HIERARCHICAL ATTENTIVE MODELING OF EARTHQUAKE SIGNALS

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we present a multi-task network for simultaneous earthquake detection and phase picking based on a hierarchical attentive model. Our network consists of one deep encoder and three separate decoders. Two levels of self-attention (global and local) are embedded into the network in a hierarchical structure that helps the neural network captures dependencies between local and global features within an earthquake waveform.The proposed model outperforms deep-learning and traditional phase-picker and detector algorithms and achieves state-of-the art performance.

1 INTRODUCTION

Earthquake signal detection and phase picking are challenging problems in earthquake monitoring. Detection refers to identification of earthquake signals among a wide variety of non-earthquake signals and noise recorded by a sensor. Seismic phase picking is the measurement of seismic arrival times of distinct seismic phases (P and S phases) within an earthquake signal that are required for estimating earthquake location. Although these two tasks share some similarities, their objectives are slightly different. Minimizing the false negative and false positive rates are the main goals in detection; however, in phase picking the focus is more on increasing the temporal resolution or precision of arrival-time picks. This is because 1 millisecond of error in determining P-wave arrivals can translate to tens of meters of error in location estimates. In the other words, phase picking is a local problem compared to the detection, which uses a more global view of the full waveform, which consists of multiple seismic phases.

All previous studies (e.g. Perol et al. (2018), Ross et al. (2018b), Zhu & Beroza (2018), Mousavi et al. (2019b), Wu et al. (2018), Ross et al. (2018a), Pardo et al. (2019), Zhou et al. (2019), Zhu et al. (2019), and Wang et al. (2019)) have approached these tasks individually using separate networks; however, these tasks are closely related to each other. In practice analysts first look at the entire waveform on multiple stations to identify consistent elements of an earthquake signal (e.g. P, S, and surface waves) with a specific ordering (P-wave always arrives before S-wave, higher frequency body waves always precede dispersive surface waves etc.) to determine whether or not a signal is from an earthquake. Then they will focus on each phase to pick the arrival times more precisely. This practice indicates the interconnection of these two tasks and the importance of contextual information in earthquake signal modeling.

Here we test the hypothesis that better representations obtained by incorporating the contextual information in earthquake waveforms will result in better models. Our expectation is that not all parts of a seismic signal are equally relevant for a specific classification task, adn that it is beneficial to determine the relevant sections for modeling the interaction of local and global seismic features. We achieve this by incorporating attention mechanism into our network. Our model contains two levels of attention mechanism, one at global level for identifying an earthquake signal in the input time series and one at local level for identifying different seismic phases within the earthquake signal.

Our model has several distinct characteristics: 1) it is the first attentive-based model specifically designed for earthquake signals; 2) it consists of both convolutional and recurrent layers; 3) with 56 activation layers, it is the deepest network that so far has been trained by seismic signals; 4) it has a multi-task architecture that simultaneously performs the detection and phase picking while

modeling the dependency of these tasks on each other through a hierarchical structure; 5) it is the first global model trained using 1.2 M waveform of local earthquakes recorded around the world.

2 RELATED WORK

Perol et al. (2018) used a network of 8 convolutional and one fully connected layers to simultaneously detect and cluster events based on three component waveforms. Wu et al. (2018) applied a densely connected network of 7 fully convolutional layers to detect laboratory earthquakes of different sizes. Ross et al. (2018b) trained a network of 4 convolutional and 2 fully connected layers using seismograms recorded in Southern California to detect short windows of P, S, and noise. Ross et al. (2018a) adopted a similar approach (3 convolutional and 2 fully connected layers) for picking P arrival times. Zhu & Beroza (2018) modified U-Net, a fully convolutional encoder-decoder network with skip connections, for an end-to-end picking of P and S phases. Mousavi et al. (2019b) used a residual network of convolutional, bi-directional Long Short Term Memory units, and fully connected layers for detecting earthquake signals in the time-frequency domain. Pardo et al. (2019) used seismograms from Northern California to train their two-stage phase picker. They used a convolutional network for a rough segmentation of phases first, and then performed a regression to pick the arrival times. Seismic data recorded from Wenchuan Earthquake in Sichuan, China were used by Zhou et al. (2019) (136 K augmented P and S waveforms) and Zhu et al. (2019) (30 K) for training deep-leaning -based detectors and pickers. While Zhou et al. (2019) used two separate networks of 8 convolutional layers and two bi-directional GRU layers for detection and picking respectively, Zhu et al. (2019) used the same network (11 convolutional and 1 fully connected layers) in a recursive manner for both detection and picking with a cost of more computational time. Wang et al. (2019) trained a modified version of VGG-16 network using 740K seismograms recorded in Japan for picking P and S arrival times.

