Vocabulary Learning Support System based on Automatic Image Captioning Technology

Mohammad Nehal Hasnine[1]*, Brendan Flanagan[1], Gokhan Akcapinar [14], Hiroaki Ogata[1], Kousuke Mouri[2], and Noriko Uosaki[3]

[1] Kyoto University, Kyoto 606-8501, Japan
[2] Tokyo University of Agriculture and Technology, Tokyo 184-8588, Japan
[3] Osaka University, Osaka 565-0871, Japan
[4] Hacettepe University, Ankara 06800, Turkey
*nehalhasnine@gmail.com

Abstract Learning context has evident to be an essential part in vocabulary development, however describing learning context for each vocabulary is considered to be difficult. In the human brain, it is relatively easy to describe learning contexts using pictures because pictures describe an immense amount of details at a quick glance that text annotations cannot do. Therefore, in an informal language learning system, pictures can be used to overcome the problems that language learners face in describing learning contexts. The present study aimed to develop a support system that generates and represents learning contexts automatically by analyzing the visual contents of the pictures captured by language learners. Automatic image captioning, a technology of artificial intelligence that connects computer vision and natural language processing is used for analyzing the visual contents of the learners' captured images. A neural image caption generator model called Show and Tell is trained for image-to-word generation and to describe the context of an image. The three-fold objectives of this research are: First, an intelligent technology that can understand the contents of the picture and capable to generate learning contexts automatically; Second, a leaner can learn multiple vocabularies by using one picture without relying on a representative picture for each vocabulary, and Third, a learner's prior vocabulary knowledge can be mapped with new learning vocabulary so that previously acquired vocabulary be reviewed and recalled.

Keywords: Artificial intelligence in education, automatic image captioning, learning context representation, ubiquitous learning, visual content analysis, vocabulary learning.

1 Introduction

1.1 The Advancement and Challenges of Vocabulary Learning using Various Technologies

In recent years, computer-assisted technologies have received considerable attention for language learning particularly to support foreign vocabulary learning. One of the

many reasons for that is, unlike available traditional teaching methods, technology-mediated learning environments offers the flexibility to learn at at-time and anyplace. With regard to the adaptation of technologies to support learning systems, Stockwell stated that the range of technologies is broad and includes courseware (commercial and self-developed), online activities, dictionaries, corpora and concordance, and computer-mediated communication (CMC) technologies[1]. It has already been observed that vocabulary with various annotation styles linking for textual meaning, audio, graphics etc.[2], intelligent language tutoring systems that included sophisticated feedback systems[3], hypermedia-enhanced learning systems[4] etc. been developed. While various technologies provide various advantages but questions remain regarding how these technologies have been used in achieving learning objectives. For example, ubiquitous learning technologies, a renowned technology to support vocabulary learning often provide the facility to learn anytime and anywhere. Also, in this kind of learning environments, 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 on which words to be learned next, a learner's engagement with the system is low, and describing each learning context for each word may be difficult etc. are often discussed.

1.2 The Roles of Learning Context in Vocabulary Learning

Generally learning context refers to the learning environment including socio-cultural-political environment where learning takes place[5]. According to Gu[5], the learning context may include educators, classmates, classroom atmosphere, family support, social and cultural tradition of learning, academic curriculum etc. Learning con-texts constrains the ways learners' approach to different learning tasks. Research also evident how meaningful context can help learners in acquiring foreign languages. To describe the importance of meaningful context, Firth, in the late 1950s, said that the complete meaning of a word is always contextual and no study of meaning independent of complete context can be taken seriously[6]. Firth also articulated that each word, when used in a new context, becomes a new word that assists us in making statements of meaning[6]. Robert Sternberg stated that most vocabulary is acquired from contexts[7]. Nagy said that, 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 [8]. Concluding therefrom, in the context of informal learning of foreign vocabulary, the role of learning contexts evident as important in enhancing vocabulary development.

