2016 IEEE 9th International Conference on Cloud Computing

Application of Convolutional Neural Network to Prediction of Temperature Distribution in Data Centers

Shinya Tashiro*, Yutaka Nakamura‡, Kazuhiro Matsuda§ and Morito Matsuoka† ¶
* Graduate School of Information Science and Technology, Osaka University, Osaka 565–0871, Japan
Email: s-tasiro@ist.osaka-u.ac.jp
‡ Graduate School of Engineering Science, Osaka University, Osaka 560–0043, Japan
Email: nakamura@is.sys.es.osaka-u.ac.jp
§ NTT Advanced Technology Corporation, Kanagawa 212–0014, Japan
Email: kazuhiro.matsuda@ane.cmc.osaka-u.ac.jp
† Cybermedia Center, Osaka University, Osaka 560–0043, Japan
¶ Advanced Telecommunications Research Institute International, Kyoto 619–0288, Japan
Email: matsuoka@cmc.osaka-u.ac.jp

***Abstract*—We propose a model for predicting the temperature distribution in data centers by using a convolutional neural network (CNN). Changes in the temperature distribution depend on the local structure of the data center, such as equipment locations and server types. Although the various physical relations in a data center were modeled as a network in our previous work, there were no mechanisms for automatically extracting the structure of the data center. The use of a CNN is a technique for learning local structure adaptively, which allows learning complicated features, such as the various physical relations in a data center. We evaluate the performance of the proposed model by using actual data from an experimental data center. The evaluation indicates that the proposed model can predict 20-minute future temperature distributions over 48 locations in 0.42 ms, with a root mean square error (RMSE) of 0.96 degrees. This accuracy is a dramatic improvement over simple linear prediction models, and the accuracy is sufficient to allow for control of air conditioners on the basis of these temperature predictions.**

Keywords*—Data center, Power consumption reduction, Temperature prediction, Convolutional neural network

I. Introduction

Because of the huge number of high-performance data centers that have been built to cope with the massive demand for online services, such as social networking services and cloud services, data center power consumption has become a serious problem [1]. Consequently, there have been studies focusing on energy-efficient data centers [2]–[6]. For example, Khuller et al. proposed an energy management system whereby all tasks are assigned to the minimum number of servers and the remaining servers are shut down [7]. In [8], Iyengar et. al. proposed a control method for air conditioners based on the temperature distribution measured in the data center in order to reduce the energy consumption of the air conditioners. There are a variety of methods for reducing the power consumption of the various equipment in a data center, such as the air conditioners, power supply units, and servers [9]–[11]. However, most existing methods do not consider mutual effects between equipment.

To reduce the total energy consumption of a data center, it is important to consider the data center as a single unit. Reducing the power consumption of some particular equipment may not necessarily reduce the overall power consumption if it results in increased power consumption by other equipment. Servers and air conditioners account for the majority of the energy consumed in a data center, and their consumption levels are also strongly interdependent (Fig. 1). Therefore, cooperative control based on modeling of the data center is essential. However, data centers are not easy to model due to the interdependent relations between the various types of equipment.

Machine learning techniques, and deep learning in particular, have attracted much attention in recent years due to successful application to various ‘actual’ problems that had been regarded as difficult for several decades due to the complexity of the problem. In [12], it was shown that deep learning surpassed traditional machine learning methods using handmade feature extraction in a famous ‘natural’ image classification task [13]. It has been suggested that deep neural networks can automatically learn the important features of a problem, such as the geometrical constraints of pixels in images. For example, a deep neural network that was trained on images of humans developed neurons that responded to the human face or body [14].

In this paper, we propose a method for predicting the temperature distribution in a data center, based on a deep neural network. Prediction models are crucial for energy-

2159-6190/16 $31.00 © 2016 IEEE
DOI 10.1109/CLOUD.2016.90

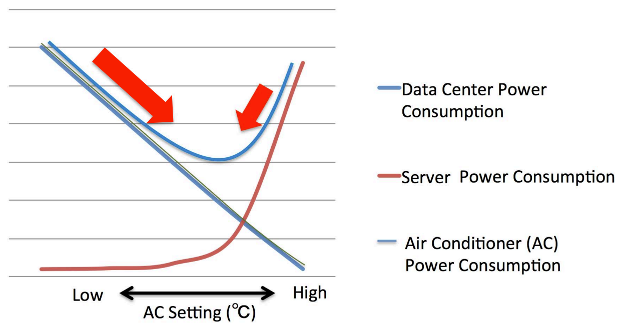


Fig. 1. Relation between power consumption and air conditioner setting. The power consumption of the data center can be reduced by turning up the air conditioner power (lower temperature setting). However, if we turn down the air conditioner power (higher temperature setting), server failure rates increase and server power consumption increases, particularly server fan power consumption.

