WaveformMAE: Self-Supervised Learning for Fault-Inversion-Free Reconstruction of Ground Motion Waveforms Across the Entire Area

S31F-0409

Borui Kang[1], Peng Wu[1], Chen Gu[1], Yichen Zhong[1]
1 Department of Civil Engineering, Tsinghua University, Beijing, China

Background and Inspiration

In recent years, deep learning methods have been increasingly used in seismic wave detection and seismic inversion. However, achieving accurate seismic waveforms for the entire region depends on the inversion of source mechanisms and subsurface velocity models. In the field of computer vision, the Masked Autoencoder [1] (MAE) has been introduced to provide improved pre-trained models. Its pre-training task involves image reconstruction, also referred to as compressive sensing. I propose that obtaining seismic waves for the entire area is analogous to an image reconstruction process. Through skillful design, the framework of MAE can be directly applied. In theory, this approach may enable the acquisition of more seismic information without the need for inversion, thereby supporting research in other seismological issues or seismic design in civil engineering.

ViT: Vision Transformer

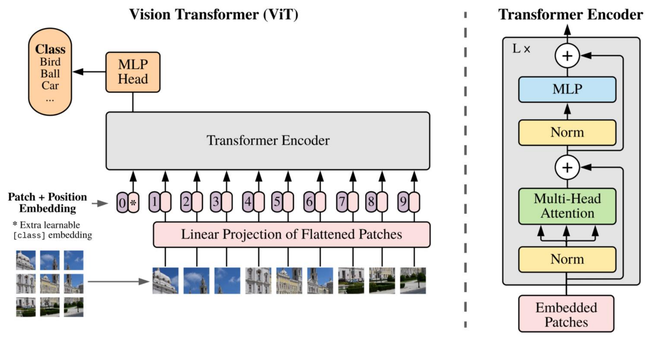


Figure 1.[2] The framework of Vision Transformer. A great deal of Computer Vision models or works have been based on ViT since its birth, including Masked Autoencoder.

Masked Autoencoder

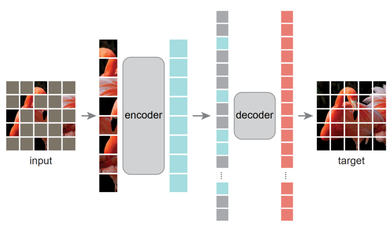


Figure 2.[1] The framework of Masked Autoencoder

The MAE approach was originally proposed for pre-training, and the pre-training process itself is an interesting and challenging task. By masking a large portion of patches (more than 75%) in an image and using a larger encoder and a smaller decoder to reconstruct the original image, remarkable results have been achieved. This pre-training strategy also proves to be beneficial for various visual tasks.

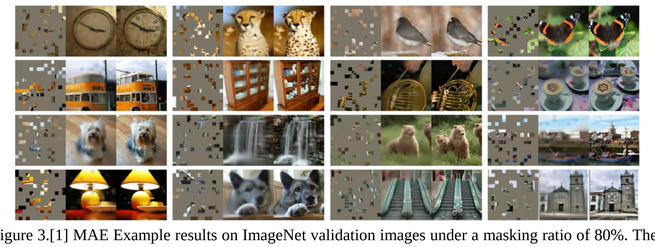


Figure 3.[1] MAE Example results on ImageNet validation images under a masking ratio of 80%. The masked image is on the left. MAE reconstruction is in the middle and the ground-truth is on the right.

Abstract

Earthquakes, especially large ones, usually pose immense hazards to human life and property safety. Understanding seismic waveforms is crucial for effective earthquake-resistant structural design. However, traditional methods for obtaining seismic waveforms at various locations require two steps, the inversion of fault conditions and forward modeling.

In this study, we propose an innovative approach with self-supervised learning and the Masked Autoencoder (MAE) as the backbone, drawing inspiration from the field of image reconstruction. In this method, ground motion waveforms all around can be reconstructed within only one step. Seismic data are preprocessed the by transforming three-channel waveforms with 256 samples (equivalent to approximately 128 seconds of data at a frequency of 2 Hz) into one-dimensional word vectors. This allows us to apply Transformer encoding.

To construct the training dataset, we manually calculate the ground motion waveform spatial distribution based on publicly available source model datasets of hundreds of earthquakes, convolved with pre-calculated Green's functions at specific area.

During the training phase, we randomly mask a significant portion (75%-90%) of the station information. The masked data is then passed through an encoder and decoder within the MAE framework, enabling us to reconstruct the seismic waveforms within a specific spatial range. For inference, when reconstructing longer seismic waveforms, we can divide them into segments and perform weighted averaging on overlapping sections to enhance the accuracy of the predictions.

This approach shows promising potential for predicting seismic waveforms across the entire spatial domain without the need for traditional fault inversion procedures. By leveraging self-supervised learning and the MAE framework, we can overcome the limitations imposed by sparse station coverage, ultimately enhancing our understanding of seismic events and improving earthquake risk mitigation strategies.

