Iranian Journal of Geophysics, Vol 19, No 2, 2025, P. 6 (Research article)
DOI: 10.30499/IJG.2024.450092.1589

Reduction of the data required for training deep learning models based on clustering of the data and its application in one-dimensional magnetotelluric inversion

Mehdi Rahmani Jevinani[1], Banafsheh Habibian Dehkordi[2*]

[1] *Ph.D student, Institute of Geophysics, University of Tehrn, Tehrn, Iran*

[2] *Assistant Professor, Institute of Geophysics, University of Tehran, Tehran, Iran*

(Received: 27 March 2024, Accepted: 25 September 2024)

Summary

Data-driven deep learning approaches have to deal with the challenge of generating large amounts of high-quality data, as well as the heavy computational cost and long training time imposed by it. Due to their ability to approximate complex nonlinear mapping functions, deep networks can be used effectively in geophysical inverse problems and better generalization can be achieved through deeper networks in many applications. In this research, an approach based on primary clustering of training data and assigning a certain percentage of each cluster to training, validation and test data has been used for data splitting. Kolmogorov Smirnov (KS) test has been applied to compare the distribution of three sets that are divided in this manner, and indicates that the training, validation and test data have the same distribution. A deep learning model based on modified U-Net architecture has been trained for one-dimensional inversion of magnetotelluric (MT) data, which is a highly non-linear regression problem. Supervised learning and back propagation error are used, and therefore, the inputs along with the corresponding outputs are given to the network in the form of training samples. For this purpose, a five-layer geoelectric model has been considered to simulate the conditions of a geothermal field. Using magnetotelluric forward modeling algorithm, the responses of this one-dimensional geoelectric model are analytically calculated in the frequency range of 0.01-100 Hz and in 13 frequencies that are uniformly distributed on a logarithmic scale, and total of 500000 sample data were generated. The thickness of the layers is variable and considered as part of the output. Pre-processing is done to scale the input and output variables before training and the network outputs are post-processed to be returned to the original interval. The mean square error (MSE) loss function and the Adam optimizer were used to train the network. Training is accomplished with a different amount of data separated by the mentioned method, and network performance is evaluated with some quantitative and qualitative criteria, including boxplots of Euclidean distance between true and predicted outputs and Nash Sutcliffe Efficiency coefficients. The trained network predicts the electrical resistivity and thickness of the layers from the new set of phase and apparent resistivity values. The results show that data splitting in this manner reduces the number of training data required to train the deep learning model by at least 50% without reducing the accuracy of the trained network. For noisy data and in more real scenarios, random separation is definitely not a suitable approach to form training, validation and test sets. In these conditions, the use of clustering is a suitable solution for equalizing the statistical distribution of the three sets and reducing the number of required data.

Keywords: Clustering, Deep learning, Inversion, Magnetotelluric.

[*]Corresponding author: bhabibian@ut.ac.ir