efficient data center management. Since heat transfers depend on the physical positions of each piece of equipment and airflow, it is expected to be beneficial to employ a learning method that is able to capture this kind of geometric constraint. In our previous work [15], we proposed a network model in which the connection structure was designed by hand to match the structure of the data center. However, there were no mechanisms for automatically extracting the structure. Since convolutional neural networks (CNNs) are able to extract the geometric structure of a target, we employ this network for the prediction.

The remainder of this paper is organized as follows. We explain how the temperature distribution in data centers is predicted in Section II. In Section III, we present the evaluation index and the evaluation results of the proposed method in terms of prediction accuracy. Section IV summarizes this paper and presents future work for research.

II. Prediction of Temperature Distribution in Data Centers

In a data center, the ICT equipment and air conditioners use the majority of the power. Broadly, increasing the temperature setting of the air conditioner reduces the power consumption of the air conditioner but increases the server power consumption, as shown in Fig. 1. Since the failure rate of servers increases as the inlet air temperature increases, the temperature of the air discharged from the air conditioner is often set to a low value without considering the energy efficiency, especially when there are insufficient human resources or a lack of cost consciousness regarding resources among the data center management [16]. A mechanism for automatically managing the data center is important for overcoming this problem.

In order to minimize the total power consumption of the entire data center, it is therefore important to consider not only reducing the power consumption of each piece of equipment independently but also applying cooperative control to the many pieces of equipment. In this paper, we propose a method for predicting the temperature distribution in a data center by using machine learning techniques. This is a necessary component for cooperative control of data centers.

Our research group built a tandem structure data center(Fig. 2) for performing demonstration experiments on high-energy-efficiency data centers; this center is called the Keihanna data center [17]. Unlike conventional data centers, the cold air flowing into the cold aisle is heated twice by server racks A and B. The exhaust air is thus hotter than in a conventional data center, and the exhaust heat can be reused, such as for air conditioning of neighboring offices. Since hot air is an efficient heat source in this energy reuse system, cooperative control of the data center equipment is important for improving the efficiency. By employing a CNN, it is expected that the geometric constraints on the heat transfer caused by the physical properties of the data center structure will be able to be automatically extracted.

A. *Proactive control in the data center*

Our research group has proposed proactive control of operational settings, such as task assignment to servers and air conditioners, as a method for reducing the total power consumption in the data center. Proactive control employs the following procedure to determine the operational settings in the data center. First, many candidate operational settings are generated. Second, the temperature distribution and power consumption of the data center for the given candidate operational settings are predicted. Finally, from among operational settings that give a tolerable temperature distribution, the one with the minimum total power consumption is selected. The selected operational settings are executed in the data center. The goal is that the total power consumption is minimized in accordance with the total task. Several equipments with each response time are implemented in the data center. Therefore, the proactive control is required to operate optimum condition to minimum the power consumption. When the process for the incoming task is scheduled, or the task pattern is obvious in advance from past cases, the proactive control is most effective.

B. *Data obtained in the data center*

A sensor system is installed in the data center to measure the state of the data center. The sensor values are recorded for training the prediction model and are also used for management of the data center during operation. The cyan boxes in Fig. 2 indicate temperature sensors placed at the air outlets of server racks A and B. Although the resolution is not particularly high, the sensor values indicate the temperature distribution of exhaust air next to each rack.

There are 150 and 206 servers installed in server racks A and B (conceptually represented by the yellow boxes in Fig. 2), and operational conditions such as the CPU usage of each server are also recorded. These values give an indication of the load distribution on the servers, i.e., the heat source distribution.

