Analysis of the trompe l'oeil recognition method using deep learning

Sayo Horie[1], **Naoki Mori**[1]
[1]Osaka Prefecture University
1-1 Gakuen-cho, Naka-ku, Sakai, Osaka, Japan
horie@ss.cs.osakafu-u.ac.jp

Abstract

Can the computer understand trompe l'oeils, which represent multiple meanings? This question is one of the most interesting questions of artificial intelligence and cognitive science field. Extending the human feeling of watching trompe l'oeils to the computer is not easy task because it is not clear what is the correct answer for trompe l'oeils recognition. It is difficult to apply a machine learning to the dataset without correct answer data. Recently, the image recognition field has been developed rapidly through deep learning; however, conventional research on image recognition has mainly focused on images that express unique content. For this reason, few studies have reported image recognition for ambiguous figures such as trompe l'oeils, which represent multiple meanings. In this research, we have proposed the method and experimental settings to understand ambiguous figures through deep learning. To confirm the effectiveness of the proposed method, computer experiments were carried out taking real ambiguous figures as examples.

Introduction

With the advent of deep learning, the field of image recognition has made rapid progress in recent years. It has already been reported that it outperforms humans in simple object recognition. In addition, the ability to recognize images is expected to improve continuously. On the other hand, the current human sensibility and cognition, which cannot be learned efficiently even by deep learning. This is due to the fact that it is difficult to quantitatively evaluate abstract concepts such as human sensibility and cognition, for which there is no clear correct answer. For this reason, conventional deep learning research has mainly dealt with problems with unique answers, such as image classification and object detection. However, problems for which the answer is not clear, such as computer-generated trompe l'oeil recognition, have been well studied and They never came. In this paper, we focus on images that can be interpreted in multiple ways, such as trompe l'oeil. Although trompe l'oeil is a typical example of such images, we define polysemous figures as those which can be interpreted with multiple meanings in this paper.

In this study, we propose an experimental framework and a method for the discrimination of polysemic figures using deep learning techniques, which have shown high performance in general object recognition. We also show the effectiveness of the proposed method through numerical experiments using actual trompe l'oeil pictures.

The structure of this paper is as follows. In Chapter , we describe our research on polysemy maps in the field of cognitive science and the elemental technology of deep learning in this study. Chapter describes the proposed method and Chapter presents the results of numerical experiments and discussion. Chapter is a summary of this study.

Previous Research

Ambiguous Figure

In this study, we define unambiguous figures as objects with unique labels, and polysemous figures as objects with multiple interpretations, such as trompe l'oeil. In the field of cognitive science, the factors that influence the interpretation of polysemous figures are considered to be gaze points and selective attention [5][6].

It has been reported that the polymorphic figure shown in Figure 1 is interpreted as a duck in 98% of cases when the participants are gazing at point 1, while it is interpreted as a rabbit in 94% of cases when the participants are gazing at point 5 [10]. However, the computational approach to the interpretation of such polysemous figures has not been sufficiently studied.

Convolutional Neural Network

Convolutional Neural Networks (CNN) [8] have been attracting attention in the field of image recognition as a deep learning method inspired by the nature of receptive fields in the visual field. In this study, we use VGG-16 [9], which has been trained on a large dataset called ImageNet [4], as a CNN.

Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) [7] is a decision visualization technique for CNNs that displays a color map of the areas that CNNs look at for classification. We consider it to be the magnitude of the contribution

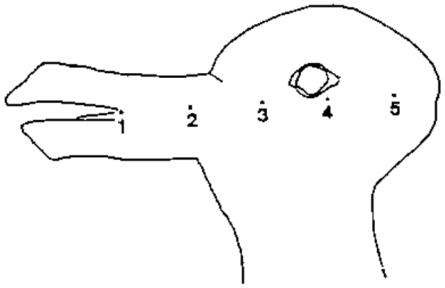

Figure 1: Attention point of rabbit duck image

Table 1: Parameter adjustment section using Optuna

Number of units in the middle layer	100 500
Dropout rate	0 1
Learning rate	0.00001 0.01

and judge it to be an important part of the classification prediction. When computing the contribution, the slope of the slope is generally used for the loss value of the prediction class in the final convolutional layer. In this study, we visualize the gradients of polysyllabic figures in order to analyze the decision-making process.

