Fake Data Generation for Medical Image Augmentation using GANs

Donghwan Kim
Dept. Electronic Engineering
Pusan National University
Pusan, Korea
dongh@pusan.ac.kr

Jaehan Joo
Dept. Electronic Engineering
Pusan National University
Pusan, Korea
jhjoo2018@pusan.ac.kr

*Suk Chan Kim
Dept. Electronic Engineering
Pusan National University
Pusan, Korea
sckim@pusan.ac.kr

Abstract—This paper uses WGAN-GP to generate fake data that can be used as augmented data for strabismus classification and analyze the results. In the introduction of this paper, the general diagnostic technique for strabismus disease is described and the diagnostic technique using deep learning is described. And the reason for generating fake data is described. Main subject describes the WGAN-GP, data set used for data generation and evaluation metrics of GAN. In the experimental result, the data generated by the GAN is visually checked, and the performance of the fake data is evaluated with the FID that is one of the evaluation metrics of the GAN. And in the conclusion, evaluation of the proposed GAN and future work are described.

Index Terms—WGAN-GP, strabismus, diagnosis, deep lear

I. INTRODUCTION

Recently, deep learning has been widely used in many fi Among them, a diagnostic technique using deep learning b on medical images will present a new paradigm in dis diagnosis. Strabismus, which is treated in this study, i ophthalmic disease in which the two eyes are not aligned. the strabismus is a disease with a good prognosis and a cure rate when it is detected early and treated at an early However, because there is a risk of blindness when strabi: treatment is neglected at early age. So early detection o disease is very important.

Diagnosis of a disease usually involves two steps. The step in diagnosis is to determine the presence or absenc the disease, and the second step is to determine the sev of the disease. Strabismus is also diagnosed with the same procedure. Generally, diagnosis of strabismus is to determine by conduction a cover test. The cover test is a diagnostic technique that covers one of the two eyes and determines the movement of the remaining eye to determine strabismus. In patients with mild strabismus, MRI(Magnetic Resonance Imaging) is sometimes used to determine the presence or absence of strabismus. However, in the case of infants and young children, it may be uncooperative during the cover test and MRI scan. Therefore, deep learning-based diagnostic techniques can be effective when diagnosing uncooperative patients.

Prior to this study, in order to determine the presence or absence of strabismus, the first stage of diagnosis, a CNN-based classifier was designed to determine the presence or

absence of strabismus [1]. Finally, the accuracy of the test set was 66.7%. However, it cannot be said to be accurate because the number of data in the data set is imbalanced. Due to the characteristics of medical data, data imbalance may occur. Various techniques are used when using imbalanced data for deep learning. To overcome the limitations of imbalanced data, this research apply a data augmentation technique using WGAN-GP.
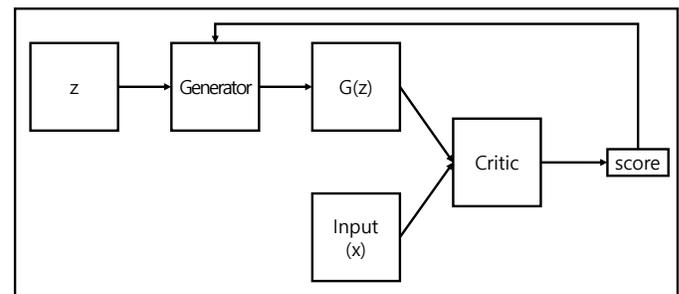
II. MAIN SUBJECT

Fig. 1. Architecture of WGAN-GP.

WGAN-GP(Wasserstein Generative Adversarial Networks - Gradient Penalty) is an upgraded version of GAN(Generative Adversarial Networks)'s loss [2]. And the WGAN-GP is more usable than the original GAN. Because WGAN-GP is stronger than the original GAN at the mode collapse problem which is the serious issue of the GAN. The GAN has two networks called a generator and a discriminator. The two networks minimize and maximize the same objective function respectively, and continue learning while maintaining and adversarial relationship [3]. As a result, the generator of the GAN that has finished learning generates fake data with a distribution similar to the distribution of the training set by using the noise vector as an input. WGAN-GP has the same architecture and operation principle as GAN. Its architecture is shown in Fig. 1. Also the task of the WGAN-GP is same as the GAN. It is generating fake data. As mentioned above, the objective function of WGAN-GP is an upgraded version of the objective function of GAN and is the same as Eq. (1)

$$L = \mathop{\mathbb{E}}_{\widetilde{x}\sim\mathbb{P}_g}[D(\widetilde{x})] - \mathop{\mathbb{E}}_{\widetilde{x}\sim\mathbb{P}_r}[D(x)]$$
$$+\lambda\mathop{\mathbb{E}}_{\widehat{x}\sim\mathbb{P}_{\widehat{x}}}[(\|\bigtriangledown_{\widehat{x}}D(\widehat{x})\|_2 - 1)^2] \quad (1)$$

