PAPER

Missing-Value Imputation of Continuous Missing Based on Deep Imputation Network Using Correlations among Multiple IoT Data Streams in a Smart Space

Minseok LEE[†a)], ***Student Member***, **Jihoon AN**[†b)], ***and*** **Younghee LEE**[†c)], ***Nonmembers***

SUMMARY Data generated from the Internet of Things (IoT) devices in smart spaces are utilized in a variety of fields such as context recognition, service recommendation, and anomaly detection. However, the missing values in the data streams of the IoT devices remain a challenging problem owing to various missing patterns and heterogeneous data types from many different data streams. In this regard, while we were analyzing the dataset collected from a smart space with multiple IoT devices, we found a continuous missing pattern that is quite different from the existing missing-value patterns. The pattern has blocks of consecutive missing values over a few seconds and up to a few hours. Therefore, the pattern is a vital factor to the availability and reliability of IoT applications; yet, it cannot be solved by the existing missing-value imputation methods. Therefore, a novel approach for missing-value imputation of the continuous missing pattern is required. We deliberate that even if the missing values of the continuous missing pattern occur in one data stream, missing-values imputation is possible through learning other data streams correlated with this data stream. To solve the missing values of the continuous missing pattern problem, we analyzed multiple IoT data streams in a smart space and figured out the correlations between them that are the interdependencies among the data streams of the IoT devices in a smart space. To impute missing values of the continuous missing pattern, we propose a deep learning-based missing-value imputation model exploiting correlation information, namely, the deep imputation network (DeepIN), in a smart space. The DeepIN uses that multiple long short-term memories are constructed according to the correlation information of each IoT data stream. We evaluated the DeepIN on a real dataset from our campus IoT testbed, and the experimental results show that our proposed approach improves the imputation performance by 57.36% over the state-of-the-art missing-value imputation algorithm. Thus, our approach can be a promising methodology that enables IoT applications and services with a reasonable missing-value imputation accuracy (80~85%) on average, even if a long-term block of values is missing in IoT environments.

key words: *missing-value imputation, deep imputation network, Internet of Things, smart space*

1. Introduction

Recently, many systems based on the Internet of Things (IoT) have been widely applied to various fields and industries such as health-care, smart buildings, smart factories, etc [1], [2]. Accordingly, many studies have been conducted on basic technologies such as context recognition [3], service recommendation [4], and anomaly detection [5], by using the data generated from IoT devices in smart spaces. These studies assume that the data generated from IoT devices are complete. However, the data generated from IoT devices can be incomplete for the following reasons: storage errors, unreliable IoT devices, unstable network status, etc. The incomplete data generated from many IoT devices may include noisy, redundant, and missing values. In particular, missing values are a very common phenomenon that causes challenges in the IoT environment. Missing values adversely affect the accuracy and reliability of IoT applications such as context awareness, real-time decision-making, as well as service recommendation and anomaly detection.

In many smart spaces with numerous IoT devices, the smart space can provide more diverse pervasive computing services to users by facilitating interactions among users and IoT devices, while the complexity of the missing-value pattern increases exponentially according to the number of IoT devices. To realize various IoT applications and services effectively, imputing missing values as accurately as possible is an indispensable prerequisite. For accurate missing-values imputation from a smart space with multiple IoT devices, the smart space system must obtain enough complete data from many IoT data streams, and these data can then be used to learn and recover the missing values by a machine learning system.

Conventionally, missing-value types can be divided into three categories according to the missing-value randomness [6]: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). With a MCAR mechanism, a probability exists that missing values are not related to any other variables exists, regardless of whether the variables are included or not included in the data or model (e.g. data are missing from a coding error). With a MAR mechanism, missing-values are related to other variables in the data. MNAR involves missing values that come neither under MAR nor MCAR mechanisms. They can also be classified depending on missing-value patterns [7]: Monotone Missing Pattern (MMP) and Arbitrary Missing Pattern (AMP). With MMP, when an individual is missing data at one time-point, values for all subsequent time-points are also missing for that individual. With AMP, there is no set structure for which variables have missing values.

Interestingly, a continuous missing-value pattern [8]

Manuscript received July 18, 2018.
Manuscript revised September 23, 2018.
Manuscript publicized November 1, 2018.

[†]The authors are with the School of Computing, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon, 34141 Korea.
a) E-mail: booo@kaist.ac.kr
b) E-mail: gjhui0191@cs.kaist.ac.kr
c) E-mail: yhlee@cs.kaist.ac.kr

DOI: 10.1587/transinf.2018EDP7257

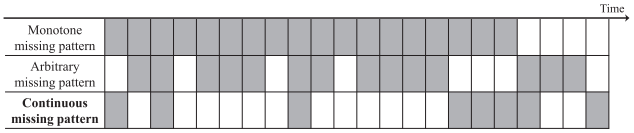


Fig. 1 Missing-value patterns. Gray parts represent observed data, whereas empty parts indicate missing values.

that has never been dealt with before in an IoT domain was found by analyzing the real IoT dataset collected from eight data streams of the IoT devices over seven months in a real smart space environment. We constructed a testbed of a smart meeting room on our campus with various IoT devices installed for context recognition such as have a meeting, watching a movie, eating a meal or taking a rest in single smart space.

As shown in Fig. 1, this pattern has the properties that combined the AMP and the MMP with the MAR mechanism. It includes characteristics that are quite different compared to the existing missing-value patterns in terms of the time in which the missing-values occur consecutively. Based on the analysis of eight IoT data streams in a smart space, we have confirmed that blocks of continuous missing-values over a few seconds and up to a few hours have approximately 40% missing rates in the total dataset. In conclusion, a *continuous missing pattern* in this paper defined as *a missing value pattern in which not only single missing value but also block of continuous missing values occur arbitrarily and repeatedly in an IoT data stream*.

Hence, a continuous missing pattern is a critical factor affecting the availability, accuracy, and reliability of IoT applications and services; yet, it cannot be addressed by the existing missing-value imputation methods [8]–[14] that use the near value of the missing-value to impute simple missing-value patterns and maximum 30% missing rate in a single data stream without considering the necessity of missing-value imputation of a continuous missing pattern in an IoT environment. This is an extraordinarily difficult problem because the time over 1 min is extremely long in a data stream. Further, the complexity of the missing pattern can be higher when considering the IoT applications that use all the data streams of multiple IoT devices in a smart space. This is because the continuous missing pattern may occur in only one data stream or it may occur in several data streams simultaneously. As a first step for the imputation of a continuous missing pattern, we consider that the continuous missing pattern may occur in only one data stream among multiple IoT data streams in single smart space.

