applied sciences

MDPI

Article

An Intelligent Ubiquitous Learning Environment and Analytics on Images for Contextual Factors Analysis

Mohammad Nehal Hasnine [1,*], Gökhan Akçapınar [2], Kousuke Mouri [3] and Hiroshi Ueda [1]

1 Research Center for Computing and Multimedia Studies, Hosei University, Tokyo 184-8584, Japan; uep@hosei.ac.jp
2 Department of Computer Education & Instructional Technology, Hacettepe University, 06800 Ankara, Turkey; gokhana@hacettepe.edu.tr
3 Academic Center for Computing and Media Studies, Kyoto University, Kyoto 606-8501, Japan; mourikousuke@gmail.com
* Correspondence: nehal.hasnine.79@hosei.ac.jp; Tel.: +81-04-2387-6070

Received: 29 October 2020; Accepted: 15 December 2020; Published: 16 December 2020

Abstract: Contextual factors in which learning occurs are crucial aspects that learning analytics and related disciplines aim to understand for optimizing learning and the environments in which learning occurs. In foreign vocabulary development, taking the notes or memos of learning contexts along with other factors, play an essential role in quick memorization and reflection. However, conventional tools fail to automate the learning contexts generation process as learners still need to take memos or e-notes to describe their vocabulary learning contexts. This paper presents the Image Understanding Project (hereafter IUEcosystem) that could produce smartly-generated learning contexts primarily in a learner's target languages. The IUEcosystem uses visual content analysis of lifelogging images as the sensor data to produce smartly-generated learning contexts that could be used as an alternative to handwritten memos or electronic notes. The IUEcosystem uses applied artificial intelligence to produce smartly-generated learning contexts. This intelligent learning environment collects a learner's learning satisfaction and interaction data and, later on, analyzes them to produce time-based notifications for enhancing retention. Furthermore, a new learning design is presented that aims to map a learner's prior vocabulary knowledge with new learning vocabularies to be learned. This learning design would help learners to review and recall prior knowledge while learning new vocabulary.

Keywords: AI-assisted system; applied AI; contextual analysis; contextual factors; image analytics; image understanding; language learning; learning context; visual contents analysis; scene analysis; vocabulary learning using authentic context

1. Introduction

1.1. The Link between Learning Context, Learning Design, and Analytics

A learning context is referred to as learning environments, including socio-cultural-political environments where a learner's learning occurs [1]. For an individual, the learning contexts expands to the surroundings, including educators, classmates, classroom atmosphere, the social and cultural tradition of learning, and academic curriculum. For humans, the problem-solving process is constrained by the learning context where the problem is encountered and tackled [1]. Therefore, there is a strong relationship between contexts and learning outcomes. Previous research has suggested that learning outcomes result from interactions between various factors, such as our emotional aspects, prior knowledge, motivation, and the contextual factors in which the learning occurs [2,3]. In learning

design, contextual factors play an essential part. To understand a learning context correctly, it is necessary to observe the physical environment's properties and characteristics where learners learn and apply the acquired knowledge. Besides, mental and emotional states, cultural norms, social situations, and relationships can be used as triggers to describe a learning context. It is challenging to establish a link between a learning design and an individual learner's contextual factors. However, contextual learning analytics could be a powerful method to establish a connection between a learning design and the contextual factors associated with learning. For instance, an app with contextual learning analytics could be embedded into the learning context when a learner learns alone without an instructor's supervision, understands the context using artificial intelligence, and visualizes the data presented in that specific context. A contextual learning analytics-enabled app could track the learning context using contextual data (such as location, time, and place), and therefore, a learner does not need to leave that particular learning context to see the analytics, thus avoiding context switching.

1.2. Vocabulary Learning and Contextual Factors

Computer-assisted language learning is a discipline that primarily concerns using computer, mobile, ubiquitous, pervasive, and multimedia technologies to support foreign language learning. Contemporary learning technologies are deeply embedded in learning design, and with the emergence of AI, computer-assisted language as a discipline continues to grow. Therefore, computer-assisted vocabulary learning synonymously with foreign vocabulary learning as a sub-domain of computer-assisted language learning has received much attention [4]. This sub-domain focuses on how foreign vocabulary can be learned using technologies, including multimedia annotations such as picture, sound, text, video, and graphics. Foreign vocabulary acquisition is considered to be a challenging task for foreign language learners, as vocabulary learning is not much emphasized in in-class activities [5]. Vocabulary is known to be learned using informal learning. As a result, many foreign language learners learn foreign vocabulary using computer and mobile-assisted learning systems. To foreign language learners, meaningful contexts are essential as those learning contexts can help them memorize new words and use them in similar contexts. Prior research to describe the necessity of learning contexts in foreign vocabulary learning suggests that- each word, when used in a new context, becomes a new word that helps us make meaning [6]. A previous study also addressed that most vocabulary is acquired from contexts [6]. Apart from explicit instruction, what a word means often depends on the context in which it is used, and people pick up much of their vocabulary knowledge from context [7]. Based on the learning contexts and learning design literature, it is clear that learning contexts are vital in enhancing vocabulary development while using informal learning approaches. Therefore, many conventional ubiquitous learning systems allow learners to take notes or memos to describe their learning contexts.

1.3. Contextual Learning Analytics

In this regard, learning analytics [8] and its sub-domains such as ubiquitous learning analytics [9], contextual learning analytics [10], and contextualizable learning analytics [11] are continuously contributing to understand and optimize learning and the environments in which learning occurs through the measurement, collection, analysis, and reporting of data about learners and their contexts [3]. Contextual learning analytics primarily aims to understand the learning context by providing awareness information to the learners. It uses analytics and artificial intelligence to present information about a specific context to a dashboard for generating awareness information. However, understanding a learner from the self-described notes is extremely complex. In general, describing the learning context automatically by a contextual learning analytics app could be challenging. Therefore, further investigation is required about learning context generation and contextual learning analytics as these are yet to be very much an under-researched area of practice and research.

1.4. Motivations

To date, technologies for foreign vocabulary learning using various annotation styles such as textual meaning, audio, graphics; intelligent language tutoring systems that include sophisticated feedback systems; and hypermedia-enhanced learning systems are developed. Moreover, various ubiquitous and mobile learning tools are developed and available to use to enhance vocabulary. While these technologies provide various advantages, questions remain regarding how these technologies have been used to achieve learning objectives. For example, ubiquitous learning technologies, a renowned technology to support vocabulary learning, often provide the facility to learn anytime and anywhere. In addition, in this kind of learning environment, previous learning experiences are used to measure one learner's current knowledge level, which can be used to learn new knowledge. However, there are certain limitations; namely, the scope of learning new vocabulary is limited, students often cannot determine which words to be learned next, low learner engagement with the system, and describing each learning context for each word may be challenging, etc., are often discussed.

Within this research scope, a publicly available dataset [12] is analyzed to understand how the learning contexts in ubiquitous learning are supported. As of now, the dataset consists of 31,420 learning logs therefore, those logs are analyzed. The dataset consists of a total of 1959 users. Out of 1959 users, the registered English-native, Japanese-native, and Chinese-native users are 325, 451, and 455, respectively. The remaining 758 users are native speakers of various languages such as Arabic, German, and French. Our analysis indicates that most of the learning logs created by the users are created without a description of the learning context. In Figure 1, the result of the log analysis is presented.