3 METHOD

Seismic signals are sequential time series consisting of different local (individual seismic phases) and more global (e.g. packages of body and surface waves or scattered waves in the coda) features. Hence, it is useful to retain the complex interaction between the local and global dependencies in an end-to-end deep learning model. Traditionally, recurrent neural networks have been used for such a sequence modeling; however, the relatively long duration of seismic signals requires some down sampling prior to the recurrent layers to manage the computational complexity. Hence, a combination of recurrent and convolutional layers has been shown to be optimal for sequential modeling of seismic signals. Building upon our previous work (Mousavi et al. (2019b)), here we introduce a multi-task network of recurrent and convolutional layers that incorporates attention mechanism as well. Attention mechanism is a method of encoding sequence data in which elements within a sequence will be highlighted or down-weighted based on their importance or irrelevance to a task.

Our neural network has a multi-task structure consisting of one deep encoder and three separate decoders composed of 1D convolutions, bi-directional and uni-directional Long-Short-Term Memories (LSTM), Network-in-Network, residual connections, feed-forward layers, transformer, and self-attentive layers. (Figure 1). More details are provided in the Appendices. The encoder consumes the seismic signals in the time domain and generates a high-level representation and contextual information on their temporal dependencies. Decoders then use this information to map the high-level features to three sequences of probabilities associated with existence of an earthquake signal, P-phase, and S-phase for each time point.

In self-attentive models the amount of memory grows with respect to the sequence length. Hence, we add a down-sampling section composed of convolutional and max-pooling layers to the front of the encoder. These down sampled features are transformed to high-level representations through a series of residual convolution and LSTM blocks. A global attention section at the end of the encoder aims at directing the attention of the network to the parts associated with the earthquake signal. These high-level features are then directly mapped to a vector of probabilities representing the existence of an earthquake signal (detection) using one decoder branch. Two other decoder branches are associated with P-phase and S-phase respectively in which a LSTM/local attention unit is placed at the beginning. This local attention will further direct the attention of the network into

local features within the earthquake part that are associated with individual seismic phases. Residual connections within each block and techniques such as network-in-networks help to expand the depth of the network while keeping the number of learnable parameters manageable.

