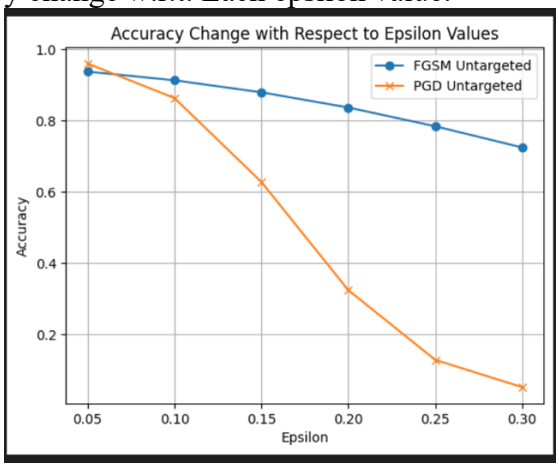
a. *Plot and analyze the accuracy change with respect to each of the epsilon values. How do the images change for each of these values? Will a higher epsilon value mean a good attack? If not, why?*

Ans) Chart for the Accuracy change w.r.t. Each epsilon value:



Firstly, the epsilon value is a measure of the amount of noise that is added to the input image to create an adversarial example. The graph shows that both FGSM and PGD can reduce the accuracy of the model as the epsilon value increases and the accuracy of the model is more sensitive to changes in epsilon value at lower epsilon values. PGD is more effective than FGSM at reducing accuracy.
For example, when the value of epsilon is set to 0.1, PGD can decrease the accuracy of the model by around 20%, whereas FGSM can only reduce it by about 10%. Additionally, the model's accuracy drops more significantly when epsilon values are between 0.05 and 0.1 than when they are between 0.2 and 0.25. The success of the attack depends on the attack method used and the value of epsilon chosen.

How images change with epsilon: The epsilon value in adversarial attacks represents the maximum allowed **Threshold** of the perturbation added to each pixel in the image. As epsilon increases, the changes to the image become **larger and more noticeable**.

Low Epsilon	Medium epsilon	High Epsilon
Small changes are made to individual pixels, often imperceptible to the human eye. Involves slight shifts in color/brightness.	Larger changes become visible, like adding noise patterns or altering specific features slightly. The image might still be recognizable but appear distorted.	Significant modifications occur, resulting in drastic changes to image content. Objects might be added/removed, colors completely shifted, or image become unrecognizable.

Does a higher epsilon mean a good attack?

While a higher epsilon allows for more significant modifications, it is not directly indicative of a good attack because of these reasons.

- Highly distorted images, with a high epsilon, can be easily detected by humans, which raises concerns about their real-world applicability. Ideally, an attack should be imperceptible to maintain deception.
- Attacks with large perturbations may not generalize to unseen examples or different models. Using smaller epsilon values can create more transferable adversarial examples.
- Not all perturbations with high epsilon guarantee successful misclassification. The optimal epsilon depends on the specific model, dataset, and attack method.

Thus, achieving an effective attack requires finding the right balance among factors like epsilon, imperceptibility, transferability, and success rate.

a. Which attack is the best? List down the pros and cons of each attack. When would you use a targeted attack vs an untargeted attack?

Ans) The choice of the best attack depends on various factors, including the specific task requirements, the robustness of the target model, and the available computational resources.
For Robustness: PGD is generally preferred due to its iterative nature, which improves robustness against defense mechanisms.
For Speed: FGSM may be chosen when computational efficiency is a priority, as it requires fewer iterations.
For Versatility: PGD offers more flexibility with both untargeted and targeted attacks, making it suitable for a wider range of scenarios.

Fast Gradient Sign Method (FGSM):

Pros:

- **Simplicity:** FGSM is straightforward to implement and understand.
  Fast: It requires only a single forward and backward pass through the network, making it computationally efficient.
- **Effective:** FGSM can achieve significant perturbation of input images, leading to misclassification.

Cons:

- **Brittleness:** FGSM perturbations are often easily detectable and can be fragile to small changes, limiting their robustness.
- **Untargeted Only:** FGSM generates perturbations to maximize the loss without considering a specific target class, making it suitable only for untargeted attacks.
- **Limited ε Sensitivity:** FGSM might not perform well with large ε values, as it may result in overly perturbed images with reduced visual similarity to the original.

RSAI - Hrithik Sagar - 2023900021

Projected Gradient Descent (PGD)

Pros:

- **Robustness:** PGD iteratively applies FGSM steps with smaller step sizes, making it more robust against defense mechanisms like gradient masking or gradient obfuscation.
- **Flexibility:** PGD allows for both untargeted and targeted attacks, offering more versatility in attack strategies.
- **Customizable:** Parameters like step size, number of steps, and ε can be adjusted to tailor the attack to specific scenarios.

Cons:

- **Computational Complexity:** PGD requires multiple iterations of FGSM, leading to increased computational overhead compared to FGSM.
- **Gradient Estimation:** PGD relies on gradient estimation, which may not always accurately represent the true gradient direction, potentially affecting the effectiveness of the attack.
- **Sensitivity to Parameters:** The performance of PGD can be sensitive to parameter choices, requiring careful tuning for optimal results.

When to Use Targeted vs. Untargeted Attacks:

- Untargeted attacks are used to cause misclassification without targeting a specific class. Examples of untargeted attacks include FGSM and PGD variants.
- Targeted attacks are used by attackers to manipulate the prediction of a model to a specific chosen class. These attacks can be more difficult to detect and mitigate effectively for the defender. FGSM and PGD-targeted variants are examples of techniques that support this strategy.

Hence, the choice between FGSM and PGD depends on the specific requirements of the attack scenario, including the desired trade-off between computational efficiency, robustness, and attack versatility.