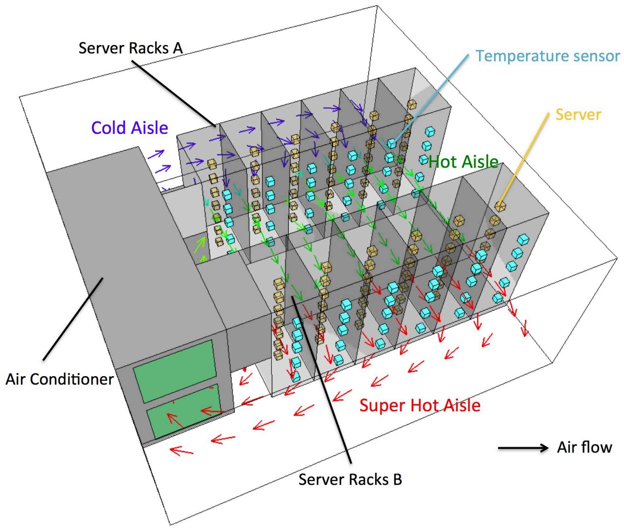


Fig. 2. Arrangement of equipment and airflows in the Keihan-na Data Center

The air conditioner temperature settings are also recorded at a sampling interval of 1 minute. The obtained data are linearly interpolated over time to fill in missing data values and to obtain a consistent sampling rate.

C. Lattice-based representation of distributions

We assume that the measured distributions can be summed into value distributions over two-dimensional planes. For the sake of simplicity, we assign a plane to each aisle and server rack, and assign nodes to the grid lattice points of each plane. The nodes on the server rack planes represent server CPU usage, while the nodes on the aisle planes represent temperature. The distributions are thus approximated into a lattice-based representation.

Index numbers are assigned to each plane, with plane 1 assigned to the cold aisle, plane 2 assigned to server rack A, plane 3 assigned to the hot aisle, plane 4 assigned to server rack B, and plane 5 assigned to the super hot aisle (represented by the yellow and blue planes in Fig. 3).

The value of node (i, j) on plane k at time step t is denoted by $x^{(k)}_{i,j}(t)$, and the value distribution of plane k is a vector of these values:

$$\boldsymbol{x}^{(k)}(t) = \ [x^{(k)}_{1,1}(t), \quad \ldots, \quad x^{(k)}_{i,j}(t), \quad \ldots, \quad x^{(k)}_{M,M}(t)]^\top, \tag{1}$$

where $M \times M$ is the number of nodes on each plane.

The values of the temperature sensors on the outlet side of server rack B are represented by:

$$\boldsymbol{y}(t) = \ [y_1(t), \quad y_2(t), \quad \ldots, \quad y_{48}(t)]^\top. \tag{2}$$

The prediction model is conceptually written as a regression model f as follows:

$$\hat{\boldsymbol{y}}(t + 20) = f(X(t)), \tag{3}$$

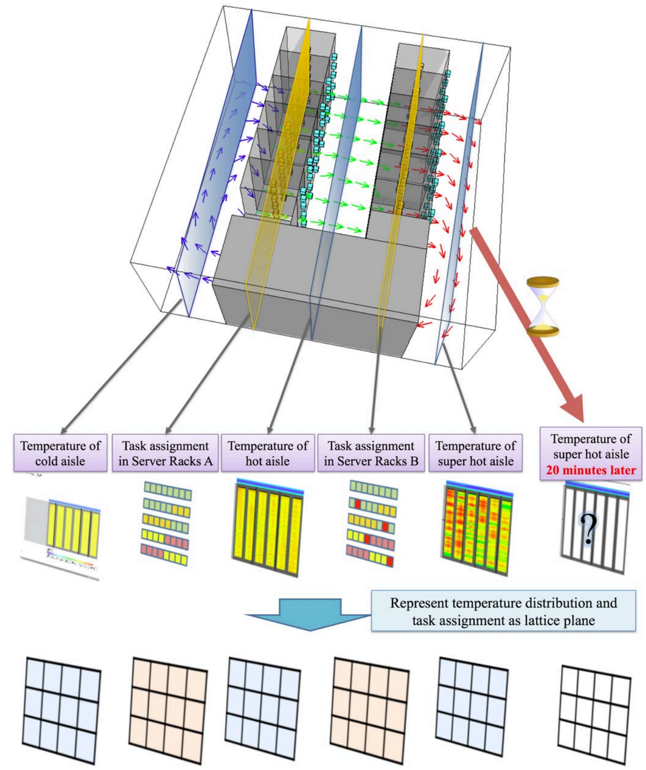


Fig. 3. Lattice-based representation of the data center

where the input $X(t)$ consists of the sensor values and the operational setting of the data center. Strictly speaking, the input $X(t)$ must include all information about the data center such as the air flow at each point and the temperature of each component, but in this research it is assumed that the sensor values and their temporal history contains enough information for making predictions. Since the temperature distribution does not change immediately, our model predicts the temperature 20 minutes in the future. The purpose of the learning is to obtain a good approximation of this function, but it is not easy since the input of the function is a high-dimensional vector and f may be a complicated non-linear function.