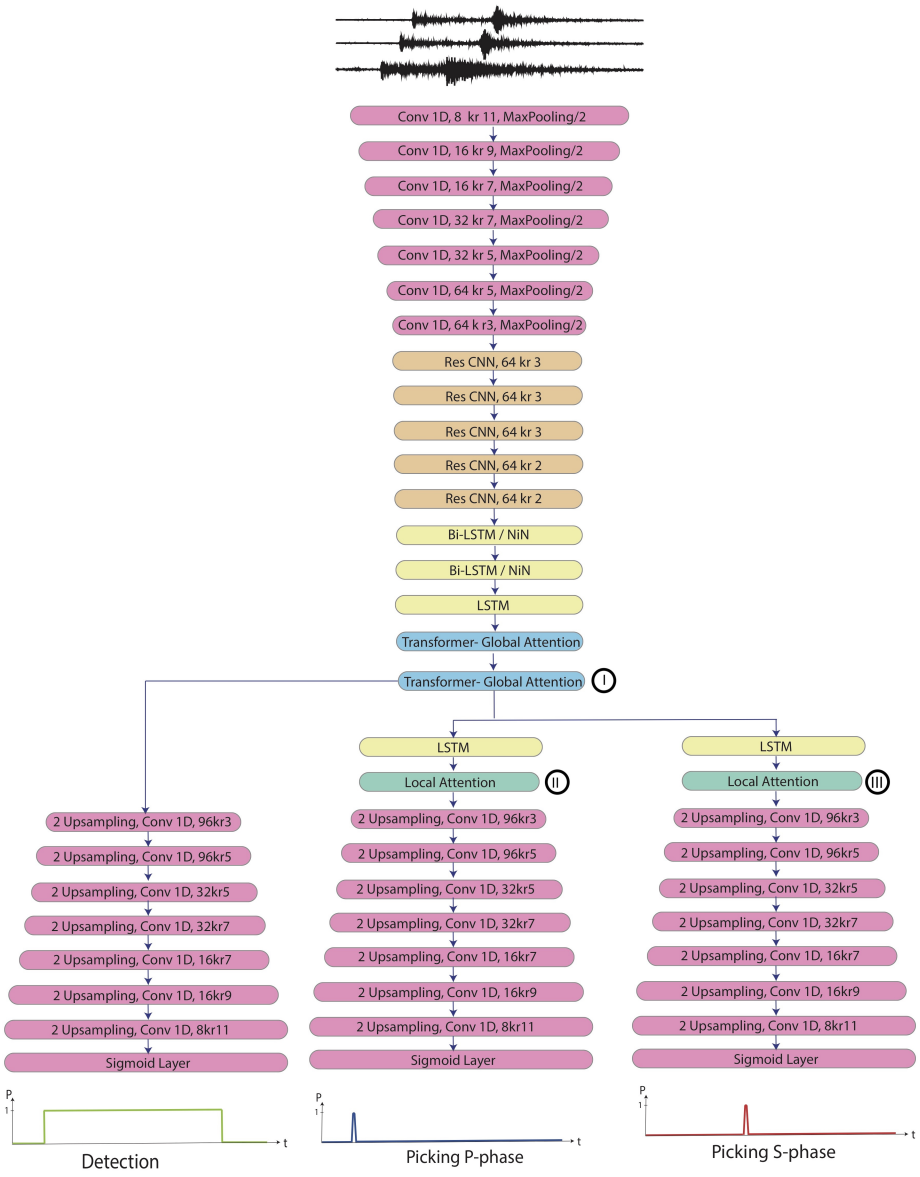

Figure 1: Our network architecture. Full details of each block are provided in the method section. The convolutional layers read as (number of kernels)kr(kernel size).

4 EXPERIMENT SETTINGS

We used the STanford EArthquake Dataset (STEAD) (Mousavi et al. (2019a)) to train this network. STEAD is a large-scale global dataset of earthquake and non-earthquake signals. Here we used 1 M earthquake and 200 k noise waveform recorded by seismic stations at local distances (up to 300 km from the earthquakes). We split the data into training (85%), validation (5%), and test (10%) sets randomly. Waveforms are 1 minute long with a sampling rate of 100 HZ and have been band-passed filtered between 1-45 HZ.

For both convolutional and LSTM units, all the weight and filter matrices were initialized with a Xavier normal initializer (Glorot & Bengio (2010)) and bias vectors set to zeros. We used ADAM optimization (Kingma & Ba (2014)) with varying learning rates, while the learning rate varied during training. The model took O(5) days to converge using 4 parallel Tesla-V100 GPUs under the tensorflow framework (Abadi et al. (2016)). We augmented the data by adding a secondary earthquake signals into waveforms, adding random noise, randomly shifting the waveform, and randomly dropping one channel with 0.3, 0.5, 0.9, and 0.3 probabilities respectively. In each batch, half of the batch are augmented versions of the waveform in the other half. Data augmentation and normalization are done simultaneously during the training on 24 CPUs in parallel. We used a dropout rate of 0.1 for all dropout layers.

5 RESULTS

5.1 EXPLORING NETWORK'S ATTENTION

The attention weights define how much of each input state should be considered for predicting each output, and can be interpreted as a vector of importance weights. By explicitly visualizing these attention weights we can see which parts of the input sequence, the neural network has learned to focus on. In Figure 2, the calculated energies (or scoring) for attention layers are presented. This is a measure of alignment or match between encoder and decoder states and is used by the decoder to decide on which parts of the source sequence to focus. High energies indicate alignments of predicted probabilities and corresponding parts of waveform.