1.3 The Aim of the Research

Earlier, we have discussed how important learning contexts are in enhancing vocabulary development but describing each learning context for each word is considered to be difficult. Therefore, this research investigates how technology can be applied to generate special learning contexts. To generate learning contexts, we used pictures that are

captured by the learners of foreign languages. We analyzed the visual contents of images in order to produce learning contexts which we refer to as special learning contexts for vocabulary learning. The assumption was that those special learning contexts will help learners in enhancing their vocabulary.

This research also attempts to find answers to the following research questions:

1. Can we build technology that provides multiple visual learning contexts for learning multiple words from one picture?
2. Can we develop a support system that will provide visual cues to highlight the word to be learned which can be related to visual mnemonics?
3. Can we apply learning analytics to learning vocabulary with the context-vs-without learning context?
4. Can we use one picture to learn multiple vocabularies?
5. Can we use learning analytics and knowledge graph to map a learner's prior vocabulary knowledge in order to recommend new vocabulary to be learned?

2 Literature Review

In recent years, the successful application of deep neural networks to computer vision and natural language processing tasks have inspired many AI researchers to explore new research opportunities at the intersection of these previously separate domains[9]. Caption generation models have to balance an understanding of both visual cues and natural language. In this present work, we used automatic image captioning technology to generate learning contexts from images.

2.1 Google's Image2Text

Google's Image2Text system is a real-time captioning system that generates human-level natural language description for any input image. In this model, a sequence-to-sequence recurrent neural networks (RNN) model for image caption generation. This system also enables users to detect salient objects in an image, and retrieve similar images and corresponding descriptions from a database[10].

2.2 O'Reilly-Show and Tell

O'Reilly-Show and Tell Model is an image caption generation models combine advances in computer vision and machine translation to produce realistic image captions using neural networks. These models are trained to maximize the likelihood of producing a caption given an input image and can be used to generate novel image descriptions[9].

2.3 Deep visual

Deep visual is a model that generates natural language descriptions of images and their regions. The deep model leverages datasets of images and their sentence descriptions for learning the intermodal correspondences between language and visual data[11].

This caption descriptor model is capable of producing state of the art results on Flickr8K[12], Flickr30K[13] and MSCOCO [14]datasets.

2.4 Microsoft Cognitive API

Microsoft Cognitive API for images' visual content analysis is an API (Application Program Interface) that returns features about visual contents found in a natural image. This API, in order to identify contents and label uses tagging, domain-specific models and descriptors[15].

We have found several recognized models and APIs that generates texts and sentences from images. Most of the studies using this technology are carried out for NLP applications, explaining frame-by-frame video clips, social media such as Facebook where faces are inferred directly from images, computer vision, robotics, machine learning etc. We have not found many recognized studies on building educational technology using this promising technology. Therefore, we aimed to use this technology to solute a recognized problem in language learning namely learning contexts generation in informal learning.

3 Overview of the Research

3.1 Background and Motivation

While previous research indicates that most of the vocabulary is learned from context[7] particularly for recalling learned vocabulary but describing learning context for each word is difficult. To begin with, we analyzed a dataset that contains ubiquitous learning logs. The aim of the analysis was to observe whether or not leaners of foreign language have difficulties in describing learning contexts while learning via ubiquitous learning tool. Therefore, we analyzed SCROLL (System for Capturing and Reminding of Learning Logs) dataset[16]. a dataset that consists of foreign language learners lifelong learning experiences (i.e. lifelogs). Logs are chronologically collected from a context-aware ubiquitous language learning system called SCROLL[16].

Fig.1. Interface of the SCROLL system

To elaborate more, in this system, a foreign language learner can create his/her own vocabulary learning materials using this system[17]. A learning material consists of a contextual image, the translation data, and the pronunciation. Moreover, the system is capable of recording a learner's vocabulary learning experience (such as the geolocation information, vocabulary knowledge, quiz, learning context, contextual image information etc.) into its server. The interface of the system is displayed in Figure 1.

