IBM Developer
SKILLS NETWORK

Hands-On Lab: Computer Vision (20 Mins)

Objective for Exercise:

- Learn about IBM's Adversarial Robustness Toolbox, and use it to mitigate simulated attacks by hackers..

Computer Vision

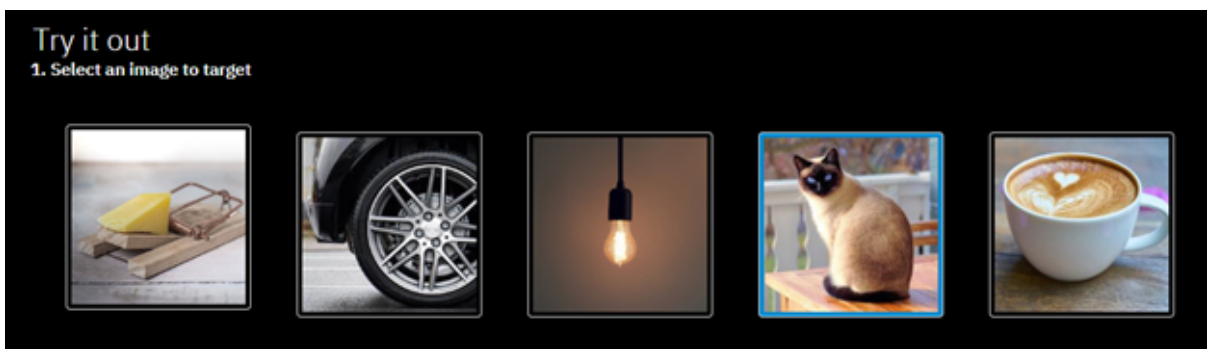
IBM Research creates innovative tools and resources to help unleash the power of AI.

Follow these steps to explore the demo:

1. Access the demo here:Your AI model might be telling you this is not a cat.

2. In the **Try it out** section, click the image of the Siamese cat.

Figure 1 - Select an image

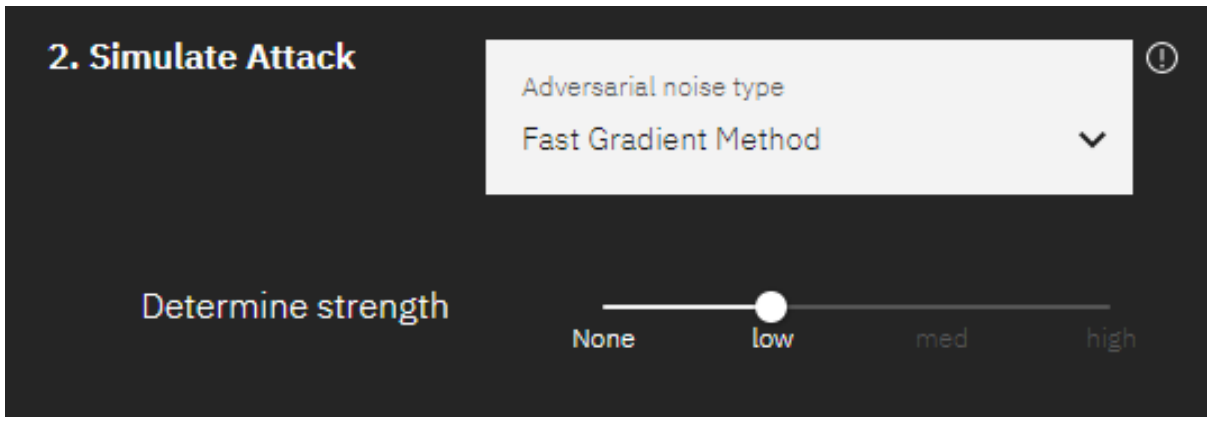


3. In the **Simulate Attack** section, ensure that no attack is selected, and that all the sliders are to the far left, indicating that all attacks and mitigation strategies are turned off.

What does Watson identify the image as, and at what confidence level? E.g. Siamese cat 92%

4. In the **Simulate Attack** section, under **Adversarial noise type**, select **Fast Gradient Method**. The strength slider will move to low.

