Augmenting Seismic Data via Generative Adversarial Network for Low-cost MEMS Sensors

Aming Wu
Kyungpook National University
Email:wuaming@knu.ac.kr

Jae-Kwang Ahn
Korea Meteorlogical Administration
Email: propjk@korea.kr

Young-Woo Kwon
Kyungpook National University
Email: ywkwon@knu.ac.kr

***Abstract*—The deep learning (DL) model's performance is related to algorithm design and depends on sufficient training data set. The lack of real seismic data set and many data polluted by dif-ferent types of noise are the key factors that restrict the breakthrough in seismology research based on machine learning approaches. Therefore, data generation technic is fundamental for data set augmentation. Because it is difficult to perceive the spatial-temporal correlation and evolution process of seismic sequences, the intelligent generation of seismic sequences is still a significant challenge. A novel deep generation model (DGM) based on a generative adversarial network (GAN) is developed to generate high-quality seismic data. The experimental results show that the model can generate realistic seismic data through autocorrelation analy-sis, error quantitative index analysis, and other evaluation methods, and the model's accuracy is over 80%.**

I. Introduction

Due to the strong uncontrollability of seismic, it has only brought many losses and disasters to people in the past few decades, which drives more and more re-searchers to develop a rapid automatic seismic monitoring system [1], [2]. Many existing seismic data from MEMS sensors contain many noise data, which will hinder the research process, primarily based on the data processing technology deep learning of seismology. This paper introduces seismic data generation technology to expand seismic data sets to solve the problem of insufficient high-quality data sets in seismological research. The key to seismic data generation is feature extraction. The seismic sequence is a high-dimensional complex data structure, which contains many implicit features; it is difficult to accurately set appropriate indicators to represent the characteristics of seismic data accurately. Previous research [3], [4] proposed to use of machine learning (ML) algorithms to extract the implic-it features of earthquake waves for seismic detection. However, these methods are based on artificial label setting to extract features. One of the main advantages of the DL is that it can automatically extract the explicit and implict features of seismic sequences, avoiding the influence of the traditional human intervention on the experimental results. 3-component earthquake acceleration data is a series of discrete measures recorded at continuous time points of seismic evolution, including different dimesional spatial-temporal distribution patterns. Therefore, the design of the generation model needs to combine the characteristics of the evolution process of the seismic sequence. While ensuring the generation of high-quality seismic data, it is also essential to improve the diversity and stability of the data generated by the model. Based on the above goal, we developed a new depth generation model based on the GAN framework to capture the time evolution relationship of different ranges and realize the accurate and stable generation of seismic data. The results show that our model can generate a variety of high-quality seismic data through visual presentation, autocorrelation analysis, and accuracy. The rest of the paper is structured as follows. Section 2 compares relevant research approaches. Section 3 explains the algorithms' theory used in our model. While Section 4 evaluates the experimental results, Section 5 concludes this paper and future work.

II. Related Work

In previous seismological studies, Li et al. [5] used GAN to train the deep learning model to perceive the first arrival P wave features, the GAN structure aims to extract and recognize the features of seismic data, but it does not attempt to complete the generation task. While Li et al. [6] proved that conditional GAN effectively augmentated seismic data set. the EarthquakeGen [7] was developed to generate seismic data and verified its rationality. Although these methods can generate seismic data, they all need a complex preprocessing process and can not achieve the diversification of generated data. In the study of seismology, the diverse and long sequence waveform, including P-wave and S-wave, is also fundamental. Therefore, in this paper, based on the GAN framework, we integrate the new algorithm theory to design a deep learning model that can stably generate various seismic data.

III. Model Design

Different distribution patterns of time series lead to different construction methods of mapping space. To complete seismic data generation, we need to find the spatial distribution pattern $P\ (t, \theta)$ according to the discrete real seismic sequence T and solve the optimal parameter combination of continuous data space:

$$\theta^* = \arg\max_{\theta} \prod_{i=1}^{N} P\left(t^i, \theta\right) \tag{1}$$

Since the earthquake sequence x in the spatial dimension $m = \{1,\ 2, 3\}$ (m represents three different dimensions of

east-west, north-south, and up-down) and time series $N = \{1, 2, ..., n\}$, θ is the parameter space that satisfies the mapping relation, the maximum likelihood function Eq (2) is used to solve the optimal parameter combination:

$$\theta^* = \arg\max_{\theta} \prod_{n=1}^{N} \prod_{m=1}^{3} P\left(t_m^N \mid t_m^{1\to n-1}, \theta_f\right) \cdot P\left(y_m^n \mid t_m^n, \theta_g\right) \tag{2}$$

