CHIN. PHYS. LETT. Vol. 36, No. 4 (2019) 044302

Source Ranging Using Ensemble Convolutional Networks in the Direct Zone of Deep Water *

Yi-Ning Liu(刘一宁)[1,2], Hai-Qiang Niu(牛海强)[1], Zheng-Lin Li(李整林)[1]**

[1]*State Key Laboratory of Acoustics, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190*
[2]*University of Chinese Academy of Sciences, Beijing 100190*

(Received 14 January 2019)

Using deep convolutional neural networks as primary learners and a deep neural network as meta-learner, source ranging is solved as a regression problem with the ensemble learning method. Simulated acoustic data from the acoustic propagation model are used as the training data. Real data from an experiment in the South China Sea are used as the test data to demonstrate the performance. The results indicate that in the direct zone of deep water, signals received by a very deep receiver can be used to estimate the range of underwater sound source. Within 30 km, the mean absolute error of the range predictions is 1.0 km and the mean absolute percentage error is 7.9%.

PACS: 43.60.Np, 43.60.Jn, 43.30.Wi DOI: 10.1088/0256-307X/36/4/044302

Acoustic source ranging in ocean waveguides has always been a concern in the field of underwater acoustics. As the transmission loss in the direct zone of deep water is less than that in the shadow zone, and the horizontal range of the direct zone increases with the depth of the receiver,[1] signals received by a very deep (near-bottom) hydrophone can be used for underwater source ranging. As Lloyd's mirror pattern exists and can be described with relatively simple expressions,[2] this pattern can be used for source ranging. A widely used method is matched-field processing (MFP),[3,4] which is achieved by matching replica fields with measurements.

With the development of statistical learning theory and computer technology, machine learning has made breakthrough progress in speech recognition, image processing and natural language processing. In the last five years, AlexNet, VGGNet, InceptionNet and ResNet[5–7] have greatly promoted the development of deep neural networks, which has made more and more fields begin to notice the potential of machine learning. Techniques of using max pooling as the pooling function and ReLU function as the activation function,[5] increasing the number of hidden layers, using small size of kernels, designing networks with branches, and using residual connections between hidden layers,[6,7] are helpful to improve the performance of the deep neural networks in image recognition.

However, publications on machine learning in underwater acoustics are still scarce. Since the 1990s, machine learning, mainly neural networks, has been used for underwater source localization,[8] and combined with the MFP to simulate the range and depth discrimination.[9] More recently, feed-forward neural networks (FNN), support vector machines (SVM), random forests (RF) and convolutional neural networks have been used for the range estimation of shallow water sound sources.[10–12] A single model has been used in the above tasks, but in fact most machine learning tasks may achieve better performance by ensemble learning such as averaging, voting and learning.[13]

In this Letter, stacking, an ensemble learning method,[14] combined with some complex deep neural networks are used for source ranging. Due to the deficient real observations, the training data are generated by the acoustic propagation model KRAKEN.[15] The experimental data received by a very deep hydrophone are used to demonstrate the performance as the test data. Tensorflow and Keras[6] are used to build networks. Firstly, we give the main process for source ranging. Secondly, we introduce the experimental environment and data preprocessing. Thirdly, we design the networks and analyze the ensemble learning method. Finally, the results of primary learners and the meta-learner are given.

The main process is shown in Fig. 1. After preprocessing, the normalized spectrograms of simulated and experimental data are used as the training and test data. Firstly, the training data are split into two subsets. The first subset is used to train the primary networks. Such networks are used to make predictions on the second subset. A new training data set is created using these predicted values as features and used to train the meta-learner. After the two steps of training, the test data set is used to examine the performance.

