Chest X-ray Segmentation from Small Dataset by Data Augmentation Using Distribution Fitting GAN

Anonymous

Abstract. Deep Learning Network requires large-scale annotated training data to perform accurate prediction. However, in medical imaging, both obtaining medical data and annotating them by expert physicians are challenging. Therefore, data augmentation is widely used to enrich the data and avoid overfitting. However, traditional augmentation methods are designed for general nature images rather medical imaging data. Generative adversarial networks (GANs) can capture the underlying distribution of the data and synthesize realistic samples, and then provides a new way to do data augmentation. In our work, we do chest X-ray data augmentation based on image-to-image translation networks. In order to better capture the distributions of the real data, we add Fréchet Inception Distance (FID) loss which calculates the Wasserstein-2 distance between real images' and synthetic images' distributions into the network. With FID loss, we can fit the generated images' distribution into real data's distribution. We call this network Distribution Fitting GAN (**DF-GAN**). We can generate more realistic synthetic images and show significant improvement on the segmentation task with small amount of data via this method.

Keywords: Generative adversarial networks (GANs) · Data Augmentation · Segmentation.

1 Introduction

Deep learning techniques require large amounts of data to train effective models for tasks such as object detection and segmentation. In medical imaging, data is not as sufficient as other computer visions fields due to privacy issues and happening probabilities. Also, annotation of medical images is very expensive and requires huge amount of time and effort from the domain experts, however, having correct labeled data is mandatory for a supervised learning task.

Generative Adversarial Networks (GANs) [7] have become a new technique to perform data augmentation. GAN is composed of a generator and a discriminator. The generator tries to generate images that can confuse the discriminator, and the discriminator tries to recognize the images from real data or synthesized by generator. Two networks train against each other, thus the generative network can generate more realistic images. Compared with traditional data augmentation methods, such as rotations, translations, reflections, and adding Gaussian

noise, the advantage of GANs is that the models can generate new synthetic data with much larger diversity.

Various GANs have been applied to generate synthetic medical imaging data for data augmentation. Noise-to-image translation models such as DCGAN [16], PGGAN [13] have been used in the medical imaging domain to perform data augmentation [6][14][8][4]. The problems with these noise-to-image translation models are that the synthesized images may not look visually plausible and generated disease features are not real for diseases. There are also image-to-image translation models like MUNIT [11], CycleGAN [24] that could generate realistic images, showing improvement in medical images classification, detection and segmentation [20][9][15]. MUNIT and CycleGAN are unsupervised learning model, so they don't require paired normal and abnormal images from the same patient. However, that also means we don't get the ground truth of new generated images if we don't annotate manually or have a precise segmentation model.

To solve the above issues, we use a pairwise image-to-image translation(Pix-2Pix [12]) as our baseline model. Pix2Pix is widely used in medical imaging data augmentation [22][2][1][18] because it adds a L1 distance loss to generate images visually close to real ones. Therefore, it can be used to generate regions around undistorted pathological areas. We leverage the limited number of mask labelled data to create a set of paired training images to fit into the training of the Pix2Pix model. Realistic images can be synthesized through Pix2Pix, but it's mainly due to the L1 distance loss' constraint; that is, the effectiveness of the synthetic data to the segmentation model can't be assured. Thus, we incorporate a new loss which based on the idea of Fréchet Inception Distance (FID) score [10] to fit the synthetic data into the underlying distribution of the real data. Fréchet Distance [5] is used to measure the distance between two normal distributions through calculating the distance of mean and standard deviation. FID score is an evaluation of the performance of GANs at image generation, where it calculates the Fréchet distance of generated images' distribution and real one. Through adding FID score as a loss into the Pix2Pix model, we can reduce the distance between two distributions and draw the generated images' distributions close the real ones. We call this method Distribution Fitting GAN (DF-GAN).