In our initial endeavor to solve the imputation of the continuous missing pattern in a smart space with multiple IoT devices, we deliberate that even if the missing values of the continuous missing pattern occur in one data stream, missing-value imputation is possible through learning together other data streams correlated with this data stream in the same smart space. Therefore, we focused on the relationship among multiple IoT streams in a smart space. To impute the missing values of the continuous missing pattern, we obtained and utilized the *correlation* among multiple IoT-device data streams. The *correlation* can be extracted from the temporal-spatial dependencies between the IoT-device data streams in a smart space, because some IoT devices directly affect other IoT stream data. For example, when a light is turned on, the brightness and door sensors generate different data compared to those before the light was turned on. We found at least two and at most six *correlations* for each IoT device using the Pearson correlation coefficient (PCC) [15] via eight IoT data streams in a smart space. For example, an IoT device in a smart space with multiple IoT devices can have a set of *correlation* as follows: {*light sensor correlated with door sensor, sound sensor, and brightness sensor*}.

Based on these insights and findings, we propose a deep learning-based model [16], called the deep imputation network (DeepIN) that uses the *correlation* information for the missing-value imputation of a continuous missing pattern from multiple IoT data streams in a smart space. Since multiple IoT devices generate various types of data, the typical long short-term memory (LSTM) [17] used for only time-series data may not be pertinent for our goal. Therefore, we customized the LSTM model to deal with various types of data in the same machine learning system. To be resilient to large blocks of consecutively missing values in multiple IoT data streams, the proposed DeepIN uses a deep network consisting of multiple LSTMs that are constructed according to the *correlation* information of each IoT device. Each LSTM is dedicated to the input data stream of each IoT device. For example, the multiple LSTMs of the DeepIN include the light sensor-LSTM, door sensor-LSTM, temperature sensor-LSTM, etc.

We evaluate the proposed DeepIN on our dataset collected from eight IoT devices over seven months in a real IoT testbed. Our experimental results show remarkable performance improvement in missing-value imputation, i.e., by 57.36% over the existing missing-value imputation algorithms as well as by 41.09% over a single-LSTM model. To confirm the validity of the *correlation*, we conducted additional experiments and showed that the imputation performance of the DeepIN is improved by approximately 24% compared to the case without the *correlation*. Based on the proposed model for the IoT, useful applications can be realized. For example, even if a particular sensor is broken, the context can be recognized or predicted. Further, anomalies can be analyzed and detected even when missing-values exist.

The remainder of this paper is organized as follows. Section 2 discusses the related works on missing-value imputation and their differences with our proposed approach. In Sect. 3, we analyze the characteristics of multiple IoT data streams in a smart space. In Sect. 4, we present the concept of *correlation* among multiple IoT data streams to handle the continuous missing pattern. In Sect. 5, we describe the deep imputation network model to impute the missing values of the continuous missing pattern in a smart space with a multiple-IoT-device environment. In Sect. 6,

we present our experiment results and evaluate the proposed model. Finally, we conclude and discuss future work in Sect. 7.

2. Related Work

In the last few decades, various fields of missing-value imputation have received significant attention [18]–[21]. In recent studies, missing-value imputation has attracted particular examination and has been proposed with various technologies including sensors, actuators, mobile devices, and wearable devices [22], [23]. Many researchers have proposed approaches to impute the missing values in various applications and services. However, only a few studies have dealt with the missing-value problem in an IoT environment. The majority have focused on a single data stream, especially the simple missing pattern. On the contrary, we focused on the missing-value imputation of continuous missing patterns in data streams from multiple IoT devices in a smart space. To the best of our knowledge, none have studied the missing-value imputation of continuous missing patterns in the IoT domain.

One of the most popular missing-value imputation techniques is the listwise deletion method. It simply removes the missing values in the dataset and uses the remaining dataset. Although this method can be implemented easily, it can lead to a significant bias at the MNAR pattern [24]. Another common method is to use the mean value of the total data to impute missing values [25]. This is the easiest imputation method to replace each missing-value, with the average value as the estimate of the missing-value. The research in [12] proposed a K-nearest neighbors algorithm (K-NN) based method. It defines for each sample or individual a set of K-nearest neighbors and then substitutes the missing value with a given value by averaging the values of its neighbors. This method was modified to improve the imputation efficiency, for example as the sequential K-nearest neighbors algorithm [13]. In [14], the fuzzy K-means algorithm was proposed for the knowledge discovery in the database. These latter imputation methods used the near value of the missing value, assuming a simple missing pattern and a low missing rate.

Other imputation algorithms have been presented for missing-value imputation. [26] proposed multivariate imputation by chained equation (MICE) for different types of data, depending on the parametric model specified separately for each variable and involving the other variables as predictors. MICE is flexible and can handle variables of varying data types. However, prior knowledge of tuning parameters or the specifications for a parametric model is necessary and only the MAR pattern is to be considered. In [11], a random forest-based algorithm called missForest was suggested for missing-value imputation in mixed-type data (e.g., categorical and continuous). Even though the algorithm allowed for non-parametric missing-value imputation on any kind of data, complicated missing-value patterns were not considered.

The research in [27] addressed a recurrent neural network (RNN) model for missing data in clinical time series data. Although the authors considered multivariate time series of observations, they had only concentrated on the diagnostic label classification. In [9], researchers proposed an RNN model to predict medical examination data for missing information imputation. This method has no assumptions for the high missing rate and the complex missing pattern.

The study [8] focused on continuous imputation of meteorological data streams. They imputed the current continuous missing values in a time series through repeated seasonal patterns. However, the resulting study was limited because it did not consider multiple streams but considered only imputation based on historical repeating patterns.