The data during the South China Sea experiment in 2016 are used to demonstrate the performance of all models. The source depth is around 120 m and the source center frequency is 300 Hz with the bandwidth 100 Hz. The receiving depth of the single hydrophone is 4152 m with the sampling rate 16 kHz. The bottom within the range of 30 km is relatively flat with an average depth of 4312 m. The sound speed profile (SSP) measured by XCTD is given in the left subplot of Fig. 2, which shows that the sound channel axis is at the depth of 1151 m and the speed is 1540 m/s at the surface and 1533 m/s at the bottom. Two-dimensional transmission loss with depth and range within 90 km is shown in the right subplot of Fig. 2, which is computed by the RAM-PE model. As shown in Fig. 2, the

*Supported by the National Natural Science Foundation of China under Grant Nos 11434012 and 11874061.
**Corresponding author. Email: lzhl@mail.ioa.ac.cn
© 2019 Chinese Physical Society and IOP Publishing Ltd

range of the direct zone with a very deep receiver can reach 30 km. In this study, the range estimation in the direct zone is investigated.

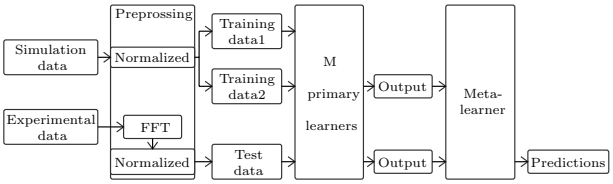


Fig. 1. The main process including data preprocessing, building, training and testing the networks.

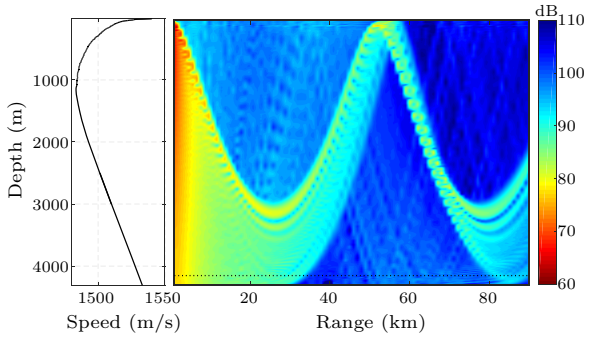


Fig. 2. The sound speed profile (left), and the two-dimensional transmission loss in the range of 90 km (right). The black dotted line shows the receiver depth, which is near the bottom.

To obtain the test data, the time-domain signals received by a single hydrophone are transformed to the normalized spectrograms before feeding into the networks. The complex pressure at frequency f obtained by performing fast Fourier transform (FFT) on the received time-domain signal is normalized according to

$$\widetilde{p}(f) = \frac{|p(f)|}{\max|p(f)|}. \tag{1}$$

Since the experimental signals are hyperbolic frequency modulated (HFM), the signal amplitudes are not constant across the frequency band. Therefore, segment normalization is performed with 10 Hz intervals. Then the normalized spectrum is sampled every 1 Hz in the frequency band 251–350 Hz to form the test data.

The training data are simulated using KRAKEN. Since the test data have a high signal-to-noise ratio (SNR) of about 35 dB, no noise is added to the training data. The environmental parameters are chosen to simulate the South China Sea experiment[16] with the bottom sound speed $c_\mathrm{b} = 1565\,\mathrm{m/s}$, the bottom density $\rho_\mathrm{b} = 1.6\,\mathrm{g/cm^3}$, and the attenuation coefficients $\alpha_\mathrm{b} = 0.3\,\mathrm{dB}/\lambda$. The source depth is from 110 to 130 m with 1 m increments. The receiver and bottom depths are 4152 and 4312 m, respectively. The source frequency is 251–350 Hz, the same as the test data. The range for the training data is from 0.1 to 30 km with 10 m increments. The comparison of simulated and experimental data is shown in Fig. 3. The result shows that the two are roughly consistent. The interference fringes in Fig. 3 are formed by Lloyd's mirror narrow band interference fringes with different frequencies. The reverse small fringes in 25–30 km may be caused by the negative waveguide invariants of the refracting sound rays.[17] These fringes can be automatically learned by neural networks and used for range prediction.