We perform our experiment on a chest X-ray Pneumothorax Segmentation dataset. We use the trained generative model to generate new pathological images, a new dataset is prepared combining synthetic images with real ones. To validate the effectiveness of our augmentation method, we train an U-Net++[23] as our segmentation model. We also conduct some traditional augmentations such as flip, distortion, crop during training U-Net++. Then we compare the dice score under various conditions to check if the segmentation accuracy increase after our augmentation.

2 Methodology

2.1 Method Overview

The purpose of our method is to synthesize new training data which is then used to augment the existing data for training a better segmentation network. We employ an idea similar to image inpainting by generating regions around the undistorted pathological area based on a pairwise image-to-image translation network. Fig.1 shows our model architecture. Besides GAN loss and $L1$ distance loss, we incorporate a FID loss to the Pix2Pix model to fit synthetic data into the real data's distribution. After training, the trained generative model is input with a disease area to synthesize new images. Then, we add synthetic images into real data to check if the data augmentation is effective through a segmentation network.

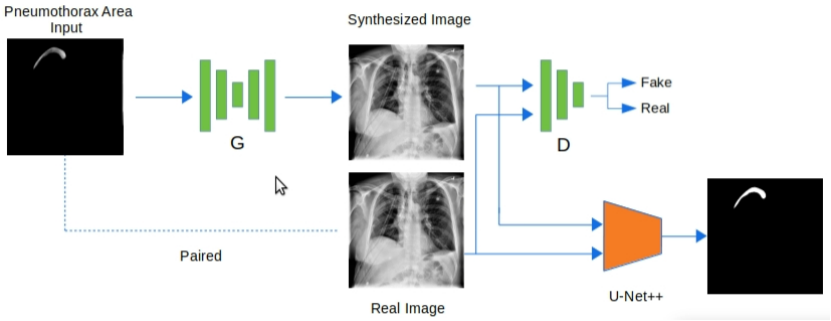

Fig. 1. Model Architecture. We use a pairwise image-to-image translation network to generate regions around the undistorted pathological area to form new diseased images and combine the same amount of synthetic data with real data as a new training dataset for segmentation model(U-Net++) to conduct segmentation task.

2.2 Distribution Fitting GAN

Image-to-image translation network learns a mapping from an input image x to a target output image y. Our goal is to synthesize new disease training images $\bar{x}$ from the original images x via the image-to-image translation network. We use Pix2Pix as our baseline model.

Loss Function The loss function for Distribution Fitting GAN (DF-GAN) is composed of three parts. The first part is the adversarial loss, which is for the discriminator D to ensure the synthetic image $\bar{x} = G(x)$ (generated by the generator G) can be distinguished from the target real image y. On the other

hand, the generator G is trained to generate more visually plausible image. D and G update iteratively against each other with the following objective function:

$$L_{GAN} = E_{x,y}[logD(x,y)] + E_{x,z}log[(1 - logD(x, G(x)))], \quad (1)$$

where G tries to minimize this objective function against D. Also, our model includes a traditional construction loss, $L1$ distance loss, to assure the synthetic images to be visually plausible:

$$L_{L1}(G) = E_{x,y}[||y - G(x)||_1]. \quad (2)$$

Different from the original Pix2Pix model, we incorporate a new loss into the original Pix2Pix model, which we called FID loss. This idea comes from Fréchet Inception Distance[10] which is a performance measure for generative models. FID score improves the drawbacks of Inception Score[3]. FID takes not only generated images but also real data into consideration. In FID, we feed data into a pretrained Inception Network[19]. We remove the original output layer and use the inception network as a feature extractor. Each image will be extracted into 2048-dimensional feature vector. We assume the distribution of real and synthetic data to be multidimensional Gaussian distributions. Then, we can calculate the Fréchet distance which also known as Wasserstein-2 distance[21] by using mean and covariance of feature vectors. We call the Fréchet distance $d(R, S)$ between the distribution R and the distribution S as the "Fréchet Inception Distance" (FID)[5], which is given by:

$$L_{FID} = d^2((m_r, C_r), (m_s, Cs)) = ||m_r - m_s||_2^2 + Tr(C_r + C_s - 2(C_r C_s)^{\frac{1}{2}}), \quad (3)$$

where the mean m_r and covariance C_r are obtained from the distribution R and the values m_s and C_s are obtained from S. With this loss function, we can calculate the Fréchet distance between real and synthetic images and minimize it to bring the synthetic data's distribution close to the real data's distribution.