We analyzed 31258 ubiquitous learning logs. The analysis indicates that 82% (13931 out of 16844), 62% (6224 out of 10034), and 94% (6648 out of 7044) learning logs that are created by English-native, Japanese-native and Chinese-native are created without learning contexts, respectively. Note that, we use the term -*native* because learners registered themselves as either of these languages as the default language. Table 2 shows the result of the log analysis. This analysis indicated that describing learning context is an area that most of the learners skipped while learning context is an essential component in vocabulary development. Hence, we aimed to develop this support system that will support learners by providing special learning contexts.

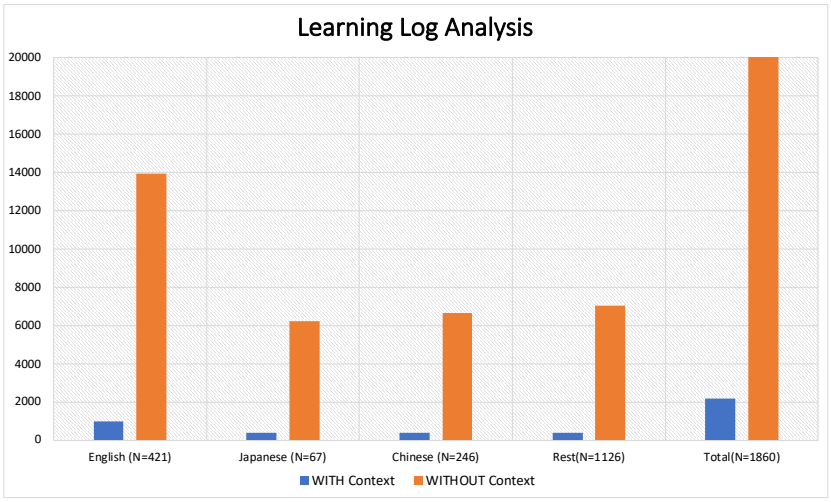

Fig.2. Result of the learning log analysis

3.2 How does Image Captioning Technology Fit into Educational Research?

Imagine a situation where a learner has previously learned words such as *sun*, *mountain*, *river,* and *sunrise*. In a new learning situation, the same learner captures a picture and uploads it into our server. The system will analyze the visual contents of the image and generate multiple special learning contexts (for instance, 1. *sunrise in the mountains*, 2. *sunrise paints the sky with pink*, 3. *sunrise offers beautiful view in mountain*) by which –the learner's prior vocabulary knowledge will be mapped with the special learning contexts. It is possible when you develop an intelligent technology that can understand the context of the picture and capable to generate the learning context automatically. We aimed to develop a system that will take pictures that are uploaded by learners as

input. Then, the system will analyze the visual contents of those pictures. And based on those visual contents, the system will produce special learning contexts.

By doing this, this study aimed to assist foreign language learners in several ways, as follows:

- Map a learner's previous knowledge with new so that he/she can review and recall the previously acquired vocabulary
- To memorize new vocabulary in context
- To acquire multiple vocabulary from the same picture
- Provide visual mnemonics that will work as visual cues to highlight that particular word

In Figure 2, mapping between a learner's prior knowledge with a new knowledge using special visual context is shown.