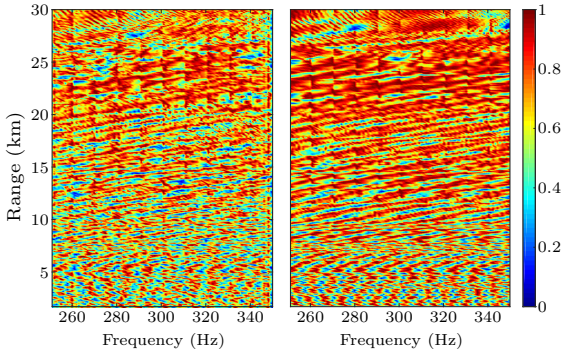


Fig. 3. Comparison of simulation data (right) and experimental data (left) by feature map of frequency and range.

One-dimensional convolutional networks (Conv-1D) are used to build the primary models. Suppose that the number of samples in the first training data subset is N_1, and the second one is N_2. Each sample has m feature frequencies. The ith input sample is denoted by $\boldsymbol{x}^{(i)} = [x_1^{(i)}, x_2^{(i)}, \ldots, x_m^{(i)}]$ and the total input data are designed as matrix $\boldsymbol{X} = [\boldsymbol{x}^{(1)}; \boldsymbol{x}^{(2)}; \ldots; \boldsymbol{x}^{(N_1)}]$.

The outputs $y^{(i)}$ of the networks are the predicted range value $r_\mathrm{p}^{(i)}$, which map the frequency spectrum to the range (i.e., regression).

The number of feature frequencies is 100. Therefore, the shape of the input layer is 100×1 and the number of neurons in the output layer is 1.

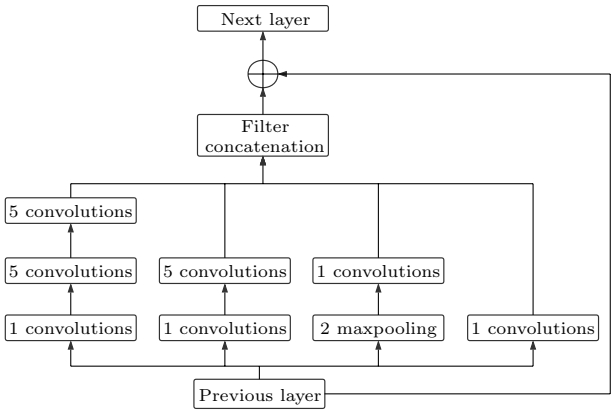


Fig. 4. Inception module with residual connection.

The hidden layers are designed by convolutional layers (kernel size 5 and stride 1) with a final fully connected layer. Inception modules and residual connections shown in Fig. 4 are used to improve the performance on the test data. Different network structures are listed in Table 1. The number of filters in each layer is 128. The ReLU function is chosen as the activation function and the Adam optimizer with the initial learning rate 0.001 is chosen to train the networks.[5] The maximum number of iterations is 100

and the batch size is $n = 32$. Dropout is used to reduce overfitting with rate 0.5. Global average pooling is used before the output layer and batch normalization is used after each of the convolution layers, which can help to improve the generalization capability.

The predicted output $y^{(i)}$ and the desired output $t^{(i)}$ are the range value of each sample. Therefore, the mean square error is chosen as a loss function

$$L_{\mathrm{MSE}}(t, y) = \frac{1}{n}\sum_{i=1}^{n}(t^{(i)} - y^{(i)})^2. \tag{2}$$

A dense connection network is used as the meta-learner. The predictions of the second training data subset predicted by M different primary networks are used to train the network. The size of the total input data is $N_2 \times M$, with the number of sample N_2 and the number of features M. The number of neurons is therefore M for the input layer, and 1 for the output layer. The hidden layers are designed to have two dense connection layers with 32 neurons.

