Suggestions of a Deep Learning Based Automatic Text Annotation System for Agricultural Sites Using GoogLeNet Inception and MS-COCO

Shinji Kawakura
Department of Technology Management for Innovation, The University of Tokyo, Bunkyo-ku, Tokyo, Japan
Email: s.kawakura@gmail.com, kawakura@motolabo.net

Ryosuke Shibasaki
Center for Spatial Information Science, The University of Tokyo, Kashiwa-shi, Chiba, Japan
Email: shiba@csis.u-tokyo.ac.jp

***Abstract*—Image recognition methodologies for use by agricultural (agri-) workers, managers, technicians, researchers, and unliving targets (e.g., harvests, agri-tools) have attracted significant interest. Currently, the most common approaches use various real-time visual analyses and recorded data-based analyses at outdoor and indoor agri-sites. However, recent Artificial Intelligence (AI)-based studies have proposed diverse automatic camera-based awareness systems with text-annotation. Although some systems have included monitoring and identification tools for the aforementioned agri-fields, their captioning abilities and accuracy levels have been insufficient for practical usage. Thus, further improvements have increased the accuracy by incorporating computing based on recent deep learning methodologies, particularly utilizing recent open services provided by huge IT companies, such as Google or Microsoft. Deep learning based analysis systems sometimes pick up on and highlight hidden, subtle points that a human may fail to notice. Thus, we develop deep learning based auto-annotating systems for Japanese small- to middle-sized indoor and outdoor agri-working sites and workers. We use visual data sets with a variety of real and common Japanese-styled agri-tools. We statistically analyze the obtained data and compare the comments obtained from experienced agri-workers. Our results confirm the systems' utility, validity, and limitations.**

***Index Terms*—automatic annotating from pictures, neural image captioning, deep learning, TensorFlow, CNN, GoogLeNet inception, MS-COCO**

I. Introduction

Agricultural (agri-) informatics researchers and manufacturers have developed diverse visual analyses-based techniques to improve the utility of existing agri-machines (e.g., automatic harvesting systems and tractors) and enhance the security level of agri-workers.

These achievements in academic and business fields have already reached sufficient levels for real-world application in agri-fields and can output a large amount of positive results [1]-[9]. Existing visual data-based analysis methods have focused on harvests, weeds, forests, farmers, and other various targets [8]-[12].

However, we focus on automatic captioning (annotation) systems [9]-[10] utilizing deep learning, such as those using recent open services provided by huge IT companies (e.g. Google, Microsoft). Currently, these are insufficient, particularly for traditional Japanese workers, for whom there are no studies.

AI-based technologies continue to be developed [11]-[23]. In light of this background and considering future expansions, we develop a visual data-based analysis and service system using deep learning to support agri-informatics researchers, manufacturers, managers, and workers.

II. Methods

A. Field

This study examines the agr-workers using agri-tools in traditional Japanese small- to middle-sized (1) outdoor farms and (2) indoor warehouses for storage and drying harvests, after consultation with real farmers.

B. Target

For the visual data used in this study, we captured and aggregated original pictures from non-specific outdoor farmlands for test data. The target subjects were agri-workers with 1-5 years' experience. Prior to the visual data collection, we consulted agri-managers and workers to mitigate the difficulties in handling dozens of samples in the farmlands. First, we captured the following three categories of picture data using a digital video camera: CANON 410f ixy (CANON Inc., Japan):

1) "Cultivating a field using a hoe" (Pictures captured workers cultivating agri-fields using a traditional Japanese hoe in the semi-standing posture in an outdoor farmland: n = 31).

2) "Harvesting vegetables" (Pictures captured workers harvesting vegetables in the sitting posture in an outdoor farmland: n = 13).

Manuscript received April 13, 2020; revised September 27, 2020.

