Vol.2021-HCI-191 No.21
2021/1/29

Reading Activity Classification Using Self-supervised Deep Learning

Md. Rabiul Islam[1,a)] Shuji Sakamoto[1] Yoshihiro Yamada[1] Andrew Vargo[1] Motoi Iwata[1]
Masakazu Iwamura[1] Koichi Kise[1]

Abstract: In this report, we propose a self-supervised Deep Learning (DL) method for reading analysis to cope the lack of labeled data issue in this domain and evaluate it on two classification tasks, reading detection using electrooculography glasses datasets and confidence estimation on answers of multiple-choice questions using eye-tracking datasets. Fully-supervised DL and support vector machines are used as comparative methods. The results show that the proposed method is always superior, especially when training data is scarce. This result indicates that the proposed self-supervised DL method is the superior choice for reading analysis tasks. The results of this study are important for informing the design and implementation of automatic reading analysis platforms.

1. Introduction

Reading analysis is essential for understanding and advancing human learning strategies because it is possible to obtain a wide variety of information from reading activities [1]. There are many different types of aspects of reading that can be analyzed and classified [2], [3]. For example, one basic classification task is reading detection, where the objective is to detect whether the user is reading or not reading [4], [5]. Other research has tackled problems like identifying the type of text the user is reading, such as reading English text or Japanese text [6]. Finally, another reading activity classification task would be a problem-solving task such as confidence estimation in answering multiple-choice questions (MCQs) [7]. In this report, we use the term reading activity to cover not only the activity of reading plain text but also problem-solving tasks completed via reading.

There are multiple ways in which to approach this endeavor. Traditional machine learning methods have achieved satisfactory results in laboratory settings where features are manually selected. This requires additional feature engineering expertise. In addition, for the outside laboratory settings (in-the-wild) studies, these methods may not produce similar results [8] due to various reasons, such as noise which obfuscates important features that need to be extracted. Deep Learning (DL) has been successfully used to solve a broad set of difficult problems in various fields [9], [10]. The key to successful DL is to prepare enough labeled samples. In most fields, the difficulty of having enough amount of labeled samples is a serious issue [11].

Lack of labeled data is also a problem for reading activity classification. Obtaining large and well-curated reading activity datasets is problematic because the annotation costs, time takes to generate a satisfactory dataset, diversity of devices, types of embedded sensors, and variations in specifications regarding sampling rates make dataset construction a challenge. For these reasons, it is very difficult to apply a fully-supervised DL method in this domain directly.

The self-supervised DL presents a potential solution to these problems. This method employs a "pretext" task to pre-train the network, before training for the task of interest (target task). Because labeled samples for the pretext task is generated without manual labeling, the network can be trained with a much larger amount of data. In general, this helps to improve classification accuracy.

In the field of human activity recognition, some researchers have attempted to employ self-supervised DL to solve the issue of the lack of labeled data and found it effective [12], [13]. For example, the method proposed by Saeed et al. [12] employs simple signal transformations such as flipping to produce the pretext task for sensor data. However, we do not know whether similar approaches can be applied to cognitive activities requiring fewer bodily movements such as the reading, which is typically captured by sensors like an eye-tracker.

This research aims to clarify how effective the self-supervised DL is at solving the lack of labeled data issue in reading activity classification. As a step toward this goal, we propose a self-supervised DL method and evaluate it for two different but related reading activity classification tasks placed at two extreme points on the reading activity spectrum. The first one is reading detection, a physical-level reading activity. The second one is confidence estimation in answering MCQs, which is an intensive cognitive level reading activity. This allows us to obtain a full picture of the effectiveness of the proposed self-supervised DL method across the reading activity spectrum. In the evaluation process, we recorded eye-movement using electrooculogra-

[1] Osaka Prefecture University, Japan
[a)] dd104006@edu.osakafu-u.ac.jp

phy (EOG) glasses for reading detection and measured eye gaze using an eye-tracker for confidence estimation. We compared the effectiveness of the proposed self-supervised DL method by training and evaluating the network for a different number of training samples per class, starting from the availability of all samples per class to 10 samples per class. We used the fully-supervised DL approach as a comparative method along with support vector machines (SVMs) as a baseline.

The results show that the proposed self-supervised DL method is superior compared to other methods at both tasks. Specifically, the proposed self-supervised DL method demonstrates better performance than the fully-supervised DL except at the largest number of training samples, where the proposed self-supervised DL method performs equally well. Although the fully-supervised DL sometimes performs worse than SVM with a smaller number of training samples, due to the impact of insufficient training samples, the proposed self-supervised DL method does not face this problem; it is always superior to SVM as well. The statistical analysis supports the above statements.

From the results, we conclude that the proposed self-supervised DL method is superior to other methods over a wide range of training samples on both tasks, and is comparable with the fully-supervised DL when the number of training samples available is high. This indicates that we can recommend the self-supervised DL method for any size of available training samples. This insight can help system designers and researchers more efficiently pursue reading activity classification.

2. Related Work

Our work relates to several active research areas, including reading detection, confidence estimation, and self-supervised DL. In this section, we describe how our work builds on these fields.

