APPLICATION OF CONVOLUTIONAL NEURAL NETWORK TO OCCURRENCE PREDICTION OF RAINFALL EVENTS IN KYOTO

SUNMIN KIM (1), TSUGUAKI SUZUKI (2) & YASUTO TACHIKAWA (3)

(1) Dept. of Civil & Earth Resources Eng., Kyoto University, Kyoto, Japan, kim.sunmin.6x@kyoto-u.ac.jp
(2) Dept. of Civil & Earth Resources Eng., Kyoto University, Kyoto, Japan, suzuki.tsuguaki.58v@st.kyoto-u.ac.jp
(3) Dept. of Civil & Earth Resources Eng., Kyoto University, Kyoto, Japan, tachikawa@hywr.kuciv.kyoto-u.ac.jp

ABSTRACT

A rainfall occurrence prediction model is developed based on the convolutional neural network algorithm, which is one of the representative machine learning algorithm in image recognition. As an image data, a spatiotemporal data array is created from the time series of related atmospheric variables from multiple ground gauge observation sites. By feeding the atmospheric data array to the CNN algorithm as an input, the algorithm is trained to classify whether there will be rain in a certain lead time, such as 30 minutes or 60 minutes ahead. The trained model shows promising results with 71% of the detection ratio and 0.40 of critical success index for 30-min of prediction lead time. The high false alarm ratio is a remaining task that should be improved in further research. This paper illustrates the basic concept of the developed model and the results from modeling tests with variant model structures and input data combinations.

Keywords: Convolutional Neural Network, rainfall prediction, atmospheric variables

1 INTRODUCTION

Rainfall prediction is an essential prerequisite for flood forecasting and warning. However, it is one of the most challenging tasks in hydrological research fields. Even though there have been extensive research efforts to achieve reasonable prediction accuracy with numerical weather prediction models (e.g., Golding, 2000), radar image extrapolation models (e.g., Kim et al., 2008), or blending those two schemes (e.g., Ganguly and Bras, 2003), quantitative precipitation forecasting (QPF) still remains one of the most difficult tasks.

Recent progress in the machine learning algorithm has attracted the attention of many researchers in variant research fields. Most of the machine learning techniques are based on the artificial neural network (ANN) algorithm, and the algorithm has been strengthened with new training techniques, such as the momentum of learning rate and stochastic gradient descent with mini-batch of training data. Improved computing power and accumulated digitized database also provide good environments for efficient ANN training. In the recent hydrological research, the ANN algorithm is once again receiving more attention for variant forecasting purposes, such as river stage prediction (Kim et al., 2018) and rainfall forecasting (Shi et al., 2015). However, as far as we have surveyed and tested, QPF with ANN is not yet available for operational forecasting, and more research is required for reliable forecasting accuracy.

There are two representative standardized deep neural networks algorithms, which are convolutional neural network (CNN) and recurrent neural network (RNN). RNN is specialized in natural language analysis, and the concept is a good fit for time series analysis in the hydrological research field. CNN is specialized in image recognition, and, unfortunately, it is yet to be fully examined in the area of hydrology. In this study, we have developed and tested the CNN algorithm for qualitative rainfall forecasting. In more detail, we have built a rainfall occurrence detection model based on the CNN algorithm, which is used to forecast whether there will be rain for a certain lead time up to 1 hour, with the input of spatiotemporal characteristics of the current atmospheric conditions.

Suzuki et al. (2018) tested the first version of the rainfall-detection model and showed promising results with 0.80 of probability of detection (POD) to detect 6 mm/hr rainfall intensity in 30 minutes of prediction lead time. However, the false alarm ratio (FAR) was rather high with 0.84, which means there were 84 false alarms among 100 times of forecasting. Considering the high prediction accuracy of 0.94 including no-rainfall events, the first version of the model was successive trial, even with a low level of critical success index (CSI) as 0.15. This paper is to illustrate our recent test results with various input options and model structures to improve the accuracy of the original detection model. In this paper, section 2 describes the basic concept of the CNN algorithm and the rainfall occurrence detection model. Section 3 provides the test results with various combinations of input data and model structures. Finally, section 4 concludes the paper by summarizing the model test results.
\[ y_{i,j,k} = F \left( \sum_{p=0}^{H-1} \sum_{q=0}^{H-1} w_{p,q,k} x_{i+p,j+q} + b_k \right) \]  

Secondly, the pooling layer is to summarize and decrease the size of extracted features by taking the maximum or average values within a given window. Taking the maximum is the most common process, known as max pooling. As shown in Equation 2, max pooling is taking only one maximum value within the given area, \( U_{s,t} \), where \( s \) and \( t \) are the vertical and horizontal size of the given area, respectively. If the given area is \( 2 \times 2 \) for \( [L \times L \times M] \) of input data, then the pooling layer provides \( ([L/2] \times [L/2] \times M) \) of output. By doing the max pooling process, only significant information is delivered to the next process, and much of the unnecessary features are eliminated to improve the model performance and to reduce computing resources.

\[ y_{i,j,k} = \max_{(i,j) \in U_{s,t}} x_{i,j,k} \]  

Thirdly, a fully connected layer is to rearrange the three-dimensional information into a one-dimensional array and connect to the output layer for the classification. The SoftMax function is often utilized as an activation function in a fully connected layer to emphasize the classification task, and cross entropy function is the most commonly used error function in the output layer in this case. More details of the CNN algorithm are described by O’Shea and Nash (2015).

