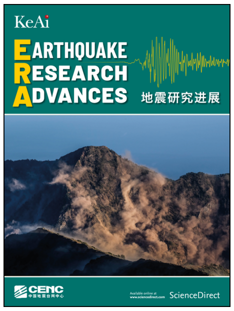


# TransQuake: A Transformer-based Deep Learning Approach for Seismic P-wave Detection

Yumeng Hu, Qi Zhang, Wenjia Zhao, Haitao Wang

PII: S2772-4670(21)00004-X

DOI: https://doi.org/10.1016/j.eqrea.2021.100004

Reference: EQREA 100004

To appear in: *Earthquake Research Advances*

# TransQuake: A Transformer-based Deep Learning Approach for Seismic P-wave Detection

HU Yumeng[3], ZHANG Qi[4], ZHAO Wenjia[2] and WANG Haitao[1]

1) China Earthquake Networks Center, Beijing 100045, China
2) Institute of Geology, China Earthquake Administration, Beijing 100029, China
3) Lanzhou Institute of Seismology, China Earthquake Administration, Lanzhou 730030, China
4) University of Science and Technology of China, Hefei 230026, China

**Recent years have witnessed the increasing research interest in developing machine learning, especially deep learning which provides approaches for enhancing the performance of microearthquake detection. While considerable research efforts have been made in this direction, most of the state-of-the-art solutions are based on Convolutional Neural Network (CNN) structure, due to its remarkable capability of modeling local and static features. Indeed, the globally dynamic characteristics contained within time series data (i.e., seismic waves), which cannot be fully captured by CNN-based models, have been largely ignored in previous studies. In this paper, we propose a novel deep learning approach, TransQuake, for seismic P-wave detection. The approach is based on the most advanced sequential model, namely Transformer. To be specific, TransQuake can exploit the STA/LTA algorithm for adapting the three-component structure of seismic waves as input, and take advantage of the multi-head attention mechanism for conducting explainable model learning. Extensive evaluations of the aftershocks following the 2008 Wenchuan $M_W$7.9 earthquake clearly demonstrates that TransQuake is able to achieve the best detection performance which excels the results obtained using other baselines. Meanwhile, experimental results also validate the interpretability of the results obtained by TransQuake, such as the attention distribution of seismic waves in different positions, and the analysis of the optimal relationship between coda wave and P-wave for noise identification.**

**Key words: Transformer; Deep learning; Seismic P-wave detection**

## INTRODUCTION

With the rapid development of seismic monitoring technology, more and more attention has been paid to the efficient detection and differentiation of microearthquakes from massive noise data, such as the intensive aftershock sequences following a destructive earthquake. Indeed, a variety of methods, such as template matching (Frank W. B. et al., 2014; Gibbons S. J. et al., 2007), similarity searching

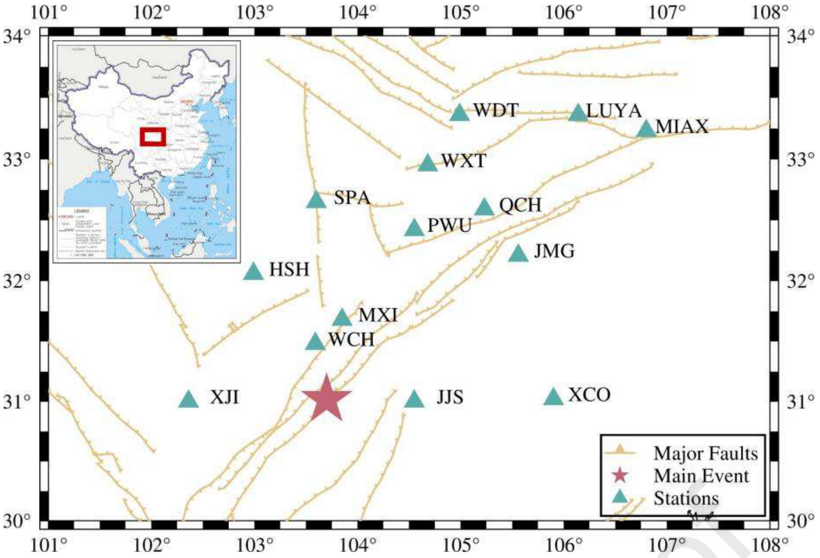


**Fig. 1** The geographical location of the study region.

(Yoon C. E. et al., 2015), and higher-order statistics (Saragiotis C. D. et al., 2000), have been developed in previous studies for addressing this challenge. Nevertheless, the most widely used technologies for seismic phase detection are developed based on the STA/LTA algorithm and/or its variants, which identify earthquake events when the ratio between the Short-Term Average (STA) and Long-Term Average (LTA) of energy function for seismic waves exceeds the pre-defined threshold (Allen R., 1982; Guo Tielong et al., 2019; Withers M. et al., 1998). However, the identification performance of all these traditional approaches is usually limited, either due to the trade-off effect between false and missing alarms or due to the that between the computational cost and time sensitivity (Liu Han et al., 2014; Reichstein M. et al., 2019).

Recently, an increasing number of researchers have focusedon developing machine learning, especially deep learning, which provides approaches for enhancing the performance of microearthquake detection. While considerable research efforts have been made in this domain, most of the state-of-the-art solutions are based on Convolutional Neural Network (CNN) structure, due to its remarkable capability of modeling local and static features (Rouet-Leduc B. et al., 2020; Zhang Xiong. et al., 2020). For example, Perol T. et al., (2018) propose an advanced CNN-based approach, CovNetquake, for earthquake detection and location. Based on CovNetquake, (Zhu Lijun. et al., 2019) propose a new approach CPIC for identifying and picking the seismic phase arrivals. However, since CNN is originally designed for pixel-based image processing rather than sequential data modeling, the global and dynamic characteristics of seismic waves cannot be fully captured by CNN-based models, which has been largely ignored in previous studies.