2.1 Reading Detection

Reading detection strategy varies depending on its purpose, and over the past years, researchers have proposed many methods for different kinds of automatic reading detection. For example, they have proposed methods for reading detection as a part of other human activities such as reading in transit [14], in office settings [15], and with talking [16] by exploring eye movements in controlled settings using traditional machine learning approaches. In another eye-based activity recognition study [17], authors detected reading with desktop activities such as search and writing by using traditional machine learning methods. Researchers used traditional machine learning methods to detect whether the user is reading or skimming [18], [19], reading or searching [4] and reading or not reading [5], [8] in laboratory settings. Recently, Ishimaru et al. [6] proposed a traditional machine learning method to classify the language of text segments, English or Japanese, read by the user by analyzing eye movement data obtained through an in-the-wild study.

However, most of the existing methods occurred in laboratory settings use traditional machine learning approaches except for some preliminary work that applies DL methods [8], [20]. Although some methods using traditional machine learning approaches produce satisfactory results in laboratory settings, they may not do so in-the-wild [6].

2.2 Confidence Estimation in Answering Multiple-choice Questions

MCQs are fundamental forms of assessing knowledge, ability, and user performance [21] and are the most popular since they are easy, quick, and offer more objective scoring. However, in this assessment, one critically common question arises: has the user answered correctly by chance or with confidence in their knowledge of the correct answer. Therefore, an assessment system must provide accurate information and give feedback. So there is a need to develop a way to estimate confidence automatically.

As a step toward this automatic confidence estimation, researchers propose some methods. Tsai et al. [22] analyzed user's visual attention spans when solving MCQs by using eye-tracking under laboratory settings and with the application of a traditional machine learning method. Yamada et al. [7] proposed a method to classify whether the user is confident or not when answering MCQs through manually selected features from the eye gaze data. All these methods occurred in laboratory settings and employed traditional machine learning approaches. This means that a significant challenge remains for in-the-wild datasets.

2.3 Self-supervised Deep Learning

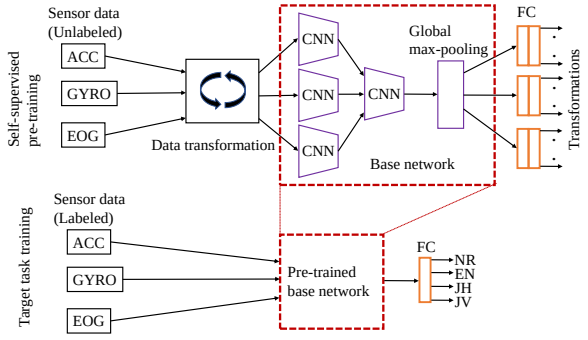
In the past decade, the development and application of DL has successfully solved many problems in the field of ubiquitous computing [23], pervasive intelligence [24], health [25], and many more. Most of the methods use fully-supervised DL approaches that need large and carefully labeled data that is feasible for use in domains such as computer vision [26] but unfeasible in others.

To overcome the innate limitations of the fully-supervised DL approaches, researchers introduced several unsupervised methods. Recently, researchers proposed a DL technique called self-supervised DL [27], [28], [29]. Self-supervised DL is now an active research approach in various domains such as computer vision and robotics [30], [31] and its achievements show that it is effective. When it comes to human activity recognition tasks, the same issue of the lack of labeled data occurs. Researchers attempted applying self-supervised DL by utilizing some simple signal transformations such as flipping and adding noise [12]. Their findings show that a self-supervised DL approach is effective in this domain.

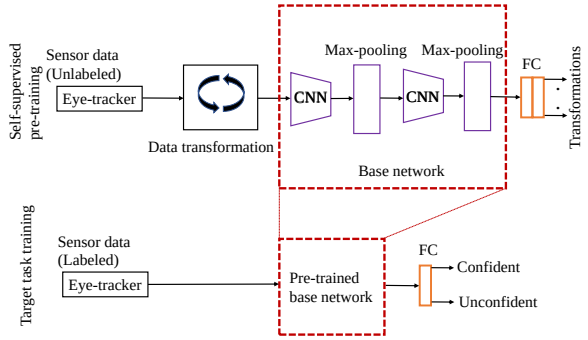
Inspired by the recent success of applying self-supervised DL to address the issue of insufficient labeled data, we set out a research agenda to explore the generalize efficacy of self-supervised DL for eye movement sensory data for physical and cognitive intensive reading tasks.

3. Proposed Method

We propose a self-supervised DL method for reading activity classification using sensor data, as shown in **Fig. 1** that consists of two stages. The first stage shown in the upper parts of **Fig. 1(a)** and **Fig. 1(b)** is self-supervised pre-training consisting of solving the pretext task, automatically applied to a large collection of unlabeled sensor data. The second stage shown in the lower



(a) Reading detection



(b) Confidence estimation

Fig. 1: Proposed self-supervised DL method for reading activity classification.

parts of Fig. 1(a) and Fig. 1(b) is target task training, i.e., training of a reading activity classification network by fine-tuning the pre-trained base network using labeled sensor data. To evaluate our method, we implemented the proposed self-supervised DL method on two very different reading activity tasks: reading detection and confidence estimation in answering MCQs. The reasons for this is that reading activities are distributed on a wide spectrum. For example, some activities are merely related to quantity of reading, such as reading periods or the number of read words. On the other hand, other activities involve the quality of reading such as understanding and confidence. To show the applicability of the proposed self-supervised DL method, we apply it to tasks which belong to these respective categories. We selected reading detection that segments reading periods from all activities and confidence estimation in answering MCQs. Another reason for selecting these two activities is that they are recorded by using two different devices: EOG glasses for reading detection, and an eye-tracker for confidence estimation. We consider that the proposed self-supervised DL method is general and effective enough if it works for both tasks using different devices. In the following, we explain details of the proposed self-supervised DL method for each task.