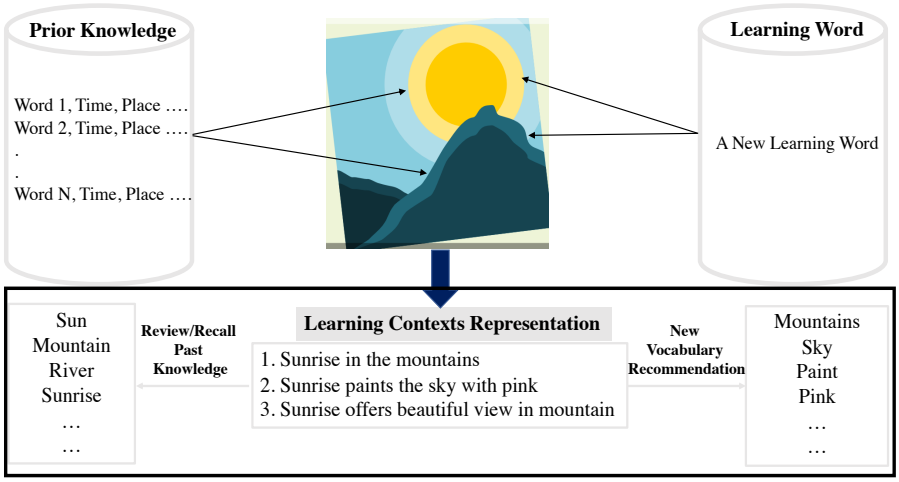

Fig. 2. Mapping a learner's prior knowledge with new knowledge using learning contexts representation

However, producing learning contexts is not an easy task. Now-a-days, artificial intelligent is gaining much popularity to build various system. Image captioning, a technology that connected computer vision and artificial intelligence is capable of generating captions from a picture. However, connecting learning analytics with computer vision and natural language processing is a quite challenging task. In this study, we hypothesized that by developing a computer program to automatically generate language description from an image may be helpful for learners to new words. Therefore, by using image captioning technology, we aimed to build an environment where learning context, multiple vocabulary acquisition and reflection of prior knowledge to learn new knowledge are supported. At this point, it can be mentioned that the adaptation of various technologies to support vocabulary learning is not new. New technology opens scopes to new research.

4 System Implementation

In this study, the Show and Tell model [18], a state-of-the-art model is trained for the image-to-word generation and special context generation tasks. In this section, we discuss about the architecture of the model along with its implementation in our server.

4.1 Architecture

The diagram below illustrates the architecture of the model. In this model, the image encoder is a deep convolutional neural network (CNN). Recently, deep CNN is widely adopted for object recognition and detection tasks. We used Inception v3 image recognition model[19] that was pretrained on the ILSVRC-2012-CLS image classification dataset[20]. In recent days, this type of neural and probabilistic framework is used for sequence modeling tasks such as language modeling and machine translation. The architecture 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. In LSTM network, words in the captions are represented with an embedding model where each word in the vocabulary is associated with a fixed-length vector representation that is learned during training[18]. Figure 3 shows the general architecture of the model.

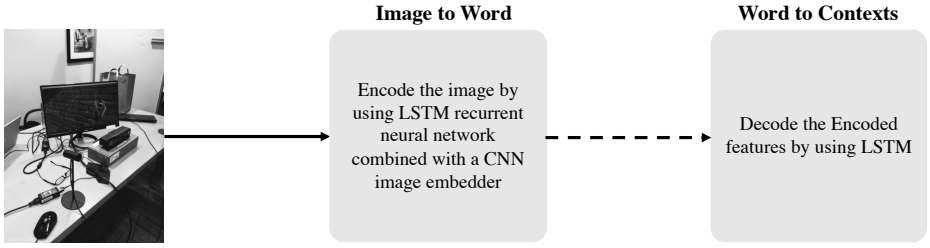

Fig. 3. General architecture of the LSTM-based Show and Tell Model

4.2 Dataset

There are several popular datasets available in computer vision such as ImageNet, PASCAL VOC and SUN etc. Each of these datasets contains different types of images that can perform well for designated tasks. For example, ImageNet was created to capture a large number of object categories, many of which are fine-grained; while PASCAL VOC's primary application is object detection in natural images; and SUN focuses on labeling scene types and the objects that commonly occur in them[14]. In this study, we trained the Show and Tell model on MSCOCO image dataset[14]. The images were in native TFRecord format. The dataset contains a total of 2.5 million labeled instances in 328k images that performs well precisely for category detection, instance spotting and instance segmentation[14]. 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. This is why we used this dataset for training our model.