To address the above challenges, in this paper, we propose a novel deep learning approach, TransQuake, for seismic P-wave detection based on the most advanced sequential model in natural language processing (NLP), namely Transformer (Vaswani A. et al., 2017). To be specific, TransQuake can exploit the STA/LTA algorithm for adapting the three-component structure of seismic waves as input, and

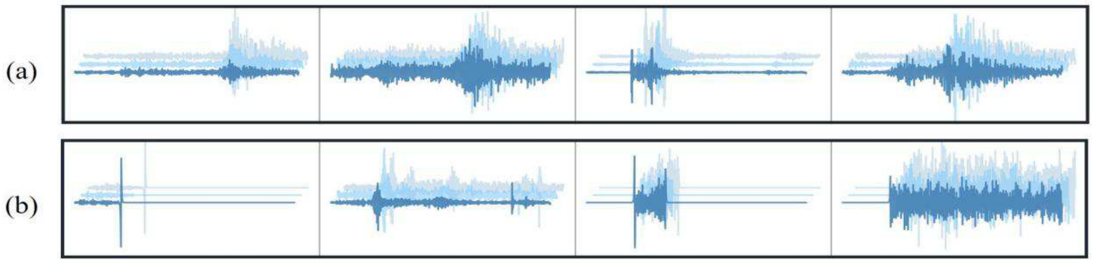


**Fig. 2** Some examples of labeled waveforms in the dataset.
(a) Positive examples; (b) Negative examples.

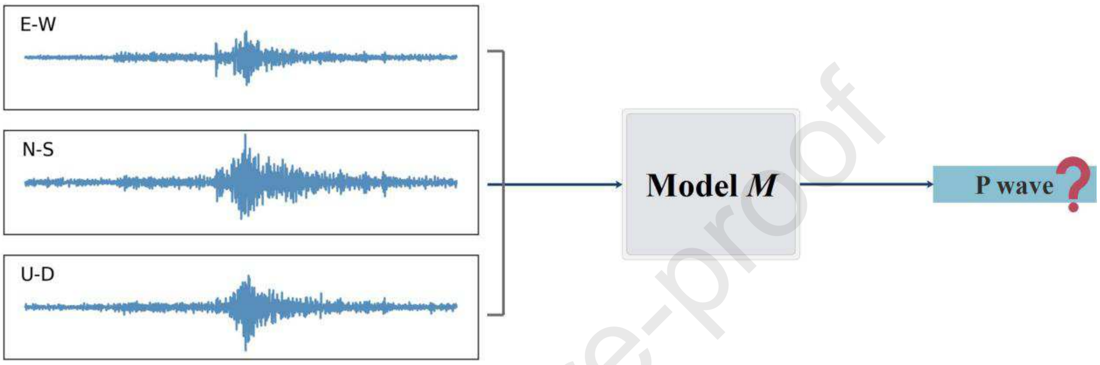


**Fig. 3** The diagram of our problem formulation.

take advantage of the multi-head attention mechanism for providing explainable model learning. Extensive evaluations of the aftershocks following the 2008 Wenchuan $M_W$7.9 earthquake clearly demonstrates that TransQuake can achieve the best detection performance which excels the results obtained using other baselines. Meanwhile, experimental results also validate the interpretability of the results obtained by TransQuake, such as the attention distribution of seismic waves in different positions, and the analysis of the optimal relationship between coda wave and P wave for noise identification.

## 1 DATA AND METHOD

### *1.1 Data*

The dataset used in this paper is the records of 100 Hz signals from the aftershocks following the 2008 Wenchuan $M_W$7.9 earthquake, which were recorded by 14 permanent earthquake stations (Fang Lihua, 2017) from July 1st to 31st. The distribution of stations and the main event are shown in Fig. 1. Each record of these stations contains 3 dimensions (3-D): vertical component, north-south component, and east-west component. Followed by the settings in (Zhang Qi et al., 2019), we filter the waves using the Bessel filter with bandwidth 2-10 Hz, and generate the negative samples by the *FilterPicker* (Lomax A. et al., 2012). Considering the different epicentral distances, we are unable to set a fixed time window that only contains a full P-wave, thus we set the time-window as 50 s to comprehensively analyze all the included waveforms. After data preprocessing, we have obtained 109 719 noise waveforms (negative samples), 9 891 aftershock waveforms (positive samples), and the shape of each waveform is structured by $1 \times 3 \times 5\,000$. Fig. 2 shows some examples of two types of waveforms. Note that, instead of directly generating the noise waveforms by background noises, here we select negative samples from those misclassified ones as aftershocks after passing through the filter. Inspired by the Adversarial Machine Learning (Goodfellow et al., 2015), which is based on the Nash Equilibrium of Game Theory, we creatively generate these misclassified waveforms as adversarial samples. In this way, our model is able to capture the critical points of earthquake detection and subsequently extract vitally inherent features of seismic waveforms. Such process overcomes the weakness of traditional detection methods such as overfitting, and provides us a more robust model (Szegedy et al., 2014).

**Table 1** Some important mathematical notations.

| Symbol | Description |
|---|---|
| $d$ | The dimension of hidden units. |
| $D$ | The dataset of waveforms. |
| $k$ | The first dimension of a wave after reshaping, representing the total position numbers. |
| $l_i$ | The label of $i$-th waveform, $l_i = \{1,0\}$ represents an earthquake. |
| $N$ | Batch size. |
| $r_i$ | The model's classification result of $i$-th wave, $r_i$ =1 represents an earthquake. |
| $S$ | The number of encoder stacks. |
| $T$ | The number of units to be reshaped in a single original channel for each new position. |
| $w_i$ | The $i$-th waveform. |