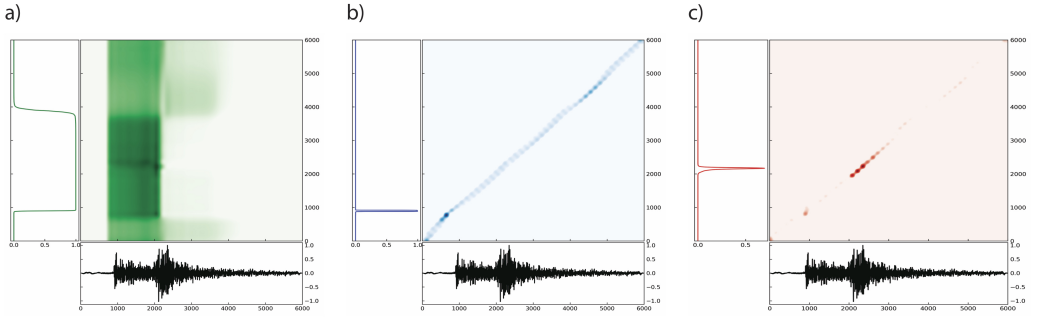

Figure 2: Input waveform (bottom boxes), output prediction probabilities (left boxes), and corresponding scoring (central boxes) for a) transformer (I in Figure 1), b) local attention for P-phase (II in Figure 1), and c) the local attention for S-phase (III in Figure 1).

In Figure 3, the outputs of each of these attention layers (summation of hidden states at all other time steps, weighted by their scoring) are presented. We can see that the network has learned to focus on different parts of the waveform at different attention levels. This highlights the most useful parts of the input waveform for each task. The shorter path through the detection decoder and its higher loss (due to longer length of label) naturally force the network to first learn to distinguish the earthquake signal within a time series. This mimics an analyst's decision making workflow. The second transformer (I in Figure 1), at the end of encoder section, mainly passes the information corresponding to the earthquake signal to the subsequent decoders. This means that the encoder learns to select which parts of the source sequence has the most important information for detection and phase picking. This information is directly used by the detection decoder to predict existence of earthquake signal in the time series. The local attention layers at the beginning of P and S decoders further focus on smaller sections, within the earthquake waveform - primarily near the arrival times of those phases, to make their predictions.

5.2 PICKING PERFORMANCE

We used more than 100 k test data to evaluate the picking performance and compare it with other methods. An example of detection and phase picking outputs are presented in Figure 4. We used 7

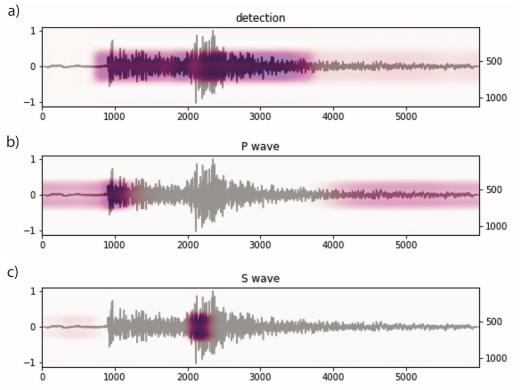

Figure 3: Input waveform overlaid by contextual information or outputs of the attention layers for a) transformer (I in Figure 1), b) local attention for P-phase (II in Figure 1), and c) the local attention for S-phase (III in Figure 1).

scores (standard deviation of error, mean error, precision, recall, F1-score, mean absolute error, and mean absolute percentage error) to measure the performance. A pick was considered as true positive when its absolute distance from the ground truth is less than 0.5 second. The measured scores for P and S picks are presented in Table 1 and 2 respectively. PhaseNet, GPD, and PickNet are three deep-leaning-based phase pickers. Kurtosis (KT), frequency band (FB), and Akaike Information Criteria (AIC) pickers are traditional (non-machine learning) algorithms. Results indicate that our approach outperforms other phase picking algorithms. Our approach increased the F-scores for P and S picks about 4%, while the recall is lower for S phase. The improvement in P-wave picks are more significant than for S-wave picks. This may be due the fact that picking S-waves is much harder and prone to more errors, which can lead to higher labeling error in the training set. We compare the detection performance with one deep-learning (DetNet) and one traditional (STA/LTA) detector in Table 3. The superior performance of the proposed method could be due to several factors (e.g. attention mechanism, depth of our network, etc). Attention mechanism helps incorporating global and local scale features within the full waveform. A deeper network can result in more discriminatory power.