We thus employ a neural network model, more properly a convolutional neural network (CNN) [18], to approximate the function as:

$$\hat{\boldsymbol{y}}(t + 20) = f_w(\hat{X}(t); w), \tag{4}$$

where w is the weight parameter vector of the CNN. Furthermore, the input to the network is defined as:

$$\begin{aligned} \hat{X}(t) = \{&\boldsymbol{x}^{(1)}(t), \boldsymbol{x}^{(2)}(t), \boldsymbol{x}^{(3)}(t), \boldsymbol{x}^{(4)}(t), \boldsymbol{x}^{(5)}(t), \\ &\boldsymbol{x}^{(1)}(t-10), \boldsymbol{x}^{(3)}(t-10), \boldsymbol{x}^{(5)}(t-10)\}. \end{aligned} \tag{5}$$

In Fig. 3, $\hat{\boldsymbol{y}}(t+20)$ is represented by the white lattice shaped plane on the right, and $\hat{X}(t)$ are represented by the other blue and yellow planes. The weight parameter is optimized to

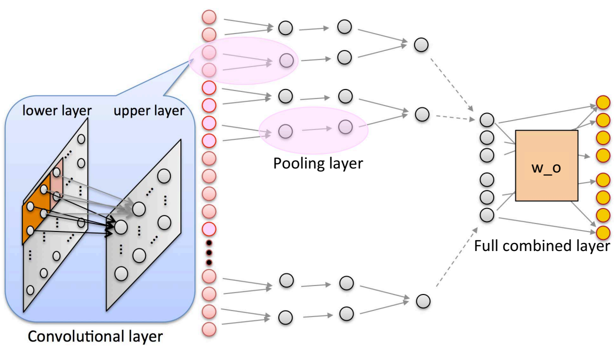


Fig. 4. Structure of CNN

minimize the error between the estimated and observed values as:

$$w = \arg\min_{w} E(w), \tag{6}$$

where

$$\begin{aligned} E(w) =& 1/T \sum_{t=1}^{T} (\boldsymbol{y}(t+20) - \hat{\boldsymbol{y}}(t+20))^2 \\ =& 1/T \sum_{t=1}^{T} (\boldsymbol{y}(t+20) - \hat{f}_w(\hat{X}(t); w))^2, \end{aligned} \tag{7}$$

and T denotes the sample size of training data, and the training data is collected from the data center in operation.

D. Convolutional neural network

Figure 4 shows the structure of CNN. CNN is a type of multi-layer perceptron (MLP). The difference between a traditional MLP and CNN is the connection structure of the network. A traditional MLP only has a small number of fully connection layers, but CNN has two types of discriminative layers with deeper layers.

The first layer is the convolutional layer. This layer has local connections and shared weights. The neurons of the successive layers then act as a filter. The network then learns features such as geometric structures through training.

The second layer is the pooling layer. This layer collects outstanding features. Because of the existence of this layer, the network is trained so that the output of the network is translation invariant.

A CNN comprises multiple repetitions of the convolutional layer and pooling layer. CNN generally have a few fully connected layers at the end that go to the network connection layer. CNN are thus able to learn various local structures automatically.

In our proposed method, the prediction model is composes of CNN that has a channel for each plane. CNN is technique that can learn local structure adaptively and can learn complicated features. Therefore, it is expected that CNN will be able to automatically learn the complicated interdependence between the nodes of the network.

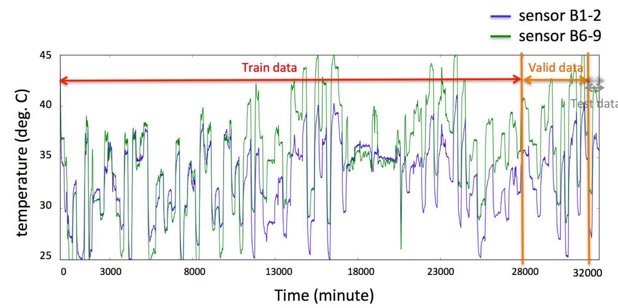


Fig. 5. Data partition

E. Structure of the CNN for temperature prediction

The input of the network consists of 8 channels of 2-dimensional arrays and the output is a 48-dimensional vector. The first layer is a convolutional layer with 32 2×2 filters. The second layer is max pooling layer with 2×2 pooling size and 2×2 pooling stride. The third layer is a convolutional layers with 16 4×4 filters. The fourth layer is max pooling layer with 2×2 pooling size and 2×2 pooling stride. The last layer is a fully connected layer with a linear activation function. The total number of layers is 5.