*1.2 Problem Formulation*

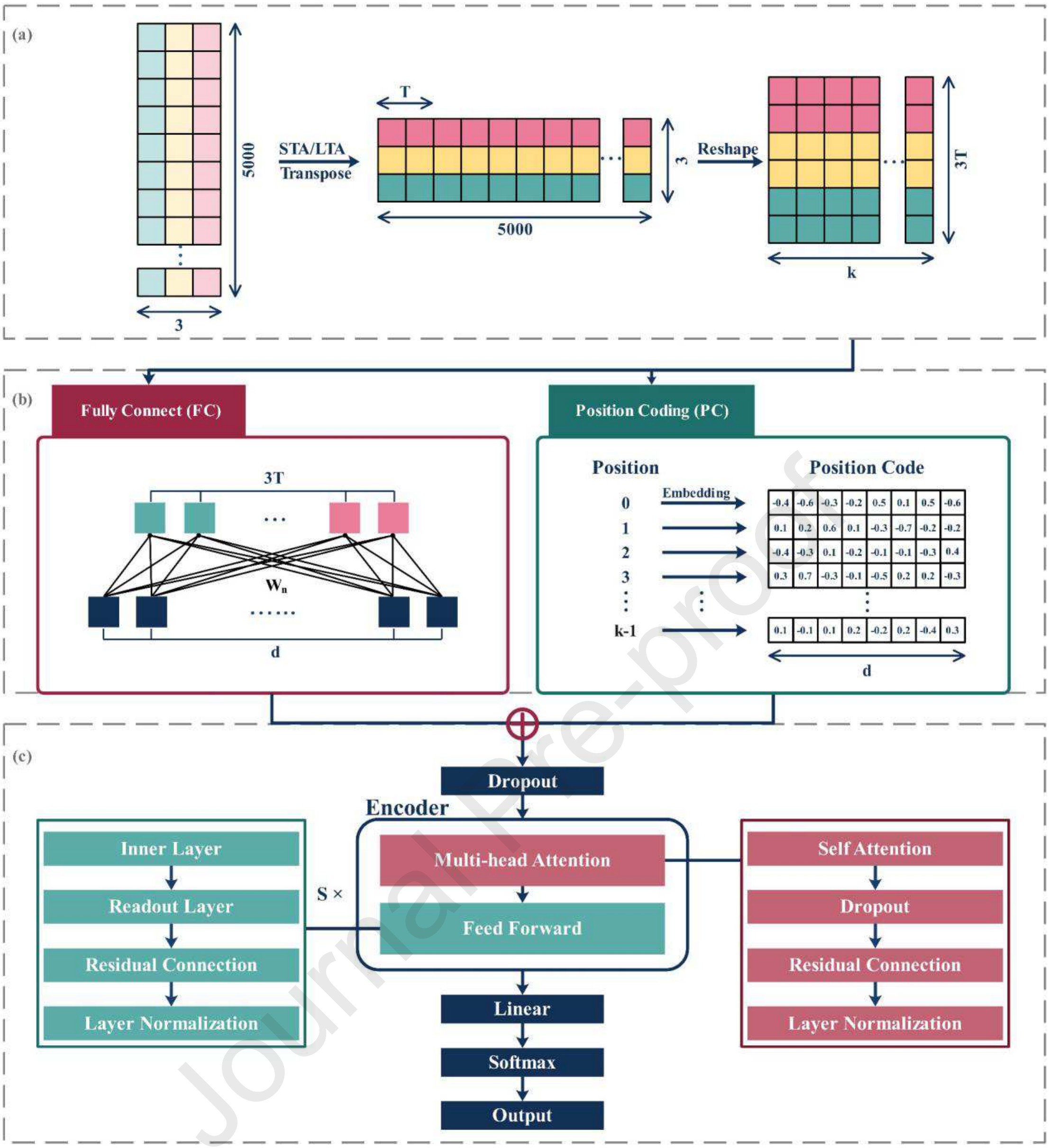
Taking 3-D waveforms in a 50 s window as the original input, our model aims to learn and output the possibility of the results of each window: whether it contains an earthquake P-wave arrival or not. Fig. 3 shows the diagram of our problem formulation, where U-D orientates vertically, N-S and E-W orientate horizontally, standing for the north-south channel and the east-west channel respectively. To formulate the problem, we use $D$ to represent a training dataset which contains $i$ waveforms, and three components of each waveform correspond to our three channels. Formally, the training dataset is expressed as:

$$D = \left\{w_1, w_2, \dots\ , w_i\right\}, \tag{1}$$

$$w_i = \{channel_1, channel_2, channel_3\}, \tag{2}$$

where $w_i$ represents a waveform and $channel_i$ means a single signal. Each $w_i$ corresponds to a label $l_i$, where $l_i = \{0,1\}$ represents a noise while $l_i = \{1,0\}$ represents an earthquake event containing P-wave arrival. Finally, we can define the problem as follows: The objective is to learn a binary-classification model $M$ from the training dataset D, which can classify a new waveform window $w_i$ through output $r_i$, where $r_i = 1$ represents an earthquake event with P-wave and $r_i = 0$ otherwise (??).

### *1.3 Model Architecture*



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -0.4 | -0.6 | -0.3 | -0.2 | 0.5 | 0.1 | 0.5 | -0.6 |
| 0.1 | 0.2 | 0.6 | 0.1 | -0.3 | -0.7 | -0.2 | -0.2 |
| -0.4 | -0.3 | 0.1 | -0.2 | -0.1 | -0.1 | -0.3 | 0.4 |
| 0.3 | 0.7 | -0.3 | -0.1 | -0.5 | 0.2 | 0.2 | -0.3 |
| 0.1 | -0.1 | 0.1 | 0.2 | -0.2 | 0.2 | -0.4 | 0.3 |

**Fig. 4** The architecture of our model, TransQuake. (a) Data preprocessing; (b) Fully connected and position code layer; (c) The full architecture of our model.

Our model TransQuake is based on the most advanced sequential model Transformer in the NLP field. The original Transformer contains two parts, namely encoders and decoders (Vaswani A. et al., 2017). Since the decoder part is designed for generative tasks rather than discriminative tasks, we only use the encoder part to establish our model. The full architecture of our model is shown in Fig. 4. Different from the original Transformer, to adapt the continuous waveform data rather than independent words as the input, we first propose a new approach to obtain the input of encoders (as shown in Fig. 4a and b). Then, with the adapted encoders of the Transformer, we further exploit a linear layer and a softmax layer to obtain the classification results, i.e., whether the input data contain a seismic P-wave or not. The related mathematical notations are summarized in Table 1.