3. Continuous Missing Pattern in a Smart Space with Multiple IoT Data Streams

As mentioned in Sect. 1, we focus on the missing-value imputation of continuous missing pattern in multiple data streams generated simultaneously from heterogeneous IoT devices installed in a smart space. Our research can also be applied to smart spaces such as smart homes and smart buildings. In view of the missing-value imputation, the diversity and complexity of missing-value patterns are different in each smart space environment. The missing-value pattern is typically relatively simple and iterative in a smart space with a single IoT device. However, more complex and various missing-value patterns can exponentially occur in a smart space where numerous IoT devices are installed. Therefore, a rigorous analysis of data streams from multiple IoT devices is needed for the missing-value imputation of IoT applications that concurrently use data streams generated from multiple IoT devices. Thus, in this section, we elaborately analyze and obtain the characteristics of multiple IoT data streams from our IoT dataset, and define the challenging requirements arising from it.

3.1 Attributes of Multiple IoT Stream Data in a Smart Space

In a smart space with multiple IoT devices, the data stream is generated from various smart IoT actuators and sensors. In our dataset, we collected over 18 million items from eight IoT devices, such as "open/close a door" contexts from smart door-lock agents (actuator) and "turn on/off a projector" contexts from projector agents (actuator). The *item* in this paper is defined as *the number of single values observed in an IoT data stream*. These values are categorical data that contain label values rather than numeric values; hence, we used a one-hot encoding form on the categorical data. On the other hand, ambient context represents environmental conditions of the smart space, and they are generated from ambient sensors in the form of numeric value type such as temperature, brightness, humidity, and sound sensors.

Within the perspective of the missing-value imputation, we describe the attributes of the IoT data stream in a smart

space consisting of multiple IoT devices as follows:

- *Simultaneous generation and integrated utilization of data streams attribute.* In a smart space with multiple IoT devices, they can concurrently generate data streams. For example, a set of IoT data streams can be generated as follows: {*humidity, temperature, brightness, door*} or {*light, humidity, projector, sound*}. It depends on the configuration of the IoT devices in the smart space.
  A set of data streams from multiple IoT devices is used together for an IoT application. For example, a combination of information from each of the sensors such as door sensor (entrance), projector sensor (turning on a projector), and presence sensor (sitting down) can be used to recognize the context, such as "have a meeting."
- *Various value types attribute.* Each data stream of the IoT devices has its own type of value. This is because a smart space is composed of heterogeneous IoT devices, e.g., door sensor (open/close), temperature sensor (26.7°), light sensor (10lux), and projector actuator (on/off, input port). For example, the data stream values generated from multiple IoT devices may take on many types such as numeric, string, binary, ternary, etc.
- *Continuous missing pattern attribute.* Data streams of multiple IoT devices contain incomplete data such as noisy data, redundant data, and missing values. Missing values account for the largest share in incomplete data. In particular, a serious missing pattern that includes high missing rates and extremely long continuous missing values is present. This is distinct from the existing missing-value pattern such as the AMP and MMP.

We consider the attributes above for the missing-value imputation in a smart space with multiple IoT devices. The *continuous missing pattern attribute* is one of the most outstanding attributes different from other studies so far on missing-value imputation. The continuous missing pattern attribute is described in detail below.

3.2 Analysis of Continuous Missing Pattern

A continuous missing pattern has characteristics that are considerably different compared to the existing missing-value patterns in terms of the time in which the missing values occur consecutively. Figure 2 shows the average missing rate per unit time and the cumulative missing rate according to continuous missing time in eight IoT data streams. As shown in Fig. 2, the continuous missing pattern has blocks of continuous missing values over a few seconds and up to a few hours. In addition, we found the continuous missing pattern in all eight IoT data streams that make up our dataset.

To determine the scope of the length of continuous missing pattern in our research, we applied the elbow method [28] that is used to find the optimal k-value in the k-means algorithm. As depicted in Fig. 2, an elbow point

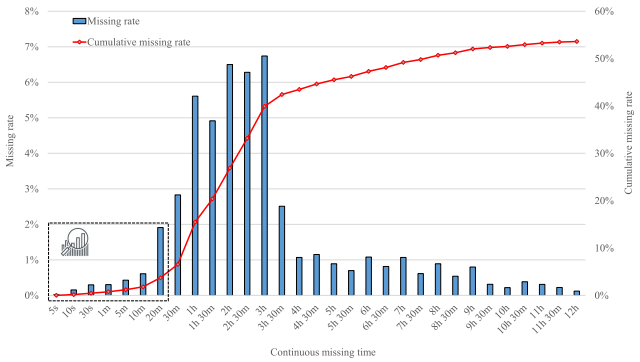


Fig. 2 Missing-values ratio of continuous missing pattern

was determined by the relationship between the period of consecutive missing and the cumulative proportion of missing data in the total data. We determined an elbow point in 3 h where approximately 40% of the total data were missing.

Formally, a continuous missing pattern is formulated as follows:

$$D = [X_1, \ldots, X_i, \ldots, X_t]^T$$
$$X_i = [x_{i1}, \ldots, x_{ij}, \ldots, x_{in}], (1 \le i \le t, 1 \le j \le n),$$

where D is a set of data streams generated from multiple IoT devices, X is a set of values on each IoT device, t is the number of IoT devices, x is the single value of each IoT data stream, and n is the size of the data stream from each IoT device. The example of continuous missing pattern can be expressed as follows:

$$D = \begin{pmatrix} 1,0,0,1,0,0,\cdots,0,1,0,1,0,\cdots,0 \\ 0,0,1,1,0,\cdots,0,0,1,1,0,1,1,0,1 \\ \vdots \\ 0,0,\cdots,0,1,0,0,\cdots,1,0,\cdots,0,1 \end{pmatrix}$$

where 0 means that the value generated from the IoT device is missing and 1 means that the value is not missing.

Consequently, the continuous missing pattern is distinct from the existing missing-value patterns in terms of continuous missing time. Based on above analysis, this is a new form of missing-value pattern combining the AMP and MMP with the MAR mechanism. This is an extraordinarily difficult problem because the time over 1 min is extremely long in a data stream. That is, since the missing-value imputation of the continuous missing pattern is not possible using only single data stream, it requires a rather considerable amount of other data streams to determine the correlation with this data stream. That is why we focused on the relationships among multiple IoT streams in a smart space.