Training Procedure Our model's architecture is based on Pix2Pix. We use U-Net [17] as our generator to capture the low frequency structure of the images and use a PatchGAN [12] to capture the high frequency details. The training procedure of our model is shown in algorithm 1. Since we assume that synthetic data is a normal distribution, we calculate the FID loss once a batch instead of once a image. In this way, the data size of synthetic data would be bigger so that the synthetic data's distribution would be closer to Gaussian. It can also reduce the training time.

3 Experiments

3.1 Dataset

The dataset used for evaluation of the our method was obtained from the 2019 SIIM-ACR Pneumothorax Segmentation Challenge on Kaggle. It contains 12089

Algorithm 1 DF-GAN. epochs = 2000, batch size = 16, save epoch freq = 50, lamda L1 = 100

for each $i \in [1, 2000]$ **do**
  Initialize model;
  Create a folder to save synthetic images;
  Load data;
  Set Count = 0;
  for each d in training dataset **do**
    Generator synthesizes a image with input d and save the image to the fake image folder;
    Count + = 1;
    update D
    Get D's gradients;
    Calculate L_{GAN};
    Update network weights;
    update G
    Get G's gradients;
    $L_{FID} = 0$;
    if Count == batch size **then**
      Calculate L_{FID} with fake image folder and training dataset;
    end if
    Calculate L_{GAN} and L_{L1};
    $G'sLoss = L_{GAN} + lamdaL1 * L_{L1} + L_{FID}$;
    Update network weights;
  end for
  Remove all the images in fake image folder;
end for

frontal-view chest X-ray images, including 2669 images with Pneumothorax mask annotation and 9420 healthy images. We use the mask annotated images to create a training set for our Pix2Pix model. Figure2 shows some samples from training dataset. Also, We downsize the original images from 1024 × 1024 to 256 × 256 for fast processing.

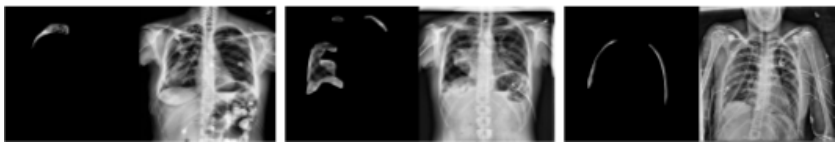

Fig. 2. Training data samples

3.2 Data Augmentation With Various Data Size

Due to the imbalance of healthy and diseased data, we mainly focus on the diseased images. Since if we add healthy images into the training dataset for segmentation model, the segmentation performance will be much better because the model mainly learns how to classify healthy images. However, what we want to observe is that if our augmentation can help the model learn to classify pathological area.

To evaluate the effectiveness of our data augmentation with various amounts of data, we train Pix2Pix model with 100, 250, 500, 1000 and 2000 images respectively and combine the same data size of synthetic images with real images as the training data for segmentation model. Figure 3 shows some samples from Pix2Pix and our DF-GAN. We use U-Net++[23] as our baseline segmentation network. Table 1 display the segmentation performance in various conditions. We also apply mathematical transformations, including rotations, reflections and cropping, to add variation into U-Net++ which denoted as Classical Augmentation in the table. We can see that when the data size is small, improvements for our GAN augmentation is significant. But, when the data size increases, the improvements diminish. We believe the reason is that while the data size increase, the segmentation model would gradually reach its limit, then it would be difficult to keep raising the accuracy even with more data. The improvement ratio with various data size is shown in figure 4.

