Discovery of Image Pixels highly Contributing to CNN regression

Yitian Li* , Nobutaka Shimada *
*Ritsumei University,Shiga,Japan

***Abstract*— If you want to output some specific information from an image, you should input the image to a network , and then output parameters that you want. But the information you want to output is only about a part of the image or just a specific object in that image.In this case, some noise elements such as other objects or the background affects the output result. Therefore, in this study, we use CNN to visualize the pixel area that contributes to the CNN regression by generating an attention map that can express where the object which we focus on exists. And then output the parameters we want at the same time.**

I. Introduction

When we want to output some information from an image. We only focus on a part of the image or a specific object. But,there is always some noise elements in the image,such as other object or the background,they can affect the output result. This study focuses on this and aims to visualize the pixel region that contributes to the CNN regression by generating an attention map that can represent where the object area exists in the image we input.

If we can reduce the influence of the nosie elements on the output result, the accuracy of the parameters we output will increase. Therefore, if the pixel area contributing to the CNN regression can be visualized, application to various fields can be expected.

CNN networks are often used in image classification and area segmentation problems. For example,as shown in Fig. 1,we input an image of an ellipse and a noisy background,and we want to extract only the area where the ellipse exists.,or we want to get the center coordinateof the ellipse.

The conventional method is used to estimate the center coordinates of the ellipse using the total pixel value of the image. But if we do not know where the ellipse is, we may have forcibly calculated from some useless region (noise, other shapes) that do not have any information. If the computational complexity is very large, we will need to much train datas.[1]

IN this study, we will construct a network as shown in Figure 2 using CNN, which has two parts. The above part is a normal supervised learning CNN network that inputs an image and outputs the corresponding parameters. The below part is a unsupervised learning network that automatically generates an attention map $W(x, y)$ of an object region which we focus on by inputting an image. In the attention map $W(x, y)$, the pixels of the region that we focus on (the ellipse area of the input image in Fig.2) are close to 1. And the useless pixels (background) close to 0. The parameters that we wan to output depend on the attention region(the pixel of the object region),

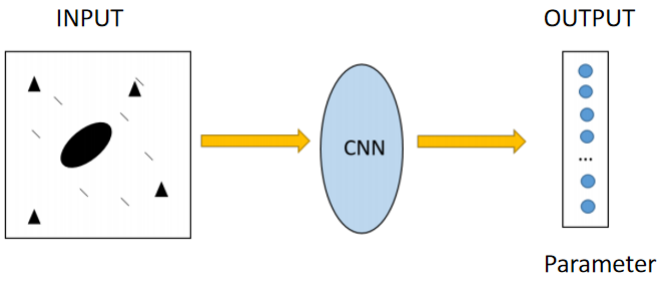

Fig. 1. Normal CNN Network

so if we change the backgroud of the image,we will still get the same output. We make a noise image N randomly, firstly multiply it by $1 - W(x, y)$ and then add the input image I. The attention region(the region of the ellipse)is not changed,so if we input this image we just made into the above CNN network, we should get the same result. If we update the weight of CNN network by minimizing the loss function,we can get the attetion map.

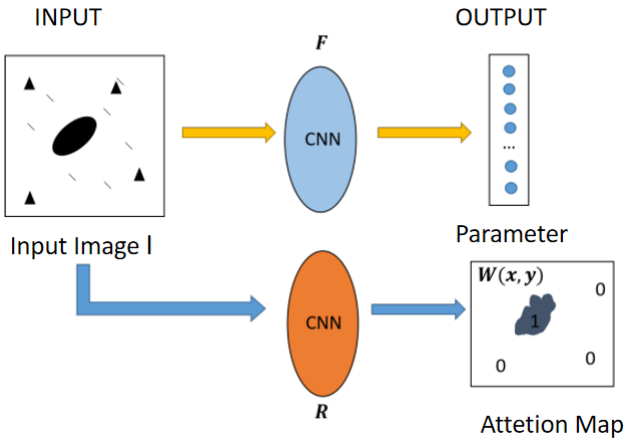

Fig. 2. Attention Map CNN Network

II. Discovery of Image Pixels highly Contributing to CNN regression

A. Network

As shown in Figure 3, This is the detailed structure of our CNN Network.