4. Missing-Value Imputation of Continuous Missing Pattern Using Correlation among Multiple IoT Data Streams

Imputing a missing value means to recover it with a good estimate that is derived from intrinsic relationships in the underlying dataset. To solve the continuous missing pattern in

Table 1 Classification of Pearson correlation coefficient

Range of r-value	Degree of correlation
$0.7 \leq \|r\| < 1$	Very strong correlation
$0.5 \leq \|r\| < 0.7$	Strong correlation
$0.3 \leq \|r\| < 0.5$	Moderate correlation
$0.1 \leq \|r\| < 0.3$	Weak correlation
$0 \leq \|r\| < 0.1$	No linear correlation

a smart space with a multiple-IoT-device-environment detailed above, we analyzed multiple IoT data streams and used their inter-relationships. Generally, the data streams generated from multiple IoT devices in a smart space may imply temporal-spatial interdependency, because some IoT devices directly affect other IoT stream data. For example, when the light is turned on, it strongly implies the presence of users in the smart space; when a door is opened by a user, the sound and presence sensors generate different data than those before the door is opened. In this regard, we define these inter-dependencies of the IoT data streams as the *correlation* of the IoT devices in a smart space.

Thus, we consider that if we can appropriately use the *correlation* concept, the missing-value imputation of the continuous missing pattern would be possible. To represent the *correlation* among multiple IoT data streams in a smart space, the PCC [15] is utilized as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

where n is the sample size of the stream data from an IoT device. Further, x_i and $\bar{x}$ are the single samples indexed with i and the sample mean, respectively, and analogously for y_i and $\bar{y}$. The PCC measure is used to determine the strength of correlation among multiple IoT data streams. From Eq. (1), the r-value is a real value between [−1 and 1]. If the values of the data streams x and y are completely the same, the r-value is +1, and if they are completely the same in the opposite direction, the r-value is −1. A value of zero means that a linear correlation does not exist between x and y.

To determine the appropriate *correlation* criteria, we used the absolute value of r in this study, because it is only necessary to confirm a *correlation* among IoT devices rather than by the positive or negative direction. Although the PCC degree is generally classified with directions, we classified the degree of *correlation* in IoT devices without directions. In addition, we used the classification criteria of the PCC, which is generally referred to in the field of statistics [29], as shown in Table 1. Based on these criteria, we determined that the r-value of 0.1 or greater is meaningful correlation information among multiple IoT data streams.

As depicted in Fig. 3, we found 16 *correlations* among multiple IoT devices based on the absolute value of r greater than 0.1. Based on the results in Fig. 3, Table 2 shows that each of the IoT devices in our testbed has a *correlation* of maximum six from at least two. The PCC was generally utilized to determine the strength of the correlation between two data streams. That is, the PCC provides the correlation of each pair of data streams. Therefore, the results shown

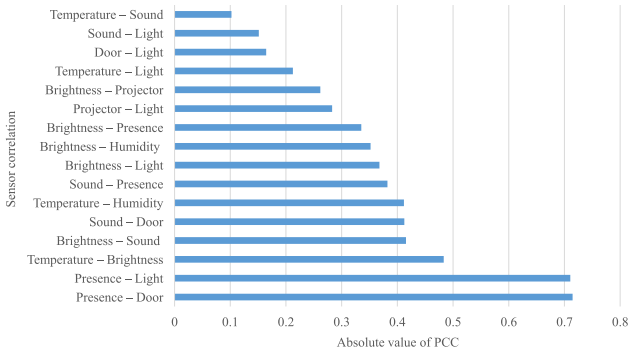


Fig. 3 Correlation among multiple IoT devices

Table 2 A set of correlation based on the PCC

IoT data stream	A set of correlation
Temperature	{Humidity, Sound, Brightness, Light}
Humidity	{Temperature, Brightness}
Sound	{Temperature, Brightness, Light, Door, Presence}
Brightness	{Temperature, Humidity, Sound, Presence, Projector, Light}
Light	{Sound, Door, Presence, Projector, Brightness, Temperature}
Door	{Sound, Light, Presence}
Projector	{Brightness, Light}
Presence	{Sound, Brightness, Light, Door}

in Table 2 can be derived because the PCC does not include causality between two IoT data streams.

Based on the results above, we deliberate that the missing-value imputation of the continuous missing pattern is possible by learning a set of data streams of IoT devices with *correlation* by using machine learning algorithms such as the deep learning algorithm. For example, the missing-value imputation of the continuous missing pattern in a brightness sensor data stream is possible through learning the set of data streams of IoT devices having the *correlation* to the brightness sensor, such as {*humidity sensor, temperature sensor, sound sensor, projector actuator, light actuator, and presence sensor*}. To utilize the *correlation* information, we designed a deep learning model for the missing-value imputation of the continuous missing pattern in a smart space with multiple IoT devices.

5. Deep Imputation Network: Deep Learning-Based Missing-Value Imputation of Continuous Missing Pattern Using Correlation

Although RNN models have been utilized in many research fields to address time-series data [30]–[32], they demonstrated weakness in sequence-inferring on large data, especially on long time-series data, and this problem is called the vanishing gradient problem [33]. Therefore, LSTM models have been suggested to solve this problem and overcome the vanishing gradient problem using the *forget gate* concept.

For accurate learning in a deep learning model, it is generally known that a large volume of data is needed. Several studies have proven that a small data volume degrades

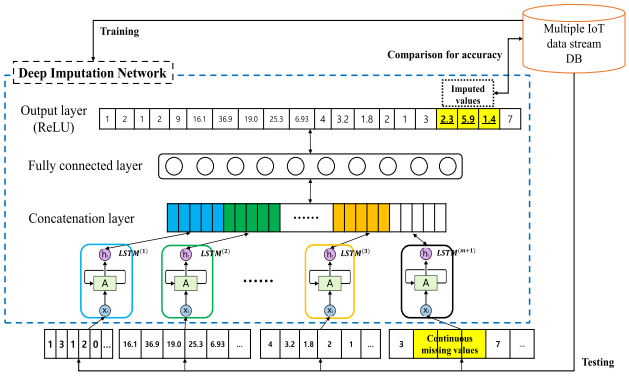


Fig. 4 Structure and overall flow of a DeepIN. m is the number of correlations for each IoT data stream. *h*, *x*, and *A* mean the output layer, input layer, and hidden layer in each LSTM module, respectively.

the performance of deep learning models [34], [35]. Similarly, in the case of extreme data status such as the continuous missing pattern defined above, many limitations will be encountered by relying on the general deep learning methods. To cope this challenging issue, we designed the model, i.e., the deep imputation network (DeepIN), that integrates the *correlation* concept into our deep learning architecture, for the missing-value imputation of the continuous missing pattern on data streams from multiple IoT devices.

