Appendix: Counterfactual-based Saliency Map: Towards Visual Contrastive Explanations for Neural Networks

1. Ablation Study

When comparing with other saliency map interpreting methods, we combine the two types of saliency maps from CCE to a unified saliency map. And define M_p for class p as $M_p = \alpha \cdot M_{\text{sup}} + (1-\alpha) \cdot M_{\text{pro}}$ and α. We performed ablation study for parameter α on 1000 images random selected from ImageNet dataset for VGG-19 model. According to the experimental results in Table 1, we chose 0.7 as the value of parameter α in the subsequent experiments.

α	Ins. ↑	Del. ↓	Diff. ↑
1	57.0968	19.6747	37.422
0.9	57.4848	19.0434	38.441
0.8	57.8177	18.461	39.357
0.7	**57.9971**	17.9553	**40.042**
0.6	57.4705	**17.6202**	39.850
0.5	56.5456	17.9865	38.559
0.4	55.0824	19.214	35.868
0.3	53.4509	20.9388	32.512
0.2	52.4718	21.6932	30.779
0.1	51.5035	22.7094	28.794

Table 1: Insertion-Deletion tests results for α. Higher insertion score (Ins.) are better and lower deletion score (Del.). The difference score (Diff. which higher is better) shows that CCE outperforms other related methods. The best records are marked in bold.

We also performed ablation experiments on number for counterfactuals(M) and hyper-parameter γ used to balance the two losses when generating counterfactuals.Refer to Table 2 for the experimental results on M and Table 3 for the experimental results of hyper-parameter γ. Considering the experimental results and time cost, we use 5 as the value of parameter M and $\gamma = 20$ in the paper.

At last we conduct the parameter study for the perturbation threshold ϵ on 1000 images random selected from ImageNet dataset on VGG-19. Refer to table 4 for the experimental results.

M	Ins. ↑	Del. ↓	Diff. ↑	time
1	57.3797	17.9267	39.4530	5m40s
2	57.3983	17.9264	39.4719	7m40s
3	57.4565	17.9264	39.5301	9m27s
4	**57.5785**	17.9261	39.6524	13m5s
5	57.5783	**17.8659**	**39.7124**	17m23s
6	57.5781	17.9260	39.6521	23m17s
7	57.5782	17.9256	39.6526	30m32s
8	57.5754	17.9542	39.6212	38m56s

Table 2: Insertion-Deletion tests results and running time for M.

γ	Ins. ↑	Del. ↓	Diff. ↑
5	53.380	19.933	33.447
10	55.378	18.926	36.452
15	56.382	18.727	37.655
20	**58.378**	**17.730**	**40.648**
25	57.875	18.233	39.642
30	57.525	18.673	38.852
35	57.385	18.373	39.012
40	57.435	18.783	38.652

Table 3: Insertion-Deletion tests results for γ.

ϵ	Ins. ↑	Del. ↓	Diff. ↑
0.5	53.380	19.933	33.447
1.0	55.378	18.926	36.452
2.0	56.382	18.727	37.655
4.0	57.875	18.233	39.642
8.0	57.875	18.233	39.642

Table 4: Insertion-Deletion results for perturbation threshold ϵ.

2. Energy-based Pointing Game

We conduct experiment on COCO dataset and follow the settings in Score-CAM. Results reported in table 5 shows our method CCE shows the best result.

Method	GBP	IG	RISE	GCAM+	SCAM	FullG	CE	APPB	CALM	CCE
Proportion	55.2	51.8	53.2	57.5	59.9	59.3	55.3	53.6	58.7	**61.1**

Table 5: Average computation cost on ResNet50 and VGG19.

3. Saliency maps for multiple objects

When the image contains multiple subjects of the same class, our proposed CCE can recognize these subjects at the same time, as shown in Fig. 1. When the image contains multiple subjects of different classes, our proposed CCE still identify these subjects separately at the same time, as shown in the Fig. 2.

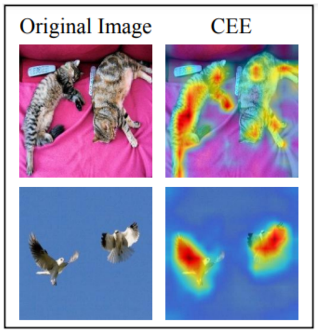


Figure 1: CCE results for images contains multiple objects of the same class.

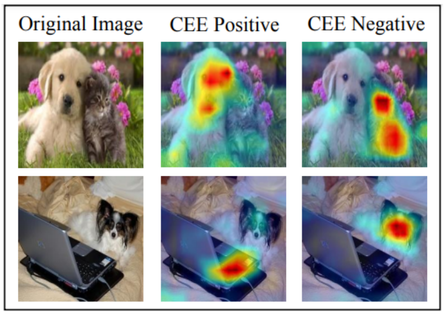
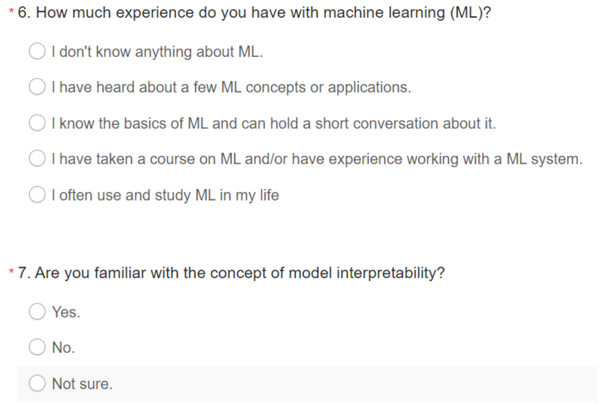


Figure 2: CCE results for images contains multiple objects of different classes.