WGAN-GP uses Eq. (1) as an objective function and limits the maximum value of the weight by giving a penalty to the gradient. By limiting the maximum value of the weight, it somewhat solves the instability of the training process and mode collapse, which are chronic issues of GAN, although not completely. And in the WGAN-based GAN, the discriminator is called a critic because it scores the input rather than calculating the probability that the input is real or fake [4].

The architecture of the original WGAN-GP is based on DC-GAN(Deep Convolutional Generative Adversarial Networks) [5]. The generator and critic of the model proposed in this paper have a RESNET(RE)architecture by adding a residual block while having a CNN-based architecture [6]. And the objective function uses the same Eq. (1) as that of the original WGAN-GP.

B. Dataset

The dataset used in this sutdy consists of a photogra which only the eye part of the patient's frontal photos The dataset consists of three classes: exotropia, esotropia normal eyes. The label for each photo is the result of cover test mentioned in the introduction. To validate data and the trained model, 10 Pusan National Univ Hospital ophthalmologists rediagnosed the patients with naked eyes and the photograph which has the correct al by cover test. As a result, the correct rate of 10 docto each data is defined as the Selection Rate(SR) in this p

TABLE I
THE NUMBER OF DATA EACH SELECTION RATE

		Class		
		Exotropia(XT)	Esotropia(ET)	Normal(NO)
SR	None	1250	451	999
	60	1175	412	959
	70	1067	367	886
	80	938	316	784

As shown in Table. I, the dataset used in this s reconstructed into a total of four datasets: a datas selection rates of 60, 70, 80, and 0. After that, each four datasets is trained with the proposed model me above.

C. Evaluation Metrics of GAN

The representative evaluation metrics of GA IS(Inception Score) and FID(Fréchet Inception Distan Both IS and FID use pre-trained weights of a CNN based network called Inception v3. And both metrics use the dis-tribution of the dataset. In particular, IS indicates how well the generated samples match the real data, and the higher the score, the better the quality of data. And FID indicates the distribution distance of the generated samples with the real

data based on KL Divergence. The closer the distance, that is, the lower the FID score, the better the data.

In the case of IS, since the diversity of the generated samples cannot be considered, even if mode collapse, which is a chronic issue of GAN, occurs, it tends to maintain a high value as long as the quality of fake data is good. On the other hand, FID considers the diversity of fake data and real data. Therefore, the generated samples in this study are evaluated by FID.

III. EXPERIMENTAL RESULTS

In this section, as in the main subject of the above, WGAN-GP of the RESNET architecture is trained on four datasets divided by selection rate. And the experimental results are discussed. 100 samples were generated for each dataset, and the top 10 and bottom 10 were selected visually. And we discuss the results by comparing the FID that compares the samples generated for each dataset with the distribution of the training set.
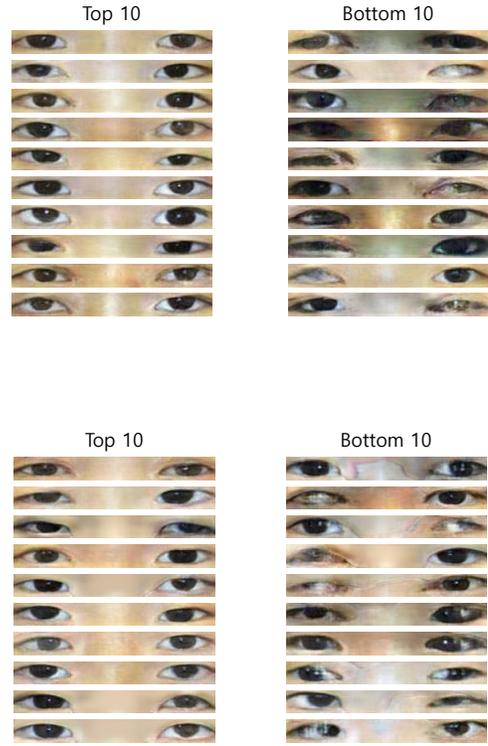

Fig. 3. Generated Samples with selection rate that is 60.

