# Interpretation Support by Extracting Time Series Classification Patterns using HMM from Text-based Deep Learning

Masayuki Ando
*R-GIRO, Ritsumeikan University*
Kusatsu, Japan
email: oh23mandou@ec.usp.ac.jp

Yoshinobu Kawahara
*Institute of Mathematics for Industry, Kyushu University, and RIKEN Center for Advanced Intelligence Project*
Fukuoka, Japan
email: yoshinobu.kawahara@riken.jp

Wataru Sunayama
*School of Engineering, The University of Shiga Prefecture*
Hikone, Japan
email: sunayama.w@e.usp.ac.jp

Yuji Hatanaka
*Faculty of Science and Technology, Oita University*
Oita, Japan
email: hatanaka-yuji@oita-u.ac.jp

***Abstract*—We developed an interpretation support system for classification patterns extracted from deep learning with texts using a hidden Markov model (HMM) and verified its effectiveness. It is well known that classification patterns by using deep learning models are often difficult to interpret the reasons derived. In the proposed system, the content of deep learning results is extracted using structure of HMM, and classification patterns are provided for the system users to interpret the learned features. The system then displays learned network structures so that anyone can easily understand the learning results. In verification experiments to confirm the effectiveness of the system, based on the learning result of deep learning classifying sentences, participants were divided into two groups. One group used the proposed system, while the other group used a system that displays words with high Term Frequency-Inverse Document Frequency (TFIDF) values. Both groups were instructed to give meanings of classification patterns peculiar to each output. The results indicate that the participants who used the proposed system were able to understand the meanings of the classification patterns of deep learning with texts better than those who used the comparison system.**

***Index Terms*—interpretation support; deep learning; text mining; text classification; data visualization**

## I. Introduction

The applications of artificial intelligence (AI) systems based on deep learning have been rapidly increasing, including image recognition, automatic driving of automobiles, automatic delivery of packages using drones, and assistance in medical diagnoses.

In the U.S., there are unmanned convenience stores that allow customers to accurately identify and pay for products simply by holding them in their hands and leaving the store without going through a cash register. A major Japanese pharmaceutical company is also collaborating with an AI research institute in the U.K. to search for new compounds using AI in an attempt to greatly improve the efficiency of drug development.

There is, however, a problem with deep learning in that the criteria for prediction and classification by learning are unknown and incomprehensible to humans. This problem is especially serious in fields such as medicine and automated driving, where the reliability and safety of learning results are important. In text processing, if it is possible for humans to understand the decision criteria of deep learning, new applications of deep learning are expected such as understanding how to write good electronic medical records on the basis of the differences between electronic medical records written by newcomers and veterans or obtaining information that can be used as hints for product development by analyzing questionnaires and reviews.

We focused on the text classification problem and considered a method to extract classification patterns including time-series information from deep learning using the likelihood calculation method of hidden Markov model (HMM). The 'classification pattern' is the feature set that contributes to the classification, and is a clue to understanding the basis for the classification criteria. This method has already been treated in previous research [1] published in ACHI 2021, which has shown a certain level of effectiveness. However, prior research did not sufficiently explain the conversion of RNN weights to HMM probability distributions, and subjects' interpreted sentences were not analyzed as a result of the experiment. Therefore, this paper describes the proposed method, including the explanation of the normalization of RNN weights, and the results of the analysis of the interpreted sentences produced by the subjects using the proposed system.

The proposed method calculates the likelihood of the classification patterns (i.e., an evaluation value indicating the importance of the classification patterns) from the trained

deep learning, just as an HMM calculates the likelihood of a given observation series. Then, we build a system for users to interpret the highly rated classification patterns. Interpretation of this classification patterns allows users to understand the rationale for the classification criteria. We believe that by constructing a system that enables even novice data analysts to interpret classification patterns, we can create an environment in which users of cloud-based machine learning application programming interfaces and individuals who wish to carry out simple text mining can easily interpret the learning results.

We provide support for interpreting classification patterns as one approach to understanding the basis for classification criteria in deep learning. Therefore, rather than improving the accuracy of classification results, we focus on how humans can understand the basis of classification criteria. In addition, this approach does not provide the users with a mechanical determination of the basis for the classification criteria. It will only provide assistance to the users in finding the basis for the classification criteria. This is because if the basis of the classification criteria is judged mechanically, a new problem arises as to whether the judgment is correct.

In this study, we considered the order in which words appear in a sentence (i.e., time-series information) to be important for understanding the basis of the classification criteria. Therefore, for preprocessing words to be learned in deep learning, we use the one-hot vector format, in which each node in the input layer has a one-to-one correspondence with a word, and for deep learning models, we use recurrent neural networks (RNNs), which can also learn time series information of words. The proposed system can also be applied to long short-term memory (LSTM) and gated recurrent units, which are extensions of RNNs. However, in this case, the proposed system does not use the gate information in the cell. An HMM [2] that uses neural networks to calculate transition probabilities has also been proposed, but it does not have a framework for interpreting the training results of RNNs. The proposed system extracts classification patterns from RNN training results by referring to the likelihood calculation method of HMM.

We discuss related work in Section II. In Section III, we describe the configuration and details of our HMM-based classification-pattern interpretation support system in deep learning networks. In Section IV, we describe the experiments we conducted to verify the effectiveness of the proposed system and conclude the paper in Section V.

## II. Related Research

AI (in this context, we refer primarily to systems that use deep learning) has been playing an increasingly important role in a wide variety of situations, such as medical treatment, image judgment in automated driving systems, and automated stock trading. With the advent of cloud-based AI [3], it has become possible to use AI easily even on personal mobile devices.

There is, however, a black box problem in deep learning. Deep learning learns information through a very complex process and can make predictions and classifications with high accuracy. However, due to the complexity of the process, it is very difficult for humans to explain the decision criteria of deep learning.

Explainable AI (XAI) [4] has been gaining attention as a research field that focuses on explaining the reliability and fairness of deep learning models and understanding the decision criteria. Research on XAI began with the need to explain what has been learned to understand and trust the behavior of deep learning models [5] [6]. In fact, research has been conducted to try to explain the behavior of the model itself, such as attempting to explain the behavior on the basis of the correlation between the data and variables in the model [7], and using counterfactual conditional statements to help users understand the behavior of the model [8]. In addition to interpreting model behavior, there are also studies that focused on the stability and reliability of model behavior, such as countermeasures against malicious data [9] or evaluating model behavior and stability by applying model behavior to different logic circuits or decision trees [10] [11].