3.1 Reading Detection

Reading detection aims to identify periods of reading from all other activities. This is implemented as a classification task; the user activities are divided into short segments and then classified into different segments of activity.

The devices employed to measure reading detection are EOG glasses that generate EOG data of eye movements, and accelerometer (ACC) and gyroscope (GYRO) data from the movement of the EOG glasses themselves. From the EOG signals, we obtained data of horizontal and vertical eye movements (EOG_H and EOG_V). ACC and GYRO data consist of x, y and z components (ACC_X, ACC_Y, ACC_Z, GYRO_X, GYRO_Y and GYRO_Z).

3.1.1 Self-supervised Pre-training

Self-supervised pre-training involves learning the representation of signal data by using a pretext task. For the pretext task, we employed the task proposed by Saeed et al. [12]; noised, scaled, rotated, negated, horizontally flipped, permuted, time-warped, and channel-shuffled. The pretext task is to recognize the transformation applied to an input signal. For ACC and GYRO data, we employ eight transformations. Because rotation is meaningless for EOG data, seven transformations excluding rotation are applied instead.

The red-dashed rectangle in the upper part of Fig. 1(a) illustrates the base network trained by the pretext task. It consists of three Convolutional Neural Network (CNN) blocks for EOG, ACC and GYRO data, a CNN block that concatenates three CNN layers, and a global max-pooling layer. Each CNN block consists of three 1D CNN layers. The numbers of units in the CNN layers are 32, 64, and 96, respectively, and the kernel sizes are 24, 16, and 8, respectively. We applied batch normalization after each CNN layer, and a dropout layer after the global max-pooling layer. We added three classifiers at the end of the base network for EOG, ACC, and GYRO data, respectively. Each classifier consists of two fully connected (FC) layers, and the numbers of units in the FC layers are 256 and 512, respectively. We use ReLU as the activation function, the softmax function as the output layer, and Adam as the optimizer.

3.1.2 Target Task Training

The final step is the target task training. For the reading detection, we have four target classes: not reading (NR), reading English text (EN), reading Japanese horizontal text (JH), and reading Japanese vertical text (JV). Japanese scripts can be written horizontally or vertically. Japanese horizontal writing is similar to English writing except there are no spaces between words, which causing different eye movements when reading the text. In the vertical writing system, characters are read from top to bottom, going right to left [32].

We use the pre-trained base network to create a reading detection network by fine-tuning the pre-trained base network using labeled sensor data with a supervised approach, as shown in the lower part of Fig. 1(a). The FC layer in the target task training has 1024 units. We use the same activation function, optimizer, and output layer as used in the self-supervised pre-training.

3.2 Confidence Estimation

Confidence estimation in answering MCQs involves classifying whether the answer is produced with confidence or not. The format for how we handled MCQs in this report is shown in **Fig. 2**. We employed an eye-tracker to describe the user's behaviors. Unlike the classification of fixed length segments in the reading detection activity, the amount of sensor data varies in this task. To cope with such an issue, researchers of human activ-

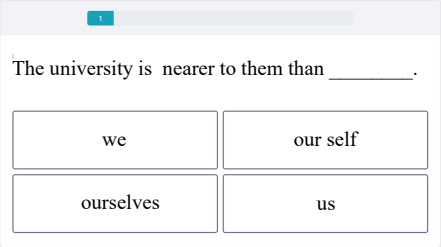
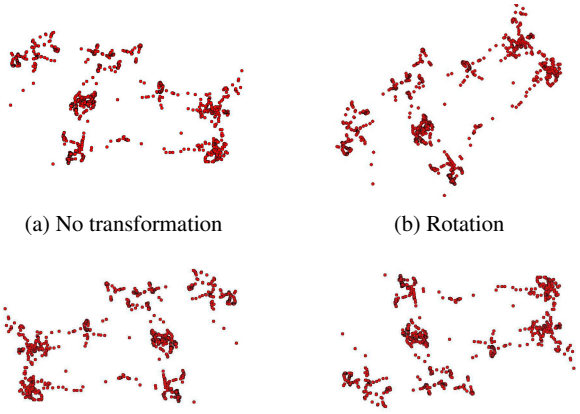


Fig. 2: MCQ format used for confidence estimation. The user is asked to select one choice to fill in the blank.



(a) No transformation (b) Rotation

(c) Reflection about y-axis (d) Reflection about x-axis

Fig. 3: Examples of transformed eye gaze image data in confidence estimation.

ity recognition transformed time-series data into images to solve the classification task using CNN [33], [34]. We also convert the eye-tracking data by plotting eye gaze graphically as shown in **Fig. 3(a)**.