We used stochastic gradient descent (SGD) to train the network with a learning rate of 0.0001 and batch size of 100. The training was stopped when $E(w)$ was not updated in 70% of the last 10 epochs.

III. Experiment

A. Dataset and evaluation environment

The performance of the proposed method was evaluated using data obtained from the data center during operation. Operational data includes the value of temperature sensors, CPU usage of the servers, and air conditioner setting temperature data. The data used was 32,640 minutes of data obtained from the data center from December 9, 2015 to December 31, 2015 at 1 minute intervals. We used the first 28,000 minutes of data as training data for the model, the next 4,000 minutes for validation, and the last 500 minutes for testing.

There are 48 temperature sensors on the outlet side of server rack B, denoted by a combination of the rack name and the sensor position from the top of the rack. For example, the second temperature sensor from the top of rack B1 is the B1-2 sensor. Rack B1 has 3 sensors, and each of the other racks have 9 sensors. Figure 5 shows the partitioning of the temperature sensor data at sensors B1-2 and B6-9. Sensor B6-9 has the highest average temperature, and sensor B1-2 has the lowest average temperature.

We implemented the proposed model using pylearn2 [19] in the python programming language running on an Ubuntu desktop 14.04 PC with 32 GB memory, a GeForce GTX TITAN X GPU, and Intel Core i7-6700K 4GHz CPU.

TABLE I. RMSE OF CNN AT EACH SENSOR

sensor location	B1	B2	B3	B4	B5	B6
1	0.97	0.70	0.94	0.76	0.68	0.74
2	0.85	0.85	0.94	0.69	0.99	0.71
3	0.94	1.28	0.68	0.69	0.93	0.89
4	-	0.96	1.01	0.43	0.97	0.97
5	-	1.00	0.76	0.68	1.14	0.85
6	-	1.03	0.92	0.58	1.12	0.53
7	-	0.87	0.87	0.86	0.72	0.68
8	-	0.72	1.54	1.35	0.90	0.78
9	-	1.13	1.30	2.44	1.63	0.98

TABLE II. RMSE OF THE LINEAR MODEL AT EACH SENSOR

sensor location	B1	B2	B3	B4	B5	B6
1	2.52	2.06	2.11	5.04	2.34	2.18
2	2.48	3.49	2.02	1.46	2.65	2.07
3	1.84	1.67	2.14	1.83	2.27	1.67
4	-	1.89	2.60	2.49	2.25	2.34
5	-	1.80	2.14	1.49	2.02	2.73
6	-	2.10	2.30	2.28	2.12	2.25
7	-	1.77	1.77	2.00	2.07	1.15
8	-	3.04	2.32	2.16	2.59	2.75
9	-	2.38	1.57	2.23	2.77	2.38

TABLE III. RMSE OF MLP AT EACH SENSOR

sensor location	B1	B2	B3	B4	B5	B6
1	0.61	0.70	0.95	0.56	0.85	0.79
2	0.51	0.53	0.61	0.51	1.05	0.78
3	0.62	0.82	0.72	0.44	0.95	0.89
4	-	0.64	0.50	0.45	0.94	0.74
5	-	0.64	0.55	0.29	1.00	0.75
6	-	0.79	0.50	0.40	0.93	0.39
7	-	0.58	0.40	0.38	0.62	0.45
8	-	0.61	1.18	0.96	0.51	0.59
9	-	1.98	0.88	2.09	1.16	0.77

B. Evaluation metric

The proposed method is compared to a linear regression model and MLP in terms of accuracy and calculation time for learning and prediction. To evaluate the model accuracy, we calculated RMSE μ_e following Eq. (8), where T_{test} is the number of predicted data.

$$\mu_e = \sqrt{1/T_{test} \sum_{t=1}^{T_{test}} (\boldsymbol{y}(t) - \hat{\boldsymbol{y}}(t))^2} \tag{8}$$