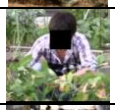
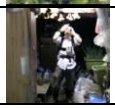
doi: 10.18178/joig.8.4.120-125

3) "Hanging harvests" (Pictures captured workers hanging harvests with changing postures (standing and sitting) in warehouses: n = 13).

Table I summarizes these categories, and their formats and sizes were uniformed and standardized before the main computation.

Next, to support price Neural Image Captioning, we used free open picture datasets from MS-COCO (Microsoft Common Objects in Context) distributed by Microsoft Inc. The datasets contain more than 300,000 text-captioned pictures, totaling about 20GB. Users can download the pictures from Microsoft websites. We also used STAIR Caption from STAIR Lab, Inc., Japan, which includes Japanese translations of English MS-COCO captions. We captured and accumulated several sets of video visual data at outdoor and indoor traditional Japanese agri-working sites, which we utilized in our experiments.

TABLE I. TARGETS, THEIR DESCRIPTION, AND THREE SETS OF SQUARE SAMPLE PICTURES

Target (Case)	Description of Target	Sample picture		
1) Cultivating a field using a hoe	Cultivating agri-fields using a hoe in the semi-standing posture in an outdoor farmland			
2) Harvesting vegetables	Harvest vegetables in the sitting posture in an outdoor farmland			
3) Hanging harvests	Hanging harvests with changing posture (standing and sitting) in a warehouse			

C. Computing

We used AVI2JPG to input and save pictures obtained using a digital video camera, and we applied recent TensorFlow-based programs to automatically create and present adequate descriptive captions for the obtained pictures. For the captioning, we coded and used python 3.6 language-based program sets, which is a common language used for deep learning.

There are currently several open Machine Learning platforms, including (1) Amazon Web Services, (2) Microsoft Azure, (3) Google Cloud Platform, and (4) IBM Bluemix. The main services of these platforms are Computer Vision Application Programming Interface (API), Text Analytics API, and Language Understanding Intelligent Service. The characteristics of these platforms are (A) users can use their high-spec development environments including GPUs, (B) the aforementioned global companies have already executed their original machine-learnings with huge amounts of data on high-powered computers, and opened the results to global users, saving time and expense (with access becoming cheaper or free), and (C) users can utilize them irrespective of their location and social status. In this study, we used Google Cloud Platform. As of June 2019, many open models were used mainly for transfer learning, including GoogLeNet's Interception-v1 to Interception-v4, ResNet-v1 series to ResNet-v2 series, VGG 16 & 19, MobileNet-v1 series and NASNet series. We selected GoogLeNet Inception-v3 (Google Inc., the U.S.) for transfer learning with Microsoft Common Objects in Context (MS-COCO). For our approach, we used these open-systems and our original picture sets because of the relatively long computational time required for machine learning.

Table I presents the classified captured pictures in the respectively named data folders. However, for the research field, in many cases, where past pictures are used for learning, and they are quite different to current trials, their analyzed characteristic points and values cannot be used directly for practical uses.