4. Details for User Study

Before the formal user study, we first investigated the participants' basic information and their understanding of machine learning and model interpretation. The demographic distribution was: man 54%, woman 46%, non-binary 1%, no gender reported 1%.The self-reported machine learning experience was 2.7 ± 1.0, between "2: have heard about..." and "3: know the basics...". The specific content of user study is shown in the figures 3-5.



* 6. How much experience do you have with machine learning (ML)?

○ I don't know anything about ML.

○ I have heard about a few ML concepts or applications.

○ I know the basics of ML and can hold a short conversation about it.

○ I have taken a course on ML and/or have experience working with a ML system.

○ I often use and study ML in my life

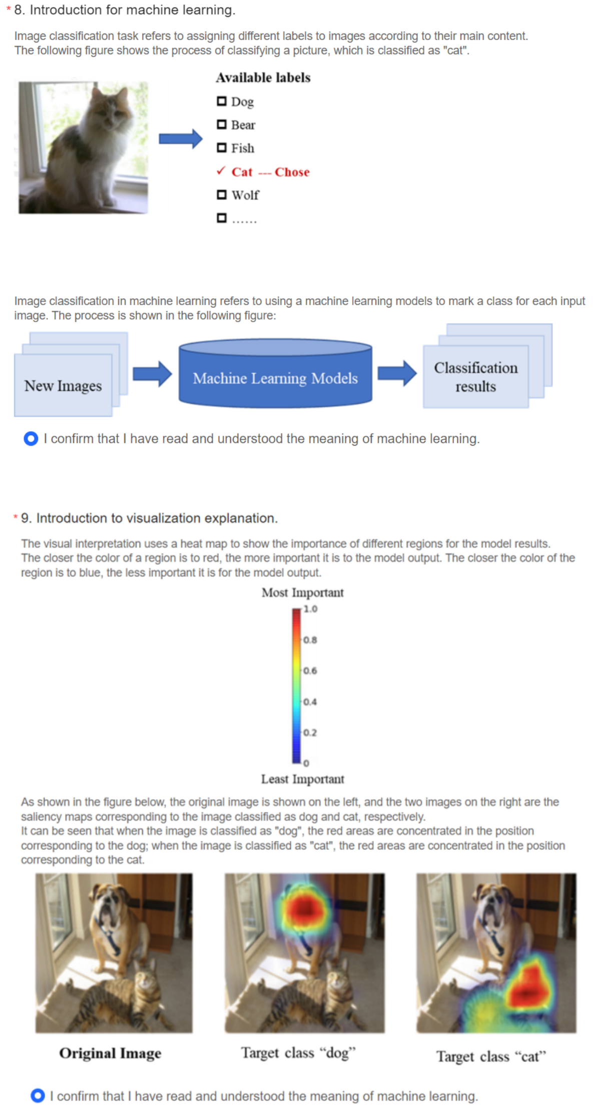
* 7. Are you familiar with the concept of model interpretability?

○ Yes.

○ No.

○ Not sure.

Figure 3: Demographics and background.



* 8. Introduction for machine learning.

Image classification task refers to assigning different labels to images according to their main content.
The following figure shows the process of classifying a picture, which is classified as "cat".

Image classification in machine learning refers to using a machine learning models to mark a class for each input image. The process is shown in the following figure:

● I confirm that I have read and understood the meaning of machine learning.