### 1.3.1 Input

Learning from the earthquake real-time monitoring system of the National Digital Seismic Network, we extract the 150 time-steps STA and the 2,000 time-steps LTA of the waveforms (Guo Tielong et al., 2019). Using STA/LTA, the data are able to generate better waveform features and model performance. It indicates that machine learning methods combining traditional physical models can complement each other and achieve better performance (Reichstein M. et al., 2019). Next, as the Transformer is proposed for NLP, its original input is a batch of words, while our input is 3D continuous-time data. To adapt to this difference, we propose a module to make corresponding changes in input processing. In NLP tasks, each word is embedded into a vector whose size equals the hidden units. Then the position code (PC) is added. As shown in Fig. 4, we reshape the waveform and follow a fully connected (FC) layer to explore the relationship between each position of vectors, and then add the PC to get the input of encoders. The processing details are as follows. Considering the combination of 3 components in each position and the suitable size of the attention model, we reshape each waveform from $3 \times 5000$ into $k \times 3T$, where $k = (3 \times 5000)/3T$. Then we parallelly feed it to FC and PC layers. In the FC layer, each neuron cell connects all new cells as:

$$y_m = \sum_{n=1}^{3T} W_n x_n \ , \tag{3}$$

where $y_m$ is the value of a new cell with $m \in \{0,1,\dots,d\}$ and $d$ denotes the dimension of hidden units; $W_n$ is the wight that should be trained; $x_n$ is the value of the original cell. After the FC layer, the last dimension of data equals the hidden units. In the PC layer, we embed each position number of reshaped data with the Xavier initializer to a vector whose length also equals hidden units (Glorot X. et al., 2010). As FC and PC are the same shapes, we are able to add two matrices corresponding elements together and take the new matrix as our input.

### 1.3.2 Encoders

Before feeding the new input into the encoders, we use a dropout layer. The dropout layer randomly drops out a certain percentage ($P$) of model neurons in training, which effectively avoids the model from relying on certain neurons and enhances the generalization ability. After all preparations, data are fed into the encoders of our model. The detailed process of the first encoder is shown in Fig. 5. Using the layer-normalization after each residual connection layer and working together with feed-forward layers, it prevents the extreme situation, balances the data, and avoid model deterioration. However, compared with CNN, the Transformer model fully relies on the attention mechanism, which is the real workhorse of our model and can model the sequential information well. Specifically, the mission of the self-attention layers is to find out the dependency between different positions in the input that can better encode the information on the current position. For example, in the NLP field, the word “he” in the sentence “The boy didn’t go to school, because he felt sick.” may pay the highest attention to the word “boy”. In this task, the P-wave

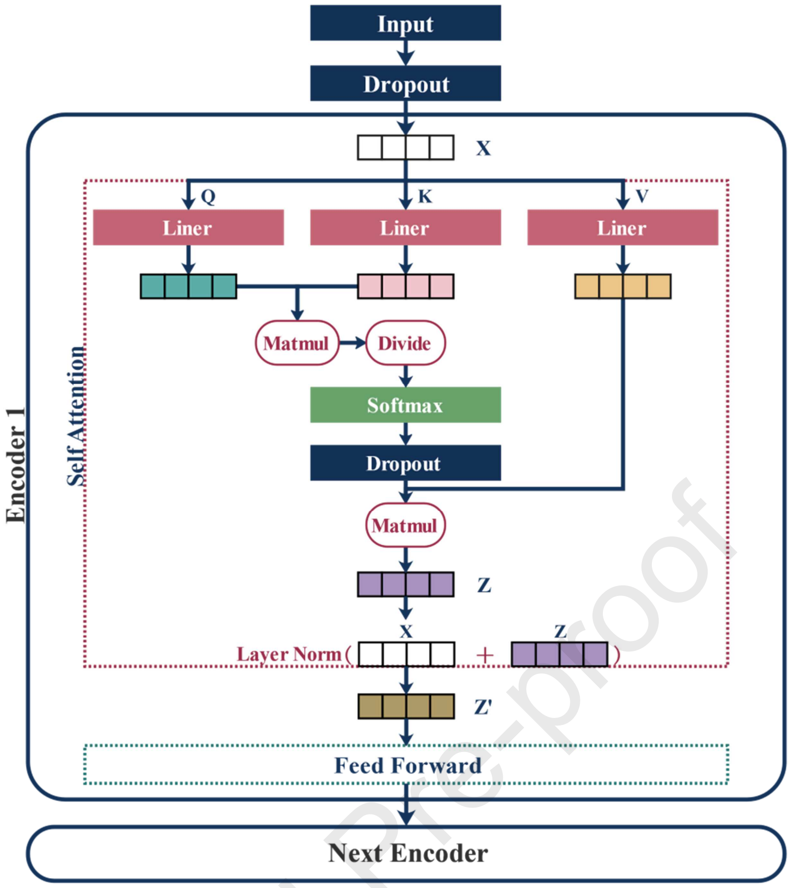


**Fig. 5** The architecture of the encoder in TransQuake.

may pay the highest attention to the S wave. Furthermore, the multi-head attention mechanism operates multiple attention in parallel and enables the model to explore more relationships than single attention which greatly strengthens the ability to receive information on different positions and decreases the training time. Take the attention mechanism as mapping a query and corresponding key-value pairs to an output, we have the core formula as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \quad , \tag{4}$$

where $Q, K, V$ are matrices packed respectively by a batch of queries, keys, values (Alammar J., 2018). Note that we explore the internal relationship among waveforms; thus, $Q = K = V$ is set before they go through their own linear layers and begin the attention calculation.

### 1.3.3 Output

To turn the vectors into a classification result, we use a linear layer, which uses the Xavier initializer and $bias = 0.1$. We then then add a softmax layer to get the possibility of each label. Finally, our model will output the result label which has a higher probability. As we mentioned before, $r_i = 0$ represents a noise wave and $r_i = 1$ represents an earthquake event with P-wave arrival.