In the following figure, the size of the gradient is visualized by using heat-maps, and the large part of the gradient is shown in red and the small part in blue.

Proposed Method

Composition of Deep Learning

In this study, we propose a problem framework for understanding polysemous graphics using existing deep learning methods, and conduct numerical experiments based on a concrete experimental method.

Data Augmentation was applied to a dataset, and Transfer Learning from the third and subsequent layers of the final convolutional layer of VGG-16, which has been trained by ImageNet, was applied to each of the tasks using the dataset, and the number of intermediate layer units, dropout rate, learning The three parameters of the rate were adjusted. Table 1 shows the parameter-adjustment interval for Optuna [3]. Finally, we visualize the basis for the CNN decision using Grad-CAM and compare them. We also call the output of Softmax functions of the final combined layer the classness of the class.

Structure of the Numerical Experiment

In this study, the following two experiments were conducted.

- Experiment 1: An experiment on discrimination between polysemous and unimodal figures by deep learning
- Experiment 2: An Experiment on the Relationship between Deep Learning and Human Cognition for Polysemous Figures

In each experiment, we analyzed the gazing points of polymorphic figures by computer using Grad-CAM.

The purpose of Experiment 1 is to confirm whether or not CNN can distinguish between polysemous figures and two types of unambiguous figures, which are the components of polysemous figures, and to analyze the basis of the decision. Concretely, we conducted the following experiments.

Experiment 1-1 Polysemous figures of landscapes and human faces, landscapes, and portraits are discriminated into three classes. Polysyllabic figures and unisyllabic figures with similar colors and feature points in the test images are extracted and compared with the basis of judgment.

Experiment 1-2 The polymorphic figures, landscapes and portraits are experimentally examined to see if the substructures of the polymorphic figures can be extracted from the unimodal figures. In this study, we confirm whether we can extract the parts that can be seen as both faces and landscapes from the landscape paintings.

The purpose of experiment 2 is to investigate the boundary of the change of perception of polysyllabic figures, which changes the human interpretation of the objects drawn by the rotation, and to analyze the basis of the decision. Specifically, the following experiments are conducted.

Experiment 2-1 Using the learned model of a horse and a frog, we discriminated between a horse and a frog in a polysemous figure that appears to be transformed into a horse/frog by rotating it by 90 degrees.

Experiment 2-2 Using the learned model, we discriminated between a rabbit and a duck in a polysyllabic figure that gradually transformed into a rabbit/duck by rotation. In addition, 16 undergraduate and graduate students aged 20 to 25 years old were shown 168 images rotated by 15 degrees to show the polysyllabic figure that appears to turn into a rabbit or a duck depending on the rotation of 7 images, and they were asked to choose either ”rabbit, duck or neither. The average value was obtained by normalizing as “ 0: duck, 0.5: neither, 1: rabbit ” and compared with the discrimination result by CNN.

Experiment 2-3 In order to compare the differences between the human gazing points identified in Experiment 2-2 and the computer's decision grounds, we created a pseudo- gaze state by applying a mask to the test image.

Dataset

We collected trompe-l'oeil images from books and the Internet and created a dataset of trompe-l'oeil images that can be interpreted as polysemous figures.

In Experiment 1, it is noted that there are many polysemic figures of landscapes and human faces, and many portraits and landscapes similar to those of polysemic figures. The polysemic figures of landscapes and human faces are defined as a class of polysemic figures based on images that the author judges to be landscapes and human faces. Only the training data for the ”polymorphic figures” class was augmented from 257 images to 2570 images. In addition, the “landscape” and “cityscape” labeled images in ”WikiArt” [2] are used as the dataset of unique figures, and the “portrait” labeled images are used as the “landscape” class, and the “portrait” class is used as the “portrait” class.

Table 2: Experiment 1-1: Conditions of the experiment

Class	3class
Epoch	22
Batch size	32
Number of Training	2570 /class
Number of Validation	36 /class
Number of tests	72 /class
Image size	200 × 200 × 3(RGB)
Activation function	Softmax
Optimization function	Adam
Loss function	cross-entropy

Table 3: Experiment 1-1: Optuna Optimization Results

Dropout rate	0.67640
Learning rate	2.2686e-05
Number of units in the middle layer	100

In Experiment 2, we focus on polymorphic figures that change the appearance of objects by rotation. The reason is that the analysis of the gazing point is expected to be easy because the number of objects drawn is a single one, and it is easy to see the transition of the gazing point by rotation of the computer.