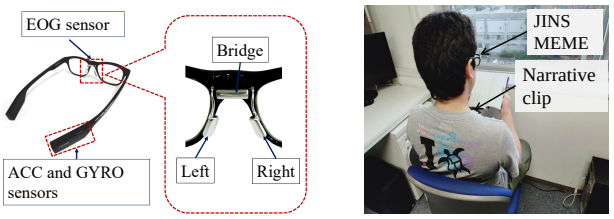
3.2.1 Self-supervised Pre-training

In the pretext task for confidence estimation, we consider three image transformations as shown in **Fig. 3(b)** to **Fig. 3(d)**. The rotation is to apply 45° anti-clockwise rotation to the original image. Reflection about x and y axes means the transformation of each pixel at (x, y) to $(x, -y)$ and $(-x, y)$, respectively.

The red-dashed box in the upper part of Fig. 1(b) shows the base network that consists of two CNN blocks and a max-pooling layer after each CNN block. Besides, we add a dropout layer after the second max-pooling layer followed by a flatten layer. Each CNN block consists of two 2D CNN layers. The numbers of units of CNN layers are 8 for the first CNN block and 16 for the second CNN block, respectively. The kernel size is 3×3 for all four CNN layers. We added a batch normalization after each CNN layer. Finally, we add a classifier consisting of two FC layers to identify the type of transformations applied, and the number of units of both FC layers is 36. We use ReLU as the activation function for all CNN and FC layers, the softmax function as the output layer, and the SGD as the optimizer. The input image size is $64\times64\times3$.

3.2.2 Target Task Training

After the pre-training, the target task training is performed by replacing the FC layers of the pre-trained network and fine-tuning the pre-trained base network using labeled eye gaze image data. We designed the confidence estimation target task as a binary classification: confident or unconfident. For the target task training, the number of units in the FC layer is 64. We used the same input image size, activation function, optimizer, and output layer used in the self-supervised pre-training task.



(a) (b)

Fig. 4: Data recording for reading detection; (a) JINS MEME EOG glasses, and (b) and (c) a user reading text documents wearing narrative clip and JINS MEME EOG glasses.

4. Data Collection

4.1 Reading Detection Datasets

We used two datasets for reading detection: a labeled dataset and an unlabeled dataset. We recorded data for both datasets using JINS MEME EOG glasses. This is an eye-wear device developed by JINS, as shown in **Fig. 4(a)**, which equips EOG, ACC, and GYRO sensors. EOG is a technique that measures the corneo-retinal standing potential between the front and the back of the human eye. The sampling rate of EOG, ACC, and GYRO sensors is 100 Hz. We describe labeled and unlabeled datasets for reading detection in detail in the following two subsections.

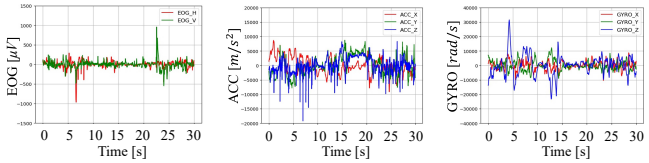
4.1.1 Labeled Dataset for Reading Detection

For the labeled dataset, we employed OPU_RD dataset, which was introduced by Ishimaru et al. [6]. Ten participants were recruited for data collection. Each participant wore the JINS MEME glasses as shown in **Fig. 4(b)** for about 12 hours a day for two days and was asked to read English documents, vertically written Japanese documents, and horizontally written Japanese documents for about 1 hour for each in a day. Participants also had a small camera called narrative clip on their clothing as shown in Fig. 4(b) to take frontal images every 30 seconds, used to label recorded data. The narrative clip was allowed to be removed in places where recording was inappropriate. Except for the above-mentioned conditions, no restrictions were imposed during data recording. Thus, the dataset can be regarded as "in-the-wild."

After recording, the EOG, ACC, and GYRO data were split into segments by using a window of size 30 seconds slid by 15 seconds. Thus segments overlap with each other. **Fig. 5** shows examples of segments. After that, each segment is labeled. EOG data sometimes suffers from bursts of noise of about several seconds due to poor contact of the EOG electrodes to the skin. We found such EOG segments and discarded.

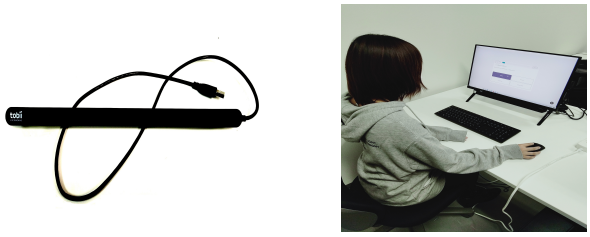
4.1.2 Unlabeled Dataset for Reading Detection

We recruited 13 Japanese university students. Each participant wore a JINS MEME device for three to eight days and read English document, horizontally written Japanese documents, and vertically written Japanese documents, or did not read anything at all. The measurement time is about 20 to 60 hours per person, and



(a) EOG (b) ACC (c) GYRO

Fig. 5: Data samples of 30 seconds segment recorded with the JINS MEME EOG glasses in reading detection.



(a) Tobii 4C eye-tracker (b) User answering MCQs

Fig. 6: Data collection environment for confidence estimation.

the total recorded time from all participants is 676 hours. In addition, we employed an unlabeled dataset that recorded EOG, ACC, and GYRO data when 39 participants attended presentations at a conference. These unlabeled datasets are also "in-the-wild." The unlabeled data totals about 1,359 hours. We also prepared unlabeled data in the same way described for labeled data except labeling. The total number of samples in the unlabeled dataset is 177,921.