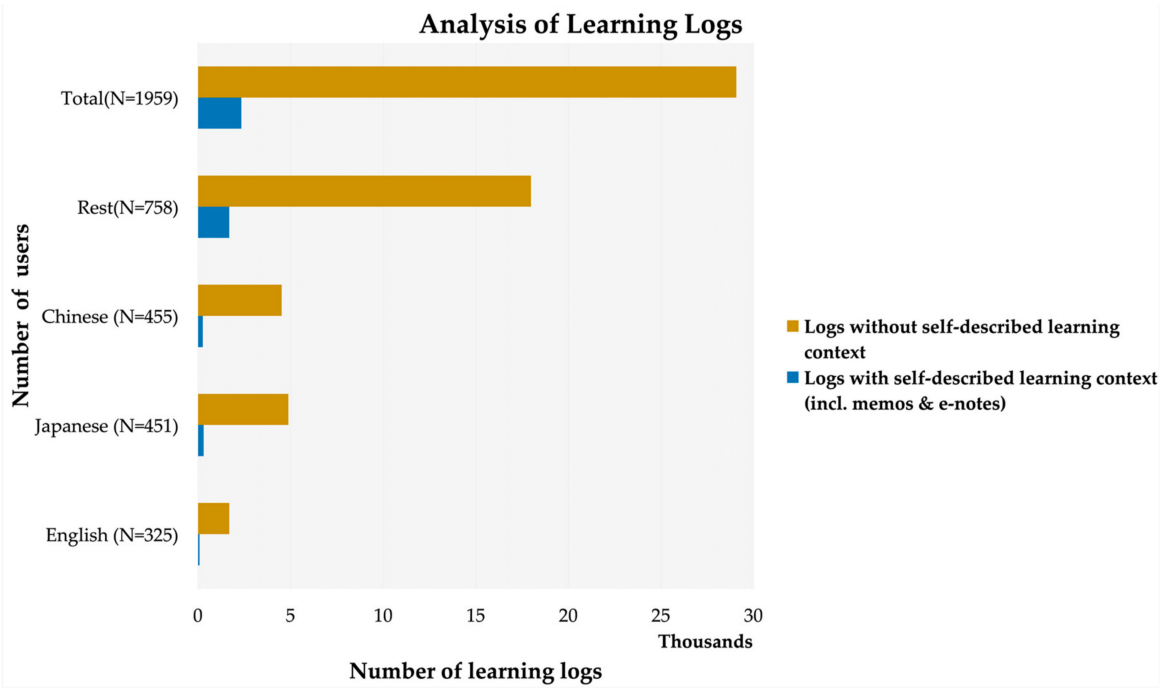


Figure 1. Result of the analysis.

The result indicated that describing the learning contexts manually (i.e., taking memos or e-notes) is a burden for the learners, and therefore it is often skipped by the learners. Our analysis contradicts the research that suggests that learning context is an essential component in vocabulary development. Therefore, to understand foreign language learners' learning contexts and support them, we aimed to develop our ecosystem. In our ecosystem, foreign language learners will be supported by smartly-generated learning contexts generated automatically by artificially intelligent (AI) models.

1.5. Contributions

Earlier, we discussed how crucial meaningful learning contexts could be in enhancing vocabulary. However, we found (referring to Figure 1) that describing the learning context each time for each word to be learned is considered time-consuming and challenging for the learner. Therefore, this research investigates how AI technologies can be applied to generate unique but meaningful learning contexts. In this study, to produce smartly-generated learning contexts, we used image understanding as the method. In this study, to clarify, images refer to pictures captured (i.e., lifelogged) by foreign language learners in real situations. This study is limited to the lifelogging images captured in a real situation either in a particular physical location (say a university or a supermarket or a station) or at a particular place in a piece of software, creating a learning challenge. Our hypothesis is in the learning design line, suggested by a previous study, whereby real performance contexts are turned into learning opportunities [13–15]. This study hypothesized that, from real-life (i.e., lifelogging) images, it could be possible to detect and understand where the learner is and provide support. Besides, we can minimize transfer distance and mimic a real situation in a context [16]. Moreover, the images captured in authentic contexts by the learners may contain rich and valuable information that could trigger their memories. It is worth mentioning that an image speaks better than thousands of words.

Taking the learning design and contextual factors into account, we analyze a lifelog images' visual contents to generate learning contexts that we refer to as the smartly-generated learning contexts for vocabulary learning. The assumption was that those smartly-generated learning contexts would help learners enhance their vocabulary. Therefore, the original contributions of this study are firstly, to introduce a new learning design; secondly, to use image understanding as the method to solve a complex problem about learning context representation; and thirdly, develop a new tool that could take the burden of taking handwritten memo or electronic notes.

2. Literature Review

Our study mainly focused on combining a context-aware mechanism with image-to-context generation and real-time analysis mobile learning (M-learning) techniques to create an intelligent ubiquitous learning environment. The IUEcosystem (Image Understanding Project) has been developed to enhance one's willingness to learn, remember learning contexts from visual clues and location, and effectively use vocabulary in similar real-life contexts. The following sections explore current ubiquitous learning applications that use context-aware learning theories (in Section 2.1), a glimpse of our previous works is outlined in Section 2.2, and AI-based image analysis models that could be used to generate learning contexts and understand the hidden contextual factors are explored in Section 2.3.

2.1. Survey on Compter-Supported Applications for Language Learning to Identify and Bridge the Gaps

The continual innovative progress of various information, communication, and multimedia technologies has led to massive developments in computer-mediated, mobile, and ubiquitous learning environments. Several learning environments have been brought light in contextual support using the context-aware mechanism and shown great promises. For example, Personalized Context-aware Recommendation (PCAR) is a learning system that uses Global Positioning System (GPS) data in combination with a personalized context-aware learning algorithm to support English language learning [17]. This novel context-aware recommendation learning mechanism can provide appropriate English learning content to learners according to their location and environment through a 3G, 4G, or wireless local area network. However, an automatic generation of learning contexts is not offered in this study. Another promising development is TANGO (Tag Added learNinG Objects) [18]. This system can support learners to learn Japanese efficiently by accessing virtual context-aware learning information in real-world situations. However, the automatic context generation function is not available in this system. Mathew Montebello's study introduced the circumstances that may play roles in designing smart ubiquitous learning environments [19]. This study suggests that social

aspects, including students' and teachers' communal presence within the classroom, physical or virtual, can support context while using technologies. In ICT-based science education, Muhammad Ashar et al. developed a tool that disrupts learning through intelligent algorithms to help the learning process be more intelligent optimal [20]. However, like other tools mentioned above, this tool does not support remembering learning contexts.

2.2. Our Previous Developments

Our research team is actively researching the line of computer-assisted language learning, learning analytics, and ubiquitous learning. We have developed several tools such as AIVAS (Appropriate Image-based Vocabulary Acquisition System) [5,21,22], SCROLL (System for Capturing and Reminding of Learning Logs) [18], WLS (Word Learning System), and LFO (Learn From Others) Panel [23]. AIVAS is a web-based vocabulary learning system that assists foreign language learners in creating on-demand learning materials using multimedia annotations. SCROLL is a ubiquitous tool that collects ubiquitous learning logs and analyzes them for feedback. The WLS is a native iOS app that uses the learning material creation mechanism of AIVAS. However, the WLS uses spaced repetition to reflect on learners learning. Finally, the LFO Panel is a vocabulary recommendation tool that uses learning analytics to determine authentic, partially-authentic, and word-only vocabularies. This tool uses the KNN-based profiling method to find top five partners to recommend authentic learning experiences.