As illustrated in Fig. 4, we utilized and customized the LSTM model to deal with both kinds of categorical and numeric data in the same machine learning system for the missing-value imputation of the continuous missing pattern in extremely long time-series data generated from multiple IoT devices. The output from our proposed model represents the imputed value. Our proposed DeepIN is designed to learn the correct values using the many-to-many approach when it receives each IoT stream data as the input data. To impute the missing values of the continuous missing pattern in multiple IoT data streams, the DeepIN consists of multiple LSTMs, one fully connected layer, one output layer, as shown in Fig. 4. The LSTM generate real-valued feature vectors from given IoT data stream. All the output vectors generated from the LSTMs are concatenated into one vector, which moves to the fully connected layer. Each output node value in the output layer is represented the imputed values of missing values the given data stream belongs through fully connected layer and rectified linear unit (ReLU) [36] layers.

The hyperbolic tangent function is used for the activation function of both the LSTMs and the fully connected layers since it in general provides better performance in LSTM learning than the sigmoid function [37]. For imputing the missing value for each output node, we use the ReLU activation function in the output layer of DeepIN. Therefore, given multiple IoT data streams included with continuous missing pattern, DeepIN calculates the imputed value of missing values as the most accurate value.

The DeepIN is constructed for each IoT device having a *correlation*, and the configuration of the LSTMs in the DeepIN can be varied corresponding to the *correlation* information of each IoT device. Specifically, each LSTM is dedicated to each data stream generated from other IoT devices that have *correlations* with the IoT device having the continuous missing pattern in its data stream. Each LSTM module is designed in a many-to-many method that is suitable for the purposes of missing value inference. Thus, the DeepIN comprises $m+1$ LSTM modules when its IoT device has m *correlations*. The m means the number of *correlations* for each IoT device. For example, the presence sensor has four *correlations* with the {*light actuator, door sensor, sound sensor, and brightness sensor*}; thus, the DeepIN structure for the presence sensor includes five LSTM modules. Therefore, the structure of the DeepIN is highly flexible in adaptively configuring the LSTMs according to the status of the IoT devices in a smart space. In other words, the proposed DeepIN is suitable to learn the correlation among multiple IoT data streams by adjusting the number of LSTM modules in DeepIN according to the correlation information of each IoT data stream. In addition, the input length of the LSTM determines the size of the model, in which a larger size would degrade the accuracy and effectiveness of learning. Using multiple dedicated LSTMs, our proposed model can maintain the learning accuracy on extremely long data streams from multiple IoT devices, because each LSTM has a shorter input length. In conclusion, it is resilient to continuous missing values because each LSTM module in DeepIN is dedicated to only one IoT data stream.

The objective function in DeepIN is defined as the imputation errors on the given IoT data stream, which is minimized by the training phase. Single IoT data stream are given as input to each LSTM in DeepIN. The real-valued vectors feature-extracting the input data are generated by LSTM. All the outputs generated by each LSTM are concatenated into a vector, which is forwarded into the fully connected layer. Fully connected layer is utilized into our proposed model to obtain the output of imputed value. Subsequently, the output layer is the imputed missing values contained within the real-valued vector, which is the ReLU activation function in the output layer fitted from a given concatenated vector. In the training step, the mean-squared error (MSE) is used as a loss function in the DeepIN. The errors are propagated from the output layer into the fully connected hidden layer and the weights of the hidden layer are updated by hyperbolic tangent activation function. The back-propagated error gradients are distributed into the top hidden node errors of each LSTM. Then, the weights of the LSTMs are updated via the back-propagation through time (BPTT) [38]. In the test step, missing values of continuous missing pattern is imputed by providing the previous data as sequential inputs for already trained LSTM module.

Consequently, the whole system for the missing values of the continuous missing pattern in a smart space with multiple IoT devices is accomplished via the *correlation* information of each IoT device and the DeepIN model. Through this architecture, we can impute the missing values of the continuous missing pattern efficiently.

6. Experiments and Evaluation

6.1 Dataset and Experimental Setup

To evaluate the proposed DeepIN model, data from multiple IoT devices were collected over seven months from our IoT testbed. We built a smart space with various IoT devices in the campus meeting room. Eight IoT devices were installed in a smart space, as illustrated in Fig. 5. The devices included in the our testbed were two actuators (Light, Projector) and six sensors (Temperature, Humidity, Brightness, Sound, Presence, Door). All sensors generated the data every 3 s and the actuator devices were controlled by a board PC, e.g., the Raspberry Pi.

Table 3 lists the specifications of our dataset for the experiments. The records presented in Table 3 contain 18.7 million items from January 2016 to July 2016. The raw data were generated by multiple IoT devices and stored in a MongoDB. The data of the user activity context were generated from many kinds of actuators. For example, when people performed activities for presentation in the testbed, the projector actuator generated values such as "Turn on" and the light actuator generated values such as "On". The data of the space context, such as temperature, humidity, sound, and brightness, were sensed directly from ambient sensors.

As preprocessing, the input data stream were formatted as 30 days long for each LSTM in the DeepIN. We eliminated rare noisy and redundant data. The values of actuators are categorical data that contain label values rather than numeric values and can be expressed as [1, 1, 0, 2, 0, ...] by a vector. The fixed parameters of DeepIN for the experiment are learning rate and minibatch size, which are

Fig. 5 IoT testbed

Table 3 IoT dataset specification

IoT devices	Data type	Value example	# of data
Temperature	Real value	26.7°	4,233,619
Humidity	Real value	52.9%	3,987,412
Sound	Real value	124.5dB	2,120,684
Brightness	Real value	76lux	4,981,356
Presence	Binary	1(No user)/ 2(Present)	789,107
Door	Binary	0(Outward) / 1(Inward)	700,362
Light	Binary	0(Off) / 1(On)	1,610,081
Projector	Binary	0(Off) / 1(On)	295,196

0.002 with Adam optimization method [39] and 100, respectively. Each LSTM in DeepIN has one hidden layer with 200 nodes. Fully connected layer also has one hidden layer with 500 node size. These values are determined by many experimental experiences. The data in the dataset were randomly selected into 90% training and 10% test sets. The DeepIN model was implemented based on TensorFlow [40] using a single GPU, GTX1080 with 8GB VRAM.