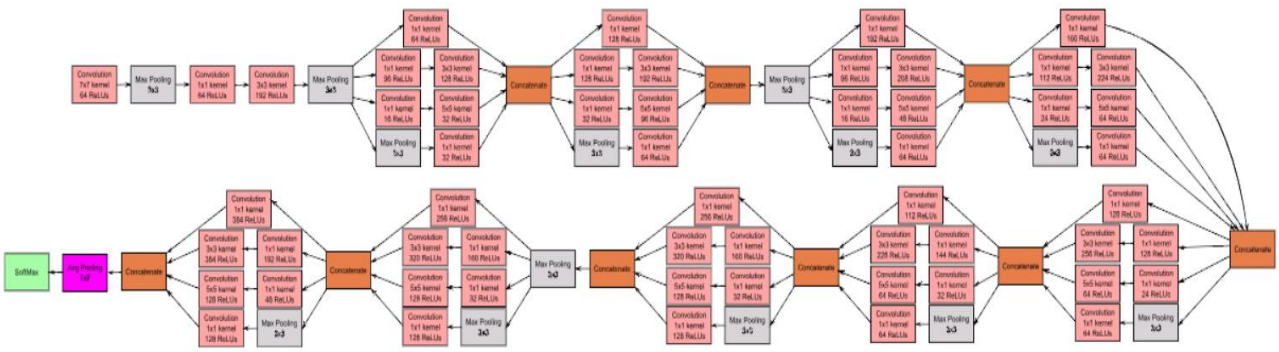


Figure 1. Architectures of different evaluated ConvNets. (Purple boxes indicate that the layers from the features were extracted where ConvNets was used for feature extraction).

According to current academic trends and past results, our methodology is adequate in the agricultural informatics field. Fig. 1-Fig. 6 show the flow of this study's computing as follows. Fig. 1 presents the architectures of the different ConvNets evaluated in this work, Fig. 2 shows the architecture of the Inception module, Fig. 3 shows the flow of the created practical model., Fig. 4 shows the flow of the training model, Fig. 5 shows the flow of the main model, and Fig. 6. presents the flowchart for the computing model. The practical study process comprised the following four steps:

1) Obtain pictures and movie data from the target area farmlands,

2) Analyze the data using our programs (we have confirmed the diverse adequacies of functions in advance),

3) Check the recognition accuracy, and

4) Discuss the validities of annotation data with an agri-worker with seven years' experience. Concerning the deep learning model, we focus on Recurrent Neural Network (RNN)-based analyses.

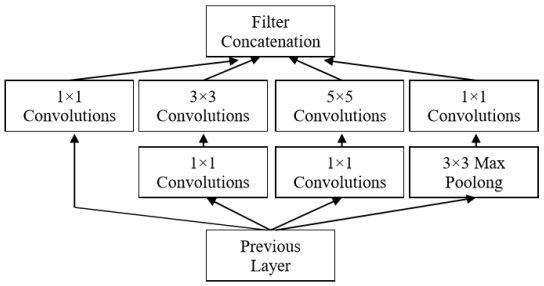


Figure 2. Architecture of inception module.

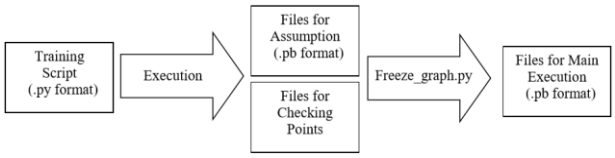


Figure 3. Flow of the created practical model.

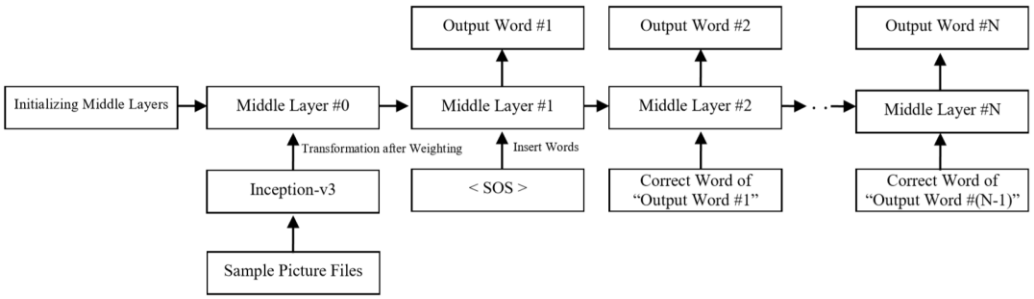


Figure 4. Flow of the training model.