Nevertheless, all our applications require manual input (handwritten and type-based) from the learning context. In other words, a user of our systems must describe the learning context by adding a memo to the system. There is no mechanism in our tools that could support our users through an automatic context suggestion. Therefore, we introduce the ecosystem of the Image Understanding Project. In this project, we produce smartly-generated learning contexts as an alternative to handwritten memos or electronic notes. With this project, we aim to support foreign language learners by understanding their contextual factors such as where learning takes place, visual contents of their lifelog images that could describe a context the most, and emotions with artificial contexts. Using this intelligent ubiquitous learning environment, we aim to connect the dots among other existing tools.

2.3. Survey on Automatic Image Captioning Models

Google has developed a multimodal image captioner called Image2Text [24]. The Image2Text system is a real-time captioning system used to generate a human-level natural language description for any input image. In the Image2Text model, a sequence-to-sequence recurrent neural networks (RNN) model is used for caption generation. This system also enables users to detect salient objects in an image and retrieve similar images and corresponding descriptions from a database. The O'Reilly's Show and Tell Model [25] is a well-known image caption generation model that combines the advances in computer vision and machine translation for producing realistic image captions. This model uses neural networks to process input images. These models are trained to maximize the likelihood of producing a caption from an input image.

The Show and Tell Model can be used to generate novel image descriptions. Microsoft Cognitive API [26], a Microsoft cognitive service, is a popular API for understanding natural images. This API can be used as a service for images' visual content analysis. It returns features about visual contents found in a raw image in the form of sentences. This API uses domain-specific models and descriptors for identifying content and label uses tagging. Another promising caption generation model is deep visual [27]. The deep visual model generates natural language descriptions of images by analyzing the regions of the images. The deep model leverages datasets of images and their sentence descriptions for learning the intermodal correspondences between language and visual data [27]. The automatic image captioning technology brought a new dimension to computer vision and Natural Language Processing (NLP)-related fields. However, as of yet, this technology is not much explored in

the context of education. Therefore, we aim to use this technology as an applied AI to build a smart computer-assisted language learning environment and conduct contextual analytics research.

3. Materials and Methods

3.1. System Architecture

The IUEcosystem is designed primarily for foreign language learners in describing their learning contexts automatically. Currently, this technology is a featured application of our previous works. Unlike conventional ubiquitous language learning systems where place, time, and handwritten memos are used as the contextual clues, the newness of this proposed technology is the uses of lifelogging images as the primary contextual clue. We leveraged lifelog images because the visual contents of lifelogging contain strong social interaction. The analysis of social interactions in lifelogging data is of fundamental importance to understanding human behavior, and the presence of people and social interactions are consistently associated with our memory. Hence, we hypothesized that as lifelog images contain vital information, they need to be analyzed for educational use [28]. In Figure 2, the architecture of the IUEcosystem is introduced.

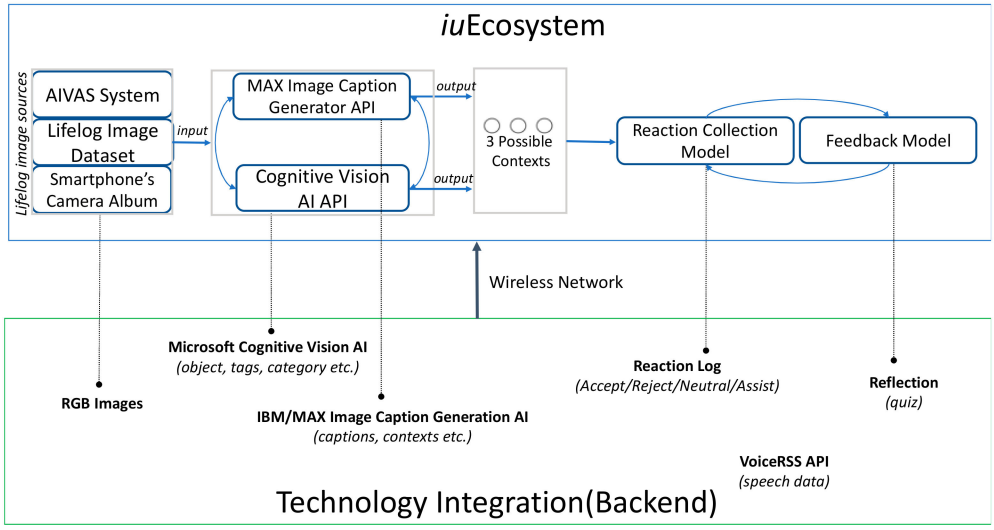


Figure 2. The architecture of the system.

There are three primary sources of getting lifelog images from the learners in this ecosystem—first, the AIVAS (Appropriate Image-based Vocabulary Acquisition System) system. The AIVAS is a web-based vocabulary learning system where a learner can either upload his/her personal image or search image using image API. The second source is a lifelog image dataset. The lifelog image dataset has over a thousand images that are collected from foreign language learners. The third source is uploading an image using a smartphone's camera function.

Once the system receives an image with its description provided by the learner, the image is directed to two AI-based image captioning models. These image captioning models are the Max image caption generator and Microsoft's cognitive vision API. After that, each image captioning model generates three possible descriptions by analyzing the scenes of the image that we address in the study as the smartly-generated learning contexts.

When the smartly-generated learning contexts are generated, they are displayed on the screen (refer to Figure 3). This allows a learner to select their favorite contexts out of six smartly-generated learning contexts. When a learner selects their favorite smartly-generated learning contexts, data are collected by the reaction collection model. In the reaction collection model, we collect the satisfaction data such as *accept* a smartly-generated learning context, *reject* a smartly-generated learning context, or put a *neutral* reaction on a smartly-generated learning context.

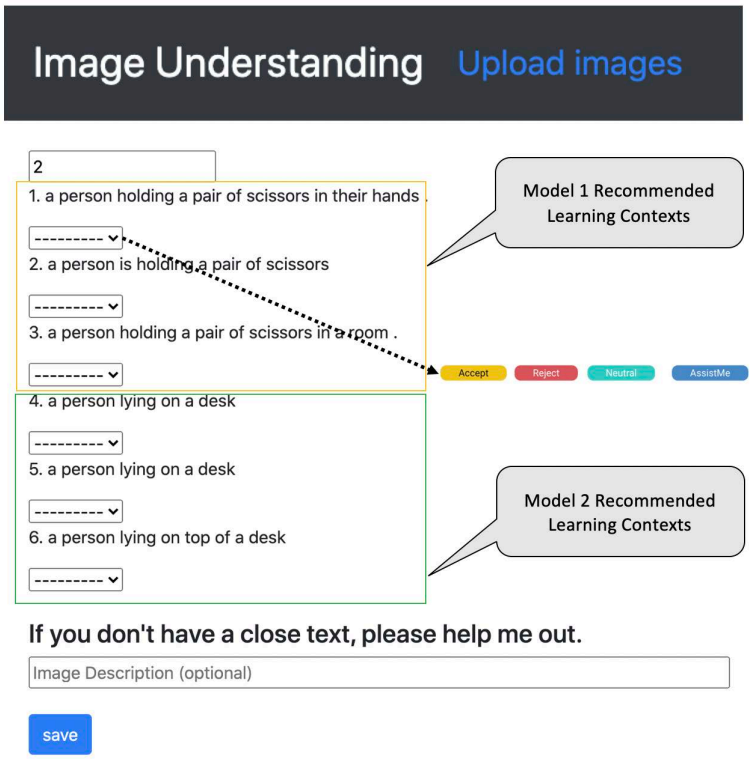


Figure 3. The user interface to collect feedback.