In Experiment 2-1, we scraped the search results of “horse illustration” and “frog illustration” from Pinterest[1] and labeled them as horses and frogs as a training/evaluation dataset. As a test data set, we used seven polymorphic figures that appear to turn into horses/frogs when rotated by 90 degrees, as shown in the upper rows of Figs. 8 and 7.

In Experiment 2-2, we scraped the search results of “rabbit illustration” and “duck illustration” from Pinterest and labeled them as rabbit and duck for training/evaluation. As a test data set, we used seven polymorphic figures that appear to gradually transform into a rabbit/duck by rotation as shown in Figs. 10 and 11.

Numerical Experiment

Experiment 1-1

Conditions of an Experiment Table 2 shows the experimental conditions and Table 3 shows the optimization results of the network parameters by Optuna.

Results of an Experiment The accuracy of the CNN discrimination of three classes of polymorphic figures, landscape and portrait was 91.2% compared to the baseline of 33.3%. Figure 2 shows the confusion matrix with the true value on the vertical axis and the predicted value by CNN on the horizontal axis. Figure 3 shows the judgment basis of the Grad-CAM computer. The results of Grad-CAM show that the Grad-CAM grounded the judgments on the whole landscape paintings, and the grounded the judgments on the human face area for the portraits. As for the polymorphic figures, we confirmed that the basis for judging them is the same as in the case of portraits, namely the human face.

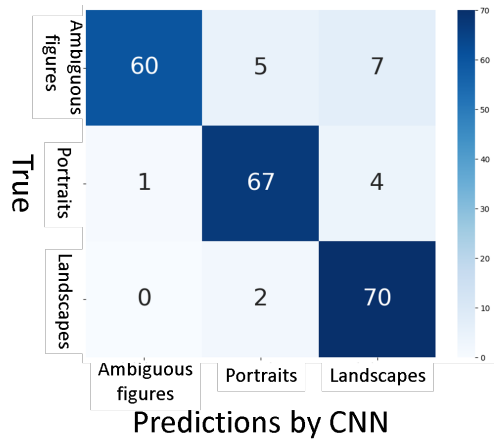

Figure 2: Experiment 1-1: Confusion Matrix

It is confirmed whether the three classes discriminating model developed in Experiment 1-1 can distinguish between similar polysyllabic figures and unisyllabic figures. A pair of images of ”polysyllabic figure” class and ”portrait” class, and a pair of images of ”polysyllabic figure” class and ”landscape” class are extracted from the test images. Similarity in color and feature points is defined as a pair of images with histogram similarity less than or equal to 0.9 and AKAZE similarity greater than or equal to 120. Figure 4 shows the basis for judging the similarity of pairs of images and the classes of ”polymorphic figures” and ”portrait” / ”landscape”. From this figure, we can see that the grounds for judging polysemy and unity of figure are completely different even for similar images, and that the judgments are accurate. In this experiment, we can say that the discrimination of the polysemy of the image was successfully accomplished.

Experiment 1-2

Conditions of an Experiment The input image was divided into 5 × 5 grids and cropped into 1 × 1, 2 × 2, 3 × 3 and 4 × 4 grids (54 choices), and all the cropped images were tested using the trained 3-class discriminator. In this experiment, images of the ”landscape” class are cut out in the above way and identified as ”portrait” class, and images of the ”polymorphic figure” class are examined in detail.

Results of an Experiment Figure 5 shows an example of a landscape image extracted from the ”landscape” class and identified as a ”portrait” class, and Figure 6 shows an example of a landscape image extracted from the ”landscape” class and identified as a ”polygenic figure” class. From 17960 landscape images with “landscape” and “cityscape” labels in WikiArt, the total number of images is 969840, of which 26990 (2.78%) are identified as ”polymorphic figure” class.