4.2 Confidence Estimation Datasets

We also use a labeled dataset and an unlabeled dataset for confidence estimation. We used the Tobii 4C pro-upgraded eye-tracker, as shown in **Fig. 6(a)**, a stationary eye-tracker whose sampling rate is 90 Hz.

4.2.1 Labeled Dataset for Confidence Estimation

We recruited 20 Japanese university students to generate the labeled dataset for confidence estimation. The data collection environment is shown in **Fig. 6(b)**. Each participant read and answered four-choice English grammatical questions on a computer screen. Right after answering each MCQ, participants were requested to assess the confidence behind their answer, which then became a label for the data. However, the labeled dataset includes a serious skew in the number of confident and unconfident answers. This is because of the differences in English ability among the participants. We did not impose any restrictions during the data collection, so that this dataset is also considered "in-the-wild."

4.2.2 Unlabeled Dataset for Confidence Estimation

We recorded the unlabeled data for confidence estimation as described above except there was no inquiry about the confidence. We recruited 80 Japanese high school students, with each participant reading and answering four-choice English vocabulary questions. The total number of samples in the unlabeled dataset is 57,460. Note that the age range of the participants and the contents of the four-choice questions are both different from that of the labeled dataset.

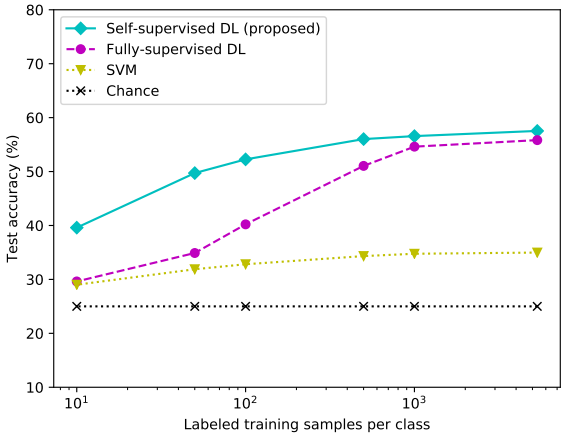


Fig. 7: Dependency of test accuracy to the number of labeled training samples per class for reading detection.

5. Experimental Results and Discussion

5.1 Reading Detection

5.1.1 Experimental Conditions

The purpose of the experiment is to evaluate the effectiveness of the proposed self-supervised DL method as compared to conventional methods; fully-supervised DL and SVM. For the fully-supervised DL, we simply use the proposed self-supervised DL method for the target task without the pre-training. For SVM, the method described in [6] is used. We used a total of ten features for SVM, the mean and variance of vertical and horizontal components of EOG data and the mean and variance of three axes components of ACC and GYRO data.

For the proposed self-supervised DL method, we first applied the pre-training by using the transformed sensor data. We applied each transformation equally so that the chance rates for the EOG data, ACC data, and GYRO data are 12.5%, 11.1%, and 11.1%, respectively. After this, the target task training was applied. The number of available labeled samples is different for each class. We simply took all 5,340 samples (the smallest number) of "reading English" and down sampled other classes to have them match in size. Thus the chance rate for the target task is 25%. The same data were also employed for training the fully-supervised DL and SVM. We changed the number of labeled training samples per class in the order of 10, 50, 100, 500, 1,000, and 5,340. All of the above methods were evaluated with the user independent Leave-One-Participant-Out cross-validation (LOOCV) approach.

5.1.2 Results of the Target Task

Fig. 7 shows the reading detection result. It describes the change of average test accuracy for the number of labeled training samples per class. From this graph, we can observe the following. First, the proposed self-supervised DL method performs best for all cases regardless of the number of training samples. The proposed self-supervised DL method is more advantageous than the fully-supervised DL when the number of labeled training samples is smaller. This indicates the effectiveness of the self-supervised DL. As compared to SVM, the fully-supervised DL performs much better when the number of labeled training samples is larger. However, this advantage disappears when the

number of labeled training samples decreases. This shows the limitation of the fully-supervised DL when a large enough number of training samples are not available. On the other hand, the proposed self-supervised DL method is always much better than SVM and is never inferior to the fully-supervised DL. In other words, we can always recommend to use the proposed self-supervised DL method.

To make sure whether the difference is significant or not, we applied the statistical analysis. The one-way repeated measures analysis of variance (ANOVA) test was first applied. The null hypothesis is that the population means of all three methods are equal and the significance level is 0.01. Then the post-hoc paired t-test was applied for the further analysis. A null hypothesis here is that the population mean of one method is equal to that of another method. We conducted three t-test experiments. Multiple comparisons for the three methods was mitigated by a Bonferroni correction. With the correction applied, significance is found if $p < 0.0033$. For the pair of the proposed self-supervised DL method and the fully-supervised DL, all but the case of 5,340 training samples, we have confirmed that the proposed self-supervised DL method is statistically significantly better than the fully-supervised DL. For the comparison with SVM, the advantage of the proposed self-supervised DL method is shown for all cases. For the comparison between the fully-supervised DL and SVM, we cannot reject the null hypothesis for the cases for smaller sample sizes (10 and 50).

By analyzing the results, we found the following tendencies. "Reading English" and "Reading horizontally written Japanese" tends to be confused, because both reading behaviors are dominated by horizontal eye movement. Among all the methods, the proposed self-supervised DL method was most effective at distinguishing these classes. Some "not reading" behaviors are classified into reading, because they contain similar behaviors by chance.