If we look at XAI research in terms of its objectives, we find that there is a large amount of research in what is called informational systems [12] [13]. An informational system is a basic method in XAI research, in which additional information is added to the output of the model, and the user can infer the validity and correctness of the AI's answer. In image processing, a method [14], [15] has been proposed for emphasizing the parts of the input image that contribute to the output. In natural language processing, however, it is difficult to apply methods used in the image processing field directly. Merely highlighting a part of the input text, as with the method called attention [16], is considered insufficient as an explanation of the classification criteria since it remains unclear what type of learning is going on inside the deep learning process.

Our aim was to develop a system to support the interpretation of classification criteria on the basis of the classification patterns including the time-series information of words by using an RNN as a deep learning model, taking the text classification problem as an example. In addition, this research exists as one of the XAI approaches, but as mentioned in the introduction, it does not provide a machine-determined basis for deep learning classification criteria. It extracts classification patterns as clues for understanding the basis of the classification criteria and encourages users to interpret them.

## III. Interpretation Support System for Classification Patterns from Deep Learning Networks Using HMM

In this section, we describe the configuration and details of our system for supporting the interpretation of classification patterns using HMMs in deep learning networks for text-based classification tasks.

### A. System Configuration

The configuration of the proposed system is shown in Figure 1. A set of texts with correct labels is used as training data,

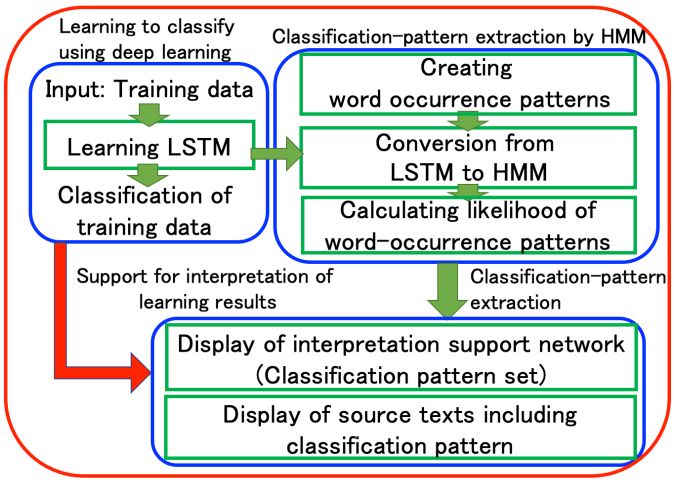


Fig. 1: System configuration

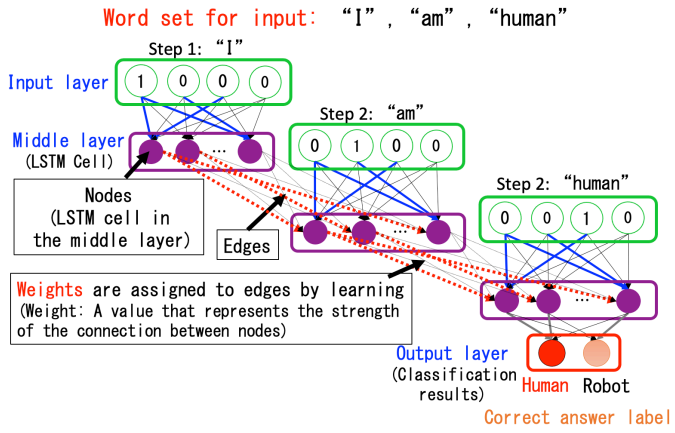


Fig. 2: Example of weighting by using RNN

and an RNN is used to classify them. The trained weighted network is then converted into an HMM, and the likelihood of word-occurrence patterns in the text set (source text) used for training is calculated. Finally, the word-occurrence patterns with the highest likelihood are displayed on the interface as classification patterns, and the user interprets the classification patterns. At this time, the user can arbitrarily set the number of classification patterns to be displayed. The system also has a source-text-display function that allows the user to refer to the source text to better understand the meaning of the classification patterns.

### *B. Training an RNN*

We used RNNs, which are generally used to learn ordering patterns in time series data, and considered situations in which they are applied to the problem of classifying text sets (Figure 2). In Figure 2, an RNN with one intermediate layer is used as an example. In an RNN, words are trained in turn at each time step, and the information is saved in the middle layer. At this time, the nodes that fire in the middle layer change with each time step, and this change corresponds to a change in state in the HMM. The reason for using an RNN is that we considered not only the type and frequency of words but also the time series of word occurrences to be important features in sentences. Also, unlike most deep learning research, we did not aim to achieve high classification accuracy but to build a system that encourages interpretation of deep learning networks, which are generally difficult to interpret.

In RNNs, a set of texts with correct labels is used as input, and the word vectors of each text (nouns, verbs, and adjectives in the text are represented by 0 and 1, respectively) are used to learn the edge weights in an RNN with one intermediate layer so that the classification accuracy is high. In this study, we extracted the weight set from the learned weighted network and applied it to an HMM to extract the classification patterns.

To interpret the learned classification patterns, it is assumed that proper training has been carried out. For this reason, we assume that the network has been trained by deep learning so that the classification accuracy of the test data in the 10-fold crossover test during training or the test data different from the training dataset is at least 90% and that the network does not contain large errors.

### *C. Creating Word-occurrence Patterns*

To improve the interpretability of the classification patterns by making them closer to the actual text, the proposed system uses the word patterns that actually appear in the text set used for training the RNN as the observation series to be fed to the RNN converted to HMM (see Section III-D for details). In this case, all word patterns that satisfy the following conditions are used as candidates. The length (number of words) of the word-occurrence patterns can be set arbitrarily by the user. However, the length of each pattern cannot be set individually.

- The words in a word pattern are the nouns, verbs, and adjectives in the source text (adjectives may be omitted in experiments).
- The words in the word-occurrence pattern are only those words that appear in at least 1% of the sentences frequency.
- The order of words in a word pattern should be based on the actual order of words in the source text.

The reason for these conditions is that we aimed to promote the interpretation of patterns that are typical among classification patterns (i.e., those with a large amount of applicable data). Therefore, even if we extract classification patterns consisting of particles or infrequently used words, it is difficult to enable interpretation of typical patterns. There is also a possibility of misinterpretation when interpreting patterns of word sequences that do not appear in the source text. To solve this problem, we use nouns, verbs, and adjectives as word-occurrence patterns, and only words with sentence frequencies above a certain threshold. The order of the words in word-occurrence patterns is also based on the time series of the actual words.