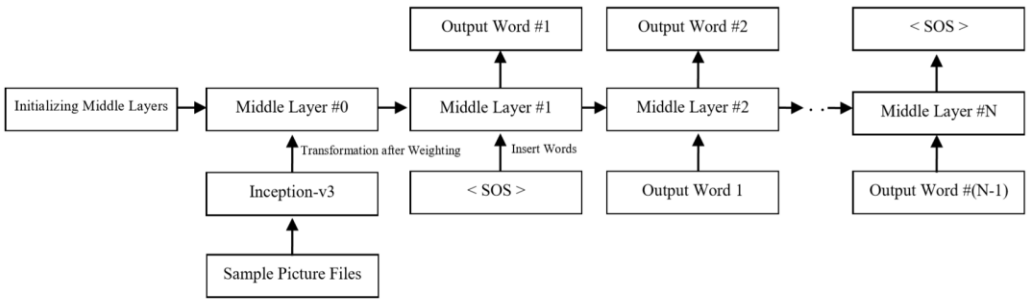


Figure 5. Flow of the main model.

1. Reading and inputting picture files

1.1 Import "glob" and "TensorFlow" into programs

1.2 Obtain the path of the picture files (.jpeg format)

1.3 Read and change these picture files into binary formats

1.4 Register obtained byte-styled dataset into variables "key.value"

1.5 Register and write data into "Example" class as "key.value" format and "list" style

2. Reading and inputting "TF (TensorFlow) Record" format data files

2.1 Execute "TF (TensorFlow) Record" format parse

2.2 Iterate dataset, obtaining factors from iterator

2.3 Execute "Tensor," and show picture images

3. Creating program for dataset-formatting

3.1 Read correct data of "STAIR Captions"

3.2 Recognize picture IDs, and make (1) dictionaries and (2) paths of files for captioning according to the IDs

3.3 Separate datasets into "Training dataset," "Validation dataset," and "Testing datasets"

3.4 Make a dictionary based on the "Training dataset"

3.5 Read pictures and write records into TFRecord (1 record consists of 1 picture and 1 caption)

4. Transfer learning using "Interception-v3"

4.1 Write and output Interception-v3

4.2 Load Interception-v3 on the graph of TensorBoard

4.3 Create a trained model (set up systems for the next transfer learning)

5. Developing and setting the caption generating model

5.1 Import modules (glob, TensorFlow, Image, and numpy)

5.2 Install Python Image Library and dictionary file

5.3 Set calculation graph for the deductions (inferring) and training

5.4 Make input data and "TFRecord" parsing function

5.5 Process image data (decoding, resizing, etc.), and change datasets into batches

5.6 Check picture adequacy input trained Interception-v3, and unite to fit to RNN

5.7 Prepare and initialize functions and files for net execution

5.8 Start training, and write executed data into TensorBoard

5.9 Change result data into row of words using aforementioned dictionary

5.10 Check data, release and close several resources

6. Deduction (inferring) of picture data and showing captions using trained model

6.1 Reset (refreeze) graph

6.2 Read dictionary file and pb format file, and load it on the default graph

6.3 Obtain "Tensor"

6.4 Output the result of the earlier proposed picture captions

6.5 Input picture data, and make and present captions

Figure 6. Algorism of the computing model.

Considering similar past trials [16]-[23], in this study, we made the size of the jpeg pictures uniform at 299×299 pixels (from the system's template), and we performed a layer-oriented deep learning-based analysis. In recent years, diverse methodologies have been proposed for such deep learning-based computing targeting various visual data. We surveyed and used the latest TFRecords' (TensorFlow Records') data manipulation framework and various peripheral programs, libraries, and packages because of their successful application in past research.

III. Results

In Table II presents three caption outputs for three cases, and their ratios: ((Numbers of output captions)/(Numbers of all data))*100. The statistical results are calculated according to past studies [16]-[23].