## 2 RESULTS AND DISCUSSION

**Table 2** Definition of the confusion matrix for model evaluation.

| Output \ Label | Positive (earthquake) | Negative (noise) |
|---|---|---|
| Positive (earthquake) | True Positive (TP) | False Positive (FP) |
| Negative (noise) | False Negative (FN) | True Negative (TN) |

In our experiments, we split the dataset into 2 contiguous parts: the first 5/6 samples are collected in July for training, and the last 1/6 are prepared for validation (1/12) and test (1/12). Since

the dataset is severely imbalanced (i.e., much more negative samples than positive samples), the data-driven model will prefer the majority class, which greatly decays the performance of machine learning models. To balance the training data, we produce additional earthquake waveforms by adding zero-mean Gaussian noise (i.e., set signal-to-noise ratio randomly between 20-80dB) to the original positive waveform (Perol T. et al., 2018). After that, the number of positive and negative samples in the training dataset is equal (Zhou Zhihua et al., 2005).

In the training process, considering the possibility of label error and the robustness of our model, we use the label smoothing algorithm for our model (Szegedy C. et al., 2016). Specifically, given a smoothing rate $\varepsilon$, a uniform distribution appears over labels, the label is reformulated as

$$l_i' = (1-\varepsilon)\, l_i + \frac{1}{K}\ , \tag{5}$$

where $K$ is the number of label channels. Here, since the original label $l_i$ is equal to $\{0{,}1\}$ or $\{1{,}0\}$, $K = 2$, and $\varepsilon = 0.1$; thus, the reformulated positive label $l_i' = \{0.95{,}0.05\}$, while the negative one is equal to $\{0.05{,}0.95\}$. Besides, we utilize the $l2$-regularized cross-entropy loss function of $\lambda = 10^{-3}$ to quantify the gap between label and output (Ng A. Y., 2004). To narrow the gap, we optimize the model parameters using Adam Optimizer with $beta\,1 = 0.9$, $beta\,2 = 0.98$, and $epsilon = 10^{-8}$ (Kingma D. P. et al., 2014). Meanwhile, the dropout rate is also set to 0.1.

### 2.1 Evaluation Metrics

In our experiments, we evaluate our approach with four widely-used evaluation metrics, namely Accuracy, Precision, Recall, and F1 Score. In particular, the definition of each evaluation metric is listed as follows:

$$Accuracy = (TP+TN)/(TP+TN+FN+FP), \tag{6}$$

$$Precision = TP/(TP+FP), \tag{7}$$

$$Recall = TP/(TP+FN), \tag{8}$$

$$F1 = 2/(1/Precision + 1/Recall), \tag{9}$$

where the definitions of TP, TN, FN, and FP are shown in Table 2. Indeed, the Accuracy quantifies the percent of true samples; the Precision focuses on how many positive samples our model judges are correct; the Recall focuses on the percentage of positive samples that can be detected correctly; and the F1 score aims to strike a balance between Precision and Recall, which encourages us to find the earthquake examples accurately and completely.

**Table 3** The setting of model parameters in our experiments.

| $N$ | Head | $d$ | Learning rate | $T$ | L2 | $S$ | Drop out | Use STA/LTA | Augment strategy |
|---|---|---|---|---|---|---|---|---|---|
| 128 | 8 | 512 | $5\times 10^{-8}$ | 50 | 0.005 | 4 | 0.1 | yes | Gaussian noise |

**Table 4** The overall performance.

| Method | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| Logistic Regression | 0.499 | 0.073 | 0.504 | 0.127 |
| Support Vector Machine | 0.697 | 0.079 | 0.300 | 0.126 |
| RNN (LSTM) | 0.857 | 0.163 | 0.234 | 0.192 |
| Random Forest | 0.879 | 0.259 | 0.352 | 0.298 |
| CNN (CPIC) | 0.932 | 0.521 | 0.739 | 0.611 |
| CNN (ConvNetquake) | 0.946 | 0.649 | 0.581 | 0.613 |
| MSDNN | 0.952 | 0.678 | 0.638 | 0.658 |
| MSDNN+Multi-task Learning | 0.954 | 0.683 | 0.667 | 0.675 |
| **TransQuake** | **0.956** | **0.712** | **0.667** | **0.689** |

**Table 5** Performance with Different Time Window.

| Length (s) | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| 20 | 0.953 | 0.714 | 0.583 | 0.642 |
| 30 | 0.956 | 0.740 | 0.609 | 0.669 |
| 40 | 0.953 | 0.685 | 0.656 | 0.670 |
| 50 | 0.956 | 0.712 | 0.667 | 0.689 |

### *2.2 Baselines*

To better validate the effectiveness of our model, we have compared it with numerous state-of-the-art machine learning models that are widely used in the seismic field. Specifically, we first select some traditional machine learning models including *Logistic Regression*, *Support Vector Machine* (Cortes C. et al., 1995), and *Random Forests* (Breiman L., 2001). Then, since TransQuake is a deep learning-based sequential model, we also select some advanced deep learning methods as baselines.

- *Recurrent Neural Network (RNN)* is a classic neural network structure for sequential data modeling (Rodriguez P. et al., 1999). In our experiments, we choose the most advanced RNN-based model Long short-term memory (LSTM) as the baseline.
- *CNN* structure is widely applied for seismic detection, due to its remarkable capability of modeling local and static features. Among existing CNN models, we choose two advanced models, ConvNetquake and CPIC (Zhu Lijun et al., 2019), as baselines. Specifically, ConvNetquake is the first CNN model proposed for earthquake detection. In our experiments, we add the

number of 1d-convolution layers from 8 to 11, and other parts are set following the criterion proposed by Perol T. et al., (2018).