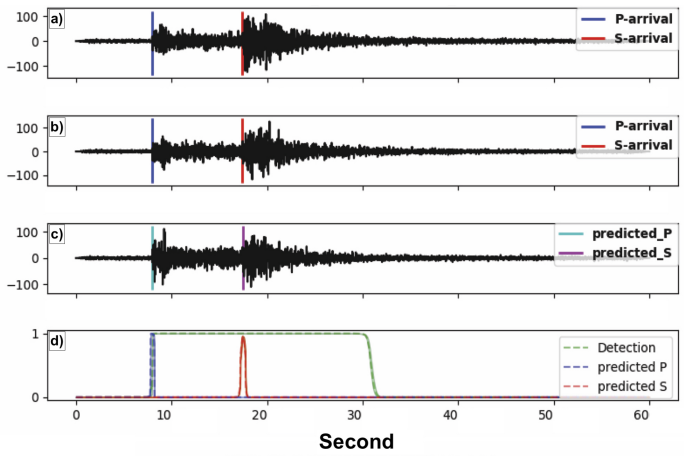

Figure 4: An example of 3 channel input waveform and ground truths (a, b, and c) and prediction outputs of three decoders.

Table 1: P-phase picking. μ and σ are mean and standard deviation of errors (ground truth - prediction) in seconds respectively. Pr, Re, and F1 are precision, recall, and F1-score respectively. MAE and MAPE are mean absolute error and mean absolute percent error respectively.

Model	μ	σ	**Pr**	**Re**	**F1**	**MAE**	**MAPE**	**Reference**
This Study	**0.00**	**0.03**	**0.99**	**0.99**	**0.99**	**0.92**	**0.15**	-
PhaseNet	-0.02	0.08	0.96	0.96	0.96	6.83	0.64	Zhu & Beroza (2018)
GPD	0.03	0.10	0.81	0.80	0.81	8.29	1.03	Ross et al. (2018b)
PickNet	**0.00**	0.09	0.81	0.49	0.61	6.94	1.56	Wang et al. (2019)
KT	-0.03	0.09	0.94	0.79	0.86	8.02	0.89	Saragiotis et al. (2002)
FB	-0.01	0.08	0.95	0.82	0.88	6.74	0.74	Lomax et al. (2012)
AIC	-0.04	0.09	0.92	0.83	0.87	8.50	0.93	Maeda (1985)

Table 2: S-phase picking. μ and σ are mean and standard deviation of errors (ground truth - prediction) in seconds respectively. Pr, Re, and F1 are precision, recall, and F1-score respectively. MAE and MAPE are mean absolute error and mean absolute percent error respectively.

Model	μ	σ	**Pr**	**Re**	**F1**	**MAE**	**MAPE**	**Reference**
This Study	**0.00**	**0.11**	**0.99**	**0.96**	**0.98**	**8.51**	**0.50**	-
PhaseNet	-0.02	**0.11**	0.96	0.93	0.94	8.86	0.51	Zhu & Beroza (2018)
GPD	0.03	0.14	0.81	0.83	0.82	10.42	0.80	Ross et al. (2018b)
PickNet	-0.02	0.13	0.83	0.55	0.66	9.88	2.94	Wang et al. (2019)
KT	-0.10	0.13	0.89	0.39	0.55	11.17	0.91	Saragiotis et al. (2002)
FB	-0.06	0.12	0.91	0.40	0.56	9.95	0.73	Lomax et al. (2012)
AIC	-0.07	0.15	0.87	0.51	0.64	11.63	1.92	Maeda (1985)

Table 3: Detection performance. Pr, Re, and F1 are precision, recall, and F1-score respectively.