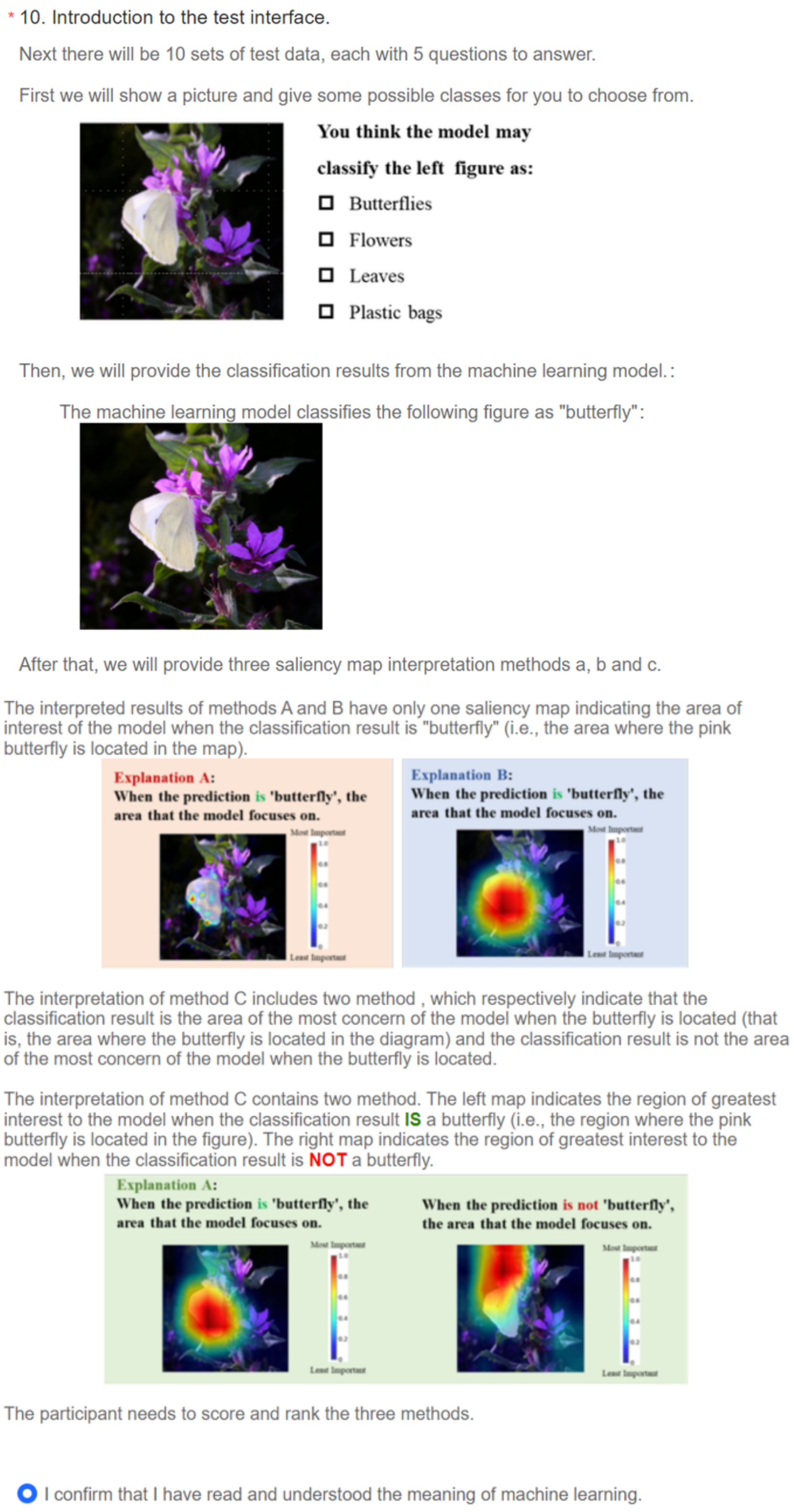
* 9. Introduction to visualization explanation.

The visual interpretation uses a heat map to show the importance of different regions for the model results.
The closer the color of a region is to red, the more important it is to the model output. The closer the color of the region is to blue, the less important it is for the model output.

As shown in the figure below, the original image is shown on the left, and the two images on the right are the saliency maps corresponding to the image classified as dog and cat, respectively.
It can be seen that when the image is classified as "dog", the red areas are concentrated in the position corresponding to the dog; when the image is classified as "cat", the red areas are concentrated in the position corresponding to the cat.

● I confirm that I have read and understood the meaning of machine learning.

Figure 4: Introduction to machine learning and visualization explanation.



* 10. Introduction to the test interface.

Next there will be 10 sets of test data, each with 5 questions to answer.

First we will show a picture and give some possible classes for you to choose from.

You think the model may classify the left figure as:

□ Butterflies

□ Flowers

□ Leaves

□ Plastic bags

Then, we will provide the classification results from the machine learning model.:

The machine learning model classifies the following figure as "butterfly":

After that, we will provide three saliency map interpretation methods a, b and c.

The interpreted results of methods A and B have only one saliency map indicating the area of interest of the model when the classification result is "butterfly" (i.e., the area where the pink butterfly is located in the map).

Explanation A:
When the prediction is 'butterfly', the area that the model focuses on.

Explanation B:
When the prediction is 'butterfly', the area that the model focuses on.

The interpretation of method C includes two method , which respectively indicate that the classification result is the area of the most concern of the model when the butterfly is located (that is, the area where the butterfly is located in the diagram) and the classification result is not the area of the most concern of the model when the butterfly is located.

The interpretation of method C contains two method. The left map indicates the region of greatest interest to the model when the classification result **IS** a butterfly (i.e., the region where the pink butterfly is located in the figure). The right map indicates the region of greatest interest to the model when the classification result is **NOT** a butterfly.

Explanation A:
When the prediction is 'butterfly', the area that the model focuses on.

When the prediction is not 'butterfly', the area that the model focuses on.

The participant needs to score and rank the three methods.

● I confirm that I have read and understood the meaning of machine learning.

Figure 5: Introduction to the test interface.