The biggest advantage of using CNN compared to the conventional ANN for the image recognition task is the efficiency of data processing with a small number of parameters. The CNN algorithm extracts specific features from the input image by focusing on a limited region of the image using filters, whereas the conventional ANN algorithm processes all the input information at once with a huge number of parameters. For example, if the input image has \( 1000 \times 1000 \) of pixels, the conventional ANN needs \( N \) times of \( 10^6 \) number of parameters in the first hidden layer, if the layer has \( N \) neuron nodes. Again, \( NXM \) of parameters are necessary if the second hidden layer has \( M \) neuron nodes. However, CNN only needs filter size times filtering numbers of parameters (e.g., 10 times of filtering with \( 3 \times 3 \) of filter = 90 parameters) for one convolution, no matter the input image size.
The CNN algorithm has been developed for efficient image recognition, and once it is trained, the algorithm can recognize the information by extracting necessary features from the input image. However, the application of the CNN algorithm would not be restricted to image recognition only. If any data can be arranged in the two-dimensional or three-dimensional data array, and the array has a certain meaning of information, CNN can extract the meaningful information in a sense of feature extraction.

2.2 Modeling Concept for Rainfall Occurrence Detection

Rainfall is generated when precipitable water in the air is lifted up and condensed under the necessary conditions. Air lifting has occurred when cold air is located in the upper atmosphere and warm air with moisture is located in the lower atmosphere and thus where the unstable vertical air column is located. This unstable vertical air column can be triggered by the frontal effect of atmosphere, orographic effect of air mass, heat island effect in urban area, massive atmospheric movements caused by typhoon, and so on. However, even the warm air moisture is lifted up into the upper atmosphere. As a complicated further process is necessary to be fall down to the ground as rainfall, such as cloud formation, rain drop formation, and droplet development, which are not easy to observe and difficult to simulate, even with very fine numerical weather prediction model.

In this study, we have developed and tested a new concept of a rainfall forecasting model based on the CNN algorithm. It is able to build a rainfall forecasting model to detect rainfall occurrence for a certain lead time by feeding the ground-observed atmospheric variables (e.g., rainfall amount, temperature, wind speed) as an input data and training the CNN algorithm with the given answer as to whether there is rainfall in a certain lead time (e.g., 30 minutes later or 1 hour later).

Spatiotemporal data of a certain atmospheric variable from the surrounded observation points for the target area can be formulated in a two-dimensional data array by attaching the time series of one atmospheric variable from multiple observation points. If there are multiple atmospheric variables available, by overlaying two-dimensional spatiotemporal data for several atmospheric variables, three-dimensional data array can be formulated (see Figure 2). This data array is the record of atmospheric movements for the target area, and there must be a signal of specific atmospheric movements to trigger rainfall some time later. If the signal is successfully captured in a sense of feature map in the CNN algorithm, the algorithm can be utilized as a rainfall-detection model for the trained forecasting lead time. The modeling concept has been tested in the Kyoto region, Japan, firstly, and the model showed promising results with its first version (Suzuki et al., 2018). This paper illustrates the sensitivity analysis on the developed model with various input combinations and model structures.

2.3 Input Data and Model Structure

In this study, six atmospheric variables are utilized as input data into the CNN algorithm, which are rainfall amounts, temperature, wind speed, u and v components of wind direction, and sunshine ratio for every 10-minute interval. Observation points for those input data are Mita, Miyama, Sonobe, Nose, Kyoto, Otsu, Hirakata, and Osaka, as shown in Figure 2. Time series of the input data is 80 minutes for the most recent data; thus, (8 points) X (8 time series) X (6 variables) of three-dimensional input data are prepared for each time step. The data set was prepared for the summer season in Kyoto from July to September within 2008 and 2017. The data of 2008~2015 were utilized for the training of the model, and the data of 2016 and 2017 were utilized for testing and validation of the model.