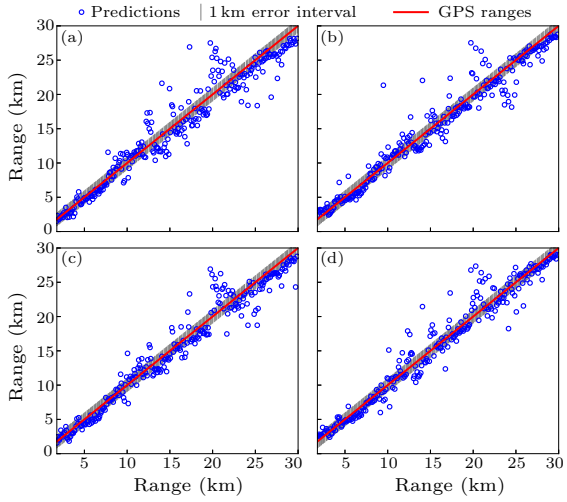


Fig. 5. Range predictions by (a) simple Conv-1D, (b) Conv-1D with residual connections, (c) Conv-1D with inception modules, and (d) Conv-1D with both inception modules and residual connections.

Suppose that the number of test data samples is N_{t}, the true range of each sample is r, and the predicted range is r_{p}. Three measures are used to quantify the prediction performance of all models, named as the mean square error (MSE), the mean absolute error (MAE) and the mean absolute percentage error (MAPE)

$$\begin{aligned} E_{\mathrm{MSE}} &= \frac{1}{N_{\mathrm{t}}}\sum_{i=1}^{N_{\mathrm{t}}}(r^{(i)} - r_{\mathrm{p}}^{(i)})^2, \\ E_{\mathrm{MAE}} &= \frac{1}{N_{\mathrm{t}}}\sum_{i=1}^{N_{\mathrm{t}}}|r^{(i)} - r_{\mathrm{p}}^{(i)}|, \\ E_{\mathrm{MAPE}} &= \frac{100}{N_{\mathrm{t}}}\sum_{i=1}^{N_{\mathrm{t}}}\left|\frac{r^{(i)} - r_{\mathrm{p}}^{(i)}}{r^{(i)}}\right|. \end{aligned} \tag{3}$$

Range predictions by different primary networks are shown in Fig. 5, and the final results as well as the absolute errors by the ensemble network are shown in Fig. 6. The performance of all models is listed in Table 1. Parts (i)–(v) correspond to the results of simple convolutional neural networks (CNN) (i), CNN with residual connections (ii), CNN with inception modules (iii), CNN with both residual connections and inception modules (iv), and the ensemble model (v). As can be seen from Table 1, part (iv) gives the best performance among the primary models, indicating that the complex structure with residual connections and inception modules are helpful to improve the performance. Using the stacking method, the prediction performance on MAPE decreases from 9.7% to 8.9%.

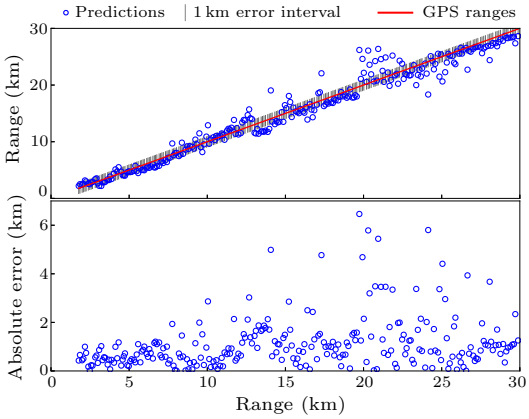


Fig. 6. Range predictions and the absolute errors by the ensemble network.

Table 1. MAPE, MAE, MSE statistics of CNN predictions.