5.2 Confidence Estimation

5.2.1 Experimental Conditions

The purpose of the experiments is the same as the reading detection and we employed the same methods for comparison. In the self-supervised pre-training, for each image, we selected one of four transformations, including no transformation, as shown in **Fig. 3** and applied it. Because each transformation was selected equally, the chance rate of the pre-training was 25%.

After the pre-training, we created the target task network of confidence estimation by fine-tuning the pre-trained base network, as shown in the lower part of Fig. 1(b). In this case, we used the labeled data. Although we applied LOOCV for reading detection, it is not appropriate for confidence estimation due to the seriously skewed distribution of confident and unconfident labels. Thus we took a different approach for data preparation for training and testing as shown in **Fig. 8**. The chance rate of the classification is 50%.

In the case of the fully-supervised DL, we trained the network using only the labeled data. We used SVM as a baseline method using basic statistical features such as mean and variance. We calculated and used four features from one sample (one MCQ);

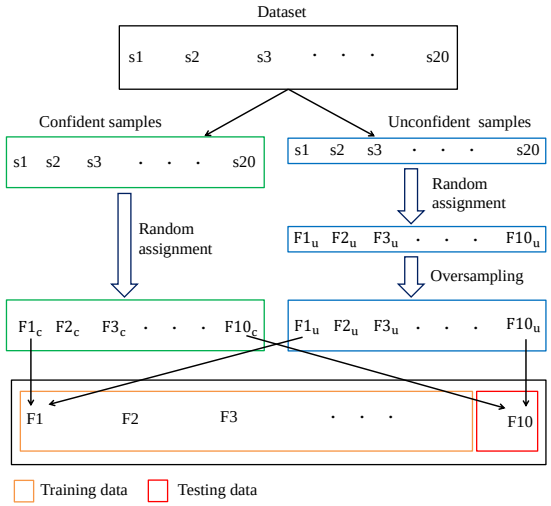


Fig. 8: Ten-fold cross-validation for confidence estimation.

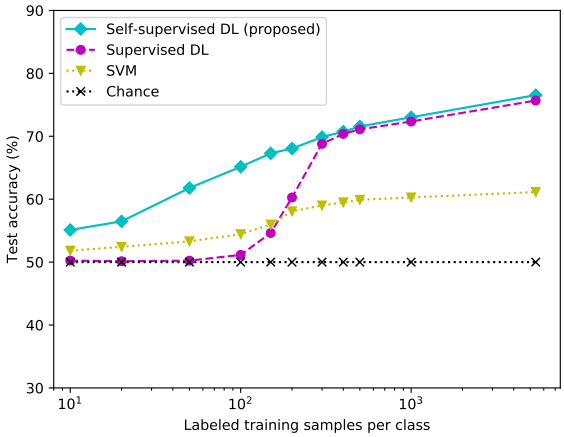


Fig. 9: Dependency of test accuracy to the number of labeled training samples per class for confidence estimation.

means and variances of the two axes of the eye gaze to classify data samples. All methods were also evaluated by the ten-fold cross-validation. We changed the number of labeled training samples per class in the order of 10, 20, 50, 100, 150, 200, 300, 400, 500, 1,000, and 5,382.

5.2.2 Results of the Target Task

Fig. 9 shows the results of the target task. Tendencies similar to the reading detection results were observed. The proposed self-supervised DL method performed the best regardless of the number of labeled training samples. The performance of the fully-supervised DL dropped when the number of labeled training samples was insufficient. With confidence estimation, SVM was not always worst, though performance was limited even with a larger number of labeled samples. From these results, we can always recommend to use the proposed self-supervised DL method in the case of reading detection.

We also applied the statistical tests for the results of confidence estimation. From the one-way repeated measures ANOVA test, we have confirmed that there exists at least one population mean is different from the rest. From the results of the post-hoc paired t-test, we have confirmed the following: For the comparison between the proposed self-supervised DL method and the fully-

supervised DL, we found significant differences except for the cases of 400, 500, and 1,000 training samples. For the comparison between the proposed self-supervised DL method and SVM, all cases show a significant difference. For the comparison between the fully-supervised DL and SVM, we could not reject the null hypothesis for the case of 150 training samples per class, but it was rejected for all other cases. For the cases with the larger number of training samples, the fully-supervised DL worked better than SVM, and for the case with the smaller number of training samples, the opposite conclusion is held. From these results of statistical analysis, we can also confirm that our discussion of the performance comparison we made above has been supported.

By analyzing the results, we found that for all the methods predictions were not biased with the larger number of training samples, but biased with the smaller number of training samples except for the case of the proposed self-supervised DL method.

6. Conclusion

Automatic recording and reading behavior analysis allow users to examine their reading habits which can help with the development of reading strategies. Methods that use classical machine learning approaches and handcrafted features may achieve good results in laboratory settings, but may not obtain satisfactory results in-the-wild. DL methods that can solve this issue that requires a large-sized labeled dataset. However, a large-sized labeled data collections are difficult to obtain. As a step towards tackling this issue, we have proposed a self-supervised DL method. We evaluated the effectiveness of the proposed self-supervised DL method by selecting two reading activities that explore physical reading and confidence, respectively. We evaluated both tasks with the proposed self-supervised DL method, the fully-supervised DL, and SVM.