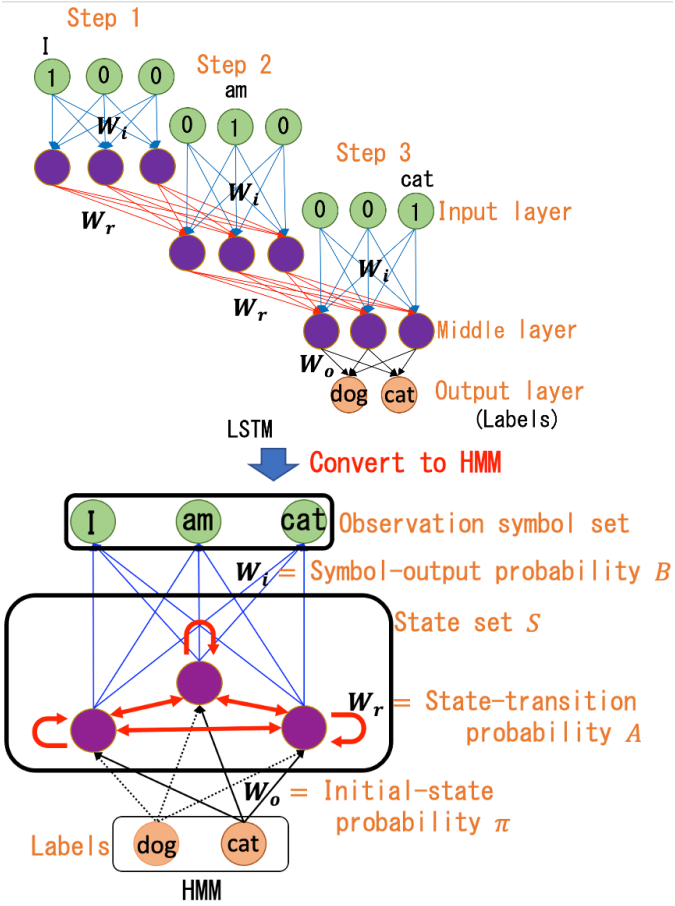


Fig. 3: Interaction between RNN and HMM

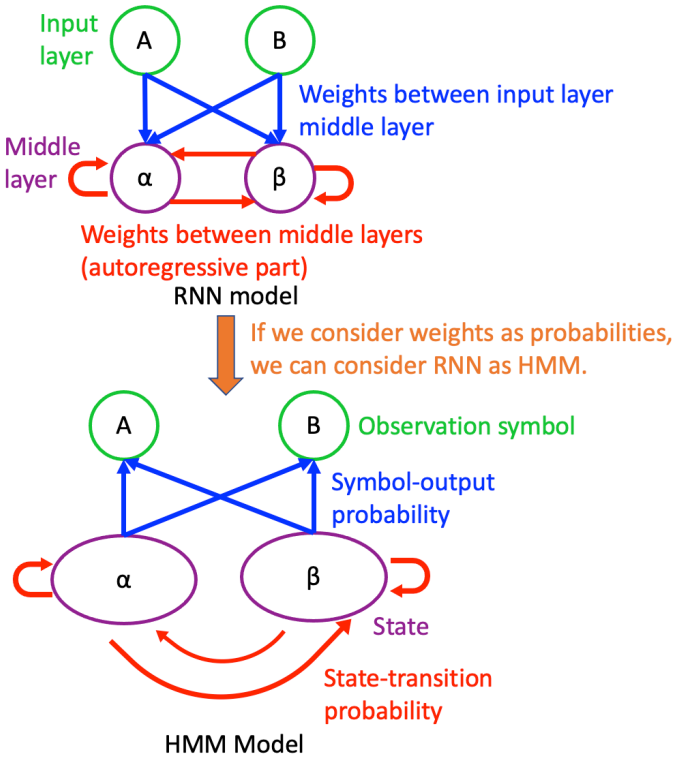


Fig. 4: Comparison of RNN and HMM

### D. Conversion of RNN to HMM

An HMM is a non-deterministic finite state automaton model with two processes: a state and an observation symbol (output). When the state is stochastic, the observation symbol is output in a stochastic manner. An HMM can calculate the likelihood, which is the value of how plausible the change in the observation symbol is. Therefore, by applying an RNN to an HMM, we can express how much a certain word-occurrence pattern contributes to the output of the RNN in terms of likelihood. Therefore, by transforming the RNN into an HMM, we can express how much a given word occurrence pattern contributes to the output of the RNN in terms of likelihood. Although we use the word 'transform' here, we are not actually changing the structure of the RNN. We are fitting the components of the RNN to the HMM so that the likelihood calculation method of the HMM can be applied to the RNN.

In the proposed system, the weighted network obtained by training an RNN is transformed into an HMM to estimate the likelihood of a word-occurrence pattern (Figure 3). The reason for using HMM is that, as shown in Figure 4, there are similarities in the structure of RNN and HMM, and we believe that the method of calculating the likelihood of HMM can be applied to RNN as well. The input layer node of the RNN is the observation symbol set of the HMM, and the middle layer node is the state set $\boldsymbol{S}$. Similarly, the weight set $\boldsymbol{W_r}$ between the time series of the intermediate layer (by recursive processing) is the state-transition probability $\boldsymbol{A}$, and the weight set $\boldsymbol{W_i}$ between the intermediate layers of the input layer is the symbol-output probability $\boldsymbol{B}$. Let the set of weights between the middle and output layers $\boldsymbol{W_o}$ be the initial state probability $\boldsymbol{\pi}$ ($\boldsymbol{\pi}$ depends on the destination to be selected at that time).

However, the weight set of the RNN does not satisfy the condition of probability. Therefore, for the weights $w_i (1 \leq i \leq N)$ that make up the weight vector $\boldsymbol{w}$ between certain intermediate layers ($N$ is the number of elements in $\boldsymbol{w}$), if the weights are negative, they are set to 0 and $w_i'$ (Equation (1)), and the value $w_i''$ is normalized so that the sum of the weights is 1 (Equation (2)).

$$w_i' = \max\{0, w_i\} \tag{1}$$

$$w_i'' = \frac{w_i'}{\sum_{s=1}^{N} w_s'} \tag{2}$$

From the above, we can treat the weight set of the RNN as the $\boldsymbol{A}$ and $\boldsymbol{B}$ of the HMM by normalizing each weight of the weight vector to sum up to 1 using Equation (2) for the weight set between each layer of the RNN.