	Number of layers	E_{MAPE} (%)	E_{MAE} (km)	E_{MSE} (km^2)
i	3	14.7	1.5	4.4
	7	11.2	1.3	4.6
ii	7	10.5	1.2	3.5
	9	11.1	1.3	3.8
iii	8	11.2	1.5	3.9
	15	10.0	1.3	3.2
iv	15	9.7	1.1	3.0
v		8.9	1.1	2.4

Figure 6 shows that the absolute errors between the range 19 and 23 km are relatively high. Degradation of the localization performance at these distances may be caused by the mismatch of the bottom parameters. Therefore, more complex bottom parameters are chosen to generate additional training data. A double-layer bottom model is used to increase training set with $c_{\mathrm{b1}} = 1548\,\mathrm{m/s}$, $\rho_{\mathrm{b1}} = 1.53\,\mathrm{g/cm^3}$, $\alpha_{\mathrm{b1}} = 0.2\,\mathrm{dB}/\lambda$, $H = 35\,\mathrm{m}$, $c_{\mathrm{b2}} = 1600\,\mathrm{m/s}$, $\rho_{\mathrm{b2}} = 1.69\,\mathrm{g/cm^3}$, $\alpha_{\mathrm{b2}} = 0.25\,\mathrm{dB}/\lambda$, where H is the thickness of sediment.

Table 2 gives the performance of all networks trained by the new training set D_1 and D_2. Here D_1 is generated by a double-layer bottom model, and D_2 is generated by both a single-layer bottom model and a double-layer bottom model. As the networks with residual connections and inception modules have the better performance, the module shown in Fig. 4 is used in all networks given in Table 2. All the details of primary models and meta-learner are the same as

the previous models. Part (i) gives the results trained by D_1, and part (ii) by D_2. Part (iii) is the results by the new ensemble network. As can be seen from Table 2, D_1 has no significant improvement compared to the previous training set but D_2 has.

Table 2. MAPE, MAE, MSE statistics of CNN predictions. Data simulated by the single-layer bottom model and the double-layer bottom model are used to train these networks.

	Number of modules	Number of layers	E_{MAPE}(%)	E_{MAE} (km)	E_{MSE} (km^2)
i	2	15	10.4	1.3	3.8
	4	29	10.5	1.3	3.6
ii	2	15	10.4	1.3	3.3
	3	22	9.7	1.2	2.9
	4	29	9.6	1.3	3.2
iii			7.9	1.0	2.4

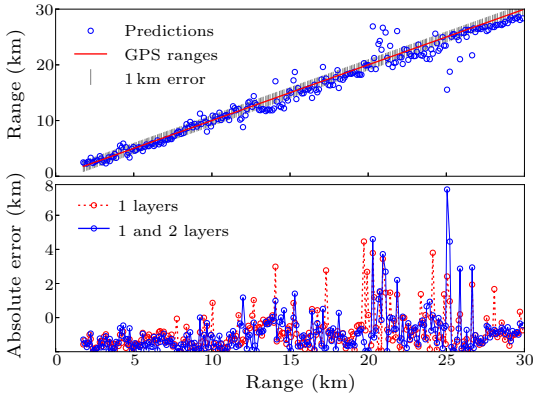


Fig. 7. Range predictions using more training data with statistical errors $E_{\mathrm{MSE}} = 2.4\,\mathrm{km}^2$, $E_{\mathrm{MAPE}} = 7.9\%$ and $E_{\mathrm{MAE}} = 1.0\,\mathrm{km}$, and the comparison of the absolute errors.

Range predictions by the new ensemble network using more training data and the comparison of the absolute errors are shown in Fig. 7. The results show that the prediction performance on MAPE decreases from 8.9% to 7.9%, and the proportion of predictions with the absolute error less than 2 km increases from 88% to 91%. It is demonstrated that one potential approach to mitigate the effects of the environmental uncertainty and to improve prediction performance is to use multiple sets of simulation data by varying the bottom parameters.