The proposed self-supervised DL method for reading detection consists of two stages. In the first stage, we trained the network by solving pretext tasks automatically applied to the unlabeled data for representation learning. In the second stage, we created the reading detection target task network by fine-tuning the pre-trained base network using labeled data. Confidence estimation followed a similar process.

From the experimental results, we have confirmed that the proposed self-supervised DL method performs the best for both reading activity classification tasks compared to the fully-supervised DL method and SVM for all cases of the numbers of training samples. Therefore we can always recommend to use the proposed self-supervised DL method regardless of the available number of training samples.

Future work includes further improvement in accuracy of the proposed self-supervised DL method by introducing other sensors, as well as its application for other reading activity classification tasks using various sensors.

Acknowledgments This work was supported in part by the JST CREST (Grant No. JPMJCR16E1), JSPS Grant-in-Aid for Scientific Research (20H04213), Grand challenge of the Initiative for Life Design Innovation (iLDi), and OPU Key-project.

References

[1] Rayner, K., Pollatsek, A., Ashby, J. and Jr, C. C.: *Psychology of Reading*, Psychology Press, New York, NY, USA, 2nd edition (2012).

[2] Martínez-Gómez, P. and Aizawa, A.: Recognition of Understanding Level and Language Skill Using Measurements of Reading Behavior, *Proceedings of the 19th International Conference on Intelligent User Interfaces*, IUI '14, New York, NY, USA, ACM, p. 95–104 (online), DOI: 10.1145/2557500.2557546 (2014).

[3] Okoso, A., Toyama, T., Kunze, K., Folz, J., Liwicki, M. and Kise, K.: Towards extraction of subjective reading incomprehension: Analysis of eye gaze features, *CHI 2015 - Extended Abstracts Publication of the 33rd Annual CHI Conference on Human Factors in Computing Systems: Crossings*, Vol. 18, ACM, pp. 1325–1330 (online), DOI: 10.1145/2702613.2732896 (2015).

[4] Campbell, C. S. and Maglio, P. P.: A Robust Algorithm for Reading Detection, *Proceedings of the 2001 Workshop on Perceptive User Interfaces*, PUI '01, New York, NY, USA, ACM, p. 1–7 (online), DOI: 10.1145/971478.971503 (2001).

[5] Landsmann, M., Augereau, O. and Kise, K.: Classification of Reading and Not Reading Behavior Based on Eye Movement Analysis, *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, UbiComp/ISWC '19 Adjunct, New York, NY, USA, ACM, p. 109–112 (online), DOI: 10.1145/3341162.3343811 (2019).

[6] Ishimaru, S., Maruichi, T., Landsmann, M., Kise, K. and Dengel, A.: Electrooculography Dataset for Reading Detection in the Wild, *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, UbiComp/ISWC '19 Adjunct, New York, NY, USA, ACM, p. 85–88 (online), DOI: 10.1145/3341162.3343812 (2019).

[7] Yamada, K., Kise, K. and Augereau, O.: Estimation of Confidence Based on Eye Gaze: An Application to Multiple-Choice Questions, *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, UbiComp '17, New York, NY, USA, ACM, p. 217–220 (online), DOI: 10.1145/3123024.3123138 (2017).

[8] Ishimaru, S., Hoshika, K., Kunze, K., Kise, K. and Dengel, A.: Towards Reading Trackers in the Wild: Detecting Reading Activities by EOG Glasses and Deep Neural Networks, *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, UbiComp '17, New York, NY, USA, ACM, p. 704–711 (online), DOI: 10.1145/3123024.3129271 (2017).

[9] Krizhevsky, A., Sutskever, I. and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, *Commun. ACM*, Vol. 60, No. 6, p. 84–90 (online), DOI: 10.1145/3065386 (2017).

[10] Ronao, C. A. and Cho, S.-B.: Human activity recognition with smartphone sensors using deep learning neural networks, *Expert Systems with Applications*, Vol. 59, pp. 235–244 (online), DOI: 10.1016/j.eswa.2016.04.032 (2016).

[11] Roh, Y., Heo, G. and Whang, S. E.: A Survey on Data Collection for Machine Learning: a Big Data - AI Integration Perspective, *CoRR*, Vol. abs/1811.03402 (online), available from ⟨http://arxiv.org/abs/1811.03402⟩ (2018).

[12] Saeed, A., Ozcelebi, T. and Lukkien, J.: Multi-Task Self-Supervised Learning for Human Activity Detection, *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, Vol. 3, No. 2 (online), DOI: 10.1145/3328932 (2019).

[13] Haresamudram, H., Beedu, A., Agrawal, V., Grady, P. L., Essa, I., Hoffman, J. and Plötz, T.: Masked Reconstruction Based Self-Supervision for Human Activity Recognition, *Proceedings of the 2020 International Symposium on Wearable Computers*, ISWC '20, New York, NY, USA, ACM, p. 45–49 (online), DOI: 10.1145/3410531.3414306 (2020).

[14] Bulling, A., Ward, J. A. and Gellersen, H.: Multimodal Recognition of Reading Activity in Transit Using Body-Worn Sensors, *ACM Trans. Appl. Percept.*, Vol. 9, No. 1 (online), DOI: 10.1145/2134203.2134205 (2012).