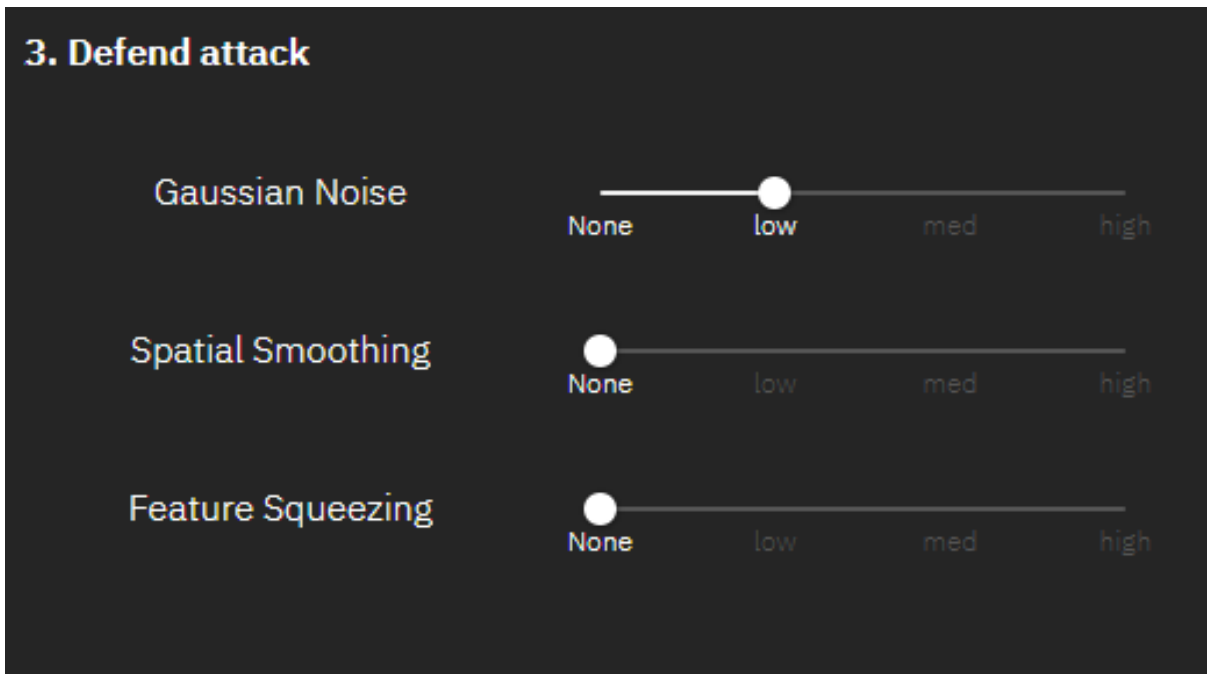
Figure 2 - Select an attack and level



What does Watson identify the image as now, and at what confidence level?