From Figure 5, we can see that the image identified as a portrait reflects the face area of the person depicted in the landscape painting. On the other hand, Figure 6 shows that in the image identified as a polymorphic figure, both the land-

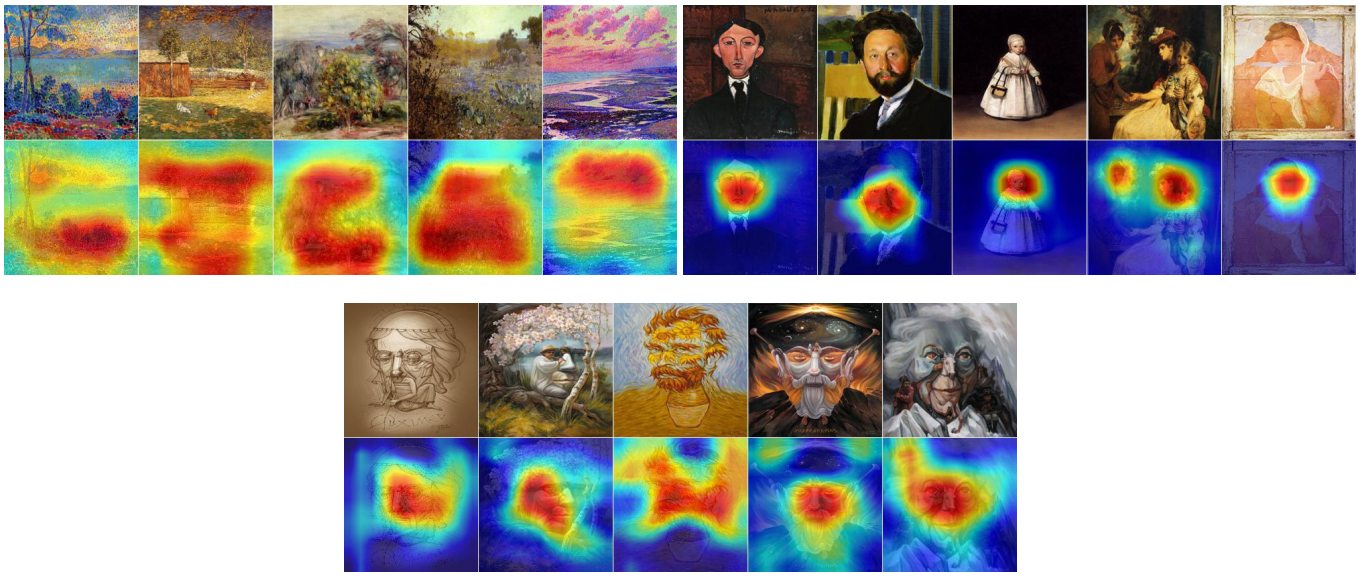

Figure 3: Experiment 1-1: Example of Grad-CAM results

scape and the face are reflected in the image. Therefore, it can be said that we were able to extract "the face reflected in the landscape" and "the polymorphic figure seen in both the landscape and the face" from the landscape image.

Experiment 2-1

Conditions of an Experiment Table 4 shows the experimental conditions and Table 5 shows the optimization results of the network parameters by Optuna.

Results of an Experiment Figure 7 shows an example of the basis of the Grad-CAM judgments for a horse-dominant image. Figure 8 shows an example of the grounded decision result of the Grad-CAM for a frog-dominated image. From the Grad-CAM results, we found that the horse-dominant participants gazed at the ears and mane in the horse-dominant image, and the frog-dominant participants gazed at the boundary between the frog's feet and the water surface in the frog-dominant image. For polymorphic figures that appear to turn into a horse/frog when rotated by 90 degrees, we succeeded in making the computer recognize the change in object recognition with more than 97

Experiment 2-2

Conditions of an Experiment The experimental conditions are the same as those in Experiment 2-1 shown in Table 4. Table 6 shows the optimization results of the network parameters by Optuna.

Results of an Experiment Figure 9 shows a scatterplot of the results of human questionnaires on the vertical axis and CNN recognition results on the horizontal axis. The correlation coefficient between the results of the questionnaire and the computer recognition is 0.772. Therefore, we can see that there is a strong correlation between the human recognition results and the computer recognition results for polysemous figures.