[15] Bulling, A., Ward, J. A., Gellersen, H. and Tröster, G.: Eye Movement Analysis for Activity Recognition Using Electrooculography, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 33, No. 4, p. 741–753 (online), DOI: 10.1109/TPAMI.2010.86 (2011).

[16] Ishimaru, S., Kunze, K., Tanaka, K., Uema, Y., Kise, K. and Inami, M.: Smart Eyewear for Interaction and Activity Recognition, *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '15, New York, NY, USA, ACM, p. 307–310 (online), DOI: 10.1145/2702613.2725449 (2015).

[17] Srivastava, N., Newn, J. and Velloso, E.: Combining Low and Mid-Level Gaze Features for Desktop Activity Recognition, *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, Vol. 2, No. 4 (online), DOI: 10.1145/3287067 (2018).

[18] Kelton, C., Wei, Z., Ahn, S., Balasubramanian, A., Das, S. R., Samaras, D. and Zelinsky, G.: Reading Detection in Real-Time, *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, ETRA '19, New York, NY, USA, ACM, (online), DOI: 10.1145/3314111.3319916 (2019).

[19] Biedert, R., Hees, J., Dengel, A. and Buscher, G.: A Robust Realtime Reading-Skimming Classifier, *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '12, New York, NY, USA, ACM, p. 123–130 (online), DOI: 10.1145/2168556.2168575 (2012).

[20] Copeland, L., Gedeon, T. and Mendis, S.: Predicting reading comprehension scores from eye movements using artificial neural networks and fuzzy output error, *Artificial Intelligence Research*, Vol. 3, No. 3, p. 35 (online), DOI: 10.5430/air.v3n3p35 (2014).

[21] Haladyna, T. M.: *Developing and Validating Multiple-choice Test Items*, Routledge, New York, NY, USA, 3rd edition edition (2015).

[22] Tsai, M.-J., Hou, H.-T., Lai, M.-L., Liu, W.-Y. and Yang, F.-Y.: Visual Attention for Solving Multiple-Choice Science Problem: An Eye-Tracking Analysis, *Computers & Education*, Vol. 58, No. 1, p. 375–385 (online), DOI: 10.1016/j.compedu.2011.07.012 (2012).

[23] Hammerla, N. Y., Halloran, S. and Plötz, T.: Deep, Convolutional, and Recurrent Models for Human Activity Recognition Using Wearables, *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, AAAI Press, p. 1533–1540 (online), available from ⟨https://www.ijcai.org/Proceedings/16/Papers/220.pdf⟩ (2016).

[24] Sarhan, S., Nasr, A. A. and Shams, M. Y.: Multipose Face Recognition-Based Combined Adaptive Deep Learning Vector Quantization, *Computational Intelligence and Neuroscience*, Vol. 2020 (online), DOI: 10.1155/2020/8821868 (2020).

[25] Nonaka, N. and Seita, J.: Data Augmentation for Electrocardiogram Classification with Deep Neural Network, *arXiv:2009.04398 [eess]*, (online), available from ⟨http://arxiv.org/abs/2009.04398⟩ (2020).

[26] Toshev, A. and Szegedy, C.: DeepPose: Human Pose Estimation via Deep Neural Networks, *2014 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 1653–1660 (online), DOI: 10.1109/CVPR.2014.214 (2014).

[27] Amis, G. P. and Carpenter, G. A.: Self-supervised ARTMAP, *Neural Networks*, Vol. 23, No. 2 (online), DOI: 10.1016/j.neunet.2009.07.026 (2010).

[28] Liu, X., v. d. Weijer, J. and Bagdanov, A. D.: Exploiting Unlabeled Data in CNNs by Self-Supervised Learning to Rank, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, No. 8, pp. 1862–1878 (online), DOI: 10.1109/TPAMI.2019.2899857 (2019).

[29] Agrawal, P., Carreira, J. and Malik, J.: Learning to See by Moving, *2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE, pp. 37–45 (online), DOI: 10.1109/ICCV.2015.13 (2015).

[30] Gidaris, S., Singh, P. and Komodakis, N.: Unsupervised Representation Learning by Predicting Image Rotations, *CoRR*, Vol. abs/1803.07728 (online), available from ⟨http://arxiv.org/abs/1803.07728⟩ (2018).

[31] Owens, A., Wu, J., McDermott, J. H., Freeman, W. T. and Torralba, A.: Ambient Sound Provides Supervision for Visual Learning, *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, Vol. 9905, Cham, Springer, pp. 801–816 (online), available from ⟨https://doi.org/10.1007/978-3-319-46448-0_48⟩ (2016).

[32] Wikipedia: Horizontal and vertical writing in East Asian scripts, (online), available from ⟨https://en.wikipedia.org/w/index.php?title=Horizontal_and_vertical_writing_in_East_Asian_scripts&oldid=984358336⟩ (accessed Oct 29, 2020.)

[33] Wang, Z. and Oates, T.: Imaging Time-Series to Improve Classification and Imputation, *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, AAAI Press, p. 3939–3945 (online), available from ⟨https://www.ijcai.org/Proceedings/15/Papers/553.pdf⟩ (2015).

[34] Hatami, N., Gavet, Y. and Debayle, J.: Classification of Time-Series Images Using Deep Convolutional Neural Networks, *CoRR*, Vol. abs/1710.00886 (online), available from ⟨http://arxiv.org/abs/1710.00886⟩ (2017).