Building Datasets Using k-Means Clustering and Evaluation of Image Classification with Convolutional Neural Network

Yuichi Miyata[1]* , Yoko Uwate[2] and Yoshifumi Nishio[3]
Tokushima University, Tokushima 770-8501, Japan
* *E-mail:* y.miyata@tokushima-u.ac.jp

1. Introduction

In Japan, the 5G environment which is a wireless communication technology, has been established. It is easy to obtain information and data by connecting networks in all fields. As a result, we can collect a lot of data such as images and videos on the Internet. However, when artificial intelligence (AI) learns, it is necessary to select effective data from these many data and use it more efficiently. We often see drones used for aerial photography and transportation of goods in various fields. The reason for this is the spread of AI and Internet of Things (IoT) and the ability to buy drones at a low price.

We expect these to become more and more familiar in the future. In the agricultural field, drones are used in the field to increase crop yield. We analyze the images taken by the drone using AI. This is effective for examining growth conditions and spraying agricultural chemicals [1], [2]. In addition, it is expected to be applied to the growth of trees and the calculation of the number of trees as well as agricultural products in the field of forestry. Furthermore, due to the increase in wild animals, the current situation is that the damage to crops and planted tree seedlings has become serious. For this reason, the study has been conducted on managing the number of wild animals using drones to protect the ecosystem [3]. To solve these problems, there is a limit in human hands, and the use of drones is indispensable. However, the battery of the drone itself and the camera mounted on the drone are limited. Therefore, it is important to use the collected image data effectively when learning with AI. We use clustering method when constructing a dataset for learning. Clustering is a method of classifying data having similar features. Therefore, we can classify images which have similar characteristics such as color and shape by using clustering.

In this study, we classify image data using k-means method, and construct datasets of various combinations using classified clusters. This dataset is trained by Convolutional Neural Network (CNN), and the datasets before and after clustering are evaluated with learning accuracy. In addition, we propose a dataset with higher quality by changing the ratio of images classified by clustering to create datasets.

2. Proposed method

Clustering algorithms are generally used in an unsupervised learning. K-means clustering is a method commonly used to automatically partition a dataset into *k* groups. It is a kind of non-hierarchical clustering and is one of the most widely used methods for clustering [4], [5]. K-means clustering divides the data into arbitrary *k* clusters by finding the center of the cluster that minimizes the evaluation function in Eq. (1).

$$f(\{C_k\}) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (\bar{x}_k - x_i)^2 \qquad (1)$$

K-means clustering is an algorithm for finding an appropriate cluster center and is often used for its merit of low calculation cost. Furthermore, we convert 3D array of input image data is converted to a 2D array data.

CNN used in this study consists of input layers, convolutional layers, max pooling layers, fully connected layers and output layers. Each layer detects the image features and learns. There are two layers in convolutional layers, max pooling layers, fully connected layers for each one. Its function is to progressively reduce the spatial size and length output from convolutional layers and reduce the amount of parameters and calculation in the network. Training and test images are compressed 128×128 pixels. We define as the training steps = 200. We learn these by using CNN and investigate the accuracy of color cut out images and original images.

The images for datasets are collected by entering a general name into the search word in the image search of google. Figs. 1 and 2 show an image of an animal (deer) and an image of the background only. In this time, we collected 200 images of animals and 200 background images.

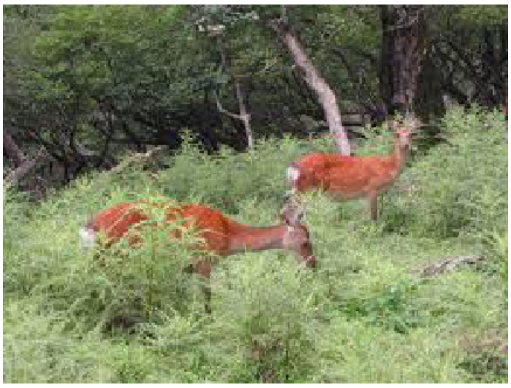

Fig.1 Image with object.

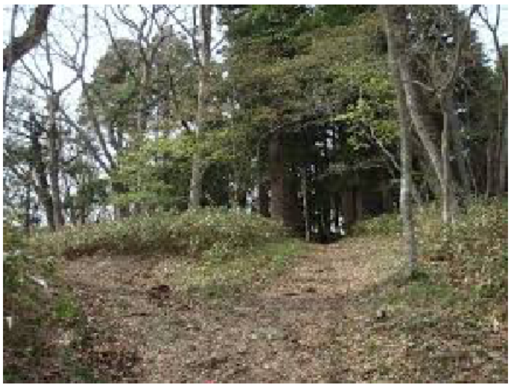

Fig.2 Background image.

3. Results and discussion

We classify each of the images of the animal and the images of the background into 2 classes using k-means clustering. Images of the animal are divided into 110 and 90 images. Images of the background are divided into 117 and 83 images. One with a large number of clustered images is referred to as cluster A and C, and the other with a small number as cluster B and D. We construct the following dataset to compare the effects of clustering. Datasets were created in various combinations using the images before and after clustering as follows. In addition, 20 images other than learning data were prepare as test data and used in common for each dataset.

In Fig. 3, we show the training accuracy and step of each dataset on CNN. The horizontal axis is step and the vertical axis is training accuracy. Each datasets have different number of images. We found that the influence of training accuracy was small even if the number of training images was different.