Data and Methods

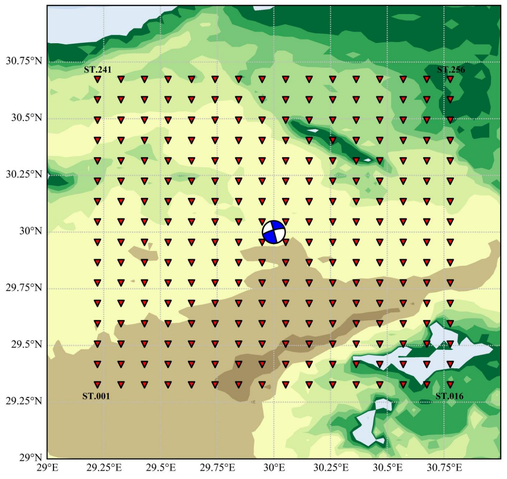


Figure 4. An example of synthetic data distribution. The center is located at 30°N, 30°E, with a spacing of 10 km.

As depicted in Figure 4, the 16*16 grid conveniently forms 16*16 patches, with each point representing a three-dimensional seismic wave sequence of 128s at 2Hz (128*2*3). This configuration is equivalent to the number of three-channel pixels in a 16*16 small patch (16*16*3), making the seismic waves across the entire region suitable for constituting an image in the framework of Masked Autoencoder (MAE).

We utilized the Colosseo earthquake scenario generator from the Pyrocko software package to generate data for preliminary experiments. The calculations were facilitated using an existing Green's function library under the crust2_2hz_dl one-dimensional velocity model. We created a station grid with a center at 30°N, 30°E, with a 10 km spacing. The source mechanism solutions were randomly sampled, and 1000 forward simulations were performed to train a reconstruction model under the local velocity model.

In the subsequent phases of our work, to fully harness the potential of the transformer model, it is imperative to generate a more extensive dataset. This involves producing data based on different velocity models at various locations, enabling the training of a generalized seismic wave reconstruction model.

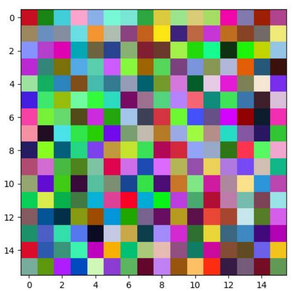


Figure 5. A seismic waveform at a station is equivalent to a patch in a 16*16 image.

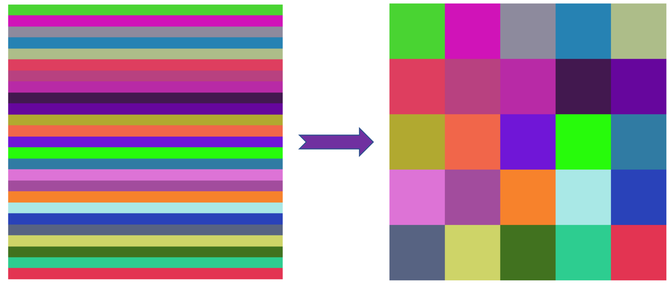
The seismic wave array information for the entire region can be compressed into a single image, divided into 16*16 patches. Each patch is represented by a color block indicating the placement of seismic wave sequences received by a station.

Efficient Implementation

We can implement WaveformMAE based on the source code of MAE and the generated seismic wave data. Of course, a slight refactoring of the network is also feasible.

Initially, we organize the generated seismic wave data as a [3,16,16,16,16] numpy array. However, this does not place the waveform of each seismic station into a small patch. Simply put, we reshape the original data into a [3,16,16,16,16] array or tensor, swapping the second 16 in the time series with the first 16 in the spatial sequence (the third and fourth dimensions). Then, we restore the reshaped array or tensor to the shape of [3,256,256]. We have provided an illustration based on a 5*5 grid to explain the efficient implementation.

During the training process, the well-organized [3,256,256] data can be directly treated as an image and loaded into the MAE framework for self-supervised learning. In the inference process, sparse data is manipulated using the same dimensional exchange and inference workflow. After the inference is complete, the data is restored to the original seismic wave data organization through the same dimensional exchange. This process results in the reconstruction of seismic wave data for the entire region.



(a) The initial waveform tensor (b) The transformed waveform tensor

Figure 6. A illustration for data transformation. In these pictures, seismic waveform at a station is equivalent to a patch in a 5*5 image. In real training, the grid is 16*16.

Expected Results

In the short term, we aim to train the WaveformMAE model to reconstruct seismic waves for a specific regional area. In the long term, our goal is to feed a large number of forward simulation results based on velocity models from various locations globally. This will enable WaveformMAE to learn the ability to reconstruct seismic waves for the entire region based on any sparse station configuration. This approach eliminates the need for inversion, streamlining the seismic modeling process effectively.

References

[1] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16000-16009).
[2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.