

Interpretability Evaluation Test

Dear Participant,

Thank you for participating in this study. In this experiment, you will observe how an AI model makes predictions and assess its decision-making process. The test contains three sub-tests for three different datasets. For each sub-test, you will face a learning phase for each session, and also an evaluation phase for the whole sub-test.

- ***Learning Phase:*** You will see examples with AI predictions and, in some cases, explanation heatmaps as well.
- ***Evaluation Phase:*** You will assess whether the AI can correctly predict the label for new samples.

Please fill out the following questionnaire as you proceed.

Sub-test 1: ImageNette dataset

The AI is learned to make predictions for 10 classes (tench, English springer, cassette player, chain saw, church, French horn, garbage truck, gas pump, golf ball, parachute). For sub-test 1, you will be provided with only one class sample for each session. Please note that the AI can have different strategies in dealing with samples of different categories.

Session 1 How does the AI predicts golf balls?

Learning Phase

You will be shown images of Golf Ball along with the AI predictions.

- If applicable, an explanation heatmap will accompany each prediction. For an explanation heatmap, **the important (positive) areas are for making the decision are colored with red, while the unimportant (negative) areas are colored with blue.**
- Focus on understanding how the AI is **making decisions**.

1. Sample 1:

AI's prediction: Correct.

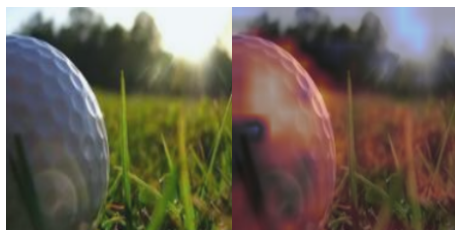


Image

AI's Explanation

2. Sample 2:

AI's prediction: Correct.



Image

AI's Explanation

3. Sample 3:

AI's prediction: Wrong.



Image

AI's Explanation

4. Sample 4:

AI's prediction: Wrong.



Image

AI's Explanation

5. Sample 5:

AI's prediction: Correct.



Image

AI's Explanation

Session 2 How does the AI predicts Churches?

You will be shown images of Churches along with the AI predictions.

- If applicable, an explanation heatmap will accompany each prediction. For an explanation heatmap, **the important (positive) areas are for making the decision are colored with red**, while the unimportant (negative) areas are colored with blue.
- Focus on understanding how the AI is **making decisions**.

Learning Phase

1. Sample 1:

AI's prediction: Correct.



Image

AI's Explanation

2. Sample 2:

AI's prediction: Correct.

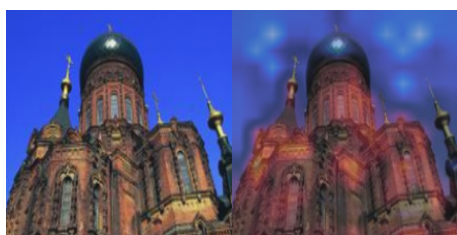


Image

AI's Explanation

3. Sample 3:

AI's prediction: Correct.



4. Sample 4:

AI's prediction: Wong.



Image

AI's Explanation

5. Sample 5:

AI's prediction: Wong.




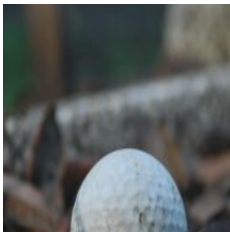




Image

AI's Explanation

Evaluation Phase

Based on what you observed in the learning phase, decide if the model can correctly predict the label for each test sample.

| Test Sample | Will the AI Predict Correctly? (Yes/No) | Test Sample | Will the AI Predict Correctly? (Yes/No) |
|---|---|--|---|
|  | |  | |
|  | |  | |
|  | |  | |
|  | |  | |
|  | |  | |

Sub-test 2: Pets dataset

The AI is learned to make predictions for 10 classes 37 categories of pets. For sub-test 2, you will be provided with only one class sample for each session. Please note that the AI can have different strategies in dealing with samples of different categories.

Session 1 How does the AI predicts Samoyeds?

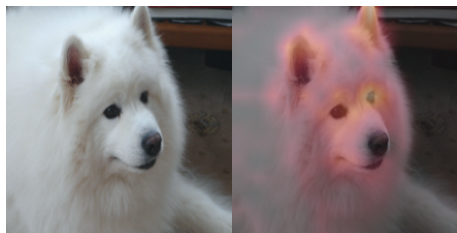
Learning Phase

You will be shown images of Samoyed along with the AI predictions.

- If applicable, an explanation heatmap will accompany each prediction. For an explanation heatmap, **the important (positive) areas are for making the decision are colored with red, while the unimportant (negative) areas are colored with blue.**
- Focus on understanding how the AI is **making decisions**.

6. Sample 1:

AI's prediction: Correct.



Image

AI's Explanation

7. Sample 2:

AI's prediction: Correct.



Image

AI's Explanation

8. Sample 3:

AI's prediction: Correct.



Image

AI's Explanation

9. Sample 4:

AI's prediction: Wrong.



Image

AI's Explanation

10. Sample 5:

AI's prediction: Correct.



Image

AI's Explanation

Session 2 How does the AI predicts Russian Blue?

You will be shown images of Russian Blue along with the AI predictions.

- If applicable, an explanation heatmap will accompany each prediction. For an explanation heatmap, the important (positive) areas are for making the decision are colored with red, while the unimportant (negative) areas are colored with blue.
- Focus on understanding how the AI is making decisions.