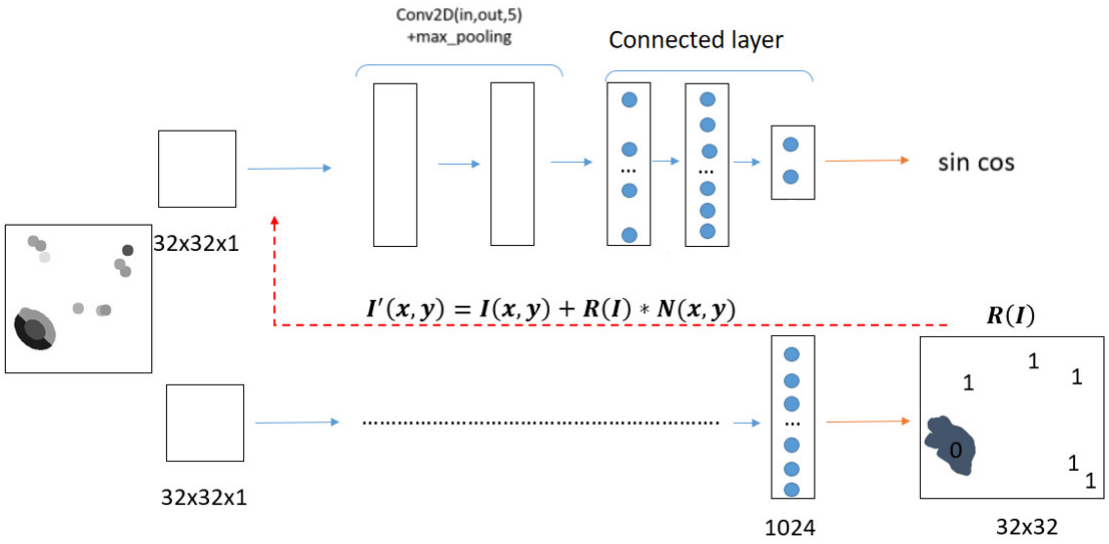

Fig. 3. Detailed structure of our CNN Network

1) Training: We update the weight by minimizing the loss function that the two parts of the network. Here, the loss function has three components.

$$L(i) = a|F(I_i) - P(I_i))|^2 + b|F(I_i') - P(I_i)|^2 + c|Area(R(I_i)) - S_R)|^2 \quad (1)$$

$$L_{total} = \sum_i L(i) \quad (2)$$

Equation 3 is the mean square error between the pretected center coordinates F (I) and supervisory signal P (I).

$$L = |(F(I) - P(I))|^2 \quad (3)$$

In Equation 4, firstly ,we make an attetion map R(I) using the input image I. And R (I) is actually 1- W (x, y). In the attention map W (x, y), the pixels of the region that we focus on (the ellipse area of the input image in Fig.2) are close to 1. And the useless pixels (background) close to 0. The parameters that we wan to output depend on the attention region(the pixel of the object region), so if we change the backgroud of the image,we will still get the same output. As shown in equation 4,we use image I and attention map R to make a image called I'.

Then we minimize the mean square error of the estimated center coordinates F (I') and supervisory signal P (I) as shown in equation 3.

$$L = |(F(I') - P(I))|^2 \quad (4)$$

$$F(I') = F(I) + R(I)N(x,y) \quad (5)$$

The ratio 1-S_R of the object area is limited using the integral value of the entire weighted image $R(I)$ is shown in Equation 5. The area of the ellipse is about 0.05 and the intergral value of S_R is about 0.95.The total number of pixels of images are $32 \times 32 = 1024$.

$$L = |(Area(R(I)) - S_R)|^2 \quad (6)$$

$$Area(R(I)) = \frac{\int R(I)\, dxdy}{1024} \quad (7)$$

III. Experiment

In this section, we will test our network in several experiments.

A. Dataset1(ellipse+cirlcle noise)

The training dataset(ellipse+circle noise) is shwon as Figure 4 ,we want to extract the region where the ellipse exists.And we wan to get the inclination of the ellipse.We have 8000 images for training . We mark the supervisory signal with a green, and

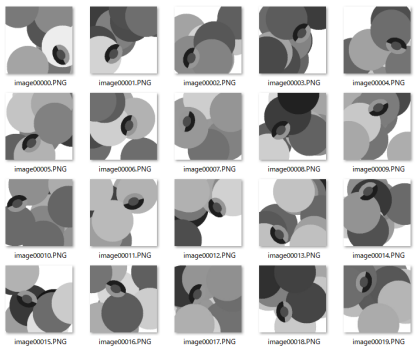

Fig. 4. Dataset1(ellipse+cirlcle noise)

mark the estimated coordinate with a blue. The attention map is a grayscale image with a value range of 0-255, the dark area is the ellipse area.

The result of 3000 test data, the average loss is 13[degree]. As shown in Fig 5,according to the weight image, we extract the region of the ellipse corectly.