4.3 Details on Hardware and Training Time

In Table 1, we summarize the hardware details and the training time require to train the model.

Table 1. Table captions should be placed above the tables.

Heading level	Example
GPU	ZOTAC GAMING GeForce RTX 2080 Ti AMP ZT-T20810D-10P
Time	10 to 14 days
Libraries/Packages	Bazel, Python, Numpy, punkt
Natural Language Toolkit	NLTK
Dataset	MSCOCO

4.4 Advantages of the Model

In this present study, we used this model because: Firstly, this neural network-based model is capable of viewing a natural image and generating reasonable contexts in plain English. Hence, it is readable without further interpretation. Secondly, the model encodes an image into a compact representation, followed by a recurrent neural network that generates a corresponding sentence. Thirdly, the deep recurrent architecture of the model preforms well in limited computation power. Finally, the performance of the model improves as the amount of data increases.

5 Result

We tested the performance of our system on the natural pictures that are uploaded by foreign language learners while using SCROLL system. To demonstrate the result, we choose three random pictures associated with three learning logs from SCROLL dataset. Each of the three logs contained a picture and the context described by the log creator. We used the pictures to generate learning contexts using our system.

First (refer to Figure 4a), for the word *lunch box* created by a Japanese-native, the learner described the learning context as *食べる(ate,* in English*)*. While our system generated top-three learning contexts are: (1) a lunch box with a variety of food items (p=0.000009), (2) a lunch box with a variety of food and a drink (p=0.000001), and (3) a lunch box with a variety of vegetables and a sandwich (p=0.000000).

Second (refer to Figure 4b), for the word *computer*, the learner described the learning context as: *this is a computer running Oculus Rift for language research.* In contrary, our system represented the learning contexts as: (1) a laptop computer sitting on top of a desk (p=0.008150), (2) a laptop computer sitting on top of a wooden desk (p=0.007242), and (3) a laptop computer sitting on top of a table (p=0.001166). We noticed that the Oculus device was not detected but the objects like desk and table were identified.

Third (refer to Figure 4c), in learning the word *carpet*, the learner described its context as: *買った*(which mean *bought* in English). After analyzing the visual contents, our system represented its contexts as: (1) a living room filled with furniture and a window (p=0.001004), (2) a living room filled with furniture and a fire place. (p=0.000599), and (3) a living room filled with furniture and a window (p=0.000045).

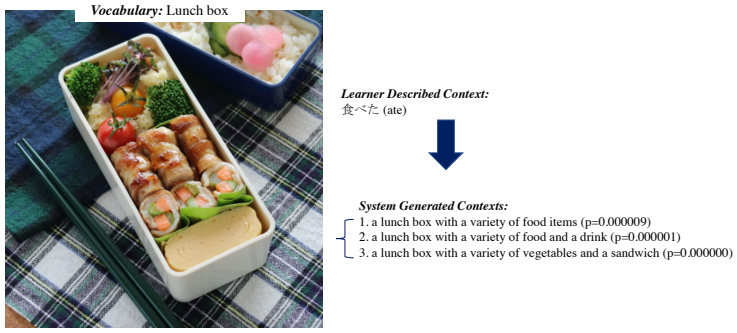

Fig. 4(a). Sample result (1)

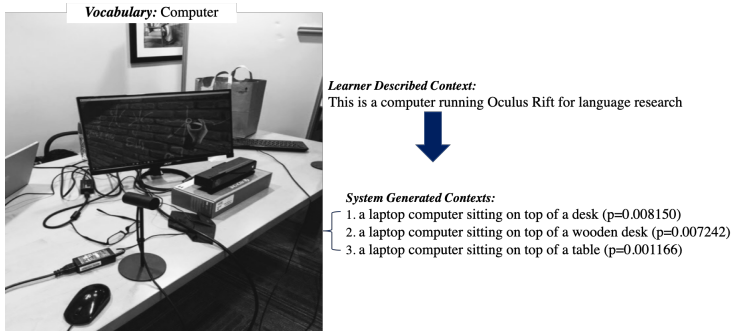

Fig. 4(b). Sample result (2)