The standard type of CNN structures is adopted for the model composition with two convolutional layers, two pooling layers, and one fully-connected layer. For the convolutional process, two types of filter size (3X3 and 5X5) were employed for each convolutional layer. The input data was prepared in the three-dimensional array.
and 5X5) were tested with four different filtering numbers (8, 12, 24, and 32 times). The neuron size in the fully connected layer was also tested with several options such as N, N/2, and N/4, where N is the data size at the second pooling layer. The summary of the model setup and tested model structure is shown in Table 1, and the sensitivity analysis results with the variant model combinations will be illustrated in Section 3.2.

In our previous study, we tested three different prediction lead times for Kyoto station, which are 10, 30, and 60 minutes, and three different thresholds to decide rainfall occurrence, which are 0.5, 1.0, and 1.5 mm per 10 minutes. In this paper, only the results for 30 minutes of prediction lead time with 1.0 mm/10min of the threshold are described to illustrate the sensitivity analysis results on model structures and input data combinations.

### Table 1. Details of Model Structure and Training Setup.

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>ITEM</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PREDICTION TARGET</strong></td>
<td>Rainfall threshold (mm/10min)</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>Lead-time (min)</td>
<td>30</td>
</tr>
<tr>
<td><strong>MODEL STRUCTURE</strong></td>
<td>Convolution filter size</td>
<td>3X3, 5X5</td>
</tr>
<tr>
<td></td>
<td>Convolution filtering number</td>
<td>8, 12, 24, 32</td>
</tr>
<tr>
<td></td>
<td>Pooling window</td>
<td>2X2</td>
</tr>
<tr>
<td></td>
<td>Neuron size in F.C. layer</td>
<td>N, N/2, N/4</td>
</tr>
<tr>
<td><strong>TRAINING OPTION</strong></td>
<td>Training Scheme</td>
<td>ADAM</td>
</tr>
<tr>
<td></td>
<td>Mini-batch size</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Epoch number</td>
<td>100</td>
</tr>
</tbody>
</table>

For the evaluation of prediction results, accuracy ratio (ACC), critical success index (CSI), probability of detection (POD), and false alarm ratio (FAR) were adopted as evaluation criteria. CSI is the correct prediction ratio among the sum of model predictions and observed events, whereas POD is the ratio of correct prediction among observed events and FAR is the ratio of false prediction among the model predictions. These three criteria exclude the correct prediction for no-rainfall events, to emphasize the model performance for the rarely occurred rainfall events, and thus it is often utilized in the evaluation of the atmospheric prediction model. ACC is the overall ratio of correct prediction including rainfall and no-rainfall events.

### 3 RESULTS AND DISCUSSIONS

#### 3.1 Sensitivity Analysis on Model Structure

We have tested several combinations of filter size and filtering numbers to examine the sensitivity on the model accuracy. For two convolution layers, 8, 12, 24, and 32 times of filtering were applied with 3X3 and 5X5 filters, and N, N/2, N/4 of neuron size were tested in the fully connected layer, where N is the data size at the second pooling layers. Table 2 summarizes the test results from the best five model structures. In the table, F1 and F2 stand for the filtering number of convolution layer 1 and 2, respectively, and FC stands for the neuron size in the fully connected layer. The best result was achieved with the model structure of 12 times of filtering for both convolution layers with 5X5 of the filter size, as well as N/4 of the neuron size for the fully connected layer.

### Table 2. Test Results from the Best Four Model Structures.

<table>
<thead>
<tr>
<th>Filter size</th>
<th>F1</th>
<th>F2</th>
<th>FC</th>
<th>ACC</th>
<th>POD</th>
<th>FAR</th>
<th>CSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>5X5</td>
<td>12</td>
<td>12</td>
<td>N/4</td>
<td>0.98</td>
<td>0.71</td>
<td>0.53</td>
<td>0.40</td>
</tr>
<tr>
<td>5X5</td>
<td>32</td>
<td>12</td>
<td>N/2</td>
<td>0.98</td>
<td>0.71</td>
<td>0.53</td>
<td>0.39</td>
</tr>
<tr>
<td>3X3</td>
<td>32</td>
<td>24</td>
<td>N/4</td>
<td>0.97</td>
<td>0.74</td>
<td>0.57</td>
<td>0.38</td>
</tr>
<tr>
<td>5X5</td>
<td>12</td>
<td>12</td>
<td>N/2</td>
<td>0.97</td>
<td>0.70</td>
<td>0.57</td>
<td>0.36</td>
</tr>
<tr>
<td>3X3</td>
<td>32</td>
<td>12</td>
<td>N/4</td>
<td>0.97</td>
<td>0.72</td>
<td>0.58</td>
<td>0.36</td>
</tr>
</tbody>
</table>
The ACC value presenting the overall accuracy is 0.98, and the CSI index is 0.40, which is a very good prediction accuracy and is improved compared to our first version. POD is 0.71, meaning that the model detects 71% of the observed events successfully, whereas the FAR index is 0.53, meaning that 53% of model prediction was falsely predicted. As shown in Table 2, the best model structures do not show a clear tendency for filter size and filtering number, especially in the first convolution layer. With the both 3X3 and 5X5 of filter, model shows reasonable prediction accuracy. A slight tendency in the model performance was noticed with the neuron size in the fully connected layer. In most cases, a smaller number of neurons at the fully connected layer provides improved prediction accuracy and for a smaller number of filtering in the second convolution layer. It seems that the input data should be fully examined in the first convolution layer no matter what the filter size is, and the extracted features should be summarized as the data pass through the second convolution layer and especially at the fully connected layer.