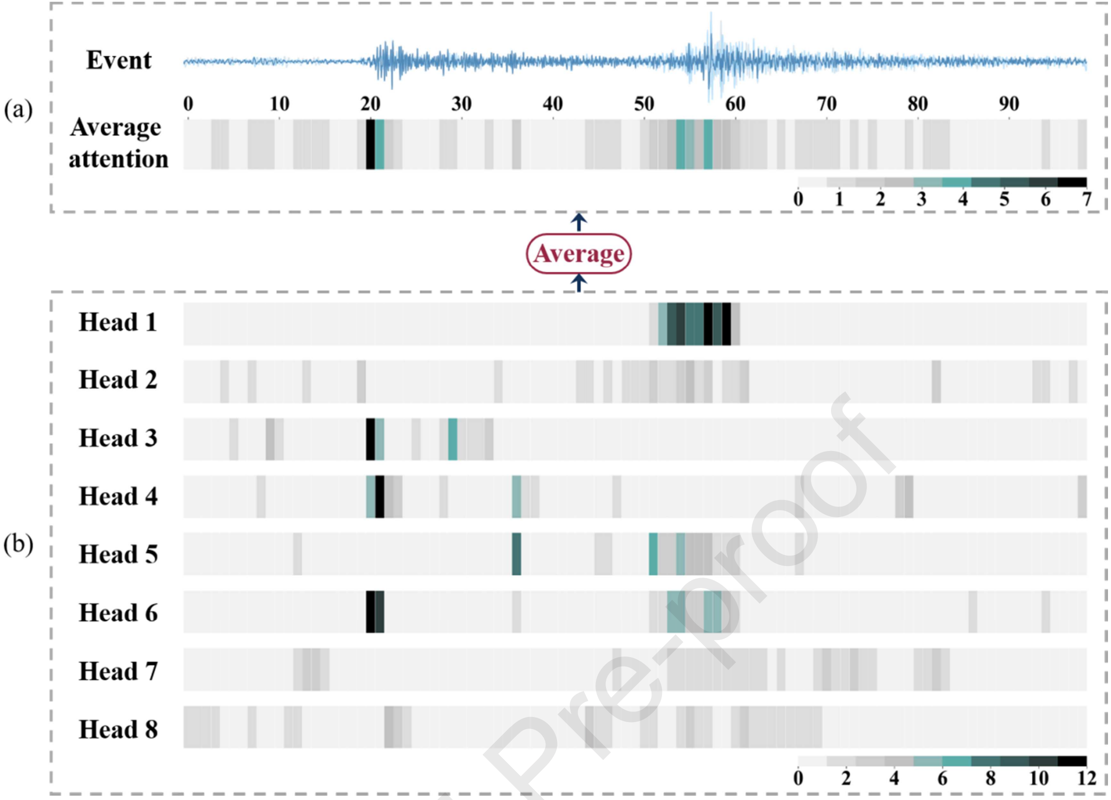


**Fig. 6** The process of generating an average attention-simplified view. (a) Average attention-simplified view; (b) 8-head attention-simplified view.

- MSDNN and MSDNN+Multi-task Learning are two advanced deep learning models for earthquake detection. Particularly, MSDNN combines the advantage of both CNN and LSTM, while MSDNN+Multi-task Learning can take advantage of the homologous earthquakes to enhance MSDNN with additional clues (Zhang Qi et al., 2019).

### *2.3 Overall Performance*

After using the validation dataset for fitting our model, we can obtain the best settings of model parameters as shown in Table 3.

The overall performance comparisons of different models on the test dataset are shown in Table 4. The results suggest the following conclusions. First, our approach TransQuake consistently outperforms all baselines. Particularly, in terms of the F1 score, the performance of TransQuake is 259% higher than RNN, 131% higher than Random Forest, 12.7% higher than CPIC, 12.4% higher than ConvNetquake, 4.7% higher than MSDNN, and 2.1% higher than MSDNN+Multi-task learning, respectively. Notably, as mentioned before, since the negative samples in our dataset are noises misclassified as aftershocks by low threshold *FilterPicker* rather than the simple background noises, the difficulty of our task is much higher than previous studies (Lara-Cueva R. et al., 2017; Perol T. et al., 2018; Rouet-Leduc B. et al., 2019),

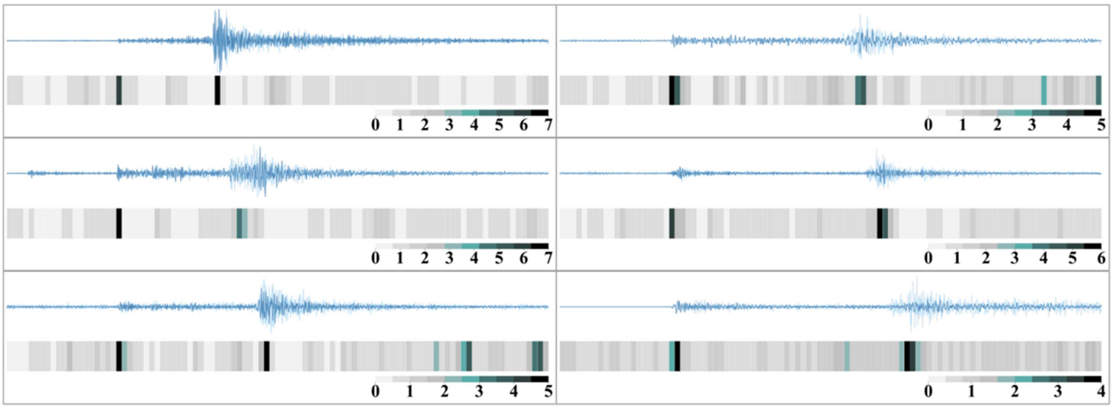


**Fig. 7** Average attention-simplified views of different waveforms.

reasonably resulting in slightly lower performance with weaker baselines compared with those obtained by the original paper (using background noises as negatives). In turn, it validates the difficulty of our task and the discrimination of our model.

*2.4 Discussion of the Length of the Time Window*

As shown in Table 5, we set 4 different time windows from 20s to 50s. Generally, metrics, especially F1, become better with the increasing of the time window length , indicating that the information besides the P wave also contributes to detection. Given the efficiency, we set the time window as 50s in the paper instead of increasing the length. In addition, short time windows of our model still perform better than most state-of-the-art models. Thus, it is possible to use our model in situations with different time requirements.

