL. Engstrom et al. showed that deep CNN models can be forced to do misclassification by attacking with some easy geometric transformations. This phenomena of performance degradation has been noticed for several dataset including MNIST, CIFAR10, and ImageNet (26% drop for MNIST, 72% drop for CIFAR10, and 28% drop for ImageNet)[16]. In another research, Ian Goodfellow and his fellow authors introduced to Fast Gradient Sign Method using a maxout network to generate adversarial examples which produces 89.4% misclassification with 97.6% confidence [17]. Li et al. in their work compared the performance of original testing data with a new type of adversarial examples [18]. Su et al. invented one pixel attack where they have shown that changing a single pixel of input images can cause 70.97% of misclassification [19]. Zajac and his fellow researchers have focused on attacking the border of images [20]. These examples, and many others, have demonstrated DL models are not robust to the attacks, which is harmful to the applications especially in health area.

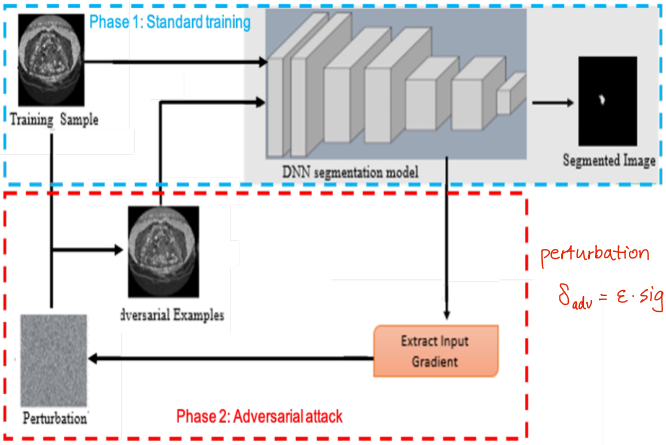


Fig. 1. Workflow of proposed approach

perturbation

$\delta_{adv} = \varepsilon \cdot \text{sign}(\nabla_x)$

with loss function

3. METHODOLOGY

The proposed methodology poses a heavy contrast to the traditional augmentation techniques. The effectiveness of adversarial training in the form of noise or augmentation search is a new concept that has not been widely tested and understood. The flow diagram of figure 1 aims to represent the whole procedure: 1) Train model with original dataset for segmented image localization, 2) Attack the model with adversarial attack technique that will find model architecture and input gradients to craft noises and create new adversarial examples, 3) Adversarial examples will be augmented to the original samples to create new training set that will be fed to the model for training which can produce better segmentation result and make model robust against attacks. The segmentation model and attack techniques has been described briefly in following sections.

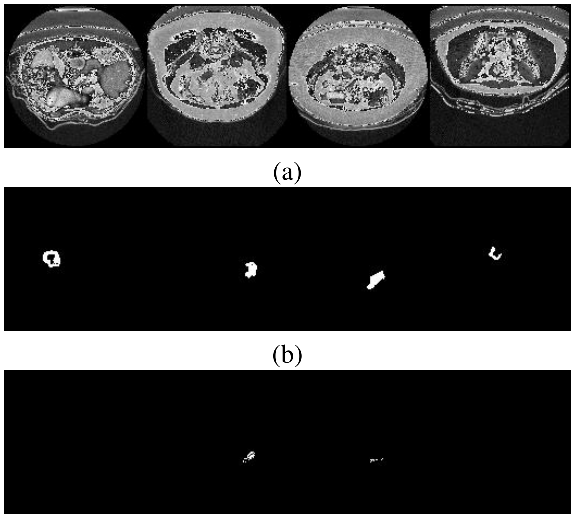
3.1. U-Net

→ A Network designed for medical image classification

U-Net Model

对称模型

The U-Net model, a semantic segmentation model, was developed by Olaf Ronnerberger [12]. The model is built of a contraction/encoder path and an expanding/decoder path. Encoder path is nothing but a pile containing convolutional and maxpooling layers that emphasizes on the context or 'What' information of image where decoder path summons some transposed convolutional layers pursuing the location or the 'Where' information. In the encoder path, the size of the images starts to gradually decrease with increasing depth such as starting from 128 × 128 × 3 to 8 × 8 × 256. In the