TABLE II. Targets and Output Captions

Target (Case)	Output Caption #1		Output Caption #2		Output Caption #3	
1) Cultivating a field using a hoe	"A worker batting a substance on the ground. There are some scattered things."	13.5 %	"A standing man leaning on a bar."	13.1%	"A standing man holding a bar."	12.1%
2) Harvesting vegetables	"On wild ground, a sitting man arranging goods."	9.5%	"A sitting man setting garbage on the floor"	8.2%	"This person is suffering from disease and struggling."	3.9%
3) Hanging harvests	"In the dark room, this person is hanging cleaned clothes."	8.1%	"In the dark room, this person is exercising."	5.8%	"A man is squatting repetitively."	4.4%

IV. Discussion

Table II presents the qualitative and quantitative features of the three cases. In this study, we were unable to obtain statistically sufficient volumes of visual data. For the data, we observed the limitation concerning the range of ratio values from 3.9% to 13.5%. Case 3 had the lowest average ratio (the average of 8.1%, 5.8%, and 4.4% is 6.1%), which may be because it is a rather specific, rare agri-working situation, and these pictures had the darkest and least-vivid visual colors. By contrast, Case 1 had the highest average ratio (the average of 13.5%, 13.1%, and 12.1% is 12.9%). Specifically, relating to the system's construction, it was difficult to obtain comments that incorporated a combination of people, tools, and backgrounds. Additionally, we could not definitely determine whether the set of tools was adequate for judging. Judgements made by an agri-worker with five years' experience suggested that these pictures could be understood using common sense, but the computational outputs of the system were likely to be wrong as they incorrectly referenced housework, exercise, or other daily movements. We believe the system could be made more efficient.

V. Conclusion and Future Tasks

In this study, we constructed and demonstrated a recent TensorFlow and deep learning based visual data analyzing and captioning system targeting outdoor and indoor agri-site worker picture data. The three cases were (1) cultivating agri-fields using a hoe in the semi-standing posture, (2) harvesting vegetables in the sitting posture, and (3) hanging harvests with changing postures: standing and sitting. We captioned our originally captured picture files considering various future practical systematic usages and presented message captioning and ratios. Our successive future works will provide further validation for varieties of detected targets and background conditions. We will also check the system accuracy, long-term performance, and appropriateness for other patterns or databases. In future, we will present the results to agricultural system developers not only for agri-workers and agri-managers, but also for security guards. We hope that progressive promising methodologies will be widely applied to real agri-working sites.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Shinji Kawakura and Ryosuke Shibasaki made the plan; Shinji Kawakura conducted the research, analyzed the data and wrote the paper; Shinji Kawakura and Ryosuke Shibasaki had approved the final version.

References

[1] N. Zhu, X. Liu, Z. Liu, K. Hu, Y. Wang, J. Tan, and Y. Guo, "Deep learning for smart agriculture: Concepts, tools, applications, and opportunities," *International Journal of Agricultural and Biological Engineering*, vol. 11, no. 4, pp. 32-44, 2018.

[2] S. Sladojevic, M. Arsenovic, A. Anderla, D. Culibrk, and D. Stefanovic, "Deep neural network based recognition of plant diseases by leaf image classification," *Computational Intelligence and Neuroscience*, pp. 1-12, 2016.

[3] R. Wang, J. Zhang, W. Dong, J. Yu, C. J. Xie, R. Li, and H. Chen, "A crop pests image classification algorithm based on deep convolutional neural network," *Telkomnika*, vol. 15, no. 3, pp. 1239-1246, 2017.

[4] F. J. Rodr ́guez, A. Garc ́a, P. J. Pardo, F. Ch ́avez, and R. M. Luque-Baena, "Study and classification of plum varieties using image analysis and deep learning techniques," *Progress in Artificial Intelligence*, vol. 7, no. 2, pp. 119-127, 2018.

[5] F. Femling, A. Olsson, and F. Alonso-Fernandez, "Fruit and vegetable identification using machine learning for retail applications," arXiv preprint arXiv:1810.09811, 2018.

[6] S. W. Chen, S. S. Shivakumar, S. Dcunha, J. Das, E. Okon, C. Qu, and V. Kumar, "Counting apples and oranges with deep learning: A data-driven approach," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 781-788, 2016.