*2.5 Case Study: The Visualization of Multi-head Attention*

As a crucial part of our model, the attention mechanism can help the model pay more attention to the important information and avoid noise interference, providing us a way to understand the judgment logic behind the model. Here, we use $T = 50$, $k = 100$ positions, which means each position corresponds to a 0.5s time window of the waveform. To appropriately show their attention for different purposes, we make two views of attention maps: one consists of 100 rectangles and briefly presents the focuses (attention-simplified view), the other is a 2D matrix with size $100 \times 100$ which helps us further explore the inner relationship (attention-full view). Using attention-simplified views, Fig. 6 shows an example of the distribution of attention scores. The darker the color, the higher the attention score (i.e., more important for model judgment). It can be seen that our model spontaneously focuses on the P wave and S wave which are important for earthquake waveforms. Specifically, Fig. 6(b) shows how our multi-head attention mechanism works. The multi-head attention uses different attention distributions to focus on different important positions (e.g., Head 1 focuses on S wave and Head 3 focuses on P wave). This kind of structure can make the attention mechanism more precise and comprehensive. Then, Fig. 7 shows some

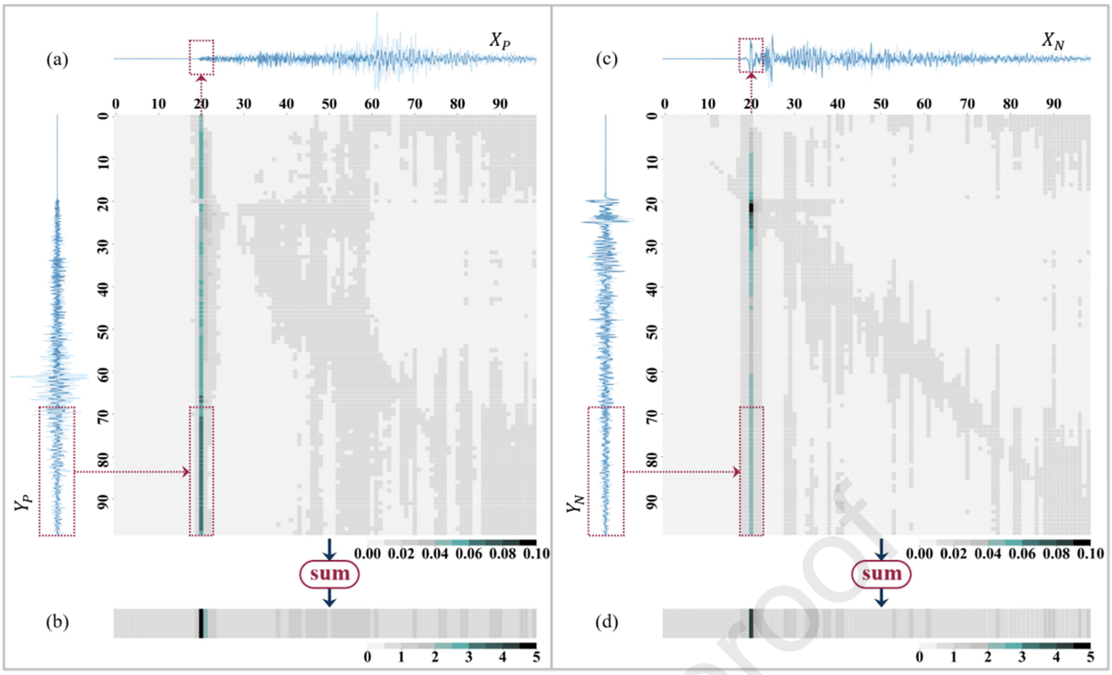


**Fig. 8** Two views of average attention map of all TP examples and all TN examples. (a) attention-full view of TP; (b) attention-simplified view of TP; (c) attention-full view of TN; (d) attention-simplified view of TN.

attention maps of different waveforms, where we can find that our model always accurately focuses on the key positions (i.e., P wave and S wave), no matter how the waveforms change.

Furthermore, in order to analyze the effect of the attention mechanism on waveforms of different types in detail, we plot the attention of all TPs (600 aftershocks) and that of all TNs (8967 noises) respectively in Fig. 8. In attention-simplified views (i.e., Fig. 8(b) and (d)), they both highly focus on the P wave and exhibit a slight difference in the attention distribution of the S-wave. In attention-full views (i.e., Fig. 8(a) and (c)), they not only present the high attention position but also tell the positions are paying attention to it (???) (i.e., the latent correlation between different positions of waveforms). In the attention-full view, each square denotes the attention weight of $Y$ to $X$. As shown in Fig. 8(c), TN (noise) has more diagonal focuses, which means it usually changes its focus for different waves rather than builds stable relationships with other positions. Another important difference is shown in the red boxes in Fig. 8(a) and (c): earthquakes own their most stable relationship between the coda wave and P-wave while noises do not (it meaninglessly focus on the diagonal of their first motion). Besides, the high weights of S-wave and coda wave paying to P-wave in Fig. 8(a) further validate that the existence of the S-wave and coda wave contributes to identifying the P-wave, especially the coda wave.

## 3 Conclusion and Prospection

In this paper, we propose a novel deep learning approach, TransQuake, for enhancing the performance of seismic P-wave detection. Specifically, TransQuake is designed on the basis of the most advanced deep sequential model, namely Transformer, integrating both STA/LTA algorithms and multi-head attention mechanisms. Compared with other deep learning models in relevant studies, TransQuake can effectively model the globally dynamic information contained within seismic waves. In addition, it shows high interpretability for result investigation. We have conducted extensive experiments on the aftershocks following the Mw7.9 Wenchuan earthquake using a number of state-of-the-art baselines. The experimental results clearly validate the effectiveness of TransQuake in terms of both seismic P-wave detection and result interpretation, such as the attention distribution of seismic waves in different positions, and the analysis of the optimal relationship between coda wave and P wave for noise identification. Indeed, the architecture of TransQuake can be easily adapted to multidimensional data from various fields. Therefore, the future works will focus on the integration of different data resources (e.g., geomagnetic and geothermal data) into our model for seismic detection.

## REFERENCES

Alammar J. The illustrated transformer [J]. *The Illustrated Transformer–Jay Alammar–Visualizing Machine Learning One Concept at a Time*, 2018, 27.

Allen R. Automatic phase pickers: Their present use and future prospects [J]. *Bulletin of the Seismological Society of America*, 1982, 72(6B): 225-242.

Breiman L. Random forests [J]. *Machine learning*, 2001, 45(1): 5-32.