Learning Phase

6. Sample 1:

AI's prediction: Correct.



Image

AI's Explanation

7. Sample 2:

AI's prediction: Correct.



Image

AI's Explanation

8. Sample 3:

AI's prediction: Wong.

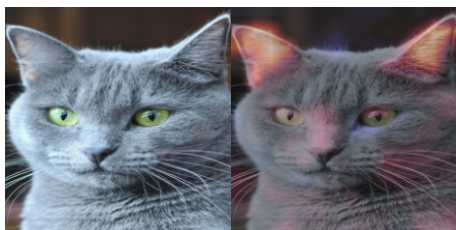


Image AI's Explanation

9. Sample 4:

AI's prediction: Correct.



Image AI's Explanation

10. Sample 5:

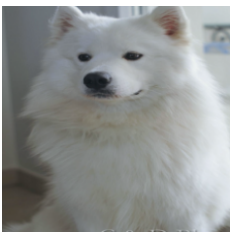
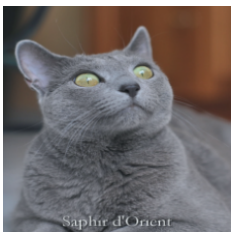
AI's prediction: Wong.





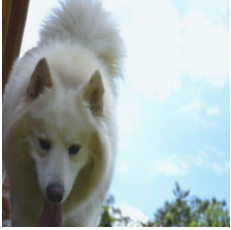

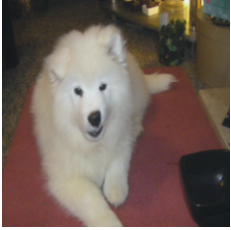
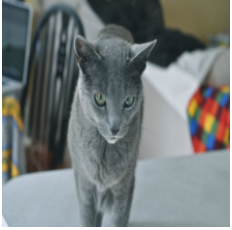


Image AI's Explanation

Evaluation Phase

Based on what you observed in the learning phase, decide if the model can correctly predict the label for each test sample.

| Test Sample | Will the AI Predict Correctly? (Yes/No) | Test Sample | Will the AI Predict Correctly? (Yes/No) |
|---|---|--|---|
|  | |  | |

| | | | |
|--|--|---|--|
|  | |  | |
|  | |  | |
|  | |  | |
|  | |  | |

Sub-test 3: MURA dataset

The AI is learned to decide whether an X-ray study is abnormal or not. For sub-test 3 you will only need to finish a single session task. *You are required to be familiar with X-ray study knowledge.*

Session 1 How does the AI predicts medical images?

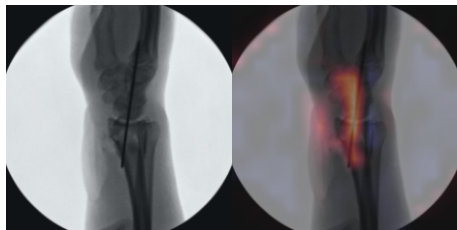
Learning Phase

You will be shown images of medical images along with the AI predictions.

- If applicable, an explanation heatmap will accompany each prediction. For an explanation heatmap, **the important (abnormal) areas are for making the decision are colored with red, while the unimportant areas are colored with blue.**
- Focus on understanding how the AI is **making decisions.**

1. Sample 1:

AI's prediction: Correct.

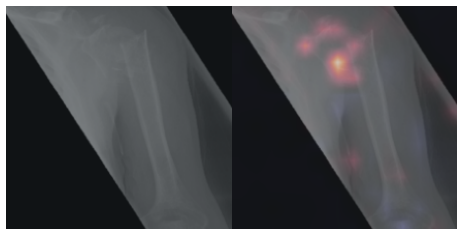


Image

AI's Explanation

2. Sample 2:

AI's prediction: Correct.

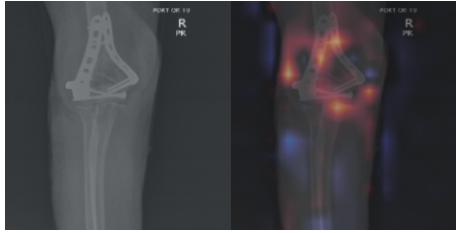


Image

AI's Explanation

3. Sample 3:

AI's prediction: Correct.



Image

AI's Explanation

4. Sample 4:

AI's prediction: Correct.

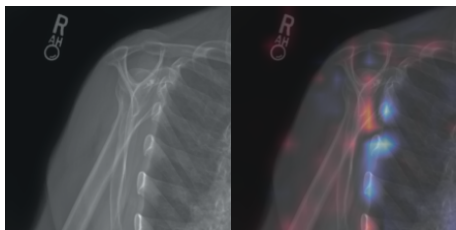


Image

AI's Explanation

5. Sample 5:

AI's prediction: Wrong.










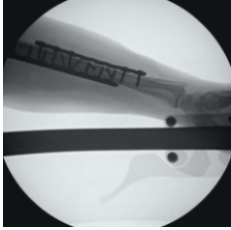


Image

AI's Explanation

Evaluation Phase

Based on what you observed in the learning phase, decide if the model can correctly predict the label for each test sample. You are also provided with the ground truth label of these images.

| Test Sample | Will the AI Predict Correctly? (Yes/No) | Test Sample | Will the AI Predict Correctly? (Yes/No) |
|---|---|--|---|
|  | |  | |
|  | |  | |
|  | |  | |
|  | |  | |
|  | |  | |

Thank you for completing the questionnaire. Your responses will help us better understand how explanation methods influence human evaluations of AI models!!!