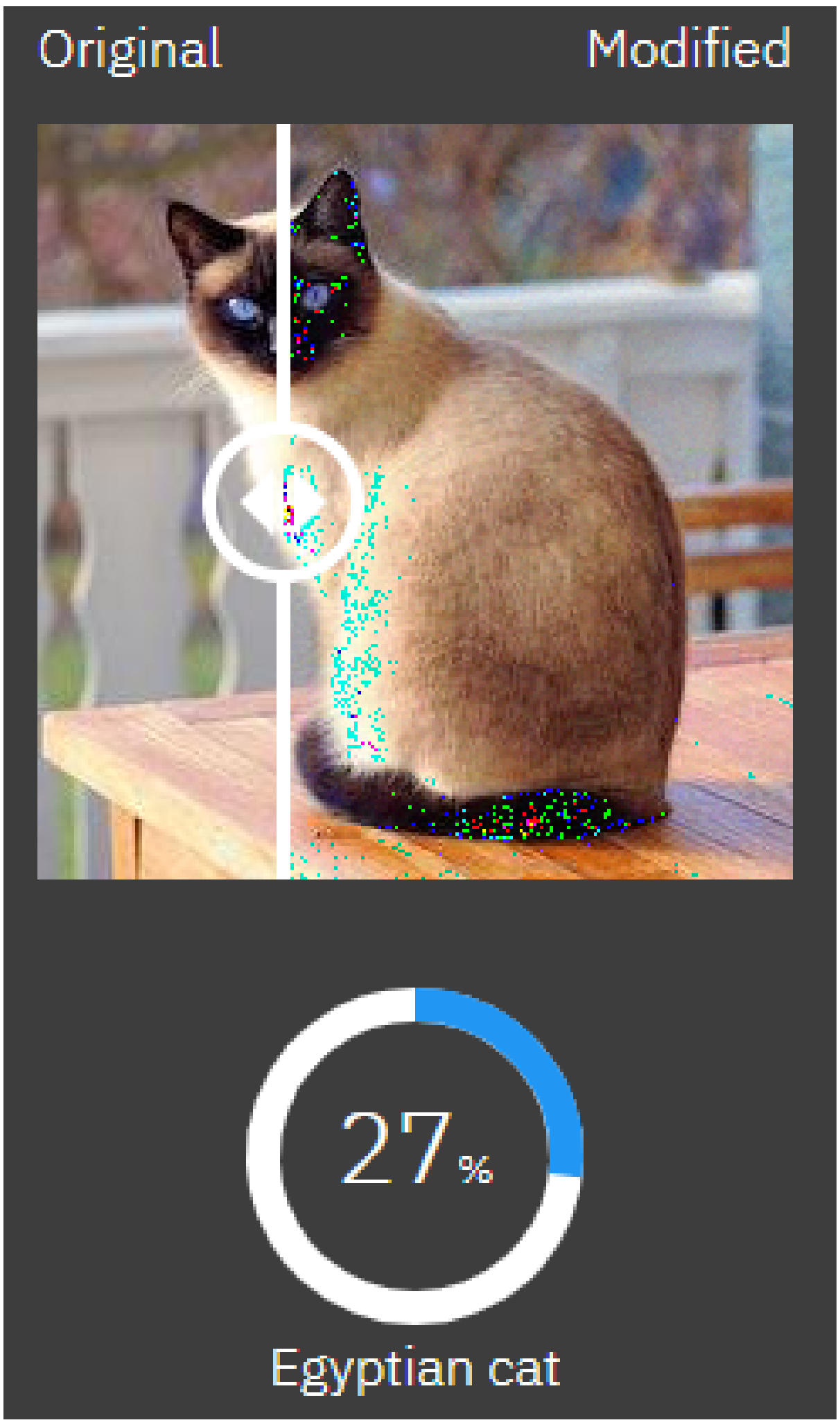
5. In the **Defend attack** section, move the **Gaussian Noise** slider to low.

Figure 3 - Mitigate the attack



6. What does Watson identify the image as now, and at what confidence level? Did the image recognition improve?

Figure 4 - View the results



Note that you can use the slider on the image to see the original and modified images.

7. Move the **Gaussian Noise** slider to **medium**, and then to **high**. For each level, note what Watson identifies the image as, and at what confidence level. Did the image recognition improve?

8. Move the **Gaussian Noise** slider to **None**.

9. In the **Defend attack** section, move the **Spatial Smoothing** slider to **low**. What does Watson identify the image as now, and at what confidence level? Did the image recognition improve?

10. Move the **Spatial Smoothing** slider to **medium**, and then to **high**. For each level, note what Watson identifies the image as, and at what confidence level. Did the image recognition improve?

11. Move the **Spatial Smoothing** slider to **None**.

12. In the **Defend attack** section, move the **Feature Squeezing** slider to **low**. What does Watson identify the image as now, and at what confidence level? Did the image recognition improve?

13. Move the **Feature Squeezing** slider to **medium**, and then to **high**. For each level, note what Watson identifies the image as, and at what confidence level. Did the image recognition improve?

14. Which of the three defenses would you use to defend against a Fast Gradient Attack?

Optional:

If you have time, use the same techniques to explore the other methods of attack (Projected Gradient Descent and C&W Attack) and evaluate which method of defense works best for each. If you want, try a different image.

Use the Discussion Forum to talk about the attacks and mitigation strategies with your fellow students.

Author(s)

Rav Ahuja

Changelog

Date	Version	Changed by	Change Description
2020-08-27	2.0	Anamika	Migrated Lab to Markdown and added to course repo in GitLab