The feedback model analyzes the satisfaction data that are collected by the reaction collection model. In generating time-based notification (refer to Figure 4), the feedback model primarily uses the information of the 'accepted' smartly-generated learning context(s). When the feedback model generates a time-based notification, it also analyzes the image's EXIF (Exchangeable Image File Format) information, such as the location where the image was taken and the time that image was taken. Finally, the feedback model generates a notification and deliveries learning material for the learner.

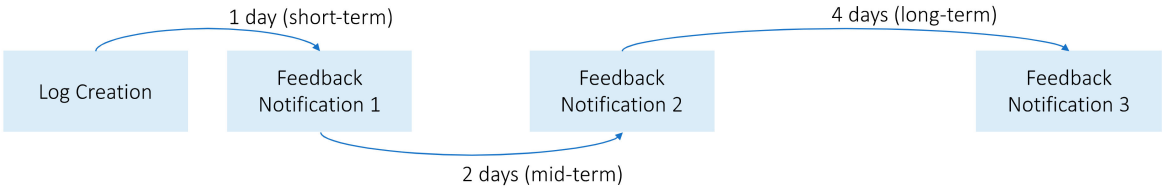


Figure 4. The method for time-based notification.

3.2. *Analytics on Lifelogging Images: Fitting into Language Learning Research*

In this study's scope, we aimed to understand how analytics on lifelogging images or image understanding using AI-models for contextualization fits into the computer-assisted language learning discipline. To do so, we use an example of a learning scenario, for instance, in a learning situation where a learner knows words such as *breakfast* and *sandwich*. Afterward, the same learner captures a picture in a new learning situation and uploads it into the IUEcosystem' server. The system will analyze the visual contents of the image and produces smartly-generated learning contexts (which can be exampled as, 1. *A plate of snacks*, 2. *Combination of sandwich, spoon, and soup*, 3. *A breakfast plate with sandwich, plate, and spoon*) by which the learner's prior vocabulary knowledge can be triggered. Those words can be mapped with the smartly-generated learning contexts represented by the system. In addition, the learner could learn new words such as *spoon*, *plate*, and *snacks*. In this work, we aimed to develop the IUEcosystem in a way that will accept the pictures that are uploaded by the registered

users. Then, the system will analyze the visual contents of those pictures. Furthermore, based on those visual contents, the system will produce smartly-generated learning contexts.

By doing so, some educational advantages could be achieved, such as:

i. The smartly-generated learning contexts will in mapping a learner's prior vocabulary knowledge with new knowledge so that they can review and recall the previously acquired vocabulary [29];
ii. The smartly-generated learning contexts will help a learner to memorize new vocabularies;
iii. The smartly-generated learning contexts may help in acquiring multiple vocabularies from a single learning experience [29];
iv. The smartly-generated learning contexts may provide the visual mnemonics that will work as the visual cues to highlight that particular word (i.e., the target word) in a specific learning context.

In Figure 5, we present the learning design of the IUEcosystem. The learning design of the IUEcosystem is based on three technologies, namely a Learning Context Representation Model (LCRM), a Learning Context Analytics (LCA), and a Picture-assisted Quiz System (PQS). The first technology, the LCRM, is used to analyze the scenes of the lifelogging images. The LCRM generates smartly-generated learning contexts. The second technology, the LCA, analyzes the smartly-generated learning contexts for producing a bag of words that contains many new vocabularies. We assume this will enhance the learning scope and increase learning curiosity. The third technology, the PQS, is used to generate picture-based quizzes so that the learner can review their knowledge after a certain period.

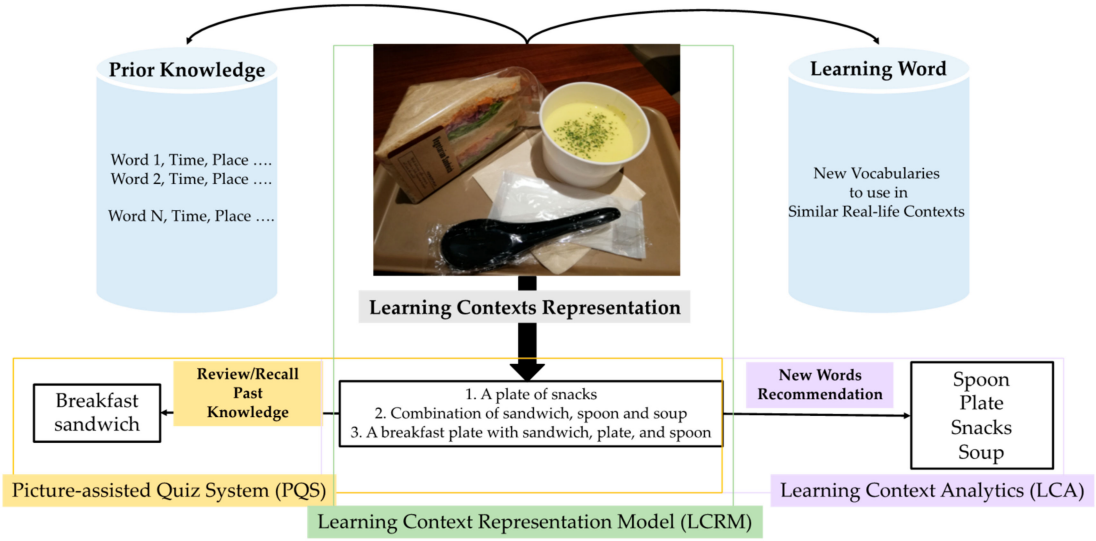


Figure 5. The learning design.

Although computer vision and natural language learning research have advanced, producing meaningful learning contexts is not easy. Besides, connecting learning analytics with computer vision and natural language processing is quite a challenging task. Leveraging analytics on lifelogging images using image captioning technology, we aimed to build an environment where learning context and multiple vocabulary acquisition and reflection of prior knowledge to learn new knowledge are supported. To our best knowledge, no recognized research in computer-assisted language learning literature has used scene analysis or image analytics for understanding learning contexts for foreign vocabulary learners. Therefore, we believe this paper will bring newness and opens the scopes to new research.

3.3. Real-Life Scene Analysis Using the Pre-Trained Models

In the literature review, we articulated several promising frameworks that could analyze scenes of natural images. In this study, we used two pre-trained neural and probabilistic caption generation

frameworks, namely, Max Image Caption Generator and Microsoft Cognitive Vision AI, for generating descriptions from images. Microsoft Cognitive Vision AI is an AI service of Microsoft, whereas Max image caption generator is an open-source project by IBM. Max image caption generator uses the deep architecture of the Show and Tell Model [25]. In this section, we discuss how lifelogging images are analyzed in this study.

3.3.1. Formula

Recent machine translation mechanisms have demonstrated great success in achieving results by directly maximizing the correct translation probability. This is achieved by end-to-end style for both training and interference [25]. In this model, recurrent neural networks are used to encode the variable length input into a fixed dimensional vector. It uses fixed dimensional vector representation to decode it to the sentences (i.e., captions). In Show and Tell, at first, the following formula (where I = input image, S = an image's correct transcription, and θ = the parameters of the model) is used to directly maximize the probability of the correct description from an input image [3].