TABLE I: Examples of extracted classification patterns

| Likelihood Rank | Extracted Classification Patterns |
|---|---|
| 1st | "Fresh cream" → "Frozen" → "Potato starch" |
| 2nd | "Fresh cream" → "Potato starch" → "Frozen" |
| 3rd | "Strawberries" → "White bean jam" → "Potato starch" |
| 4th | "Brush" → "Potato starch" → "Frozen" |
| 5th | "Chin" → "White bean jam" → "Strawberry" |

### *E. Likelihood Estimation of Word-occurrence Patterns from Trained Weighted Networks Using HMM*

This section describes the calculation of the likelihood of the set of word-occurrence patterns created in Section III-C. For the RNN weighted network converted to HMM in Section III-D, the observation sequence (the word-occurrence pattern described above) is input to $\boldsymbol{O} = \{o_1, o_2, ..., o_T\}$ ($T$ is the length of the observation sequence, i.e., the length of the word-occurrence pattern), and the number of states (the number of intermediate layer nodes) is $N$ (the state number is $i, j$), $\boldsymbol{A}$ is given by Equation (3), $\boldsymbol{B}$ is given by Equation (4), and $\boldsymbol{\pi}$ is given by Equation (5).

$$\boldsymbol{A} = \{a_{ij} | a_{ij} = P(s_{t+1} = j | s_t = i)\}(1 \leq i, j \leq N) \quad (3)$$

$$\boldsymbol{B} = \{b_{ij}(o_t) | b_{ij}(o_t) = P(o_t | s_{t-1} = i, s_t = j)\} \quad (1 \leq i, j \leq N, 1 \leq t \leq T) \quad (4)$$

$$\boldsymbol{\pi} = \{\pi_i | \pi_i = P(s_0 = i)\}(1 \leq i \leq N) \quad (5)$$

When there is a word-occurrence pattern $\boldsymbol{O}$ for a destination $x$, the initial state probability is denoted as $\boldsymbol{\pi}_x$, and the likelihood $P(O|\boldsymbol{\pi}_x, \boldsymbol{A}, \boldsymbol{B})$ is calculated using the following equation.

$$\begin{aligned} P(\boldsymbol{O}|\boldsymbol{\pi}_x, \boldsymbol{A}, \boldsymbol{B}) &= \sum_{all \boldsymbol{S}} P(\boldsymbol{S}|\boldsymbol{\pi}_x, \boldsymbol{A}, \boldsymbol{B}) P(\boldsymbol{O}|\boldsymbol{S}, \boldsymbol{\pi}_x, \boldsymbol{A}, \boldsymbol{B}) \\ &= \sum_{all s_0 ... s_T} \pi_{x s_0} a_{s_0 s_1} b_{s_0 s_1}(o_1) \cdot a_{s_1 s_2} b_{s_1 s_2}(o_2) \cdot \\ &\quad ... \cdot a_{s_{T-1} s_T} b_{s_{T-1} s_T}(o_T) \end{aligned} \quad (6)$$

Finally, the likelihood is calculated for all word-occurrence patterns using Equation (6), and the word-occurrence patterns are extracted as classification patterns that contribute to classification in the order of increasing likelihood.

### *F. Interpretation-support-network Display*

The extracted set of classification patterns in the previous section, which are strongly connected to the classification destination, is displayed as an interpretation support network with the proposed system. In this network, words are displayed as orange nodes (① in the Figure 5) and the time-series relationships between words are displayed as blue arrows (②)

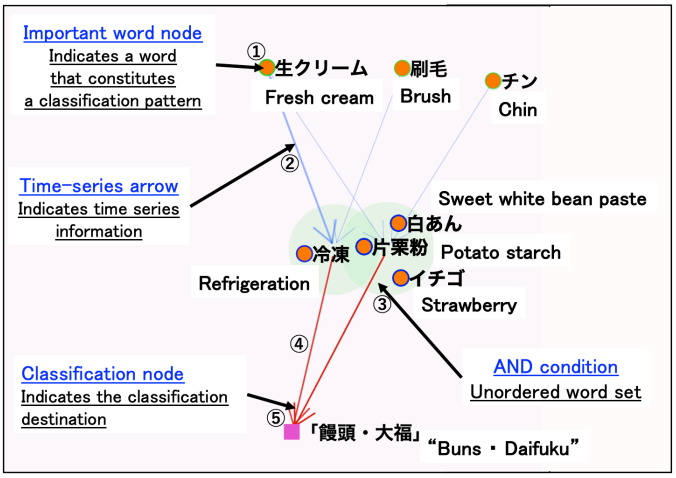


Fig. 5: Example of displayed interpretation support network

between nodes to make it easier to understand the words and the time-series relationships between words in the classification pattern. The magnitude of the likelihood is indicated by the thickness of the arrows. Furthermore, nodes with arrows in both directions are considered to have a weak time-series relationship and displayed as a single group in the green area (③). To indicate which classification pattern belongs to which destination, a red arrow (④) connecting the purple node (⑤) that displays the destination name and the last word node of the classification pattern is displayed.

The interpretation support network displayed from a set of texts on how to make five types of Japanese sweets (collected from Cookpad [17]) is shown as an example in Figure 5. The user first selects the node at the bottom of the interface where the name of the classifier (in this case, "Buns and Daifuku") is displayed for interpretation. The system first extracts the classification patterns for the selected destination name in the user's desired number in the order of likelihood. The extracted classification patterns are shown in Table I. Next, an interpretation support network is displayed, with the words of the extracted classification patterns as nodes and the time-series relations between the words as arrows. Finally, by looking at the interpretation support network, the user can find out what words and time-series relationships between words contribute to the selected classification destination and interpret the patterns. At this time, the user can use interpretation support functions such as displaying and examining any classification pattern from a group of classification patterns, and a source text display function that displays the meaning of words in the classification pattern (details are described below).

It is important to note that the classification patterns of the selected classifiers shown in the interpretation support network are only the characteristics of the selected classifiers compared with other classifiers and not the general characteristics of the word. For example, in the example shown in Figure 5, if you find the pattern "Fresh cream" → "Frozen" and "Potato

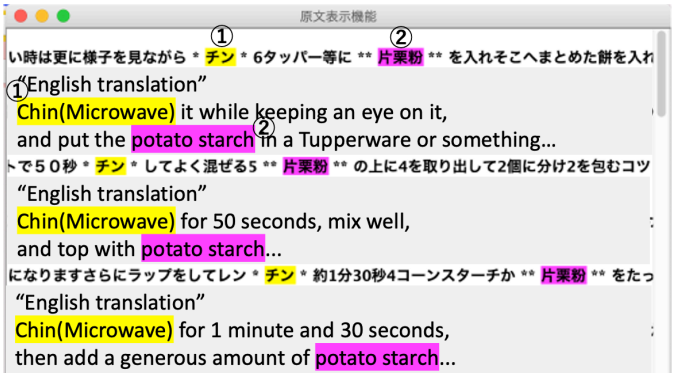


Fig. 6: Example of the source text (selecting the words “Chin” and “Potato starch” for the text about “Buns and Daifuku”)

starch”, you can interpret that the difference from the other four Japanese sweets is the way of making the“Buns and Daifuku”, such as “mixing fresh cream with frozen potato starch” or “mixing fresh cream with potato starch and freezing”. Finally, by interpreting each Japanese sweets classification pattern in the same way, one can understand the basis of the classification criteria for classifying the five Japanese sweets based on the interpretation.