6.2 Performance Measure and Comparison

To measure imputation performance, we used the normalized root mean square error (NRMSE) metric. The normalization of the RMSE is as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i^{true} - x_i^{imputed})^2}{n}} = y \tag{2}$$

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}} \tag{3}$$

where n is the sample size of the IoT data stream, X_i^{true} and $X_i^{imputed}$ are the true value and the imputed value, respectively, and y_{max} and y_{min} are the maximum and minimum values of RMSE, respectively. A good performance leads toward a value of 0 and a poor performance leads toward a value of approximately 1. Compared to the RMSE, the NRMSE is less affected by the data scale from the heterogeneous IoT devices.

We compared our proposed approach with five existing missing-value imputation algorithms, as well as a single LSTM model. To apply the missing values of the continuous missing pattern to the existing missing-value imputation methods, the imputation performance was measured by increasing the consecutive missing time. Furthermore, we validated the effects of the *correlation* concept on the DeepIN model.

6.3 Imputation Performance

The goal of our proposed method was to impute the missing values of the continuous missing pattern as accurately as possible using correlation information for a DeepIN learning system in a smart space with multiple IoT devices. To evaluate the performance of our method, first, we compared the imputation accuracy of the DeepIN model with other conventional imputation algorithms and single LSTM model. Second, we tested the extent to which the correlation information affected the imputation performance of the DeepIN.

Figure 6 presents the imputation performance of the DeepIN compared to other missing-value imputation algorithms. Based on the result in Fig. 2, the criteria of continuous missing time was determined to up to 3 h, as mentioned above in Sect. 3. The experimental results used the mean value of the imputation performance of eight IoT data

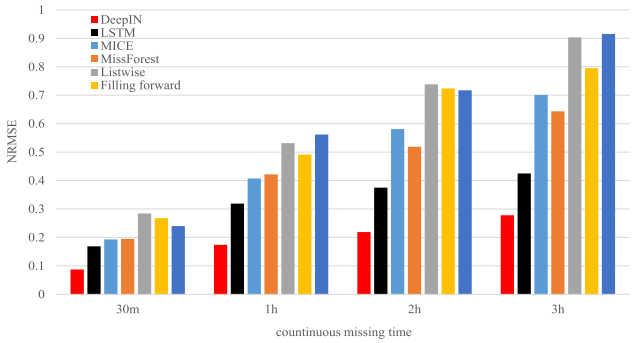


Fig. 6 missing-value imputation performance comparison between the DeepIN and the other algorithms

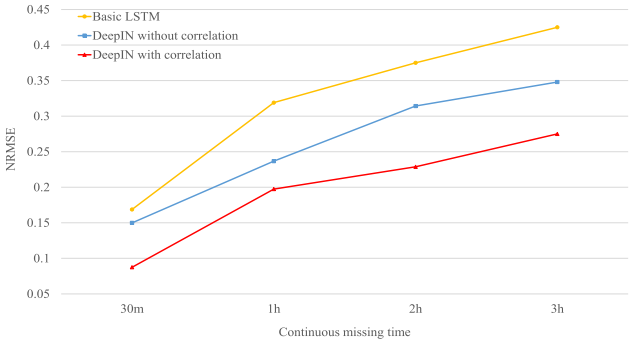


Fig. 7 Effects of correlation in DeepIN

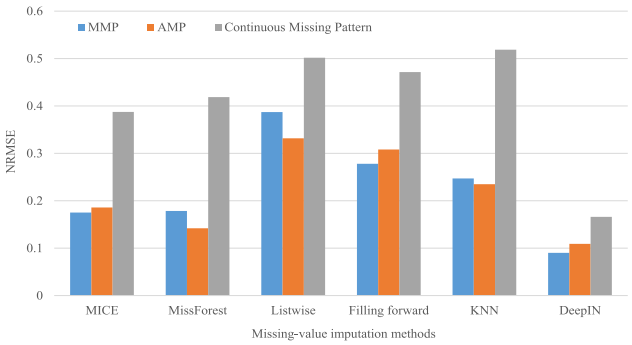


Fig. 8 Comparison of imputation performance according to missing patterns at 20% missing rates

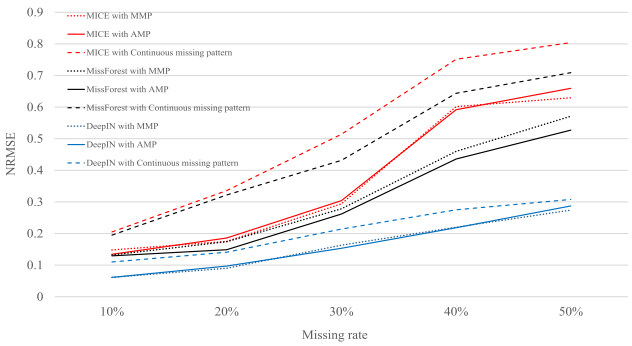


Fig. 9 Comparison of imputation performance curves according to missing patterns and missing rates

stream cases, with a continuous missing pattern on one device in each instance. As shown in Fig. 6, the DeepIN shows a 57.36% performance improvement compared to the missForest algorithm, which shows the best imputation performance among the existing algorithms. By comparing the single LSTM model with the DeepIN, our model demonstrated approximately 41.09% performance improvement. These results indicate that our DeepIN model outperforms in the imputation performance on the missing values of the continuous missing pattern generated from multiple IoT devices. From Fig. 6, we also found that the longer the period of continuous missing, the lower is the imputation performance in all methods. This is due to the proportional relationship between consecutive missing periods and high missing rates. However, the DeepIN maintains the best imputation performance while the continuous missing time is increased.