The training data are simulated based on one SSP of water column, one receiver depth and different source depths of 110–130 m. Therefore, the networks trained by such training data can only be used under the above conditions. The reason for considering only one SSP of water column is that the SSP is relatively stable for this deep water experimental data. When the sound speed of the water column changes drastically, such as the sound speed in a shallow water environment or an environment with internal waves, more training data generated from different SSPs of water column may be required.

The computational loads of each epoch are compared on Intel i7-7700 CPU. The processor base frequency is 3.60 GHz. The CPU time for each primary model is less than 392 s, and for meta-learner is less than 56 s. Therefore, the CPU time of the ensemble network mainly depends on the number of primary models.

In summary, in the direct zone of deep water, the signals received by a very deep hydrophone are used to estimate the ranges of underwater sources over long distances via ensemble deep neural networks. The ranging task is posed as a regression problem and solved by the one-dimensional convolutional neural networks with different structures. An ensemble learning method is proposed to integrate such networks. The results on the experimental data show that within the range of 30 km, the networks with inception modules and residual connections have the best performance on test data among all the primary networks. Using the stacking method, the ensemble network has the performance with $E_{\mathrm{MSE}} = 2.4\,\mathrm{km}^2$, $E_{\mathrm{MAPE}} = 8.9\%$ and $E_{\mathrm{MAE}} = 1.1\,\mathrm{km}$. For the uncertain environment, multiple sets of simulation data with different bottom parameters can be used to train the networks, and the final model has better performance with $E_{\mathrm{MSE}} = 2.4\,\mathrm{km}^2$, $E_{\mathrm{MAPE}} = 7.9\%$ and $E_{\mathrm{MAE}} = 1.0\,\mathrm{km}$.

We thank all the staff for their help in the experiment of South China Sea in 2016.

References

[1] Duan R, Yang K D, Ma Y L and Lei B 2012 *Chin. Phys. B* **21** 124301
[2] McCargar R and Zurk L M 2013 *J. Acoust. Soc. Am.* **133** EL320
[3] Michalopoulu Z H and Porter M B 1996 *IEEE J. Oceanic Eng.* **21** 384
[4] Dosso S E and Wilmut M J 2012 *J. Acoust. Soc. Am.* **132** 2273
[5] Krizenvsky A, Sutskever I and Hinton G E 2012 *Int. Conf. Neural Inf. Process. Syst.* (Nevada Am. 2–5 Dec 2012) **60** 1097
[6] Szegedy C, Liu W et al 2015 *Comput. Vision Pattern Recognit.* (Boston, USA 8–10 June **2015**) 1
[7] He K M, Zhang X Y et al 2016 *Comput. Vision Pattern Recognit.* (Las Vegas, USA 27–30 June **2016**) 770
[8] Steinberg B Z, Beran M J, Chin S H et al 1991 *J. Acoust. Soc. Am.* **90** 2081
[9] Ozard J M, Zakarauskas P and Ko P 1991 *J. Acoust. Soc. Am.* **90** 2658
[10] Niu H Q, Reeves E and Gerstoft P 2017 *J. Acoust. Soc. Am.* **142** 1176
[11] Niu H Q, Ozanich E and Gerstoft P 2017 *J. Acoust. Soc. Am.* **142** EL455
[12] Huang Z Q, Xu J, Gong Z X et al 2018 *J. Acoust. Soc. Am.* **143** 2922
[13] Zhou Z H 2016 *Machine Learning* (Beijing: Tsinghua University Press) chap 8 p 171
[14] Breiman L 1996 *IEEE Int. Workshop Mach. Learn. Signal Process.* **24** 49
[15] Jensen F B, Kuperman W A, Porter M B and Schmidt H 2011 *Comput. Ocean Acoustic* (New York: Springer) 2nd edn chap 5 p 337
[16] Wu S L, Li Z L and Qin J X 2015 *Chin. Phys. Lett.* **32** 124301
[17] Harrison C H 2011 *J. Acoust. Soc. Am.* **129** 2863