C. Evaluation result

1) Accuracy of prediction: Tables I and II respectively summarize the RMSE for temperature predicted by CNN, Linear model and MLP. The average RMSE is 0.93 degrees in Table I, 2.24 degrees in Table II, and 0.74 degrees in Table III. CNN thus offers a dramatic improvement over a simple linear prediction model. Furthermore, according to the operational data for the air conditioners at the data center, the power consumption of an air conditioner increases by 8% when its setting temperature is increased by 1 degree. Consequently, the RMSE values of the CNN model are small enough to allow for controlling the air conditioner based on temperature predictions. Also, CNN does not improve accuracy in comparison with MLP. This is because interdependent relations is not so complicated in the test bed data center. However, when we predict temperate distribution in the data center that have more equipments and sensors, and more complicated placement, it is assumed that accuracy of CNN improve over MLP.

In Tables I, the accuracy of the sensors at the bottom of the rack tends to be low. There is an air vent under the racks in the data center, which results in an inflow of outside air to the bottom of the rack. The accuracy is likely low because the model does not consider the influence of this.

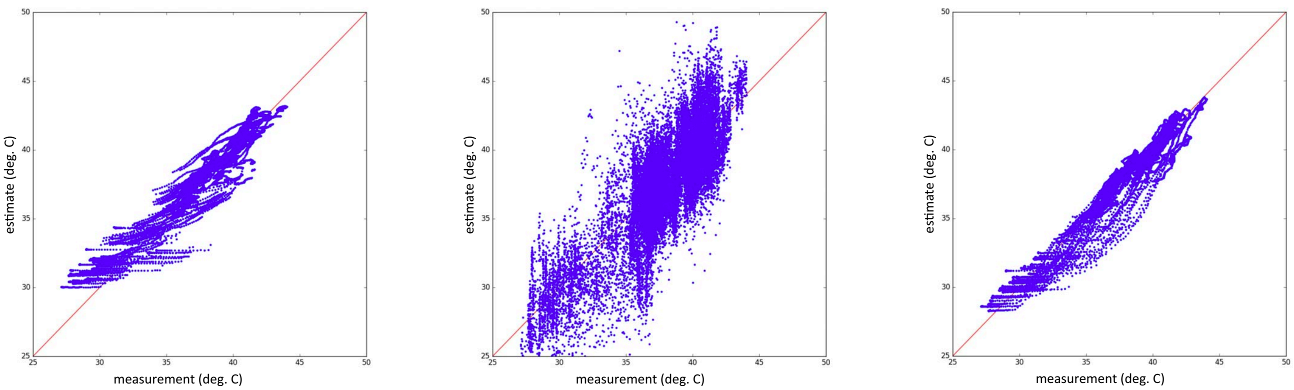
Figure 6 shows scatter diagrams of predicted and measured temperatures for the CNN and linear model. As Fig. 6 shows, the CNN offer dramatic improvement over the linear prediction model. However, there is a tendency for the low predicted values to be spread out horizontally, indicating that the model is not able to learn all features of the temperature changes. The accuracy of the model could be improved by changing network structure and training more features.

2) Calculation time for learning and prediction: The calculation time of CNN for training using the 28,000 minutes of data, and for prediction using 500 minutes of data, which corresponds to conducting 500 predictions for 48 temperature sensors, were around 4 min 49 sec and 2.1 sec, respectively. The calculation times were small because the network structure is considerably simplified. Because there is generally a trade-off between accuracy and calculation time, the most suitable structure for the model needs to be decided according to the situation. We do not need to consider calculation time for learning when we perform our proactive control of operational settings, because we collect data and train using these data for constructing prediction model beforehand. From these result, the proposed model is able to make predictions in a few seconds. We therefore conclude that our proposed method can be used for real-time temperature distribution prediction of data centers.

IV. CONCLUSION AND FUTURE WORK

We proposed a model for predicting the temperature distribution in a data center based on CNN for energy efficient data center management. Experimental results showed that our model offers accurate prediction with the proposed model able to predict the temperature distribution within 0.93 degree of RSME with several msec calculation time. For a data center energy management system, this accuracy is expected to be sufficient for elaborate control, and the calculation cost is low enough to allow for repeating the prediction many times as necessary for sampling-based control methods [20].

In future work, we plan to improve the prediction accuracy



(a) Prediction accuracy of CNN

(b) Prediction accuracy of the linear model

(c) Prediction accuracy of MLP

Fig. 6. Prediction accuracy

of the proposed model by searching for a more efficient network structure. We also intend to use this model for reducing the power consumption in actual data centers.