Datasets

[Dataset 1] : Original images (400 images)

[Dataset 2] : Randomly extracted (200 images)

[Dataset 3] : Cluster A and C (227 images)

[Dataset 4] : Cluster B and D (173 images)

[Dataset 5] : Cluster A and D (193 images)

[Dataset 6] : Cluster B and C (207 images)

[Dataset 7] : Cluster A , B 50% + Cluster C , D 50% (200 images)

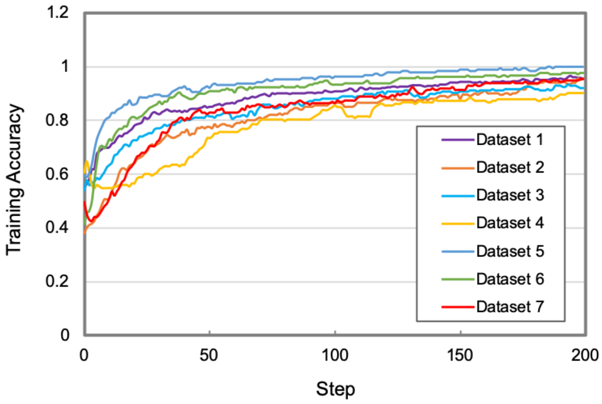

Fig.3 Line chart showing Training accuracy of each datasets.

Table.1 Average of training and test accuracies of each datasets

Datasets	1	2	3	4	5	6	7
Training accuracy [%]	95.1	93.4	91.0	89.7	96.3	96.7	94.2
Test accuracy [%]	88.7	86.7	81.1	70.0	67.8	67.8	90.4

Train each CNN using each constructed dataset and compare the learning accuracy before and after clustering. Table 1 shows the learning accuracy and test accuracy for each dataset. According to Table 1, the learning accuracy is almost 100% for all datasets. In test accuracy, variation was observed for each dataset. Dataset 1 that was not clustered had 400 images used for learning, so high test accuracy was obtained. Since dataset 2 has half the number of images of dataset 1, the test accuracy is lower than that of dataset. Next, the test accuracy of the datasets subjected to clustering is compared. Test accuracy of dataset 3 using images of clusters A was the highest, and test accuracy was low for datasets of other combinations. Furthermore, although the number of images used for each dataset was different, dataset 4 had fewer images than dataset 5, but test accuracy was higher for dataset 4. Therefore, there is no correlation that accuracy increases as the number of images increases, and it may be dependent on the image to some extent. When datasets 2 and 7 constructed before and after clustering using the same number of images were compared, test accuracy was higher for dataset 7. As a result, it was found that the diversity of features in the image by clustering as preprocessing affects the learning accuracy of the network.

Table.2 Processing time and learning efficiency of each datasets

Datasets	1	2	3	4	5	6	7
Processing time [s]	866.2	433.2	450.9	315.7	429.4	439.6	432.7

Next, the time taken for learning is compared to evaluate the effectiveness of the dataset. Table 2 shows the learning time and learning efficiency for each dataset. From Table 2, it can be seen that the more images used in the dataset, the longer it takes to process. We found that the processing time depends on the number of image data used for learning.

4. Conclusion

We used k-means clustering to construct datasets that achieve high accuracy with few images. We classified images into two classes for each feature by k-means clustering. From these images, we constructed 10 datasets and trained using CNN.

In this study, the k-means method was used to classify the image data into two for each feature, and various datasets were constructed using the divided images. Each dataset was trained by CNN, and the effect on learning accuracy before and after clustering was investigated. Some datasets after clustering had lower test accuracy than before clustering, but datasets using 50% each of the two clusters obtained higher test accuracy than before clustering. This dataset has the highest learning efficiency calculated from the processing time, and can be said to be a higher quality dataset. This study suggested that not only colors and shapes but also features that cannot be judged by the human eye are classified by clustering and used as learning data to obtain a higher quality data set. As a future topic, we will classify the features more finely by clustering and investigate which features in the image contribute greatly to learning when they are trained by CNN.

References

[1] J. Rebetez, H.F. Satizabal, M. Mota, D. Noll, L. Buchi. “Augmenting a convolutional neural network with local histograms - A case study in crop classification from high-resolution UAV imagery”, ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016.

[2] Camiel R. VerschoorPascal MettesKitso EpemaLian Pin KohSerge Wich. “Nature Conservation Drones for Automatic Localization and Counting of Animals”, European Conference on Computer Vision, ECCV 2014: Computer Vision - ECCV 2014 Workshops pp 255-270 2014.

[3] Koichi Nakajima Katsuhiko SEO Shinji Kitagami Tetsuo Siotsuki Hisao Koizumi. “Development of Unmanned Aerial Vehicle Systems for Monitoring and Prevention of Bird and Animal Damage and its evaluation”, Information Processing Society of Japan, IPSJ SIG Technical Report, Vol.2015-MBL-76 No.10 2015.

[4] V. Birodkar, H. Mobahi, and S. Bengio. “Semantic redundancies in image-classification datasets: The 10% you don't need”, ArXiv, 1901.11409, 2019.

[5] Takashi Onoda, Miho Sakai, Seiji Yamada. “Experimental Comparison of Clustering Results for k-means by using different seeding methods”, The 25th Annual Conference of the Japanese Society for Artificial Intelligence, 2011.