[7] A. Dutta, J. M. Gitahi, P. Ghimire, and R. Mink, "Weed detection in close-range imagery of agricultural fields using neural networks," *Publikationen der DGPF*, band 27, pp. 633-645, 2018.

[8] K. Acharya and A. Pandya, "Scene description from images to sentences," *International Research Journal of Engineering and Technology*, vol. 4, no. 6, pp. 1302-1306, 2017.

[9] S. Das, L. Jain, and A. Das, "Deep learning for military image captioning," in *Proc. IEEE 21st International Conference on Information Fusion*, 2018, pp. 2165-2171.

[10] T. X. Dang, A. Oh, I. S. Na, and S. H. Kim, "The role of attention mechanism and multi-feature in image captioning," in *Proc. the 3rd International Conference on Machine Learning and Soft Computing*, 2019, pp. 170-174.

[11] K. Nogueira, O. A. Penatti, and J. A. D. Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognition*, vol. 61, pp. 539-556, 2017.

[12] Y. Hua, L. Mou, and X. X. Zhu, "Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 149, pp. 188-199, 2019.

[13] B. Wang, L. Ma, W. Zhang, and W. Liu, "Reconstruction network for video captioning," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7622-7631.

[14] D. Zhao, Z. Chang, and S. Guo, "A multimodal fusion approach for image captioning," *Neurocomputing*, vol. 329, pp. 476-485, 2019.

[15] I. Ram ́ez, A. Cuesta-Infante, J. J. Pantrigo, A. S. Montemayor, J. L. Moreno, V. Alonso, and L. Palombarani, "Convolutional neural networks for computer vision-based detection and recognition of dumpsters," *Neural Computing and Applications*, pp. 1-9, 2018.

[16] Analyze an image. Microsoft. [Online]. Available: https://www.microsoft.com/cognitive-services/en-us/computer-vision-api/

[17] Image2Text: A multimodal caption generator. Microsoft. [Online]. Available: https://www.microsoft.com/en-us//research/publication/image2tet-a-multimodal-caption-generator/

[18] Show and tell: A neural image caption generator. [Online]. Available: https://research.google.com/pubs/pub43274.html

[19] K. Saitou, *Zero Kara Tsukuru DeepLearning*, Tokyo, Japan: Ohmesha Inc., 2016.

[20] K. Saitou, *Zero Kara Tsukuru DeepLearning 2*, Tokyo, Japan: Ohmesha Inc., 2018.

[21] A. Gulli and S. Pal, *Chokkan Deep Learning ― Python×Keras De Aidea Wo Katachi Nisuru Recipe*, Tokyo, Japan: Ohmesha Inc., 2018.

[22] S. To, *Genba De Tsukaeru PyTorch Kaihatsu Nyumon*, Japan: SHOEISHA Inc., 2018.

[23] T. Niimura, *TensorFlow De Hajimeru DeepLearning Jissou Nyuumon*, Tokyo, Japan: Impress Inc., 2018.

Shinji Kawakura was born in Toyama Pref., Japan on July 14, 1978. He received Ph.D. in Environmentology from University of Tokyo, Bunkyo-ku, Tokyo, Japan in 2015; B.A. in Control System Engineering from Tokyo Institute of Technology, Meguro-ku, Tokyo, Japan in 2003; M.A. in Human-Factor Engineering, from Tokyo Institute of Technology, Meguro-ku, Tokyo, Japan in 2005.
His careers include systems engineering, research for private companies, and development and verification of sensing systems for outdoor agricultural workers.
Dr. Kawakura is with Department of Technology Management for Innovation at the University of Tokyo/Bunkyo-ku, Tokyo, Japan. He is Committee member of ICEAE and ICBIP.

Ryosuke Shibasaki is with Department of Socio-Cultural and Socio-Physical Environmental Studies, The University of Tokyo/Kashiwa-shi, Chiba, Japan. He is Dr. in Engineering. He is also Professor at the Center for Spatial Information Science, University of Tokyo.