Figure 3 shows the event numbers of the testing result and validation result from the model with the best CSI. The model detected 277 rainfall events successfully among 388 observed events (= 111+277); thus, POD was 0.71 (= 277/388) for the testing data, and the POD value had slightly decreased with the validation data of 0.65 (= 143/78+143)). FAR of the testing was 0.53 (= 308/277+308), and it slightly increased to 0.66 (= 274/143+274)) with the validation data. However, even with the high FAR values, the model successfully recognizes more than 95% of no-rainfall events correctly, and thus the ACC value is still high at 0.97 for the validation. Focusing on the rainfall events detection accuracy, CSI for the validation result decreased slightly from 0.40 (testing result) to 0.29 (= 143/[78+143+274]).

The developed model shows promising results with very high CSI considering the simplicity of the model structure and very small amounts of input data with basic atmospheric variables only from eight ground gauge stations. The current version of the developed model is unable to carry out quantitative precipitation forecasting (QPF); however, the model can be further improved with multiple output nodes to classify several ranks of rainfall amounts so as to simulate the QPF process. Further research is underway to improve the prediction accuracy with a low level of false alarm ratio.

3.2 Sensitivity Analysis on Model Structure

Another name for the data-driven model, such as ANN, is the black-box model, because we cannot understand that the process occurred in the model. It is not based on a physical concept or algorithm, and thus it is true that many researchers are not comfortable utilizing the data-driven model to understand and simulate the natural phenomena. We are introducing one of the representative data-driven models based on the CNN algorithm; however, our purpose is not limited to only developing an efficient model but also attempts to understand the most significant factors in rainfall forecasting.

To learn the most significant input factors in the developed model, a blank-input test was carried out. Here, the blank-input test is a type of model sensitivity analysis used to check the model accuracy difference when a certain piece of information is missing. As shown in Figure 4, it is able to check the accuracy changes when each input data is intentionally missing. If the model accuracy is significantly changing when a certain input data is missing, the missing data are believed to have an important role in the model. Figure 3 shows the results from the blank-input test for observation points and input variables.

For the blank-input test, four representative model structures were selected and noted as combination 1 to combination 4 in the figure. Those four model structures have the same filter size with 3X3 of the filter and the same neuron size as N/4 of neurons in the fully connected layer. The differences are in filtering numbers at the convolution layers as (8, 16), (8, 32), (32, 8), and (32, 16), which are for checking the various model structures with a limited number of combinations. Note that the prediction accuracy of those four models is very stable with 0.29, 0.32, 0.31, and 0.33 of CSI during the validation.
As shown in Figure 4, no significant effect was found in the selection of observation points. Even without the information at the prediction target point, which is Kyoto in this case, the model performance had not decreased significantly. Only limited negative effects were there when the data from Kyoto station and Nose station were missing; however, the effects were negligible, even in those cases.

A noticeable accuracy decrease was found when the rainfall information was missing in the input data, and this means that the rainfall information has an important role in the model compared to the other atmospheric variables. When the rainfall information was intentionally removed from the original input data set, the model accuracy significantly decreased with a CSI of 0.10, mostly because of lower POD compared to the original accuracy. However, when the model is trained without rainfall information from the beginning, the model performs relatively well and shows reasonable prediction accuracy. It means that the rainfall information in the input data set is serving an important role among atmospheric variables, but the model also performs well if the model is properly trained with other atmospheric variables.

4 CONCLUSIONS

A rainfall occurrence prediction model was developed based on the CNN algorithm by feeding the spatiotemporal information of atmospheric variables around the target area. Only with the basic atmospheric observation data, such as rainfall amount, temperature, and wind speed, it is able to detect rainfall occurrence in 30 minutes of forecasting lead time. The developed model performs very well by detecting 65% of observed rainfall events having more than 1 mm/10 min of intensity. High level of false alarm ratio (0.66 of FAR in the validation results) is the remaining job to improve model accuracy.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number 17K18903 and Korea Environment Industry & Technology Institute (KEITI) through Water Management Research Program, funded by Korea Ministry of Environment (2019002640011).

REFERENCES