Figure 10 shows an example of the groundbreaking results of rabbit-dominant image by Grad-CAM. Figure 11 shows an example of decision basis for a duck-dominant image by Grad-CAM. The results of Grad-CAM showed that the gazing points in rabbit-dominant images were concentrated around the base of the ears, and the gazing points in duck-dominant images were concentrated in the eyes to the body. In a previous study on human gaze points [10], it was reported in Figure 1 that a duck was perceived as a duck while a rabbit was perceived as a rabbit while gazing at point 1. On the other hand, the computer Grad-CAM results show that people perceive point 4 as a duck when they are gazing at it, while they perceive point 2 as a rabbit when they are gazing at it. A comparison between the computer Grad-CAM results and a previous study [10] on humans revealed that the human gaze point and the computer's basis for judgment are different.

Experimrnt 2-3

The difference between human and computer vision We consider the difference between humans and computers. The first difference is that the area around the human gazing point appears gradually blurred, while the entire screen of a computer is uniformly visible. Therefore, we thought that by applying the heat map by Grad-CAM to the image as a mask, we could create a human-like state. The second difference is that a person sees the same image as different objects depending on the transition of the gazing point, while a computer can capture the same image in only one way. Therefore, we can consider it to be human-like if the results of the computer test change depending on the way the mask is applied.

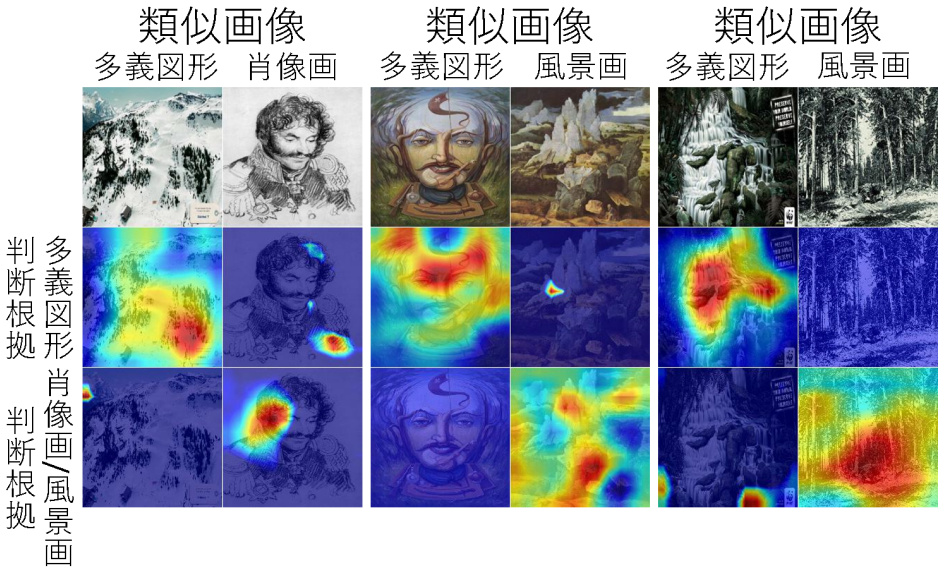

Figure 4: Experiment 1: Identification of similar images

Figure 5: An example of an image cut out of a landscape and identified as a portrait

Figure 6: An example of an image cut out of a landscape and identified as a polysemous figure

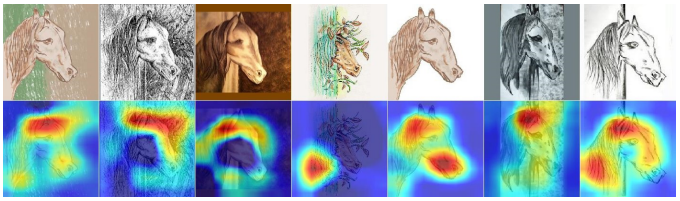

Figure 7: Experiment 2-1: Grad-CAM results for horse-dominant images

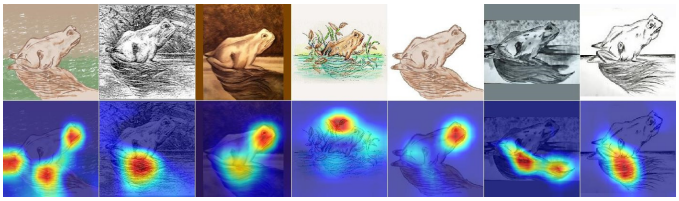

Figure 8: Experiment 2-1: Grad-CAM results for frog-dominant images

Table 4: Experiment 2-1, 2-2: Conditions of the Experiment

Class	2 class
Epoch	300
Batch size	32
Number of Training	800 /class
Number of Validation	100 /class
Number of tests	100 /class
Image size	200 × 200 × 3(RGB)
Activation function	Softmax
Optimization function	Adam
Loss function	cross-entropy