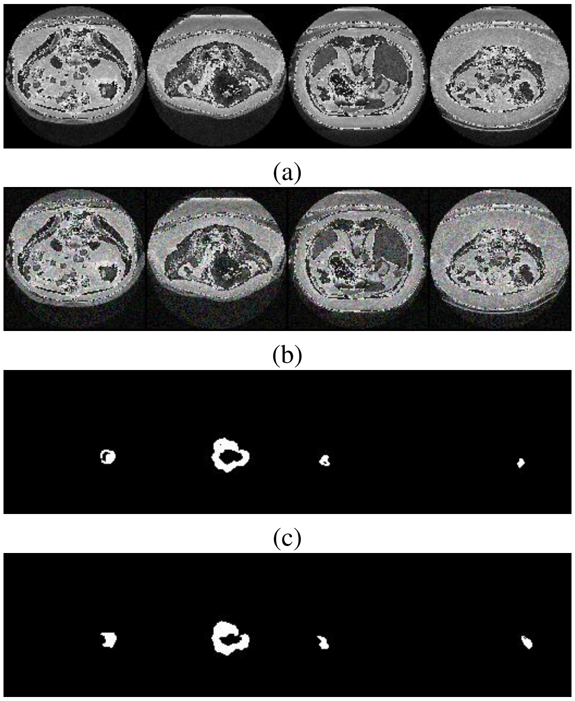


(a)

(b)

(c)

Fig. 2. Performance of UNet without any augmentation : (a) Original input, (b) Ground truth/mask, (c) Predicted mask



(a)

(b)

(c)

(d)

Fig. 3. Performance of UNet after adversarial augmentation by FGSM : (a) Original input, (b) Adversarial image for augmentation created by FGSM, (c) Ground truth/mask, (d)Predicted mask after augmentation

decoder path, the size of the image gradually increases with decreasing depth such as from $8 \times 8 \times 256$ to $128 \times 128 \times 1$. For a specific localization, a concatenation takes place at each level between the encoder's feature maps with the decoder's transposed convolution. This characteristic gives the architecture a U shape. The strength of U-Net is that the network only contains fully convolutional layers but no dense layer which enables model to accept images of any size as input.

3.2. Fast Gradient Sign Method

The most popular adversarial attack, Fast Gradient Sign Method (FGSM) has been proposed by Ian J. Goodfellow [17]. The method works as follows: Suppose $x \& y$ are the input images and corresponding labels respectively, the weights of the network is represented by θ and the loss function is $L(\theta, x, y)$. First, the gradient of the loss function for input image is calculated. The ∇ operator takes out the derivative of the loss function. It gives a shape matrix including width, height, channels, and the tangent slopes. But most importantly only the sign of the slopes is needed to know when the pixel values increase or decrease. Then, these slope signs are multiplied by a very tiny imperceptible perturbation value ϵ . The resultant pattern is then added to the original inputs to create respective adversarial examples. The mathematical representation of this procedure is as follows:

$$x_{adv} = x + \epsilon * sign(\nabla_x L(\theta, x, y)) \qquad (1)$$

3.3. Inverse Fast Gradient Sign Method

The motivation for this approach came from the definition of FGSM. From the gradient of the loss function, we can assume how the changes occur in the input image. FGSM approach targets to adjust the input to a maximum loss using backpropagated gradients rather than adjusting weights to decrease the loss. Inverse FGSM (InvFGSM) adopts an inverse approach of adjusting input to a minimum loss using gradient which definitely helps to gain better performance. For adversarial noise in the FGSM approach, only the sign of the gradient matters. As the sign of the derivatives provides adversarial noises which are added with input to generate adversarial examples, the inverse of the sign provides new type of adversarial noise that will work positively aiding in loss minimization. Data augmentation with these kind of adversarial examples with positive noises seem to be useful in enhancing model performance. The mathematical way to represent this would be as follows:

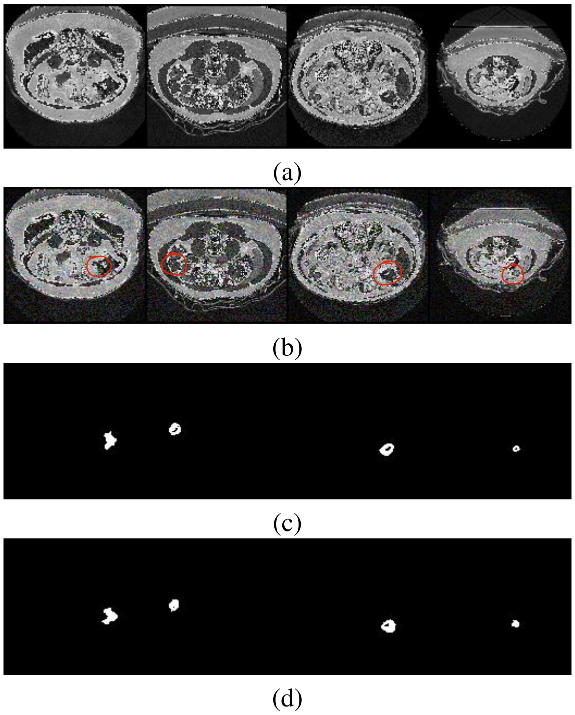
正向 (InvFGSM)
方法会生成对
梯度有利的noise
↑
一定吗?

$$x_{adv} = x - \epsilon * sign(\nabla_x L(\theta, x, y)) \qquad (2)$$

4. EXPERIMENTAL ANALYSIS

4.1. Dataset Description

For this experiment, we have used a dataset of colon cancer provided by Beijing Tsinghua Changgung Hospital and Peking Union Medical College Hospital. This dataset con-


(a)

(b)

(c)

(d)

Fig. 4. Performance of UNet after adversarial augmentation by inverse-FGSM : (a) original input, (b) adversarial image for augmentation created by inverse-FGSM, (c) Ground Truth/Mask, (d)predicted segmented image after augmentation

tains 1285 CT scan images along with same number of mask images of desired localization of segmentation. The original images are of size $512 \times 512 \times 3$ and resized into $128 \times 128 \times 3$ for faster processing when feeding into the model with the support of Jittor platform [21].

4.2. Data Augmentation by Adversarial Training

For experiment, the parameters of model training without data augmentation and with data augmentation have been set as: Loss Function = Categorical Cross-Entropy, Optimizer = SGD, Learning Rate = 0.1, Momentum = 0.99, Batch-size = 16. From table 1, the difference between the resultant Intersection over Union before and after augmentation can be noticed. The U-Net model performed quite well giving about 80% correct segmentation till epoch 30. Data augmentation helped this performance even better. UNet along with FGSM performed approximately 3% better than U-Net alone. But as suspected, Inverse FGSM gave a boost to this performance even further showing about 6% raise. The figures 2,3,4 give indication to this phenomena as well. Augmentation with adversarial machine learning attack techniques can help making model more robust. This phenomena has been shown in table 2. We employed a range of random perturbation starting from 0 to 0.2 to check the robustness of the model. It can be noticed how increasing noise affect the model. Even Noise of $\epsilon = 0.2$ pulls down the segmentation performance up to 35%. Though adversarial training definitely helped the models to regain their performance and made the model robust. For adversarial training, the perturbation was kept to 0.1 as, visual imperceptibility is another aim along with good performance and higher rate of perturbation makes adversarial images visually corrupted.

Table 1. Performance of models over increasing epochs

Models	IoU over epochs			
	0^{th}	10^{th}	20^{th}	30^{th}
UNet	0.4955	0.4986	0.4983	0.8084
UNet + FGSM	0.4987	0.7839	0.8129	0.8394
UNet + Inverse-FGSM	**0.6159**	**0.8264**	**0.8643**	**0.8986**

Table 2. Effect of attack on model robustness with increasing perturbation(ϵ)

Attacking model	Adversarial training	$\epsilon = 0.0012$	$\epsilon = 0.118$	$\epsilon = 0.1176$
FGSM	before	0.7860	0.7246	0.6028
	after	**0.9071**	**0.8242**	**0.7720**
Inverse-FGSM	before	0.7749	0.5841	0.5467
	after	**0.8724**	**0.8729**	**0.8477**

5. CONCLUSION

In this paper, we tried to apply a deep learning model for cancer cell segmentation from colon cancer dataset and also introduced a new approach for data augmentation. Limited data causes the overfitting problem when used on deep learning models, that's why data augmentation is considered to be a crucial step. Though it cannot be an ultimate backup for limited data where any class sample is absent, it can help to deduct overfitting by generating a large number of similar data. We experimented with adversarial machine learning attack techniques as a new data augmentation approach that proved to be beneficial in improving the segmentation performance for this dataset. Though the conventional aim of adversarial machine learning is related to model robustness, this work certainly gives light to the fact that it can be used with careful tuning for data augmentation tasks too. However, analysis is going on if this helps all the models for various dataset universally but it should be considered as a good start to a new approach of data augmentation using adversarial machine learning.