### G. Function for Displaying Source Text of Classification Patterns

To interpret the classification patterns, it is difficult to understand the actual context in which the words were used from the word information alone. For this reason, the source-text-display function shows how the words in the classification pattern are actually used in the text used for training.

By selecting a word (clicking on the node) on the interpretation support network, the user can see the sentence that contains the word in the source text. Selected words are highlighted. However, for ease of viewing, we limit the number of words displayed to ten before and ten after the selected word per sentence. Up to two types of words can be selected, in which case all sentences between the words are displayed. The order in which the “selected words” occur in the displayed sentence is based on the order in which the user selected the words. Figure 6 shows an example of the source-text display of the classification pattern for five different ways of making Japanese sweets when the text “Buns and Daifuku” is used as the classification destination and the words “Chin (meaning the sound of a microwave signaling it has finished cooking in Japanese)” (① in the Figure 6) and “Potato starch” (②) are selected in order.

## IV. Experiment to Verify Effectiveness of Proposed System

In this section, we describe the experiment we conduced to verify whether participants without extensive knowledge of deep learning can interpret the classification patterns on the basis of the word-occurrence patterns output with the proposed system.

### A. Experimental Procedure

The experiment consisted of the three tasks listed in Table II, in which the participants were asked to interpret the classification patterns of sentences classified into the “output labels” specified for each task. To make the interpretation easier for the participants and facilitate the analysis of the interpretation results, we set an interpretation objective for each task. The “Tsundere” described in Table II refers to a girl who is cold and demanding at first meeting or in public, but sometimes shows kindness. In addition, “Deredere”, discussed below, refers to girls who show sweet or tender feelings toward a specific person throughout. Other details of the data used for the task are described in Section IV-B. The experiment was conducted involving 16 undergraduate and graduate students who had no extensive knowledge of deep learning, and they were divided into two groups: one using the proposed system and the other using a comparison system. The experiment was conducted without changing the members of the group because we focused only on the validity of the interpretations given by the participants to the source-text data without asking about the quality of the interpretations, which may be greatly influenced by the participants’ personalities and ways of thinking.

We used a system that extracts words specific to a specified output label by the Term Frequency-Inverse Document Frequency (TFIDF) value of the expression as the comparison system. We compared the difference between the interpretation based on the characteristic words and their combinations with the comparison system and the interpretation based on the time series of the words with the proposed system.

We asked the group using the proposed system to find words that contribute to classification (one word, combinations, and time series) using the proposed system. We asked the group using the comparison system to find the words that contributed to the classification by looking at a list of words arranged in order of TFIDF value. The TFIDF value $i$ of a word in a source text is obtained as shown in Equation (7), where $i$ is the word in the text. In addition, the source-text-display function can be used in the comparison system.

While participants may have prior knowledge of the text being tested, when interpreting, participants should consider whether the interpretation actually applies to the source text rather than their own knowledge. Therefore, we believed that the presence or absence of prior knowledge would have little effect on the experimental results.

$$\begin{aligned}\text{TFIDF}_i = {} & \text{sentence frequency for word } i \\ & \times(\log(\frac{\text{output labels num}}{\text{DF value for word } i}) + 1) \end{aligned} \tag{7}$$

The following steps of the experimental procedure were done by both groups. The number of classification patterns displayed with the proposed system was set to five, consisting of three words, in order of increasing likelihood. The number of words displayed with the comparison system was set to 15 to match the proposed system.

TABLE II: Experimental tasks given to participants and interpretive objectives

| Title | Content | Purpose of Interpretation |
|---|---|---|
| Task 1 "Character dialogues": Output label "Tsundere" | Classify the lines of characters in anime and manga with unique characteristics: Ask the students to interpret the characteristics of the lines of characters with "Tsundere" characteristics. | Assuming you are a novelist, find a pattern of word usage specific to the "Tsundere" character for your novel and give your interpretation of it. |
| Task 2 "Consumer electronics reviews": Output label "useful" | Classification of reviews about popular consumer electronics on Amazon: Ask students to interpret the characteristics of reviews with a large number of "this review was useful". | Assuming you are a reporter introducing home appliances, find the patterns of word usage specific to "helpful reviews" about popular home appliances and give your interpretation of them. |
| Task 3 "Game reviews": Output label "useful" | Categorize reviews of popular game software on Amazon: Ask students to interpret the characteristics of reviews with a large number of "this review was useful". | Assuming that you are a reporter introducing a game software, find the pattern of word usage specific to "helpful reviews" and give your interpretation of it. |

TABLE III: Deep learning data for each task

| | Character dialogues | Consumer electronics reviews | Game reviews |
|---|---|---|---|
| Number of study texts | 1500 | 3108 | 4419 |
| Number of characters per text (average) | 40 | 244 | 455 |
| Input layer nodes | 510 | 916 | 1809 |
| Intermediate layer nodes | 10 | 10 | 15 |
| Output layer nodes | 3 | 3 | 3 |
| Classification accuracy | 98.7% | 99.2% | 96.7% |

Step 1 Select the output labels to be interpreted: In the "Character dialogues" task, we targeted the lines of characters classified as "Tsundere". For the "Consumer electronics reviews" and "Game reviews" tasks, we included reviews with a rating of 4 or higher and a "Usefulness" rating of 10 or higher.

Step 2 Read the "Purpose of Interpretation" corresponding to each selected output label to understand its content.

Step 3 For the selected output, display the "Interpretation support network" and find ten features (one word, combinations, time series order, etc.) that may contribute to the output.

Step 4 Ask the user to devise an interpretation of the features in accordance with the "Purpose of Interpretation" using the source-text-display function.