Fig. 2, Fig. 3, Fig. 4 and Fig. 5 are the top 10 and bottom 10 of the generated samples from the dataset composed by selection rate, respectively. Although more than half of the dataset are photos of strabismus patients, most of the top 10 do not show strabismus with the naked eyes. Also, there is no big difference in the image quality of the top 10 by selection rate. However, in the case of bottom 10, it can be seen that the color tends to be slightly better as the selection rate increases.
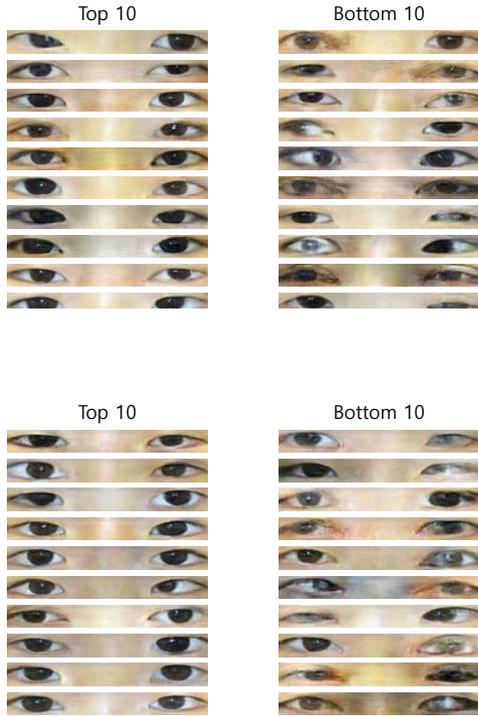
Top 10   Bottom 10

Top 10   Bottom 10

Fig. 5. Generated Samples with selection rate that is 80.

These generated samples are compared by the FIDs in the next section.

B. FID of Generated Samples

TABLE II
FIDs Each Selection Rate

	All(SR:0)	SR:60	SR:70	SR:80
All	2.185	2.297	1.971	1.630
Top 10	8.575	8.377	10.734	10.716
Bot 10	9.785	8.606	8.542	8.167

The FIDs shown in Table. II came out as a result. As expected, the FID tends to decrease as the selection rate increases. This is because the higher the selection rate, the higher the probability that the dataset has information about strabismus and better quality of samples. Also, in the case of bottom 10, as visually confirmed in Fig. 2, Fig. 3, Fig. 4 and Fig. 5, the higher the selection rate, the lower the FID. However, in the case of top 10, the higher the selection rate, the higher the FID tends to be. FID considers the diversity of samples as mentioned above. However, 10 photos are not enough to consider diversity. Therefore, an exceptional result would have been obtained.

IV. CONCLUSIONS & FUTURE WORKS

At the section of FID of Generated Samples, you can see the results by the FIDs. The larger SR, the smaller FID. By [8], Augmenting data at the only discriminator, it can lower FID. In this paper, augmenting data was used at the only discriminator. But you can see the FID score is good at high SR. So consisting useful data is important at the GAN also.

As can be seen from the experimental results, the generated samples by the proposed methods were confirmed. In both top 10 and bottom 10, the quality did not change significantly according to the selection rate. In this study, data augmentation using GAN was performed because of the lack of ET data. However, most of the generated samples are difficult to observe strabismus with the naked eyes. Therefore, in future work, data with strabismus will be generated based on Conditional GAN. In addition, a CNN-based classifier will be used to verify the augmentation effect of the generated samples.

REFERENCES

[1] Kim, D., Joo, J., Zhu, G., Seo, J., Ha, J., & Kim, S. C. (2021, April). Strabismus Classification using Convolutional Neural Networks. In 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC) (pp. 216-218). IEEE.

[2] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. (2017). Improved training of wasserstein gans. arXiv preprint arXiv:1704.00028.

[3] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.

[4] Arjovsky, M., Chintala, S., & Bottou, L. (2017, July). Wasserstein generative adversarial networks. In International conference on machine learning (pp. 214-223). PMLR.

[5] Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[7] Pan, Z., Yu, W., Yi, X., Khan, A., Yuan, F., & Zheng, Y. (2019). Recent progress on generative adversarial networks (GANs): A survey. IEEE Access, 7, 36322-36333.

[8] Zhao, Z., Zhang, Z., Chen, T., Singh, S., & Zhang, H. (2020). Image augmentations for GAN training. arXiv preprint arXiv:2006.02595.