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_{(I,\, S)} \log p(S|I;\theta) \tag{1}$$

After that, the following chain rule is applied to determine the joint probability over S_0, S_1, ... , S_N (where N = length, (S,I) = training pair, and $t-1$ is expressed by a fixed length hidden state or memory h_t).

$$\log p\ (S|I) = \sum_{t=0}^{N} \log p(S_t|I,\ S_0, \ldots, S_{t-1}) \tag{2}$$

Next, the recurrent neural network uses the following formula (where f = a non-linear function) to update the memory h_t.

$$h_{t+1} = f\ (h_t,\ x_t) \tag{3}$$

Finally, in order to determine the best output for f function, LSTM (Long-Short Term Memory) architecture is used.

3.3.2. Scene Encoding and Decoding Process Using LSTM Architecture

LSTM (Long-Short Term Memory) neural and probabilistic framework is used for solving complex problems such as language modeling and machine translation. The pre-trained model of the Show and Tell model uses the Long Short-Term Memory (LSTM) network, which is trained as a language model conditioned on the image encoding. We also used an Inception-v3 image recognition model [30] that was pre-trained on the ILSVRC-2012-CLS image classification dataset. Figure 6 shows the architecture of the model.

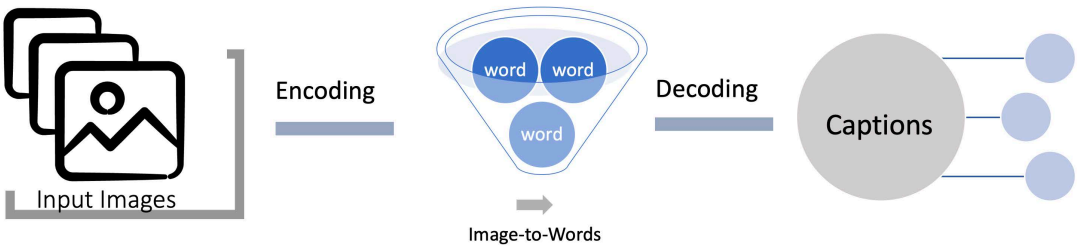


Figure 6. The architecture of the LSTM (Long-Short Term Memory)-based Show and Tell model.

In the scene encoding and decoding process, the architecture of LSTM removes the last layer from the loaded model for internal information representation, as a conventional architecture of the model is used to predict a classification for an image. Our model is less focused on classifying images, but we are interested in an image's internal representation right before a learning context is

generated. Therefore, we used the output of the last layer's output to represent the learning contexts. In our architecture, Inception-v3 is used as this image recognition model is one of the most popular convolutional neural network models for recognizing objects from natural images. Google develops the architecture of Inception-v3, which is the 3rd version in a series of deep learning convolutional architectures. The architecture [30] of the Inception-v3 is made up of symmetric and asymmetric building blocks, including convolutions, average pooling, max pooling, concats, dropouts, and fully connected layers. In addition, loss is computed via softmax.

3.3.3. Dataset

In this study, the pre-trained model is trained on the Microsoft Common Objects in COntext (MSCOCO) image dataset [31]. The dataset contains images of complex everyday scenes containing common objects in their natural context. Hence it performs well on pictures that are captured in daily-life. We used this dataset as this dataset is a standard benchmark for sentence-based image descriptions. It contains the highest number of labeled real-life images for computer vision training compared to other recognized image sets. Moreover, in the line of contextualization research, this dataset and Show and Tell's neural network architecture is used by recognized research [28]. A comparison of image sets is given in Table 1.

Table 1. A comparison of standard benchmark image sets for sentence-based image description.

Dataset	Total Images	Details
MSCOCO	330 K	>200 K labelled
ImageNet	14 M *	1000 categories
Google's Open	9 M *	6000 categories
CIFAR-10	60 K*	10 classes
Flickr30K	30 K*	158 K captions

* not all images are real-life images (contains annotated images).

3.3.4. Hardware Configurations and Computation Speed

Although artificial intelligence is gaining popularity to build various learning systems, learning context representation is a complex task. It requires a high-performance computing environment to use AI-based models. Table 2 shows the current hardware configurations of the IUEcosystem. With this hardware configuration, it may take 10–14 days to train the model. Therefore, we use the pre-trained model.

Table 2. Hardware configuration.

Type	Details
CPU	Core i9-10850K
GPU	Nvidia GeForce RTX 2080Ti 11GB
Training time	10–14 days
Libraries	Numpy, Python 3.7.0, Django
Toolkit	NLTK
Dataset	MSCOCO

The computation time depends on several factors such as the dimension of the image, aspect ratio of the image, the complexity in the scenes, and categories of the detected objects. Furthermore, the computation time may vary based on internet speed and the users' hardware configuration. With the current hardware configuration (refer to Table 2), the pre-trained image captioning models of IUEcosystem can instantly generate smartly-generated learning contexts. As the system generates results instantly, it helps the users to generate learning logs without much delay. However, depending on the internet connection and other factors, there may be some delays in generating smartly-generated learning contexts.

4. Development

This section describes the development of the system along with the user interfaces. First, in Section 4.1, we demonstrate how to sign-up to the system. After that, how to create a log in IUEcosystem is shown in Section 4.2. Next, in Section 4.3, we exhibit how we collect interaction data from the learning contexts provided by the learning context representation models. Finally, in Section 4.4, we illustrate how the system sends a time-based notification to the learners.

4.1. Sign-Up and Log-In

In Figure 7, we show the sign-up page of the system. Using this sign-up page, we collect demographic data from the user and analyze them later. The demographic data we collect include age, foreign language(s) to study, purpose, cultural background, interested culture, primary study location, and gender. We also collect the username and email address for log-in purposes.

Figure 7. The user interface of the log-in page.

4.2. Image to Learning Contexts Generation

Figure 8 is the interface where a learner can create a log. For creating a log, a user is asked to input vocabulary (including a word, a phrase, or a sentence) together with the lifelog image to analyze. Once an image is uploaded with its corresponding vocabulary (i.e., description of the image), the system asks the user to choose one of the two AI models to analyze the image's scenes. A user can choose either or both of the AI models to analyze the picture. After the scene analysis process is completed, the system returns the analysis results to the user in sentences. At present, to describe each input image, each of our AI-based models returns three smartly-generated learning contexts in a learner's target languages. The system is customizable, so it is possible to produce the smartly-generated learning contexts in a learner's native language.

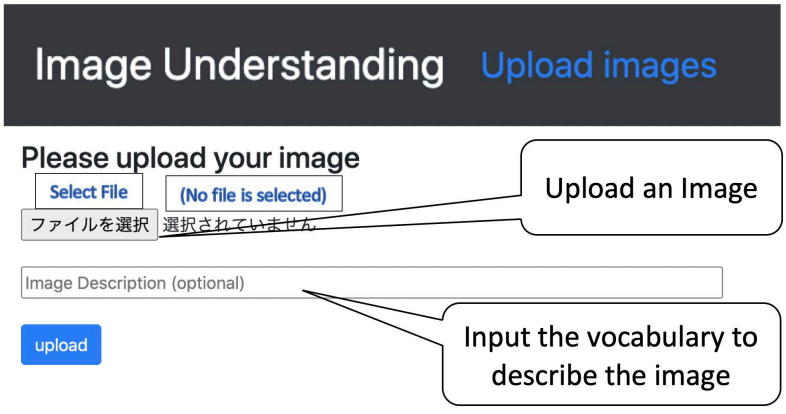


Figure 8. The user interface to create a vocabulary learning log and to analyze an image.