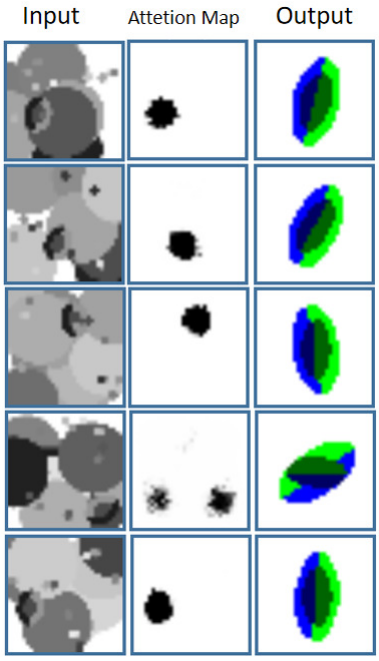

Fig. 5. Result 1(ellipse+cirlcle noise)

B. Dataset2(two ellipses+point noise)

The training dataset(two ellipses+point noise) is shwon as Figure 6 ,we want to extract the region where the ellipse exists.And we wan to get the Intersection coordinate of major axis of two ellipses.We have 3000 images for training. We mark

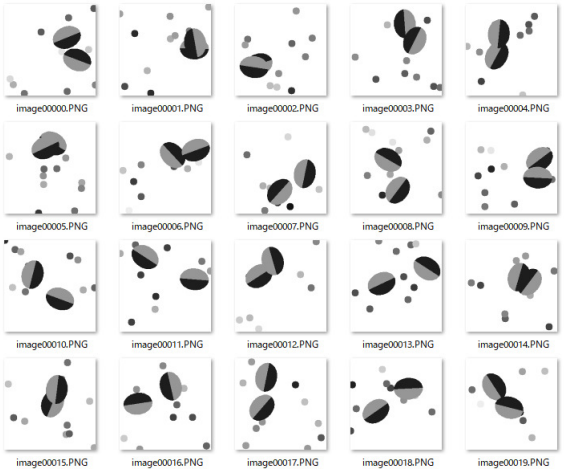

Fig. 6. Fataset 2(two ellipses+point noise)

the supervisory signal with a green, and mark the estimated coordinate with blue. The attention map is a grayscale image with a value range of 0-255, the dark area is the ellipse area. The result of 100 test data, the average loss is 2.5[pixel]. As shown in Fig 6,according to the weight image, we extract the region of the ellipse corectly.

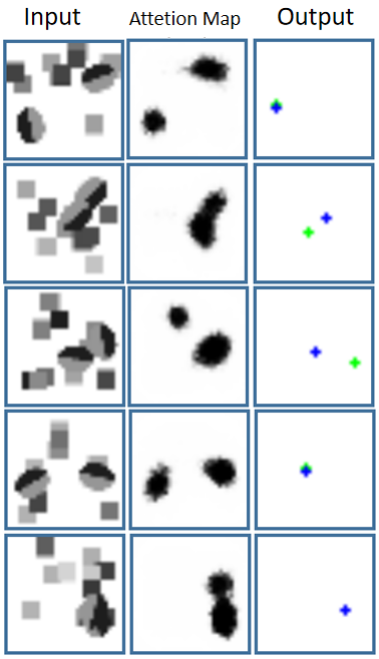

Fig. 7. Result 2(two ellipses+point noise)

C. Dataset3(two ellipses+real object noise)

The training dataset(two ellipses+real object noise) is shwon as Figure 8 ,we want to extract the region where the ellipse exists.And we wan to get the Intersection coordinate of major axis of two ellipses.We have 8000 images for training. We mark

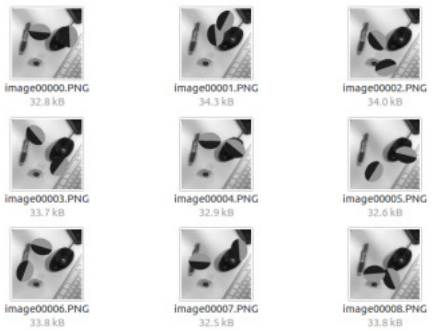

Fig. 8. Dataset 3(two ellipses+real object noise)

the supervisory signal with a green, and mark the estimated coordinate with blue. The attention map is a grayscale image with a value range of 0-255, the dark area is the ellipse area. The result of 3000 test data, the average loss is 3.28[pixel]. As shown in Fig 9,according to the weight image, we extract the region of the ellipse corectly.

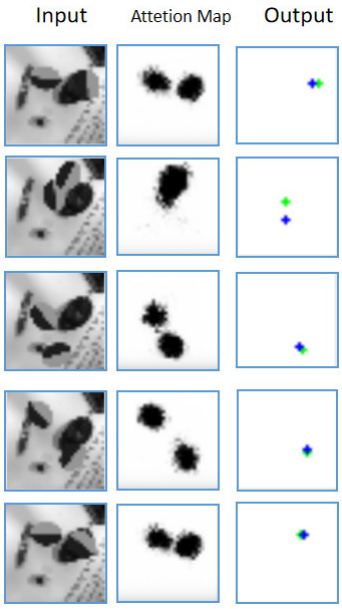

Fig. 9. Result 3(two ellipses+real object noise)

IV. Conclusion

In this study, we aim to visualize the pixel region that contributes to the CNN regression by generating an attention map that can represent where the object area exists in the image we input.By testing 3 kind of datasets, we can say that we extract the region that we want corectly, and at the same time,we can also output the parameters we want.

References

[1] Yujian Zhao,Nobutaka Shimada ”Automatic acquisition and recall of gaze areas and operation procedures related to object manipulation”