### *B. Details of Experimental Data and Deep Learning Model Used*

Table III lists the deep learning data for each task used in this experiment. For the task "Character dialogues," we used a total of 1,500 dialogues with the characteristics of "Tsundere," "Deredere", and "Normal" characters, 500 each from the "Tsundere bot," "Deredere bot," and "Normal Character dialogues bot" on Twitter. For each Twitter bot, we used the top bot accounts (data-acquisition date: July 10, 2020) when we searched for "Tsundere Twitter bot," "Deredere Twitter bot," and "Character dialogues Twitter bot". For the task "Consumer electronics reviews," we used a total of 3108 reviews from the top 50 "popular consumer electronics" on Amazon [18]: 1036 each of "Useful" (4 stars or more and 10 or more "Useful people"), "Useless" (4 stars or more and 0 "Useful people"), and "Low-rated" (2 stars or less) reviews. The reason the "4 stars or more" reviews were used was that it was thought that there were some meaningful reviews and some not so meaningful reviews among the same high-rated reviews, and it was intended to give an interpretation of the results of learning to distinguish them. In the "Game reviews" task, we used a total of 4419 reviews from the top 100 "Popular game software" on Amazon: 1473 each of "Useful" (4 stars or more and 10 or more useful people), "Useless" (4 stars or more and 0 useful people), and "Low-rated" (2 stars or less) reviews. For the text data used in this experiment, we excluded in advance texts that were extremely short, such as those with only one or two words, and texts with excessively unnatural Japanese. For example, in the review text, the correctness of the content was not questioned because the purpose of this experiment was to check whether the text data could be interpreted as it is.

The experiment uses an LSTM model, which is an advanced version of an RNN, to improve the accuracy of the training. The training was done using LSTM, and the middle layer was one layer. The number of nodes in the middle layer was reduced to the extent that the classification accuracy did not fall below 95%. The learning rate was 0.1, the 11- and l2-norm coefficients were both 0.0001, and the number of trainings was 50.

### *C. Experimental Results and Discussion*

First, the breakdown of the validity of the interpretations described by the participants (participant average) is shown in Figure 7. However, this breakdown was classified by one of the authors on the basis of the following definitions.

- Reasonable interpretation (reasonable): The correctness of the content can be confirmed from the source text and meets the "Purpose of interpretation".
- Interpretation that cannot be judged as either valid or not valid (unknown): The intention of the content is not clear and cannot be judged as either valid or not valid.
- Unreasonable interpretation (unreasonable): The content of the interpretation is confirmed to be incorrect or does not meet the "Purpose of interpretation".

This classification process was performed mechanically by the author based on the following procedure. In order to avoid

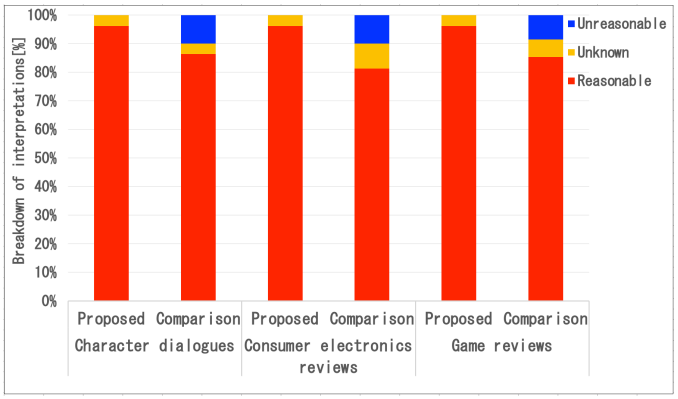


Fig. 7: Breakdown of validity of participant's interpretations (participant average)

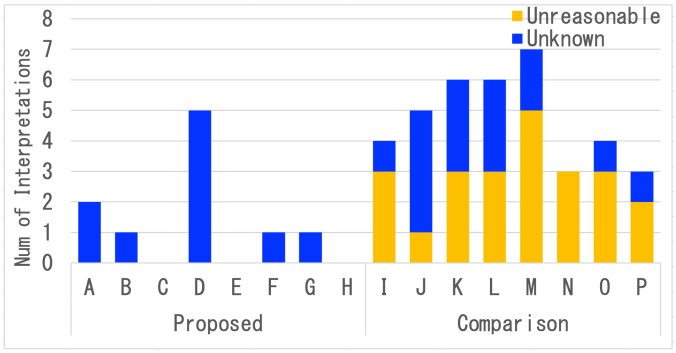


Fig. 8: Unknown and misinterpreted numbers by participant

any oversight, this process was repeated several times with a time interval between repetitions.

1) Check whether the interpretation matches the "Purpose of interpretation" in Table II. Interpretations that clearly do not meet the purpose are classified as "Unreasonable interpretations".
2) The set of source texts (ORG) in which the features (words) of interest in deriving the interpretation appear is the target of the investigation. Interpretations for which the ORG does not exist are classified as "Unreasonable interpretations".
3) If the ORG contains the content of the interpretation, it is classified as a "Reasonable interpretation"; if not, it is classified as an "Unreasonable interpretation".
4) If the meaning of the interpretation is not understood, or if there are multiple possible meanings, and it is not clear whether the interpretation is included in the ORG in 3), classify it as an "Unknown interpretation".

Table IV shows examples of interpretations that were actually classified as "Reasonable interpretation ", "Unreasonable interpretation," and "Unknown interpretation" by the above classification procedure and the reasons.

Figure 7 shows that more than 97% of the interpretations with the proposed system were classified as valid interpretations, which confirms the correctness of the proposed system. The number of interpretations that were not valid was nearly 10% in the comparison system, but 0% in the results of the proposed system. Furthermore, unknown interpretations accounted for 5 to 10% in the comparison system, but less than 3% in the proposed system. This indicates that the proposed system has clearer intentions and more valid interpretations.

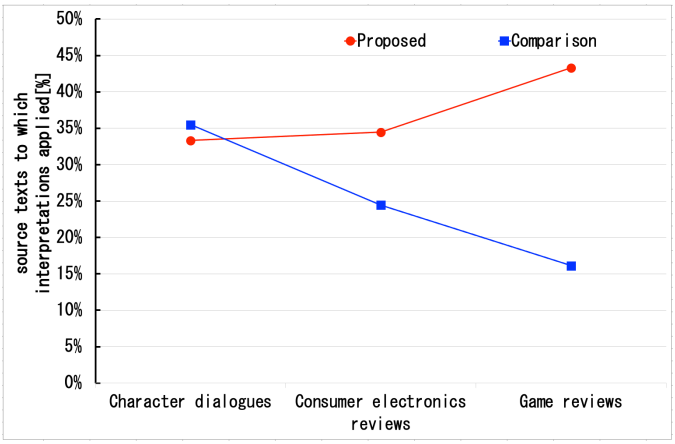


Fig. 9: Percentage of source texts to which participants' interpretations applied (participant average)

Figure 8 shows the number of "Unreasonable interpretations" and "Unknown interpretations" for each participant: A to H represent eight participants of the proposed-system group, and I to P represent eight participants of the comparison-system group.