Due to the complex evolution process of earthquake sequence, the vibration trajectory changes at different times. To use the deep learning model to extract data features based on Eq (2) automatically, we use the GAN framework to fit the distribution of real data to build the generation model. In 2014, Goodfellow [8] proposed the generative antagonism network (GAN), it was an epochmaking unsupervised learning algorithm framework (as shown in Fig.3). Its basic idea is derived from the 0-1 game theory, and it is mainly composed of a generator and a discriminator. The generator is designed to generate new data samples; the discriminator is essentially a binary classifier, which can accurately distinguish the true and false data, and both need to be continuously optimized, and its optimal objective function can be expressed as:

$$\min_{G}\max_{D} \mathrm{V(G,D)} = E_{x\sim p_t(x)}(\log(D(x))) + E_{z\sim p_g(z)}(\log(1 - D(G(z)))) \tag{3}$$

Where$(*)$ represents the generator,$D(*)$ represents the discriminator, then $_{g(z)}$ and $p_{r(x)}$ denote the distribution of the real data and generated data, respectively.

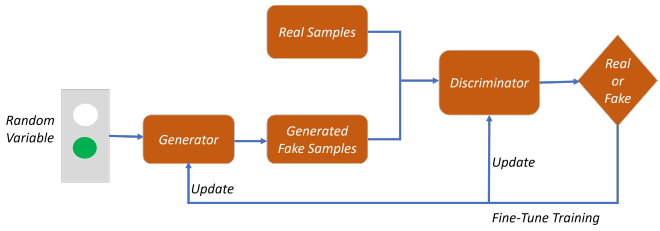

Fig. 1. System structure of different generation models

To expand the scope of seismic sequence analysis, we use LSTM to extract the sequence features of different time ranges. Simultaneously, considering that the seismic sequence results from the interaction of different spatial dimensions, we use an attention mechanism to capture the sequence features of different dimensions and use a neural network (NN) to perceive the local features of the seismic sequence. It is worth mentioning that the performance of the GAN framework depends on the design of the adversarial mechanism. In order to ensure the stability of the model, Wasserstein distance (WD) is introduced to measure the difference between generated data and real data [9].

IV. Result Evaluation

A. Visual Appearance

The detection of seismic events depends on the appearance of seismic waves, and amplitude and frequency are the most apparent characteristics of seismic waves. Therefore, they are also critical indicators to evaluate the authenticity of generated data. Fig.2 shows part of the seismic waveform generated by our model. The arrival time of the P-wave and S-wave is very similar to the real data. More importantly, the amplitude of the S-wave is more intense in the X and Y channels. Besides, in real seismic events, there are always small earthquake fluctuations such as foreshocks or aftershocks. Although it is not the critical factor of seismic detection, it can be used as an index to evaluate the diversity of generated data. Our model can generate high-quality seismic data with different amplitudes and waveforms.

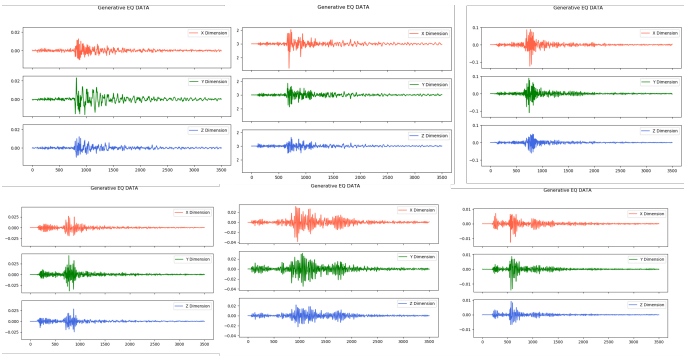

Fig. 2. System structure of different generation models

B. Autocorrelation Analysis

To further verify the data's quality, we use a scatter matrix to analyze the autocorrelation distribution. Kernel density estimation is used to observe the distribution of each channel variable. The horizontal axis corresponds to the variable's value, and the vertical axis corresponds to the density of the variable. Scatter plot: an associated scatter plot of the distribution between two variables. Pairing any two variables, one of which is the abscissa and the other is the ordinate, is used to measure the correlation of the two variables. The kernel density estimation map clearly shows that the real data and the generated data show the approximate Gaussian distribution pattern on each channel, which is significantly different from the distribution of non-earthquake data; in the scatter map, the data of any two channels X, Y, and Z are paired, and the data points are unevenly distributed and relatively scattered due to the difference of P-wave and S-wave vibration amplitude. There is a weak correlation between the two channels. The correlation between seismic data and real data is similar, but the distribution of non-earthquake data points is concentrated and even. Therefore, the scatter matrix analysis shows that our model can generate more realistic seismic data.