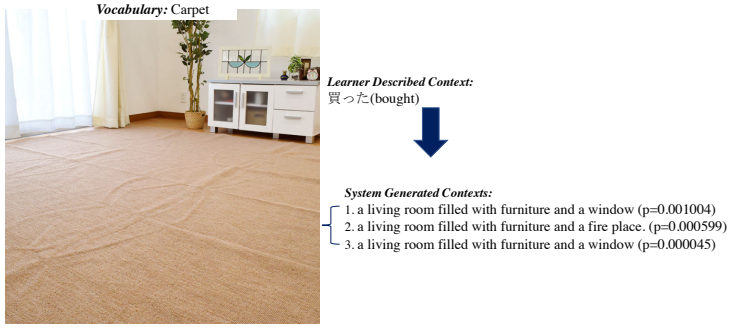

Fig. 4(c). Sample result (3)

6 Discussion

Ubiquitous learning technologies provide the facility to learn anytime and anywhere. These technologies help to capture ones learning experiences. Also, in this kind of learning environments previous learning experiences are used to measure one learner's current knowledge level, which can be used to learn new knowledge. Hence, there are certain limitations, namely the scope of learning new vocabulary is limited, students often cannot determine on which words to be learned next, a learner's engagement with the system is low, and describing each learning context for each word may be difficult etc. are often reported by many researchers. From that standpoint, this paper describes an approach to represent learning contexts by analyzing the visual contents of an image that is captured by a leaner of a foreign language. The three-fold advantages of this technology are: First, to develop an intelligent technology that can understand the context of the picture and capable to represent learning contexts automatically, Second, to develop a technology in which a leaner can learn multiple vocabulary using one picture, and Third, relating with their prior knowledge, new vocabulary can be learned.

Automatic image captioning, a technology of artificial intelligence that connects computer vision and natural language processing is used for analyzing the visual contents of an image. A neural image caption generator model called Show and Tell is trained for image-to-word generation and describing the context of an image automatically. It can be mentioned that, in recent days, the successful application of deep neural networks to computer vision and natural language processing tasks have inspired many AI researchers to explore new research opportunities at the intersection of these previously separate domains. Caption generation models have to balance an understanding of both visual cues and natural language. We trained the Show and Tell model, a state-of-the-art model is trained for the image-to-word generation and context generation tasks. In this model, the image encoder is a deep convolutional neural network (CNN). Recently, deep CNN is widely adopted for object recognition and detection tasks. Our particular choice of network is the Inception v3 image recognition model pretrained on the ILSVRC-2012-CLS image classification dataset. The decoder is a long short-term memory (LSTM) network. This type of network is commonly used for sequence modeling tasks such as language modeling and machine translation. In the Show and Tell model, the LSTM network is trained as a language model conditioned on the image encoding. Words in the captions are represented with an embedding model. Each word in the vocabulary is associated with a fixed-length vector representation that is learned during training.

We tested the performance of the system on the pictures that are captured by the learners of foreign languages while using SCROLL. We found that, in representation of learning contexts, some objects are not detected correctly. However, many new vocabularies are generated by the system because the system was capable to detect many important objects. We believe this system will assist foreign language learners in learning new vocabulary along with representing learning contexts.

7 Limitations and Directions to Future Works

The goal of image captioning is to produce sentences that are linguistically plausible and semantically truthful to the contents embedded in it [21]. Also, the original reason behind this technology is to generate simple descriptions for images taken under extremely constrained conditions[21]. However, in this study, we used this sophisticated technology to generate the learning contexts, which may be questionable to some extent. In other word, we used an image-to-word often address as automatic word captioning model to produce visual learning contexts assuming that those learning contexts will be helpful in learning foreign vocabulary. We aim to carry out an evaluation experiment in the near future to assess the learning effect of this system. In future, this model is planned integrated with a ubiquitous language learning system called SCROLL (System for Capturing and Recording of Learning Logs).