Figure 8 shows that the number of participants who gave "Unknown interpretation" was 5 in the proposed-system group and 7 in the comparison-system group, and there was no significant difference between them. This indicates that all but one of the participants gave multiple "Unreasonable interpretations". Therefore, it can be confirmed that the proposed system gave more valid interpretations regardless of individual differences.

Figure 9 shows the percentage of the source sentences that fit the interpretation given by the participants (participant average) . For each participant, the sum of the number of source texts that contain statements consistent with these interpretations is divided by the number of source texts per task (500 for the "Character dialogues" task, 1036 for the "consumer electronics review" task, and 1473 for the "game review" task), and the result is the percentage of source texts to which the interpretation applies. The results for the "Character dialogues" task were almost the same, but for the "Consumer electronics reviews" and "Game reviews" tasks, the proposed system was able to derive more interpretations that fit the source texts. For the "Game Reviews" task, the proposed system outperformed the comparison system by nearly 30%, indicating that the interpretation support network displayed with the proposed system was able to derive more typical interpretations that applied to a wider range of source texts.

Figure 10 shows the breakdown (participant average) of

TABLE IV: Examples of classification results and reasons for classification of participants' interpretations

| classification result | example interpretation | reason for classification |
|---|---|---|
| Reasonable interpretation | Task 1 "Character dialogues": After the phrase "Don't get me wrong", the character says something that negates the previous conversation | "Don't get me wrong" is found in the source text, which negates the other person |
| | Task 2 "Consumer electronics review": Nozzle performance for carpets is considered to be a feature of the article | the description "about nozzles for carpets" was found in the source text |
| Unknown Interpretation | Task 1 "Character dialogues": The attribute "Tsundere" can be assumed to be strongly related to romantic relationships | it is difficult to determine which sentences in the source text are related to romantic relationships |
| | Task 1 "Character dialogues": Tend not to make negative comments | difficult to determine whether there are any "negative comments" in the text |
| Unreasonable interpretation | Task 2 "Consumer electronics reviews": Robot vacuum cleaners are not considered to be highly rated compared with other types of vacuum cleaners, etc. | there is no indication in the source text that robot vacuum cleaners are not highly rated |
| | Task 2 "Consumer electronics reviews": It is considered to be characterized by writing about product specifications and evaluations | does not achieve the purpose of interpretation as it applies to all the source texts |

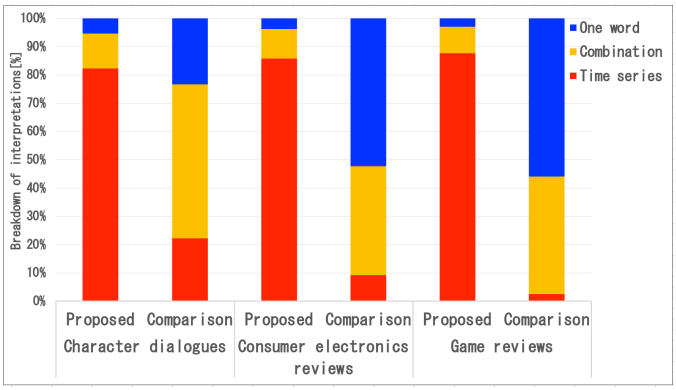


Fig. 10: Breakdown of features focused on by participants (participant average)

which features (one word, combinations, time series, etc.) the participants focused on for interpretation. However, the breakdown of the features focused on was classified by one of the authors on the basis of the following definitions.

- One word: A single interpretation is made from one word.
- Combination: A single interpretation is made from multiple words without considering the time-series relationship.
- Time series: A single interpretation is made from multiple words, taking the time-series relationship into account.

Figure 10 shows that more than 80% of the interpretations with the proposed system focused on time-series information of words. With the comparison system, only about 10% of the interpretations focused on time-series information between words, and the rest were word units and word combinations. This may be because it was easy to understand the time-series information of the words with the interpretation support network of the proposed system; thus, the participants could easily make interpretations focusing on the time-series of the words. With the comparison system, although the characteristic words of the top TFIDF were displayed, the connection between each word was unclear, and the participants often interpreted the words themselves or by combining words with similar meanings. This could be the reason why many of the participants in the comparison system led to wrong interpretations that did not fit the source texts. Therefore, we can say that the proposed system performed a typical interpretation considering the time series of words.

Finally, Table V shows the purpose and examples of interpretations that were particularly common in the proposed system and the comparison system as a trend of interpretation for each task. The words in "[" and "]" indicate words that were actually displayed in the interpretation support network of the proposed system and in the word lists of the comparison system. In the examples of features of interest, the classification of whether the features of interest is time series or not is also indicated. In addition, since all interpretations answered by the participants were in Japanese, Table V includes both Japanese sentences and their translations.

Table V shows that many of the interpretations for the tasks "Consumer electronics reviews" and "Game reviews" consider the details of the product or game, such as what the review should focus on and descriptions of what people might be interested in, rather than the content of the product itself or the game, in the proposed system. Most of these interpretations focused on features of the time-series, suggesting that attention to the features of the time-series allows for interpretations that consider the text as a whole. Conversely, the comparison system resulted in many interpretations of the characteristics of the product itself and the content (genre) of the game, based on individual words. Therefore, in the comparison system, there were many interpretations that applied to only some products and games, and the percentage of source texts to which the interpretations applied was considered to have dropped.

On the other hand, in the task "Character dialogues", many interpretations focused on patterns of time series in the proposed system and patterns of individual words in the comparison system, but in both cases, we confirmed a tendency for many descriptions of the unique expressions of the characters. This is because the average length of "Character dialogues" is only about 40 characters (about 20 words), and the full text is available in the original text display function whether the user selects a single word or multiple words, which may result in similar interpretations in both groups. The same reason can be considered for the result that the percentage of source texts to which the interpretation of the task "Character dialogues" applies was about the same in both