To validate if adding more synthetic images into the segmentation training set would keep improving the segmentation performance, we conduct flip augmentation to the diseased images and their corresponding masks, then we use these images to generate new diseased data with trained generative model. Table 2 indicates that the augmentation's effectiveness diminishes when we add more data into the training set. Considering the time cost and computation resources,

Table 1. Quantitative results for segmentation (Mean ± standard error) (R denotes real images, and S denotes synthetic images) The Synthetic data are generated from the generative model that is trained with the corresponding masks and the real images.

Method	Classic Aug.	50 R	50 R + 50 S	100 R	100 R + 100 S	250 R	250 R + 250 S
Dice	no	0.232±0.021	0.28±0.033	0.256±0.015	0.322±0.01	0.398±0.009	0.373±0.007
	yes	0.257±0.009	0.328±0.025	0.324±0.038	0.382±0.008	0.423±0.011	0.433±0.007
	Classic Aug.	500 R	500 R + 500 S	1000 R	1000 R + 1000 S	2000 R	2000 R + 2000 S
	no	0.419±0.003	0.407±0.012	0.458±0.002	0.461±0.001	0.498±0.002	0.504±0.005
	yes	0.46±0.002	0.47±0.004	0.496±0.004	0.5±0.001	0.53±0.003	0.532±0.002

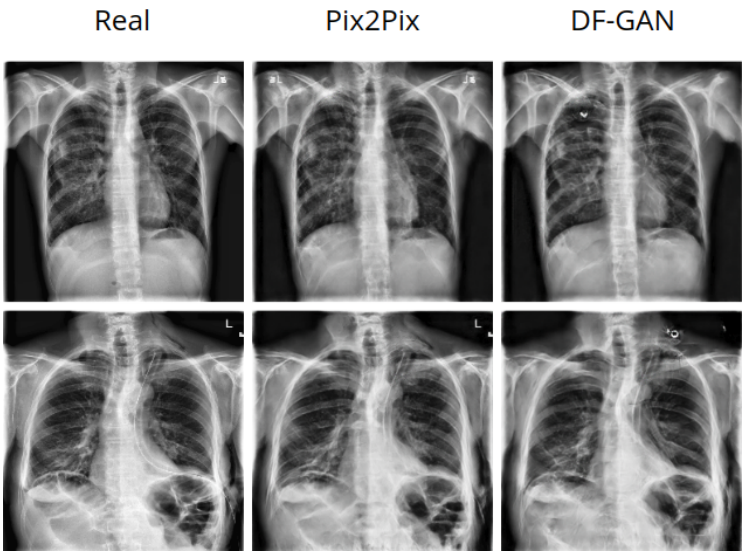

Fig. 3. Real samples and Synthetic data samples from Pix2Pix and DF-GAN

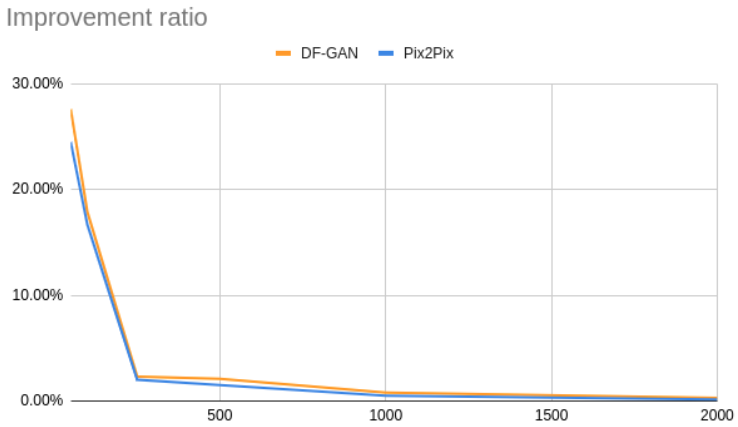

Fig. 4. Improvement ratio across various size of data

4 Conclusions

In this work, we proposed Distribution Fitting GAN(**DF-GAN**), a Pix2Pix based model to perform data augmentation for a chest X-ray pneumothorax segmentation task. We come up with this new loss **FID Loss** to capture the

Table 2. Quantitative results for segmentation (Mean ± standard error) (R denotes real images, and S denotes synthetic images)

	100 R	100 R + 100 S	100R +200 S
Dice	0.324±0.038	0.382±0.008	0.39±0.011

underlying distribution of real data and fit synthetic data's distribution into real data's distribution. We evaluate the effectiveness of our method with a segmentation network and show a significant improvement in the segmentation accuracy. Future work direction will be assuring the effectiveness of FID loss with non-Gaussian distribution and try to adjust the weights for each loss.