ACKNOWLEDGMENTS

This work was supported by a CO2 emission reduction project of the Japanese Ministry of the Environment and NICT. We thank Takasago Thermal Engineering Inc. for support with the air conditioner and aisle arrangement. We also thank Fujitsu Ltd. for supporting server management and sharing information about the characteristics of the servers.

REFERENCES

[1] J. Koomey, "Growth in data center electricity use 2005 to 2010," *A report by Analytical Press, completed at the request of The New York Times*, Aug. 2011.

[2] A. Banerjee, T. Mukherjee, G. Varsamopoulos, and S. K. Gupta, "Cooling-aware and thermal-aware workload placement for green HPC data centers.," in *Proceedings of Green Computing Conference 2010*, pp. 245–256, Aug. 2010.

[3] T. D. Boucher, D. M. Auslander, C. E. Bash, C. C. Federspiel, and C. D. Patel, "Viability of dynamic cooling control in a data center environment," *Journal of electronic packaging*, vol. 128, pp. 137–144, June 2006.

[4] A. Beloglazov, J. Abawajy, and R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing," *Future generation computer systems*, vol. 28, pp. 755–768, May 2012.

[5] B. Battles, C. Belleville, S. Grabau, and J. Maurier, "Reducing data center power consumption through efficient storage," *NetApp Technical Report*, Feb. 2007.

[6] L. Parolini, B. Sinopoli, and B. H. Krogh, "Reducing data center energy consumption via coordinated cooling and load management," in *Proceedings of HotPower 2008*, vol. 8, pp. 14–14, Dec. 2008.

[7] S. Khuller, J. Li, and B. Saha, "Energy efficient scheduling via partial shutdown," in *Proceedings of SODA 2010*, pp. 1360–1372, Jan. 2010.

[8] M. Iyengar, R. Schmidt, and J. Caricari, "Reducing energy usage in data centers through control of room air conditioning units," in *Proceedings of ITherm 2010*, pp. 1–11, June 2010.

[9] M. Norota, H. Hayama, M. Enai, T. Mori, and M. Kishita, "Research on efficiency of air conditioning system for data-center," in *Proceedings of INTELEC 2003*, pp. 147–151, Oct. 2003.

[10] C. Ge, Z. Sun, and N. Wang, "A survey of power-saving techniques on data centers and content delivery networks," *IEEE Communications Surveys Tutorials*, vol. 15, pp. 1334–1354, July 2013.

[11] D. Cavdar and F. Alagoz, "A survey of research on greening data centers," in *Proceedings of GLOBECOM 2012*, pp. 3237–3242, Dec. 2012.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of NIPS 2012 Advances in neural information processing systems*, pp. 1097–1105, Dec. 2012.

[13] "IMAGENET." available at http://image-net.org/index.

[14] Q. V. Le, "Building high-level features using large scale unsupervised learning," in *Proceedings of ICASSP 2013*, pp. 8595–8598, IEEE, May 2013.

[15] S. Tashiro, Y. Tarutani, G. Hasegawa, Y. Nakamura, K. Matsuda, and M. Matsuoka, "A network model for prediction of temperature distribution in data centers," in *Proceedings of IEEE CloudNet 2015*, pp. 251–256, Oct 2015.

[16] Yevgeniy Sverdlik, "The problem of inefficient cooling in smaller data centers." available at http://www.datacenterknowledge.com/archives/2015/12/04/the-problem-of-inefficient-cooling-in-smaller-data-centers/, Dec. 2015.

[17] T. Deguchi, Y. Taniguchi, G. Hasegawa, Y. Nakamura, N. Ukita, K. Matsuda, and M. Matsuoka, "Impact of workload assignment on power consumption in software-defined data center infrastructure," in *Proceedings of IEEE CloudNet 2014*, pp. 440–445, Oct 2014.

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, Nov. 1998.

[19] I. J. Goodfellow, D. Warde-Farley, P. Lamblin, V. Dumoulin, M. Mirza, R. Pascanu, J. Bergstra, F. Bastien, and Y. Bengio, "Pylearn2: a machine learning research library," *arXiv preprint arXiv:1308.4214*, Aug. 2013.

[20] D. Dunlap, C. Caldwell, and J. Collins, E.G., "Nonlinear model predictive control using sampling and goal-directed optimization," in *Proceedings of CCA 2010*, pp. 1349–1356, Sept 2010.