Table 5: Experiment 2-1: Optuna Optimization Results (horse/frog)

Number of units in the middle layer	100
Dropout rate	0.042160
Learning rate	1.8731e-05

Table 6: Expeeriment 2-2: Optuna Optimization Results (rabbit/duck)

Number of units in the middle layer	500
Dropout rate	0.30511
Learning rate	5.0936e-05

Method of an Experiment Based on Section , we performed Experiments 2-3. Figure 12 illustrates Experiment 2-3. We extracted images that were both rabbit-like and duck-like, i.e., both rabbit-like and duck-like, from the trained two-class discrimination model created in Experiment 2-2. Next, we multiply the CAM values of the Grad-CAM and the CAM values of the Grad-CAM, which are the basis for judging ducks and rabbits, respectively, in the extracted images by the value of each pixel in the original image. By this operation, we created a duck image and a rabbit image, both of which were masked except for the CNN decision-based region, and compared the test results of these images with those of the unprocessed images.

Results of an Experiment Twenty-five out of 1008 images, which are both rabbit-like and duck-like, were rotated by 5 degrees from 7 rightward and 7 leftward, respectively. Among the 25 images, 17 of them showed the change in the discrimination from rabbit to duck by the mask. About 68% of the images showed a change in discrimination. This suggests that the masking process based on the judgmental domain of CNN visualized by Grad-CAM is close to the human gaze state.

Conclusion and Future Tasks

In this study, we proposed a method to make computers understand polysemous figures, and showed its effectiveness and its relation to human cognition through numerical experiments.

Experiment 1-1 shows that the computer can discriminate between polysemous figures and univalent figures with 91.2

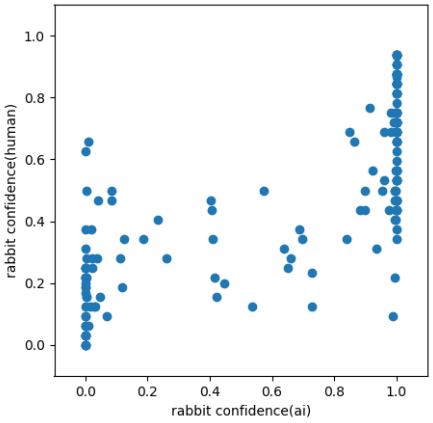

Figure 9: Experiment 2-2: Scatterplot of people and CNN identification results

% accuracy. Then, the similarity between computerized portraits and polymorphic figures in terms of gaze points was confirmed by analysis using Grad-CAM. We are able to discriminate between similar polysyllabic and unisyllabic figures. In Experiments 1-2, we are able to cut out polymorphic figures that can be seen as both landscapes and faces from the landscape painting.

In Experiment 2-1, we succeeded in the recognition of polymorphic figures that appear to turn into a horse/frog when rotated by 90 degrees with more than 97% accuracy. In Experiment 2-2, a questionnaire survey on polymorphic figures that appeared to change into a rabbit/duck shape with rotation showed a strong correlation between human recognition and computer recognition, with a correlation coefficient of 0.772. In Experiment 2-2, the results of the Grad-CAM analysis revealed that the points that humans and computers pay attention to were different when they recognized rabbit-like and duck-like behaviors. In Experiments 2-3, 68% of the images were masked according to the Grad-CAM decision-making rationale for both rabbit and duck, and 68% of the images showed a change in the recognition target, suggesting that the Grad-CAM visualization of the CNN decision-making rationale is similar to the human gaze state.

Future issues include questionnaire experiments that take into account the presentation of images and the surrounding environment, application of eye tracking technology, and automatic generation of trompe l'oeil pictures using computers.

This work was partly supported by the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (B) (Project No. 19H04184).

References

[1] ???? Pinterest. https://www.pinterest.jp/.