Acknowledgement

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (S)16H06304 and 17K12947; NEDO Special Innovation Program on AI and Big Data 18102059-0; and JSPS Start-up Grant-in-Aid Number 18H05745.

References

1. G. Stockwell, "A review of technology choice for teaching language skills and areas in the CALL literature," ReCALL, vol. 19, no. 2, pp. 105–120, 2007.
2. Y. Yeh and C. Wang, "Effects of multimedia vocabulary annotations and learning styles on vocabulary learning," Calico Journal, pp. 131–144, 2003.
3. T. Heift, "Error-specific and individualised feedback in a Web-based language tutoring system: Do they read it?," ReCALL, vol. 13, no. 1, pp. 99–109, 2001.
4. J. F. Coll, "Richness of semantic encoding in a hypermedia-assisted instructional environment for ESP: Effects on incidental vocabulary retention among learners with low ability in the target language," ReCALL, vol. 14, no. 2, pp. 263–284, 2002.
5. P. Y. Gu, "Vocabulary learning in a second language: Person, task, context and strategies," TESL-EJ, vol. 7, no. 2, pp. 1–25, 2003.
6. W. J. Ibrahim, "The importance of contextual situation in language teaching," Adab AL Rafidayn, no. 51, pp. 630–655, 2008.
7. R. J. Sternberg, "Most vocabulary is learned from context," The nature of vocabulary acquisition, vol. 89105, 1987.
8. W. E. Nagy, "On the role of context in first- and second-language vocabulary learning," Champaign, Ill. : University of Illinois at Urbana-Champaign, Center for the Study of Reading., text, Nov. 1995.
9. R. P. Ricciardelli Daniel, "Caption this, with TensorFlow," O'Reilly Media, 28-Mar-2017. [Online]. Available: https://www.oreilly.com/learning/caption-this-with-tensorflow.
10. C. Liu, C. Wang, F. Sun, and Y. Rui, "Image2Text: a multimodal caption generator," in ACM Multimedia, pp.746-748, 2016.
11. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3128–3137, 2015.

12. M. Hodosh, P. Young, and J. Hockenmaier, “Framing image description as a ranking task: Data, models and evaluation metrics,” Journal of Artificial Intelligence Research, vol. 47, pp. 853–899, 2013.
13. B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models,” in Proceedings of the IEEE international conference on computer vision, pp. 2641–2649, 2015.
14. T.-Y. Lin et al., “Microsoft coco: Common objects in context,” in European conference on computer vision, pp. 740–755, 2014.
15. “Image Processing with the Computer Vision API | Microsoft Azure.” [Online]. Available: https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/.
16. H. Ogata, N. Uosaki, K. Mouri, M. N. Hasnine, V. Abou-Khalil, and B. Flanagan, “SCROLL Dataset in the Context of Ubiquitous Language Learning,” in Workshop Proceedings of the 26th International Conference on Computer in Education, Manila, Philippines, pp. 418–423, 2018.
17. M. N. Hasnine, K. Mouri, B. Flanagan, G. Akcapinar, N. Uosaki, and H. Ogata, “Image Recommendation for Informal Vocabulary Learning in a Context-aware Learning Environment,” in Proceedings of the 26th International Conference on Computer in Education, Manila, Philippines, pp. 669–674, 2018.
18. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: Lessons learned from the 2015 mscoco image captioning challenge,” IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 4, pp. 652–663, 2017.
19. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826, 2016.
20. O. Russakovsky et al., “ImageNet Large Scale Visual Recognition Challenge,” Int J Comput Vis, vol. 115, no. 3, pp. 211–252, Dec. 2015.
21. S. Bai and S. An, “A Survey on Automatic Image Caption Generation,” Neurocomputing, pp.291-304, 2018.