We investigated the influence of the *correlation* on imputation performance. In the case of DeepIN without correlation, all IoT data streams generated from our testbed were used as input in eight LSTM modules of the proposed DeepIN. In the case of single LSTM model, all IoT data streams were constructed as one vector and used as input values. Other conditions of the proposed DeepIN for experiments, such as the number of hidden layer in multiple LSTMs, the activation function, and the optimization function, were the same as for DeepIN with correlation. Figure 7 shows the imputation performance curves of the DeepIN with *correlation*, DeepIN without *correlation* and single LSTM model. Even though the three curves present a similar trend while the continuous missing time is prolonged, Fig. 7 confirms that the *correlation* enhances the performance of the DeepIN model by approximately 24%. This implies that our design approach with multiple LSTMs learned and extracted various *correlation* information, and is feasible and efficient for the missing values imputation of the continuous missing pattern. These results prove the validity of our proposed method based on the *correlation* information among multiple IoT data streams.

6.4 Effects of Other Missing-Value Patterns

We additionally conducted experiments and compared the results for other missing-value patterns with the continuous missing pattern case. In this experiment, we investigated the influence of AMP and MMP on the imputation performance of the proposed model. Figure 8 presents the imputation performance with the AMP and MMP case by the DeepIN and other missing-value imputation algorithms. To compare the imputation performance with that of other missing-value patterns, we experimented on various missing-value imputation methods under average 20% missing rates with the AMP, MMP, and continuous missing pattern. The missing-values of the AMP, MMP and the continuous missing pattern were generated randomly via our dataset.

As illustrated in Fig. 8, all missing-value imputation methods except the DeepIN show a large imputation perfor-

mance drop in the continuous missing pattern case. We also confirmed that the DeepIN model shows the best imputation performance on existing missing-value patterns such as the AMP and MMP. Therefore, these results indicated that the proposed method achieved an imputation performance comparable with the previous studies by focusing on simple missing-value patterns and can be applied to an IoT services of a smart space with a multiple-IoT-device-environment, such as a context recognition/prediction system.

As the auxiliary results shown in Fig. 9, we compared the imputation performance curves of the three methods, MICE, missForest, and the DeepIN, which show better imputation performance than other methods. To apply the missing values of the AMP and MMP, the imputation performance was measured by increasing the missing rates from 10% up to 50% as shown in Fig. 9. The DeepIN not only maintains the imputation performance while the missing rate is increased but also shows the best imputation performance in any missing-value pattern. These results verify that the DeepIN can appropriately cope with various types of missing patterns including the continuous missing pattern.

Our evaluations were conducted using the dataset generated from a single smart space with multiple IoT devices. Although our IoT testbed includes various IoT devices, the complexity of missing-value pattern may increases according to the number of IoT devices. Despite of that, these results show that the DeepIN achieves consistent imputation performance even if the continuous missing time or missing rate increases in various missing-value pattern. Consequently, this implies that reasonable imputation performance is predicted even in other IoT environments.

7. Conclusion

Missing-value imputation is a significant and challenging problem in the IoT environment. We considered a smart space with a multiple-IoT-device environment. The continuous missing-value pattern that has never been dealt with before includes blocks of consecutive missing values over a few seconds and up to a few hours. To impute the missing values of the continuous missing pattern, we figured out the *correlation* concept among the IoT data streams. Based on the *correlation* information, we proposed a deep learning model named DeepIN for the missing-value imputation of the continuous missing pattern, which uses multiple dedicated LSTMs for feature construction and a fully connected layer. By using the *correlation*-based structure, the DeepIN not only maintains the size of the input data from extremely long data streams, but can also cope with the continuous missing-value patterns and high missing rates of heterogeneous data streams from multiple IoT devices.

We evaluated the DeepIN on a real IoT dataset, and the experimental results showed that the proposed DeepIN model dramatically improves the imputation accuracy compared to conventional missing-value imputation algorithms. Further, we confirmed the validity of the *correlation* on the imputation performance through experiments. Furthermore, we verified the effects of other missing-value patterns such as the AMP and the MMP on the imputation performance of the DeepIN. In conclusion, our proposed model achieved an imputation performance improved by 57.36% over the state-of-the-art missing-value imputation algorithm, which enabled IoT applications and services with more than 80% missing-values imputation accuracy on average, even under the long-term continuous missing-values in IoT environments.

The DeepIN still has room for improvement despite its encouraging results. Although its imputation performance is relatively good, further improvement is still desired for the time of consecutive missing of more than 3 h. In addition, it is noteworthy that the continuous missing pattern can occur in multiple data streams. Concurrent continuous missing patterns with more IoT data streams will be considered as future work.

Acknowledgments

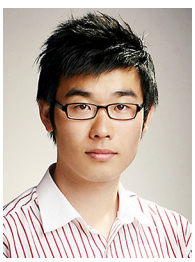
This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2016R1A2B4014688).

References

[1] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," Computer networks, vol.54, no.15, pp.2787–2805, Oct. 2010.
[2] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of things (IoT): A vision, architectural elements, and future directions," Future generation computer systems, vol.29, no.7, pp.1645–1660, Sept. 2013.
[3] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, "Context aware computing for the internet of things: A survey," IEEE communications surveys & tutorials, vol.16, no.1, pp.414–454, 2014.
[4] I. Mashal, T.Y. Chung, and O. Alsaryrah, "Toward service recommendation in internet of things," 2015 Seventh Int. Conf. Ubiquitous and Future Networks (ICUFN), pp.328–331, IEEE, 2015.
[5] S. Raza, L. Wallgren, and T. Voigt, "Svelte: Real-time intrusion detection in the internet of things," Ad hoc networks, vol.11, no.8, pp.2661–2674, Nov. 2013.
[6] D.B. Rubin, "Inference and missing data," Biometrika, vol.63, no.3, pp.581–592, Dec. 1976.
[7] Y. Dong and C.Y.J. Peng, "Principled missing data methods for researchers," SpringerPlus, vol.2, no.1, p.222, May 2013.
[8] K. Wellenzohn, M.H. Böhlen, A. Dignös, J. Gamper, and H. Mitterer, "Continuous imputation of missing values in streams of pattern-determining time series," 2017.
[9] H.G. Kim, G.J. Jang, H.J. Choi, M. Kim, Y.W. Kim, and J. Choi, "Recurrent neural networks with missing information imputation for medical examination data prediction," 2017 IEEE Int. Conf. Big Data and Smart Computing (BigComp), pp.317–323, IEEE, 2017.
[10] X. Yan, W. Xiong, L. Hu, F. Wang, and K. Zhao, "Missing value imputation based on gaussian mixture model for the internet of things," Mathematical Problems in Engineering, vol.2015, 2015.
[11] D.J. Stekhoven and P. Bühlmann, "Missforest—non-parametric missing value imputation for mixed-type data," Bioinformatics, vol.28, no.1, pp.112–118, Jan. 2011.
[12] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R.B. Altman, "Missing value estimation methods for dna microarrays," Bioinformatics, vol.17, no.6, pp.520–525, June 2001.