C. Accuracy Analysis

Although above evaluation approaches have verified the potential of our model in earthquake sequence generation, they only qualitatively reveal the quality of individually generated data, which is also a common weakness in many machine learning models. In this research, in order to further clarify

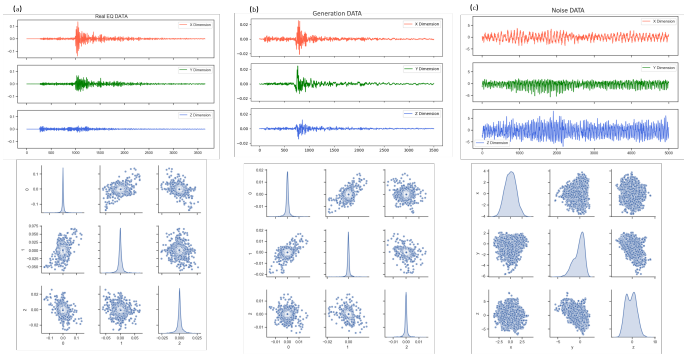

Fig. 3. Scatter matrix. (a) shows the distribution of real earthquake data, (b) denotes the generation data auto-correlation distribution, (c) is the autocorrelation distribution of noise data

TABLE I
ACCURACY IN DIFFERENT ERROR INDEXES

Error index	MSE	MAPE	WD
Accuracy	81%	87%	84%

the accuracy of our model, we use the mean squared error (MSE), the mean absolute percentage error (MAPE), and WD to measure the similarity between the generated data and the real data and analyze the accuracy of the generated model. The results denote that our model's accuracy can achieve more than 80% (Table 1).

V. Conclusion

In this paper, based on the GAN framework, we propose a new DGM, generating realistic seismic data. To test its reliability, we verify it from visual presentation and correlation. The results show that the generated data and real data have high similarity. Also, this paper only uses the qualitative evaluation scheme for analysis. In future research, we will try to combine the quantitative evaluation index to evaluate the generated data further. Simultaneously, we will also apply the generated data to the existing seismic detection scheme for verification to provide reliable data support for seismic research.

VI. Acknowlegment

This work was supported by the development of earthquake, tsunami, volcano monitoring and prediction technology (KMA-135002988)

References

[1] M. Böse, E. Hauksson, K. Solanki, H. Kanamori, Y.-M. Wu, and T. Heaton, "A new trigger criterion for improved real-time performance of onsite earthquake early warning in southern california," *Bulletin of the Seismological Society of America*, vol. 99, no. 2A, pp. 897–905, 2009.

[2] C. E. Yoon, O. O'Reilly, K. J. Bergen, and G. C. Beroza, "Earthquake detection through computationally efficient similarity search," *Science advances*, vol. 1, no. 11, p. e1501057, 2015.

[3] Q. Kong, R. M. Allen, L. Schreier, and Y.-W. Kwon, "Myshake: A smartphone seismic network for earthquake early warning and beyond," *Science advances*, vol. 2, no. 2, p. e1501055, 2016.

[4] I. Khan, S. Choi, and Y.-W. Kwon, "Earthquake detection in a static and dynamic environment using supervised machine learning and a novel feature extraction method," *Sensors*, vol. 20, no. 3, p. 800, 2020.

[5] Z. Li, M.-A. Meier, E. Hauksson, Z. Zhan, and J. Andrews, "Machine learning seismic wave discrimination: Application to earthquake early warning," *Geophysical Research Letters*, vol. 45, no. 10, pp. 4773–4779, 2018.

[6] Y. Li, B. Ku, S. Zhang, J.-K. Ahn, and H. Ko, "Seismic data augmentation based on conditional generative adversarial networks," *Sensors*, vol. 20, no. 23, p. 6850, 2020.

[7] T. Wang, Z. Zhang, and Y. Li, "Earthquakegen: Earthquake generator using generative adversarial networks," in *SEG Technical Program Expanded Abstracts 2019*, pp. 2674–2678, Society of Exploration Geophysicists, 2019.

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv preprint arXiv:1406.2661*, 2014.

[9] S. Vallender, "Calculation of the wasserstein distance between probability distributions on the line," *Theory of Probability & Its Applications*, vol. 18, no. 4, pp. 784–786, 1974.