[2] ???? WikiArt. https://github.com/cs-chan/ArtGAN/tree/master/WikiArt%20Dataset.

[3] Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; and Koyama,

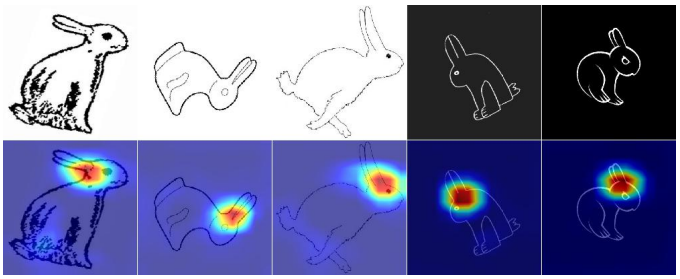

Figure 10: Experiment 2-2: Grad-CAM results for rabbit-dominant images

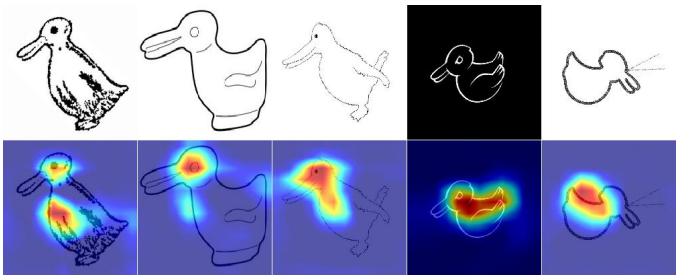

Figure 11: Experiment 2-2: Grad-CAM results for duck-dominant images

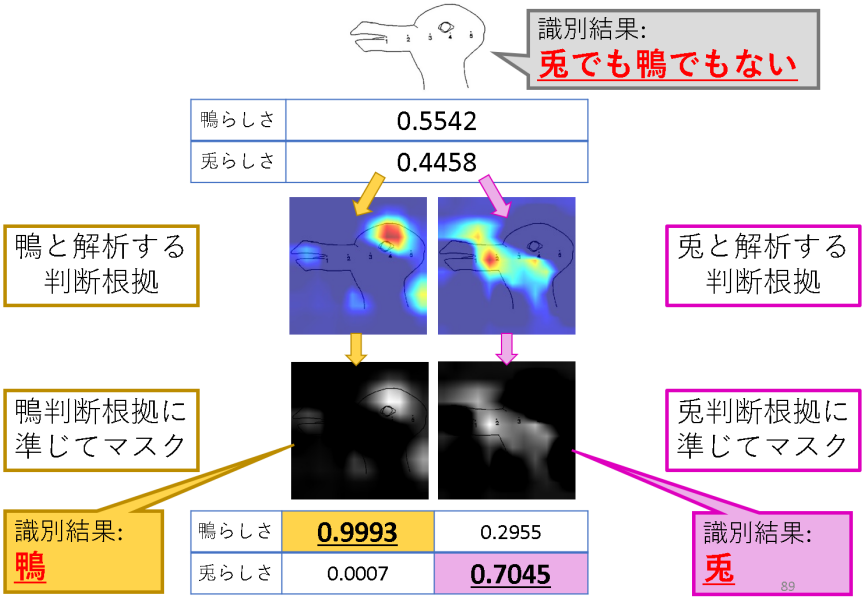

Figure 12: Mask Analysis Experiment

M. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. *CoRR* abs/1907.10902. URL http://arxiv.org/abs/1907.10902.

[4] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

[5] Kawabata N, Nobuo Yamagami, K. N. M. 1977. Visual fixation points and depth perception. .

[6] N, K. 1986. Attention and depth perception. .

[7] Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *The IEEE International Conference on Computer Vision (ICCV)*.

[8] Simard, P. Y.; Steinkraus, D.; and Platt, J. C. 2003. Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis. ISBN 0-7695-1960-1. URL http://dl.acm.org/citation.cfm?id=938980.939477.

[9] Simonyan, K.; and Zisserman, A. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556.

[10] 充史, 岸.; and 信男, 川. 1996. 局所的・大域的情報選択モデルによる多義図形の非あいまい化. テレビジョン学会誌 50(5): 594–598. doi:10.3169/itej1978.50.594.