4.3. Interaction Data Collection

In collecting learners' actual learning satisfaction data, we primarily collect four types of satisfaction data. They are *accept*, *reject*, *neutral*, and *assist me*. The *accept* data represents that the user is happy and accept a context. The *reject* data represents that the user is unhappy and does not accept a context

generated by the model. The *neutral* represents that the user is neither happy nor unhappy with a context provided by the model. Finally, the *assist me* data represents that the user is confused and needs assistance by the system to decide. The IUEcosystem's time-based notification mechanism uses these simple satisfaction data to understand the learner's emotion and motivation. With these logs, the system assesses the performance of the AI models that are used.

4.4. Time-Based Notification

At present, the system sends notifications to the learners on a specific time interval to engage themselves in learning activities. To do this, we created a simple stochastic mechanism for sending notifications. The notifications are created based on the log creation time.

The first notification is sent after one day of the log creation. The second notification for a created log is sent after three days. Finally, the third notification is sent after seven days of the log creation. After receiving the feedback notification, a learner is expected to open and review the log that they created previously. The method for a time-based notification generation is shown in Figure 8.

As stated earlier, the first notification is addressed as the short-term notification to assist the learner in short-term memory retention. The first notification is sent after 1-day of log creation. The second and the third notifications are addressed as the mid-term and the long-term notification, respectively. The second notification is sent after two days of the first notification. The third notification is delivered after four-days of the second notification. In generating these notifications, the feedback model analyzes learning satisfaction data (i.e., *accept*, *reject*, *neutral*, and *assist me*) from the reaction collection model. At present, the feedback model considers those smartly-generated learning contexts that are only accepted by the learner. In other words, smartly-generated learning contexts that are confusing (determined by the *reject*, *neutral*, and *assist me* data) are not taken into consideration. In generating time-based notifications, the feedback model also uses EXIF information, such as where and when the picture was taken.

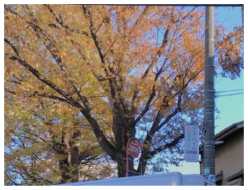
4.5. Technical Specifications

The IUEcosystem is developed using a free and open-source web framework called Django. Therefore, Django is used for both front-end and back-end development. Currently, SQLite is used as a database, however, we aim to use Firebase in the future as Firebase offers easy tools for machine learning. Sendgrid API is used for email management related tasks such as email verification and resetting passwords. Celery is used for notification scheduling. Python and HTML are used for front-end. Please refer to Table 2 for hardware-related specifications.

5. Performance Evaluation of the Models in Producing Smartly-Generated Learning Contexts

In performance evaluation, we compare the results between the two models we used to implement the IUEcosystem. We compared the results with the learner-described learning contexts. For this experiment, we analyzed five learning logs of a foreign language learner captured by the IU Ecosystem. The learner is a Japanese national enrolled in a Japanese university aged between 20–24. The learner is learning English as a foreign language. For this experiment, the learner described his vocabulary learning contexts in his native language, namely Japanese. The corresponding translation in the English language is presented in Table 3. For this evaluation, the transfer of knowledge representation is into a learner's target language(s). We assumed that knowledge representation in the target language would be more effective in acquiring knowledge. For example, for a learner who wishes to learn English as the target language, the English language's knowledge transfer would help the native language. The knowledge transfer is highly integrated with the learning design. In our system, the knowledge transfer of image captioning entirely depends on the APIs that we used to develop the framework.

Table 3. A comparison of standard benchmark image sets for sentence-based image description.

Image	Vocabulary	Learner-Described Context	Model 1-Generated Context (Probability)	Model 2-Generated Context (Probability)
	Word to be learned: 肉 (meat) Additional vocabulary: 卵 (egg), おわん(bowl), お皿 (plates)	定食屋のランチセット Lunch set at a set restaurant (in English)	A table topped with plates of food and cups of coffee (0.00077)	A pot that is sitting on a table (0.62)
	Word to be learned: 猫 (Cat) Additional vocabulary: 木 (wood), 土 (soil), 網 (net), レンガ (brick)	猫のいる公園 Park with cats (in English)	A black and white cat sitting on top of a wooden bench (0.00179)	A bear that is standing in the grass (0.80)
	Word to be learned: パンケーキ (pancake) Additional vocabulary: バナナ (banana), ブルーベリー (blueberry), コーヒーカップ (coffee cup)	フルーツパンケーキ Fruit pancakes (In English)	A white plate topped with different types of food (0.00087)	A plate of food on a table (0.97)
	Word to be learned: 木 (tree) Additional vocabulary: 標識 (sign), 看板 (signboards), 電柱 (utility poles)	道路沿いの植木 Trees along the road (In English)	A clock on a pole in front of a tree (0.00009)	A close up of a tree (0.96)
	Word to be learned: 人 (man) Additional vocabulary: 公園 (park), 建物 (building), 木 (tree), 空 (sky)	広い公園 Large park (In English)	A large clock on the side of a building (0.00047)	A building that has a sign on the side of a road (0.92)

For creating a log, the learner inserted vocabulary together with the image to represent the word. We instructed the learner to include a description of the learning context in the system. We also instructed the learner to include additional vocabularies that he wishes to learn using one image. Therefore, in Table 3, 'word to be learned' refers to the main vocabulary that the learner wishes to learn, while 'additional vocabulary' refers to the auxiliary vocabularies the learner wanted to learn using the same image. In Table 3, we present the results of human-described learning contexts in comparison with AI-inspired models.

We conducted a *t*-test to compare the results between the two models we used to develop the IU Ecosystem. The *t*-test is based on the probabilities of the models for each of the images. The result indicated that Model 2 significantly outperformed Model 1 ($p < 0.0001$). The mean of Model 1 and Model 2 was 0.0007 and 0.85, respectively. The SD of Model 1 and Model 2 was 0.0006 and 0.14, respectively. Based on the result, we yield that Model 2, which is based on Microsoft's service, offers higher confidence in describing the scenes of the natural images captured by language learners for vocabulary learning.

6. Featured Application

The research and development of the IU Ecosystem is a featured application for our previous works described in Section 2.2. At this stage, our existing tools, namely AIVAS [5,21], SCROLL [12,18], and LFO Panel [23] do not support learners in the automatic generation of learning contexts. Therefore, we developed IUEcosystem as a featured application to our existing systems. We aim to use this application to understand foreign language learners' behavior in terms of context descriptions, note-taking behaviors, quiz generation using the AI-generated contexts, memory retention, and other contextual factors.

7. Discussion

In language learning, learning contexts play a crucial part in a learner memorizing new vocabularies. While using computer-assisted learning systems, it is suggested that learners take short notes of their learning contexts, typically in electronic notes or memos. These self-described notes help them understand their learning processes in retention and reflection, including where, when, and how the new knowledge was acquired. In addition, a self-described memo is considered to be a personal record of a learner's authentic learning experience. Research domains such as contextual learning analytics, contextualizable learning analytics, and context-awareness are continuously trying to understand a learner's learning behavior and learning processes by proposing innovative learning designs. Furthermore, ubiquitous learning logs are analyzed using machine learning and AI-based methods to uncover information on learning contexts and contextual factors. As learning processes occur in physical and digital spaces, contextual learning analytics could be a powerful approach to establishing a connection between a learning design and the contextual factors associated with a particular learning context. Contextual learning analytics apps aim to understand the learning contexts and contextual factors such as emotion, motivation, and reflection data, together with conventional contextual data such as time, place, and multimedia information, in developing analytics dashboards and predictive models. Our analysis indicates that (refer to Figure 1) describing the learning context manually for each vocabulary is considered a burden and a time-consuming process. A common tendency among foreign language learners is ignoring taking the learning context's memo where they learn new vocabulary. Therefore, understanding the learning contexts in a computer-assisted language learning environment is essential.