Model	**Pr**	**Re**	**F1**	**Reference**
This Study	**1.0**	**1.0**	**1.0**	-
DetNet	**1.0**	0.89	0.94	Zhou et al. (2019)
STA/LTA	0.91	1.0	0.95	Allen (1978)

6 CONCLUSION

We propose a multi-task network for simultaneous earthquake detection and phase picking based on a hierarchical attentive model. Our network consists of one deep encoder and three separate decoders composed of 1D convolutions, bi-directional and uni-directional Long-Short-Term Memories (LSTM), Network-in-Network, residual connections, feed-forward layers, transformer, and self-attentive layers. Two levels of self-attention (global and local) are embedded in the network in a hierarchical structure that helps the neural network retaining dependencies between local and global features within an earthquake waveform. Three-component waveform time series are input to the encoder while the decoders output three sequences of probabilities associated with existence of an earthquake signal, P-phase, and S-phase for each time point. The proposed model outperforms existing deep-learning and traditional phase-pickers and detector algorithms to achieve state-of-the art performance.

REFERENCES

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-

scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pp. 265–283, 2016.

Rex V Allen. Automatic earthquake recognition and timing from single traces. *Bulletin of the Seismological Society of America*, 68(5):1521–1532, 1978.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit Das. Very deep convolutional neural networks for raw waveforms. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 421–425. IEEE, 2017.

Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*, pp. 1019–1027, 2016.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Jianshu Ji, Qinlong Wang, Kristina Toutanova, Yongen Gong, Steven Truong, and Jianfeng Gao. A nested attention neural hybrid model for grammatical error correction. *arXiv preprint arXiv:1707.02026*, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

Anthony Lomax, Claudio Satriano, and Maurizio Vassallo. Automatic picker developments and optimization: Filterpicker—a robust, broadband picker for real-time seismic monitoring and earthquake early warning. *Seismological Research Letters*, 83(3):531–540, 2012.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

Naoki Maeda. A method for reading and checking phase times in autoprocessing system of seismic wave data. *Zisin*, 38:365–379, 1985.

S Mostafa Mousavi, Yixiao Sheng, Weiqiang Zhu, and Gregory C Beroza. Stanford earthquake dataset (stead): A global data set of seismic signals for ai. *IEEE Access*, 2019a.

S Mostafa Mousavi, Weiqiang Zhu, Yixiao Sheng, and Gregory C Beroza. Cred: A deep residual network of convolutional and recurrent units for earthquake signal detection. *Scientific reports*, 9 (1):10267, 2019b.

Esteban Pardo, Carmen Garfias, and Norberto Malpica. Seismic phase picking using convolutional networks. *IEEE Transactions on Geoscience and Remote Sensing*, 2019.

Thibaut Perol, Michaël Gharbi, and Marine Denolle. Convolutional neural network for earthquake detection and location. *Science Advances*, 4(2):e1700578, 2018.

Zachary E Ross, Men-Andrin Meier, and Egill Hauksson. P wave arrival picking and first-motion polarity determination with deep learning. *Journal of Geophysical Research: Solid Earth*, 123 (6):5120–5129, 2018a.

Zachary E Ross, Men-Andrin Meier, Egill Hauksson, and Thomas H Heaton. Generalized seismic phase detection with deep learning. *Bulletin of the Seismological Society of America*, 108(5A): 2894–2901, 2018b.

Christos D Saragiotis, Leontios J Hadjileontiadis, and Stavros M Panas. Pai-s/k: A robust automatic seismic p phase arrival identification scheme. *IEEE Transactions on Geoscience and Remote Sensing*, 40(6):1395–1404, 2002.

Matthias Sperber, Jan Niehues, Graham Neubig, Sebastian Stüker, and Alex Waibel. Self-attentional acoustic models. *arXiv preprint arXiv:1803.09519*, 2018.

Tian Tan, Yanmin Qian, Hu Hu, Ying Zhou, Wen Ding, and Kai Yu. Adaptive very deep convolutional residual network for noise robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(8):1393–1405, 2018.

Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 648–656, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Jian Wang, Zhuowei Xiao, Chang Liu, Dapeng Zhao, and Zhenxing Yao. Deep-learning for picking seismic arrival times. *Journal of Geophysical Research: Solid Earth*, 2019.