Cortes C., Vapnik V. Support-vector networks [J]. *Machine learning*, 1995, 20(3): 273-297.

Frank W. B., Shapiro N. M., Husker A. L., Kostoglodov, V., Romanenko, A., Campillo, M. Using systematically characterized low‐frequency earthquakes as a fault probe in Guerrero, Mexico [J]. *Journal of Geophysical Research: Solid Earth*, 2014, 119(10): 7686-7700.

Fang Lihua, Wu Zhongliang, Song Kuan. SeismOlympics [J]. *Seismological Research Letters*, 2017, 88(6):1429-1430.

Gibbons S. J., Sørensen M. B., Harris D. B., Ringdal F. The detection and location of low magnitude earthquakes in northern Norway using multi-channel waveform correlation at regional distances [J]. *Physics of the Earth and Planetary Interiors*, 2007, 160(3-4): 285-309.

Glorot X., Bengio Y. Understanding the difficulty of training deep feedforward neural networks [C]. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics, 2010.

Goodfellow, I. J., Shlens, J., Szegedy, C. Explaining and Harnessing Adversarial Examples [J]. *Computer Science*, 2015.

Guo Tielong, Huang Zhibin, Zhao Bo. The Improvement of Earthquake Real-Time Monitoring System of Chinese National Digital Seismic Network [J]. *Earthquake Research in China*, 2019.

Kingma D. P., Ba J. Adam: A method for stochastic optimization [J]. *arXiv preprint* *arXiv:1412.6980,* 2014.

Lara-Cueva R., Benítez D. S., Paillacho V., Villalva M., Rojo-Álvarez J. L. On the use of multi-class support vector machines for classification of seismic signals at Cotopaxi volcano [C]. In: 2017 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC), 2017.

Liu Han, Zhang Jianzhong. STA/LTA algorithm analysis and improvement of Microseismic signal automatic detection [J]. *Progress in Geophysics*, 2014, 29(4): 1708-1714 (in Chinese with English abstract).

Lomax A., Satriano C., Vassallo M. Automatic picker developments and optimization: FilterPicker—A robust, broadband picker for real-time seismic monitoring and earthquake early warning [J]. *Seismological Research Letters,* 2012, 83(3): 531-540.

Ng A. Y. Feature selection, L 1 vs. L 2 regularization, and rotational invariance [C]. In: Proceedings of the twenty-first international conference on Machine learning, 2004.

Perol T., Gharbi M., Denolle M. Convolutional neural network for earthquake detection and location [J]. *Science advances,* 2018, 4(2): e1700578.

Reichstein M., Camps-Valls G., Stevens B., Jung M., Denzler J., Carvalhais N. Deep learning and process understanding for data-driven Earth system science [J]. *Nature,* 2019, 566(7743): 195-204.

Rodriguez P., Wiles J., Elman J. L. A recurrent neural network that learns to count [J]. *Connection Science,* 1999, 11(1): 5-40.

Rouet-Leduc B., Hulbert C., Johnson P. A. Continuous chatter of the Cascadia subduction zone revealed by machine learning [J]. *Nature Geoscience,* 2019, 12(1): 75-79.

Rouet-Leduc B., Hulbert C., McBrearty I. W., Johnson P. A. Probing slow earthquakes with deep learning [J]. *Geophysical Research Letters,* 2020, 47(4): e2019GL085870.

Saragiotis C. D., Hadjileontiadis L. J., Savvaidis A. S., Papazachos C. B., Panas S. M. Automatic S-phase arrival determination of seismic signals using nonlinear filtering and higher-order statistics [C]. In: IEEE 2000 International Geoscience and Remote Sensing Symposium. Taking the Pulse of the Planet: The Role of Remote Sensing in Managing the Environment. Proceedings (Cat. No. 00CH37120), 2000.

Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z. Rethinking the inception architecture for computer vision [C]. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow, I., Fergus R. Intriguing properties of neural networks [J]. *Computer Science,* 2014.Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser, Ł., Polosukhin I. Attention is all you need [C]. In: Advances in neural information processing systems, 2017.

Withers M., Aster R., Young C., Beiriger J., Harris M., Moore S., Trujillo J. A comparison of select trigger algorithms for automated global seismic phase and event detection [J]. *Bulletin of the Seismological Society of America,* 1998, 88(1): 95-106.

Yoon C. E., O'Reilly O., Bergen K. J., Beroza G. C. Earthquake detection through computationally efficient similarity search [J]. *Science advances,* 2015, 1(11): e1501057.

Zhang Qi, Xu Tong, Zhu Hengshu, Zhang Lifu, Xiong Hui, Chen Enhong, Liu Qi. Aftershock detection with multi-scale description based neural network [C]. In: 2019 IEEE

International Conference on Data Mining (ICDM), 2019.

Zhang Xiong, Zhang Jie, Yuan Congcong, Liu Sen, Chen Zhibo, Li Weiping. Locating induced earthquakes with a network of seismic stations in Oklahoma via a deep learning method [J]. *Scientific reports*, 2020, 10(1): 1-12.

Zhou Zhihua, Liu Xuying. Training cost-sensitive neural networks with methods addressing the class imbalance problem [J]. *IEEE Transactions on knowledge and data engineering*, 2005, 18(1): 63-77.

Zhu Lijun, Peng Zhigang, McClellan James, Li Chenyu, Yao Dongdong, Li Zefeng, Fang Lihua. Deep learning for seismic phase detection and picking in the aftershock zone of 2008 Mw7. 9 Wenchuan Earthquake [J]. *Physics of the Earth and Planetary Interiors*, 2019, 293: 106261.

## About the Author

HU Yumeng, born in 1996, is a master student at the Lanzhou Institute of Seismology. She is mainly engaged in seismic signal detection research by machine learning. E-mail: huyumeng19@mails.ucas.ac.cn

Corresponding author: ZHAO Wenjia, born in 1987, is a research engineer and editor of the academic journal "Seismology and Geology" at the Institute of Geology, China Earthquake Administration. She is mainly engaged in earthquake detection technology, machine learning, and big data application research. E-mail: zwj_dzdz@126.com