The conventional approaches in computer-assisted language learning environments primarily used location, time, conceptualization, knowledge, and interaction-based methods; however, the conventional approaches ignore image understanding-based approaches. One of the key advantages of using image understanding-based methods is that for humans, it is relatively easy to describe learning contexts using images because images describe an immense amount of details at a glance that

text annotations cannot do. Prevalent learning theories such as the picture superiority effect [32,33] in learning sciences also support the idea that images are a powerful tool in gaining new knowledge. Previous studies in the area of image captioning in various languages have explored that potential from different perspectives. For instance, focusing on developing a Japanese version of the MS COCO caption dataset and a generative model based on a deep recurrent architecture that takes in an image and uses this Japanese version of the dataset to generate a caption in Japanese [34] and constructing a generative merge model for Arabic image captioning based on deep RNN-LSTM and CNN models [35]. However, using these image captioning models has not been explored much in contextual learning analytics and computer-assisted language learning disciplines.

In this study, we used an image understanding-based method for overcoming the problems that language learners face in describing learning contexts while learning foreign languages using a computer-mediated informal learning environment. As a method, visual contents analysis (i.e., scene analysis) of foreign language learners' lifelog images using automatic image captioning technology is employed. Modern architectures of automatic image captioning technologies use artificial intelligence that connects computer vision and natural language processing. This study used a caption generator model called Max image caption generator pre-trained on Microsoft's COCO (Common Objects in Context) dataset and Microsoft cognitive services to build the ecosystem. Our system can understand foreign language learners' lifelog images to produce a bag of words (i.e., image-to-word) and produce three smartly-generated learning contexts automatically of an image. The system feedbacks each of the learning contexts in the form of a flashcard-like interactive gamified learning environment. A flashcard consisting of a learning context is displayed on the user interface asking for learners' reactions to the system recommended learning context associated with the vocabulary. While interacting with the flashcard, a learner can either *accept*, *reject*, *neutral*, or *assist me* in evaluating each of the smartly-generated learning contexts recommended by the IUEcosystem.

The IUEcosystem is designed to transfer the knowledge representation into a learner's target language(s). In learning design, we assumed that a foreign language learner's memo or the learning context representation in the target language would be more effective in acquiring knowledge. For example, if a native speaker of Japanese wants to learn the English language, our system will present the English language's smartly-generated learning contexts. Therefore, the knowledge transfer is highly integrated with the learning design. In this intelligent ubiquitous learning system, knowledge transfer of image captioning depends on the APIs that we used to develop the framework. To accomplish this task, first, both models (i.e., Max image caption generator and Microsoft's cognitive vision API) are trained individually into the MSCOCO dataset. Finally, the knowledge is transferred using LSTM-based image-to-words encoding and image-to-sentence decoding processes (refer to Figure 4).

At present, the system has several limitations. First, it uses a simple stochastic mechanism for time-based notification delivery. However, this may not be effective for some learners. Therefore, we aim to integrate a spaced repetition-based mechanism such as the Leitner system in the future. Second, we proposed a new learning design. However, at this point, we could not evaluate the learning effectiveness of this learning design. Third, in this paper, we could not address how the smartly-generated learning contexts could enhance vocabulary in daily life learning contexts and how these types of artificially-generated sentences could be an alternative to learning notes/memos. Lastly, the reaction collection and feedback models of the IUEcosystem are not AI-based models. Therefore, we aim to apply AI to collect learners' reactions and provide feedback to them. Our future works will address the existing limitations.

8. Conclusions

In this paper, we introduced the development of IUEcosystem. The key advantages of this platform are that- first, it allows one to capture learning logs. Second, it produces smartly-generated learning contexts by understanding and analyzing images that are captured for vocabulary memorization. Doing this reduces the learners' burden in hand-written memo taking or electronic note-taking

by typing. This paper also proposed a learning design that could be used to memorize multiple vocabularies using a single learning log. We presented the results of human-described learning contexts—vs.—two image captioning models inspired by modern AI architectures.

The system is designed in a way that it could produce smartly-generated learning contexts in a learner's target languages. When the system generates smartly-generated learning contexts in a learner's target language, those are displayed on the screen. The system lets the learner choose the most appropriate smartly-generated learning context(s). The system stores this information to the reaction collection model and later uses it to provide a time-based notification. In generating a time-based notification, the feedback model analyzes the reaction collection model's satisfaction data. In generating time-based notifications, the feedback model generates a short-term (after 1-day of log creation) notification, a mid-term (after 2-day of the short-term notification) notification, and a long-term notification (after 4-days of the mid-term notification). When a notification is generated, the system uses an image's EXIF information embedded inside the image.

This research is a featured application for our existing ubiquitous learning platforms. By developing this intelligent ubiquitous learning environment, we aim to contribute to applied AI in language learning and the area of contextual learning analytics, which is still a very much under-researched area of practice and research. Hence, in the future, we aim to integrate it with other tools. In the future, we aim to conduct evaluation experiments regarding the efficiency of the context descriptions, note-taking behaviors, quiz generation, memory retention, and contextual factors (such as emotion and learning satisfactions).

We conclude this article by discussing a number of critical privacy and ethical issues that have to be considered by learning designers, developers, and researchers in learning analytics, ubiquitous learning analytics, and contextual learning analytics. In recent days, employing analytics on learners' learning logs is cited as one of the key emerging trends while developing an intelligent ubiquitous learning environment. In other words, an analysis of learning logs relating to learners and their engagement with their learning is the foundation of a smart ubiquitous learning environment. However, such data collection and its use face several ethical challenges, including issues of location and interpretation of data, informed consent, privacy, the deidentification of data, and data classification and management [36]. For instance, in learning analytics, ethical issues fall into the following categories: (a) Location data and its interpretation; (b) collecting informed consent, privacy, and anonymization of data; and (c) the management, classification, and storage of data [36]. The ubiquity of information and communication system can even be understood at a global scale today using positioning systems, making smart devices aware of one's location. This raises specific concerns about privacy and ethical issues, including less freedom in learning as learners are closely monitored by technology, revealing personal information without knowing, concerns about data protection, face detection from pictures, etc. It can be reported that the data collected for this study were collected using a user agreement and consent, and we did not share the learner's data.