[13] K.Y. Kim, B.J. Kim, and G.S. Yi, "Reuse of imputed data in microarray analysis increases imputation efficiency," BMC bioinformatics, vol.5, no.1, p.160, Oct. 2004.
[14] D. Li, J. Deogun, W. Spaulding, and B. Shuart, "Towards missing data imputation: A study of fuzzy k-means clustering method," Int. Conf. Rough Sets and Current Trends in Computing, pp.573–579, Springer, 2004.
[15] K. Pearson, "Note on regression and inheritance in the case of two parents," Proc. Royal Society of London, vol.58, pp.240–242, 1895.
[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," nature, vol.521, no.7553, p.436, May 2015.
[17] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," Neural Networks, vol.18, no.5-6, pp.602–610, July–Aug. 2005.
[18] G. Krempl, I. Žliobaite, D. Brzeziński, E. Hüllermeier, M. Last, V. Lemaire, T. Noack, A. Shaker, S. Sievi, M. Spiliopoulou, and J. Stefanowski, "Open challenges for data stream mining research," ACM SIGKDD explorations newsletter, vol.16, no.1, pp.1–10, June 2014.
[19] A.W.C. Liew, N.F. Law, and H. Yan, "Missing value imputation for gene expression data: computational techniques to recover missing data from available information," Briefings in bioinformatics, vol.12, no.5, pp.498–513, Sept. 2010.
[20] E.L. Silva-Ramírez, R. Pino-Mejías, M. López-Coello, and M.D. Cubiles-de-la Vega, "Missing value imputation on missing completely at random data using multilayer perceptrons," Neural Networks, vol.24, no.1, pp.121–129, Jan. 2011.
[21] I.B. Aydilek and A. Arslan, "A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm," Information Sciences, vol.233, pp.25–35, June 2013.
[22] J. Xu, Y. Li, Y. Zhang, and A. Mahmood, "Wsn missing data imputing based on multiple time granularity," Int. J. Future Generation Communication and Networking, vol.9, no.6, pp.263–274, June 2016.
[23] G. Fortino, S. Galzarano, R. Gravina, and W. Li, "A framework for collaborative computing and multi-sensor data fusion in body sensor networks," Information Fusion, vol.22, pp.50–70, March 2015.
[24] R.J.A. Little and D.B. Rubin, Statistical analysis with missing data, John Wiley & Sons, 2014.
[25] J.L. Schafer and J.W. Graham, "Missing data: our view of the state of the art.," Psychological methods, vol.7, no.2, p.147, June 2002.
[26] S. Van Buuren and K. Oudshoorn, "Flexible multivariate imputation by mice," Leiden, The Netherlands: TNO Prevention Center, 1999.
[27] Z.C. Lipton, D.C. Kale, and R. Wetzel, "Modeling missing data in clinical time series with rnns," Machine Learning for Healthcare, 2016.
[28] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," J. Royal Statistical Society: Series B (Statistical Methodology), vol.63, no.2, pp.411–423, 2001.
[29] R. Taylor, "Interpretation of the correlation coefficient: a basic review," J. diagnostic medical sonography, vol.6, no.1, pp.35–39, Jan. 1990.
[30] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," Int. Conf. Machine Learning, pp.2048–2057, 2015.
[31] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," 11th Annual Conf. Int. Speech Commun. Assoc., 2010.
[32] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," 2013 IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP), pp.7378–7382, IEEE, 2013.
[33] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," Int. J. Uncertainty, Fuzziness and Knowledge-Based Systems, vol.6, no.02, pp.107–116, 1998.
[34] X.W. Chen and X. Lin, "Big data deep learning: challenges and perspectives," IEEE access, vol.2, pp.514–525, 2014.
[35] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," ICML Deep Learning Workshop, 2015.
[36] V. Nair and G.E. Hinton, "Rectified linear units improve restricted boltzmann machines," Proc. 27th Int. Conf. Mach. Learn. (ICML-10), pp.807–814, 2010.
[37] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," Int. Conf. Machine Learning, pp.2342–2350, 2015.
[38] P.J. Werbos, "Backpropagation through time: what it does and how to do it," Proc. IEEE, vol.78, no.10, pp.1550–1560, Oct. 1990.
[39] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
[40] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D.G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: A system for large-scale machine learning," OSDI, pp.265–283, 2016.

Minseok Lee is a Ph.D. student at the School of Computing at KAIST, South Korea. He received a B.E degree in Computer Engineering from the Myoungji University, South Korea in 2012 and a M.S. degree in Computer Science from KAIST, South Korea in 2014. His research interests include the mobile computing, Internet of Things, machine learning, pervasive computing, and missing-value imputation.



Jihoon An is a Ph.D. student at the School of Computing at KAIST, South Korea. He received a B.E degree in Department of Computer Science and Technology from the Tsinghua University, CHINA in 2007 and a M.S. degree in Department of Computer Science and Technology from the Tsinghua University, CHINA in 2010. His research interests include the Internet of Things and machine learning.

Younghee Lee received his B.S degree in Electronic Engineering from the Seoul National University in 1976, and a M.S degree in Electronic Engineering in 1980. Then, he received a Ph.D. degree in Computer Science from Universite de Technologie de Compiegne (UTC), France in 1984. He had worked for ETRI from 1984 until 1997, and had been the professor in ICU from 1998 to 2010 until the merger with KAIST. He is now a professor in the Computer Science department in KAIST, and his interest fields are the computer networks, overlay networks, pervasive computing and future Internet.