Yue Wu, Youzuo Lin, Zheng Zhou, David Chas Bolton, Ji Liu, and Paul Johnson. Deepdetect: A cascaded region-based densely connected network for seismic event detection. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):62–75, 2018.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 1480–1489, 2016.

Licheng Yu, Mohit Bansal, and Tamara L Berg. Hierarchically-attentive rnn for album summarization and storytelling. *arXiv preprint arXiv:1708.02977*, 2017.

Yu Zhang, William Chan, and Navdeep Jaitly. Very deep convolutional networks for end-to-end speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4845–4849. IEEE, 2017.

Yijian Zhou, Han Yue, Qingkai Kong, and Shiyong Zhou. Hybrid event detection and phase-picking algorithm using convolutional and recurrent neural networks. *Seismological Research Letters*, 90 (3):1079–1087, 2019.

Lijun Zhu, Zhigang Peng, James McClellan, Chenyu Li, DongDong Yao, Zefeng Li, and Lihua Fang. Deep learning for seismic phase detection and picking in the aftershock zone of 2008 mw7. 9 wenchuan earthquake. *Physics of the Earth and Planetary Interiors*, 2019.

Weiqiang Zhu and Gregory C Beroza. Phasenet: a deep-neural-network-based seismic arrival-time picking method. *Geophysical Journal International*, 216(1):261–273, 2018.

A VERY DEEP ENCODER

Several studies (e.g. Zhang et al. (2017), Tan et al. (2018), and Dai et al. (2017)) have shown, in end-to-end learning from raw waveforms, that employing deeper networks can be beneficial by having more expressive power, better generalization, and more robustness to noise in waveforms. Here, we build a very deep encoder that is known to be helpful in improving the performance of a sequence-to-sequence model with attention.

The encoder consists of several blocks of residual convolution layers and recurrent blocks including network-in-network connections. Convolutional layers exploit local structure and provide the model a better temporal invariance, which typically leads to better generalization. To be able to extend the depth of the network without degradation problem we use blocks of convolutional layers with residual connections(He et al. (2016)) as depicted in Figure 4.

Long-short term memory (LSTM)(Hochreiter & Schmidhuber (1997)) are specific types of recurrent neural networks commonly used for modeling longer sequences. The main element in an LSTM unit is a memory cell. At each time step, an LSTM unit receives an input, outputs a hidden state, and updates the memory cell based on a gate mechanism. Here we expand the bi-directional LSTM blocks by including Network-in-Network (Lin et al. (2013)) modules in each block that help to increase the network's depth without increasing the number of learnable parameters (Figure 2). LSTM layers prior to self-attention layers, have been shown to be necessary for incorporating positional information (Sperber et al. (2018)).