Author Contributions: Conceptualization, M.N.H.; methodology, M.N.H., G.A. and K.M.; software, M.N.H., and H.U.; validation, M.N.H., G.A., K.M. and H.U.; formal analysis, M.N.H.; investigation, M.N.H., and H.U.; resources, M.N.H., and H.U.; data curation, M.N.H.; writing—original draft preparation, M.N.H.; writing—review and editing, M.N.H., G.A., K.M. and H.U.; visualization, M.N.H.; supervision, M.N.H.; project administration, M.N.H.; funding acquisition, M.N.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: We thank Sho Okuhara (RA of this project) for helping us with the coding. The source codes can be cloned (on-demand) from Github repository: https://github.com/keroido/ImageUnderstanding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gu, P.Y. Vocabulary learning in a second language: Person, task, context and strategies. *TESL-EJ* **2003**, *7*, 1–25.
2. Dumas, D.; McNeish, D.; Greene, J.A. Dynamic measurement: A theoretical–psychometric paradigm for modern educational psychology. *Educ. Psychol.* **2020**, *55*, 88–105. [CrossRef]
3. Cukurova, M.; Giannakos, M.; Martinez-Maldonado, R. The promise and challenges of multimodal learning analytics. *Br. J. Educ. Technol.* **2020**, *51*, 1441–1449. [CrossRef]
4. Levy, M. *Computer-Assisted Language Learning: Context and Conceptualization*; Oxford University Press: Oxford, UK, 1997.
5. Hasnine, M.N.; Ishikawa, M.; Hirai, Y.; Miyakoda, H.; Kaneko, K. An algorithm to evaluate appropriateness of still images for learning concrete nouns of a new foreign language. *IEICE Trans. Inf. Syst.* **2017**, *100*, 2156–2164. [CrossRef]
6. Sternberg, R.J.; McKeown, M.G.; Curtis, M.E. *Most Vocabulary Is Learned from Context the Nature of Vocabulary Acquisition 1987 Hillsdale*; NJ/London Lawrence Erlbaum Associates: Mahwah, NJ, USA, 1987; Volume 89, p. 105.
7. Nagy, W.E. *On the Role of Context in First-And Second-Language Vocabulary Learning*; University of Illinois at Urbana-Champaign, Center for the Study of Reading: Champaign, IL, USA, 1995.
8. Siemens, G.; Long, P. Penetrating the Fog: Analytics in Learning and Education. *Educ. Rev.* **2011**, *46*, 30.
9. Mouri, K.; Ogata, H.; Uosaki, N. Ubiquitous learning analytics in the context of real-world language learning. In Proceedings of the Fifth International Conference on Learning Analytics And Knowledge, New York, NY, USA, 16 March 2015; pp. 378–382. [CrossRef]
10. Vozniuk, A.; Rodríguez-Triana, M.J.; Holzer, A.; Govaerts, S.; Sandoz, D.; Gillet, D. Contextual learning analytics apps to create awareness in blended inquiry learning. In Proceedings of the 2015 International Conference on Information Technology Based Higher Education and Training (ITHET), Lisbon, Portugal, 11–13 June 2015; pp. 1–5.
11. Shibani, A.; Knight, S.; Shum, S.B. Contextualizable learning analytics design: A generic model and writing analytics evaluations. In Proceedings of the 9th International Conference on Learning Analytics & Knowledge, Tempe, AZ, USA, 4–8 March 2019; pp. 210–219.
12. Ogata, H.; Uosaki, N.; Mouri, K.; Hasnine, M.N.; Abou-Khalil, V.; Flanagan, B. SCROLL Dataset in the context of ubiquitous language learning. In Proceedings of the 26th International Conference on Computer in Education, Manila, Philippines, 26–30 November 2018; pp. 418–423.
13. Quinn, C. Learning Integrating System and Methods. U.S. Patent Application 10/973,756, 27 April 2006.
14. Quinn, C.N. Learning at large: Situating learning in the bigger picture of action in the world. *Educ. Technol.* **2004**, *44*, 45–49.
15. Quinn, C.N.; Boesen, M.; Kelmenson, D.; Moser, R. *Designing Multimedia Environments for Thinking Skill Practice*; Report No ISBN-1-880094-06-1 Pub date Jun 93 Note 673p; Association for the Advancement of Computing in Education: Charlottesville, VA, USA, 1993; p. 428.
16. Understanding the Importance of Context in Your Learning Solutions | Litmos Blog. SAP Litmos, 9 March 2016. Available online: https://www.litmos.com/ja-JP/blog/articles/understanding-the-importance-of-context-in-your-learning-solutions (accessed on 15 September 2020).
17. Yao, C.-B. Constructing a User-Friendly and Smart Ubiquitous Personalized Learning Environment by Using a Context-Aware Mechanism. *IEEE Trans. Learn. Technol.* **2017**, *10*, 104–114. [CrossRef]
18. Ogata, H.; Yano, Y. Context-aware support for computer-supported ubiquitous learning. In Proceedings of the 2nd IEEE International Workshop on Wireless and Mobile Technologies in Education, JungLi, Taiwan, 24 August 2004; pp. 27–34.
19. Montebello, M. Smart ubiquitous learning environments. *Int. J. Educ.* **2017**, *5*, 17–24. [CrossRef]
20. Ashar, M.; Rosyid, H.A.; Taufani, A.R. Ubiquitous Learning Environment for Smart Improving Disruption Activities in Classroom on Media. *Int. J. Interact. Mob. Technol.* **2020**, *14*, 200–212. [CrossRef]
21. Hasnine, M.N.; Hirai, Y.; Ishikawa, M.; Miyakoda, H.; Kaneko, K. A vocabulary learning system by on-demand creation of multilinguistic materials based on appropriate images. In Proceedings of the 2014 e-Case & e-Tech, Nagoya, Japan, 2–4 April 2014; pp. 343–356.
22. Hasnine, M.N. Recommendation of Appropriate Images for Vocabulary Learning. Ph.D. Thesis, Tokyo University of Agriculture and Technology, Tokyo, Japan, March 2018.

23. Hasnine, M.N.; Ogata, H.; Akçapınar, G.; Mouri, K.; Kaneko, K. Closing the Experiential Learning Loops Using Learning Analytics Cycle: Towards Authentic Experience Sharing for Vocabulary Learning. *Int. J. Distance Educ. Technol. (IJDET)* **2020**, *18*, 78–98. [CrossRef]
24. Liu, C.; Wang, C.; Sun, F.; Rui, Y. Image2Text: A Multimodal Image Captioner. In Proceedings of the 24th ACM International Conference on Multimedia, New York, NY, USA, 15–19 October 2016; pp. 746–748. [CrossRef]
25. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, MA, USA, 7 June 2015; pp. 3156–3164.
26. Del Sole, A. Introducing Microsoft Cognitive Services. In *Microsoft Computer Vision APIs Distilled: Getting Started with Cognitive Services*; Del Sole, A., Ed.; Apress: Berkeley, CA, USA, 2018; pp. 1–4.
27. Karpathy, A.; Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, MA, USA, 7 June 2015; pp. 3128–3137.
28. Bolanos, M.; Dimiccoli, M.; Radeva, P. Toward storytelling from visual lifelogging: An overview. *IEEE Trans. Hum. Mach. Syst.* **2016**, *47*, 77–90. [CrossRef]
29. Hasnine, M.N.; Flanagan, B.; Akcapinar, G.; Ogata, H.; Mouri, K.; Uosaki, N. Vocabulary Learning Support System based on Automatic Image Captioning Technology. In Proceedings of the International Conference on Human-Computer Interaction, Orlando, FL, USA, 26–31 July 2019; pp. 346–358.
30. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2818–2826.
31. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European conference on computer vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
32. Paivio, A.; Rogers, T.B.; Smythe, P.C. Why are pictures easier to recall than words? *Psychon. Sci.* **1968**, *11*, 137–138. [CrossRef]
33. Paivio, A.; Csapo, K. Picture superiority in free recall: Imagery or dual coding? *Cogn. Psychol.* **1973**, *5*, 176–206. [CrossRef]
34. Miyazaki, T.; Shimizu, N. Cross-lingual image caption generation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; pp. 1780–1790.
35. Al-Muzaini, H.A.; Al-Yahya, T.N.; Benhidour, H. Automatic Arabic Image Captioning using RNN-LST M-Based Language Model and CNN. *Int. J. Adv. Comput. Sci. Appl.* **2018**, *9*, 67–73.
36. Slade, S.; Prinsloo, P. Learning analytics: Ethical issues and dilemmas. *Am. Behav. Sci.* **2013**, *57*, 1510–1529. [CrossRef]