TABLE V: Trends in participant' interpretations (A: proposed, B: comparison)

| Tasks | Purpose of Interpretation (number of items) | Examples of Noted Features | Examples of Interpretation |
|---|---|---|---|
| A: Character dialogues | Character-specific expressions (32 items) | Time series: "[ない] → [嫌い]" ("[not] → [dislike]") etc. | "好きじゃないが嫌いでもないと，相手への好意を示す際に曖昧な表現をする" ("She is ambiguous in expressing his fondness for the other person, saying, I don't like him, but I don't dislike him either.") etc. |
| A: Consumer electronics reviews | Product accessories and other details (22 items) | Time series: "[付属] の後に [充電], [パック], [ノズル] という言葉が続いている" ("[attached] followed by the words [charging], [pack], [nozzle]") etc. | "付属品についての詳しい情報が役立つ場合が多いと考えられる" ("I think more information about the attached accessories would be helpful in many instances.") etc. |
| A: Game Reviews | Interesting Game Details (24 items) | Time series: "[史上], [オープン] と続いている" ("[ever] followed by [open]") etc. | "史上最高と書くことで面白さが伝わりさらに最近人気のオープンワールドゲームという情報を入れることで興味を持たせられると考えられる" ("By writing that it's the best ever, we think it will convey the fun of the game, and by including the information that it's an open-world game, which is very popular these days, we think it will generate interest.") etc. |
| B: Character dialogues | Character-specific expressions (34 items) | One word: "[ない] が上位に上がっている" ("[Not] is rising to the top of the list.") etc. | "好きじゃないのように言葉を否定するのが特徴と考えられる" ("Like I don't like it, denying words is considered a characteristic.") etc. |
| B: Consumer electronics reviews | About the product Itself (36 items) | One word: "[明るい] の単語が頻度が高い" ("[Brighter] words are more frequent.") etc. | "ライトの明るさに関しての記事が明確に書かれているものが多い傾向にある" ("Many of the articles tend to be clearly written regarding the brightness of the lights.") etc. |
| B: Game Reviews | Genres users are looking for (21 items) | One word: "[ファンタジー] がジャンルとして出現している" ("[Fantasy ] is appearing as a genre.") etc. | "ファンタジー性をゲームに求めているユーザーが多いと考えられる" ("It is thought that many users are looking for fantasy nature in their games.") etc. |

groups.

In summary, we confirmed that the proposed system was able to derive typical and reasonable interpretations that were applicable to a wide range of source texts with a higher rate of correct answers than the comparison system. This can be attributed to the fact that the proposed system focuses on the time-series information between multiple words. We also confirmed that even in the case of short texts, such as in the "Character dialogues" task, the proposed system was able to derive typical interpretations at the same level as referring to words with high TFIDF values.

## V. Conclusion

We proposed a classification-pattern interpretation support system to classify multiple text data with an RNN that can learn the time-series relationship of words and interpret the learned network. One of the features of the proposed system is that it can easily extract the time-series information of the learned features without learning on a special model by fitting the network structure of the learned recursive deep learning to an HMM. In the verification experiment, we confirmed that the proposed system can easily lead to a reasonable interpretation that covers a wide range of content of the source text from the classification patterns including the time-series information, even for users who are not familiar with deep learning.

In the future, we would like to change the input of the RNN to a distributed representation that includes information on the relationship between words, so that the interpretation can be more focused on the meaning of the words. We aim to build an interpretation environment for more complex deep learning networks, such as Bidirectional Encoder Representations from Transformers, by obtaining data from inside and outside the training data to support the validity of the interpretation given by the user and presenting it to the user.

## References

[1] M. Ando, Y. Kawahara, W. Sunayama, Y. Hatanaka, 'Interpretation Support System for Classification Patterns Using HMM in Deep Learning with Texts', In Proceedings of the Fourteenth International Conference on Advances in Computer-Human Interactions (ACHI 2021), pp. 64-70, 2021.

[2] K. M. Tran, Y. Bisk, A. Vaswani, D. Marcu, and K. Knight, 'Unsupervised neural hidden markov models', In Proceedings of the Workshop on Structured Prediction for NLP, pp. 63-71, 2016.

[3] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J. Sohn, K. Lee, and D. Papailiopoulos. 'Attack of the Tails: Yes, You Really Can Backdoor Federated Learning', 34th Conference on Neural Information Processing Systems (NeurIPS 2020), 2020.

[4] D. Gunning, 'Explainable artificial intelligence (xAI)', Tech. rep., Defense Advanced Research Projects Agency (DARPA), 2017.

[5] A. Fernandez, F. Herrera, O. Cordon, M. Jose del Jesus, F. Marcelloni, 'Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to?', IEEE Computational Intelligence Magazine 14 (1), pp. 69-81, 2019.

[6] J. Haspiel, N. Du, J. Meyerson, L. P. Robert Jr, D. Tilbury, X. J. Yang, A. K. Pradhan, 'Explanations and expectations: Trust building in automated vehicles', Companion of the ACM/IEEE International Conference on Human-Robot Interaction, ACM, pp. 119-120, 2018.

[7] O. Goudet, D. Kalainathan, P. Caillou, I. Guyon, D. Lopez-Paz, M. Sebag, 'Learning functional causal models with generative neural networks', Explainable and Interpretable Models in Computer Vision and Machine Learning, Springer, pp. 39-80, 2018.

[8] R. M. J. Byrne, 'Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning', Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, pp. 6276-6282, 2019.

[9] X. Yuan, P. He, Q. Zhu, X. Li, 'Adversarial examples: Attacks and defenses for deep learning', IEEE Transactions on Neural Networks and Learning Systems 30 (9), pp. 2805-2824, 2019.

[10] G. Audemard, F. Koriche, P. Marquis, 'On Tractable XAI Queries based on Compiled Representations', KR Proceedings 2020 Special Session on KR and Machine Learning, pp. 838-849, 2020.

[11] Q. Zhang, Y. Yang, H. Ma, Y. N. Wu, 'Interpreting CNNs via decision trees', IEEE Conference on Computer Vision and Pattern Recognition, pp. 6261-6270, 2019.

[12] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, 'Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai', Information Fusion, vol. 58, pp. 82-115, 2020.

[13] E. Tjoa and C. Guan, 'A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI', IEEE Transactions on Neural Networks and Learning Systems 20 Oct. 2020, pp. 1-21, 2020.

[14] J. Wagner, J. M. Kohler, T. Gindele, L. Hetzel, J. T. Wiedemer, S. Behnke, 'Interpretable and fine-grained visual explanations for convolutional neural networks', Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9097-9107, 2019.

[15] C. Panigutti, A. Perotti, D. Pedreschi, 'Doctor XAI: an ontology-based approach to black-box sequential data classification explanations', Proceedings of the 2020 Conference on Fairness Accountability and Transparency, 2020, pp. 629-639.

[16] M Daniluk, T Rocktaschel, J Welbl, S Riedel, 'Frustratingly Short Attention Spans in Neural Language', ICLR, 2017.

[17] Cookpad Inc., 'cookpad', http://cookpad.com, <link> 2022.06.01.

[18] Amazon.com, Inc., 'Amazon.co.jp', https://www.amazon.co.jp, <link> 2022.06.01.