References

1. Abhishek, K., Hamarneh, G.: Mask2lesion: Mask-constrained adversarial skin lesion image synthesis. In: International Workshop on Simulation and Synthesis in Medical Imaging. pp. 71–80. Springer (2019)
2. Bailo, O., Ham, D., Min Shin, Y.: Red blood cell image generation for data augmentation using conditional generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
3. Barratt, S., Sharma, R.: A note on the inception score. arXiv preprint arXiv:1801.01973 (2018)
4. Bowles, C., Chen, L., Guerrero, R., Bentley, P., Gunn, R., Hammers, A., Dickie, D.A., Hernández, M.V., Wardlaw, J., Rueckert, D.: Gan augmentation: Augmenting training data using generative adversarial networks. arXiv preprint arXiv:1810.10863 (2018)
5. Dowson, D., Landau, B.: The fréchet distance between multivariate normal distributions. Journal of multivariate analysis **12**(3), 450–455 (1982)
6. Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. Neurocomputing **321**, 321–331 (2018)
7. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. arXiv preprint arXiv:1406.2661 (2014)
8. Han, C., Hayashi, H., Rundo, L., Araki, R., Shimoda, W., Muramatsu, S., Furukawa, Y., Mauri, G., Nakayama, H.: Gan-based synthetic brain mr image generation. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 734–738. IEEE (2018)
9. Han, C., Rundo, L., Araki, R., Nagano, Y., Furukawa, Y., Mauri, G., Nakayama, H., Hayashi, H.: Combining noise-to-image and image-to-image gans: Brain mr image augmentation for tumor detection. IEEE Access **7**, 156966–156977 (2019)
10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. arXiv preprint arXiv:1706.08500 (2017)

11. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European conference on computer vision (ECCV). pp. 172–189 (2018)
12. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
13. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
14. Madani, A., Moradi, M., Karargyris, A., Syeda-Mahmood, T.: Chest x-ray generation and data augmentation for cardiovascular abnormality classification. In: Medical Imaging 2018: Image Processing. vol. 10574, p. 105741M. International Society for Optics and Photonics (2018)
15. Malygina, T., Ericheva, E., Drokin, I.: Data augmentation with gan: Improving chest x-ray pathologies prediction on class-imbalanced cases. In: International Conference on Analysis of Images, Social Networks and Texts. pp. 321–334. Springer (2019)
16. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
18. Shin, H.C., Tenenholtz, N.A., Rogers, J.K., Schwarz, C.G., Senjem, M.L., Gunter, J.L., Andriole, K.P., Michalski, M.: Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In: International workshop on simulation and synthesis in medical imaging. pp. 1–11. Springer (2018)
19. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
20. Tang, Y.B., Tang, Y.X., Xiao, J., Summers, R.M.: Xlsor: A robust and accurate lung segmentor on chest x-rays using criss-cross attention and customized radiorealistic abnormalities generation. In: International Conference on Medical Imaging with Deep Learning. pp. 457–467. PMLR (2019)
21. Vaserstein, L.N.: Markov processes over denumerable products of spaces, describing large systems of automata. Problemy Peredachi Informatsii **5**(3), 64–72 (1969)
22. Xing, Y., Ge, Z., Zeng, R., Mahapatra, D., Seah, J., Law, M., Drummond, T.: Adversarial pulmonary pathology translation for pairwise chest x-ray data augmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 757–765. Springer (2019)
23. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support, pp. 3–11. Springer (2018)
24. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)