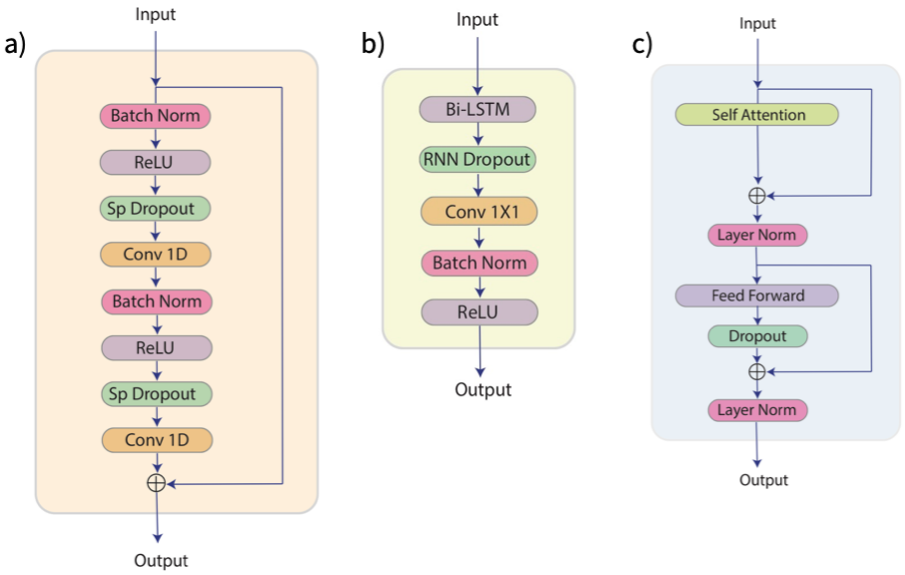

Figure 5: Individual blocks of our neural network. a) ResCNN blocks used in the encoder. Spatial dropout (Sp Dropout)(Tompson et al. (2015)) layers have been used after each ReLU activation layer which is preceded by a batch normalization (Batch Norm)(Ioffe & Szegedy (2015)). b) Bi-LSTM/NiN blocks including a Bidirectional LSTM, one convolution layer with one filter sized 1, batch normalization, ReLU activation layer. RNN Dropout is the recurrent dropout (Gal & Ghahramani (2016)). c) Multi-head self-attention block for global attention. It includes one position-wise feed-forward network and layer normalization (Ba et al. (2016)).

B ATTENTION MECHANISM

Let's represent the output of an LSTM layer by $\boldsymbol{h} = \{h_t\} \in \mathbb{R}^{n \times d_h}$ as a sequence of vector elements (a high level representation of the original input signal). We calculate the self (internal) attention as follows (Luong et al. (2015) and Yang et al. (2016)):

$$e_{t,t'} = \sigma(W_2^T[\tanh(W_1^T h_t + W_1^T h_{t'} + b_1)] + b_2), \quad (1)$$

$$\alpha_{t,t'} = \frac{exp(e_{t,t'})}{\sum_{t'} exp(e_{t,t'})}, \quad (2)$$

$$c_t = \sum_{t'=1}^{d_h} \alpha_{t,t'}.h_{t'}, \quad (3)$$

where h_t and $h_{t'}$ are hidden state representations at time steps t and t' respectively. W and b are weight matrices and bias vectors respectively.σ is the element-wise sigmoid function. $\alpha_{t,t'}$ are scalar scores (also called alignment) indicating pairwise similarities between the elements of the sequence. The attentive hidden state representation, c_t, at time step t is given by summation of hidden states at all other time steps, $h_{t'}$, weighted by their similarities to the current hidden state, $\alpha_{t,t'}$. Vector $c_t \in \mathbb{R}^{d_h}$ is a sequence of contex-aware (with respect to surrounding elements) encoding that defines how much attention will be given to the features at each time step based on their neighborhood context. This will be incorporated into a downstream task as additional contextual information to direct the focus to the important parts of the sequence rather than the less relevant parts.

Here we adopt the residual attention blocks introduced in the Transformer (Vaswani et al. (2017)) by replacing the multi-head scaled-dot product attention by the above single-head additive attention. The feed-forward layer consists of two linear transformations with a ReLU activation in between, $FF(x) = max(0,\ xW_1 + b_1)W_2 + b_2$, intended to introduce additional nonlinearities.

Our goal is to implement two levels of attention mechanisms in a hierarchical structure (Ji et al. (2017) and Yu et al. (2017)) at the earthquake full waveform and individual phase levels. A logical way to do this is to perform attention mechanisms at two levels with different temporal resolution. For instance, applying the detection attention at the high level representation at the end of the encoder and the phase attention at the end of associated decoders where higher temporal resolutions are available. However, with $O(n^2.d)$ complexity of self-attention, this is not computationally feasible for the long duration time series (6000 samples) used here. Hence, we applied attention mechanisms with global and local attentions both at the bottleneck. The attention block at the end of the encoder performs global attention, by attending to all the positions in the sequence, to learn to identify the earthquake signals within the input time series. Shortened path from this layer to the detection encoder and naturally higher detection loss make this learning easier.

Attention blocks at the beginning of phase-picker decoders perform additional local attention by attending only to a small subset of the sequence (Luong et al. (2015)), to further narrow the focus to individual seismic phases within the earthquake waveform. One LSTM layer with 16 units is applied before the first attention block at each level to provide position information (Sperber et al. (2018)).