6. REFERENCES

[1] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning.,” *Journal of Big Data*, vol. 6, no. 1, pp. 60, 2019.

[2] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. De Vries, M. J. Benders, and I. I˘sgum, “Automatic segmentation of mr brain images with a convolutional neural network,” *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1251–1261, 2016.

[3] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, “Brain tumor segmentation using convolutional neural networks in mri images,” *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1240–1251, 2016.

[4] C. Marques, A. Pospori, L. Pereira, S. Marques, P. Antunes, O. Bang, D. Webb, and P. Andr´e, “High-quality phase-shifted bragg grating sensor inscribed with only one laser pulse in a polymer optical fiber,” *2017 SBMO/IEEE MTT-S International Microwave and Optoelectronics Conference (IMOC)*, pp. 1–4, 2017.

[5] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu andE. B. Turkbey, and R. M. Summers, “Deeporgan: Multilevel deep convolutional networks for automated pancreas segmentation,” *International conference on medical image computing and computer-assisted intervention*, pp. 556–564, 2015.

[6] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, “Discriminative unsupervised feature learning with exemplar convolutional neural networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 9, pp. 1734–1747, 2015.

[7] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” *nature*, vol. 542, no. 739, pp. 115–118, 2017.

[8] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, “A survey on deep learning in medical image analysis,” *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[9] M.Paschali, S.Conjeti, F.Navarro, and N.Navab, “Generalizability vs. robustness: investigating medical imaging networks using adversarial examples,” *InInternational Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 493–501, 2018.

[10] S.C. Wetstein, C. González-Gonzalo, G.Bortsova, B.Liefers, F. Dubost, I. Katramados, L.Hogeweg, B. van Ginneken, J.P. Pluim, M. de Bruijne, and C.I. Sánchez, “Adversarial attack vulnerability of medical image analysis systems: Unexplored factors,” *arXiv preprint arXiv:2006.06356*, 2020.

[11] C. Chen, C. Qin, C. Ouyang H. Qiu, S. Wang, L. Chen, G. Tarroni, W. Bai, and D. Rueckert, ,” *In International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 667–677, 2020.

[12] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, 2015.

[13] A. A. Novikov, D. Lenis, D. Major, J. Hladůvka, M. Wimmer, and K. Bühler, “Fully convolutional architectures for multiclass segmentation in chest radiographs,” *IEEE transactions on medical imaging*, vol. 37, no. 8, pp. 1865–1876, 2018.

[14] Y. Yuan, “Automatic skin lesion segmentation with fully convolutional-deconvolutional networks,” *arXiv preprint arXiv:1703.05165*, 2017.

[15] S. Hwang and S. Park, “Accurate lung segmentation via network-wise training of convolutional networks,” *In Deep learning in medical image analysis and multimodal learning for clinical decision support*, pp. 92–99, 2017.

[16] L. Engstrom, D. Tsipras, L. Schmidt, and A. Madry, “A rotation and a translation suffice: Fooling cnns with simple transformations,” *arXiv preprint arXiv:1712.02779*, vol. 1, no. 2, pp. 3, 2017.

[17] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.

[18] S. Li, Y. Chen, Y. Peng, and L. Bai, “Learning more robust features with adversarial training,” *arXiv preprint arXiv:1804.07757*, 2018.

[19] J. Su, D. V. Vargas, and K. Sakurai, “One pixel attack for fooling deep neural networks,” *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, 2019.

[20] M. Zajac, K. Zołna, N. Rostamzadeh, and P. O. Pinheiro, “Adversarial framing for image and video classification,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 10077–10078, 2019.

[21] S. M. Hu, D. Liang, G. Y. Yang, G. W. Yang, and W. Y. Zhou, “Jittor: a novel deep learning framework with meta-operators and unified graph execution,” *Science China Information Science*, vol. 63, no. 12, pp. 222103:1–222103:21, 2020.