Re-evaluating PGD and FGSM Attacks & Extending Attacks to Visual Explanation Robustness

Chaewon Park
Computer Science
Columbia University
cp3227@columbia.edu

Jinwoo Choi
Computer Science
Columbia University
jc5669@columbia.edu

Chungil Lee
Computer Science
Columbia University
cl4159@columbia.edu

Abstract

The aim of this project is to implement two adversarial attacks on CIFAR-10 and re-evaluate each attack performance based on sensitivity analysis and per-class accuracy decomposition. In addition, we extend attacks to visual explanation and empirically show that visual explanation is robust to adversarial noise.

1 Introduction

Deep neural networks achieve high accuracy in many computer vision tasks, but a large body of work has demonstrated the threat of adversarial attacks. In this project, we implement FGSM and PGD and re-evaluate their attack success rate on CIFAR-10 by tuning hyper-parameters such as perturbation size, alpha(step size), and number of steps and observing the attack impact on each individual class. In addition, we extract attention maps on adversarial images and assess their robustness by measuring attention map deviation through dice loss.

2 Related Works

Adversarial attack is a threat imposed on neural networks to cause mispredictions, and images that trick the model are called 'adversarial examples'. These examples look just like normal images but have indistinguishable perturbations. They are constructed by assessing sensitivity (gradient) of the loss function in terms of the image and perturbing the pixels which are most likely to maximize the loss function. FGSM and PGD are two of the most well-known adversarial attack methods.

FGSM assumes that the subspace in which adversarial examples exist are continuous and by adding perturbations to the image in the direction of the gradient's sign, we can easily obtain numerous adversarial examples. This is summed up by the following expression:

$$adv_x = x + \epsilon \cdot sign(\nabla_x \mathscr{L}(\theta, x, y))$$

where ϵ is the bound for the perturbation size, $\mathscr{L}$ is the loss function with model parameters θ, input image x, and ground truth label y.

PGD is a method proposed by Madry which overcomes the downsides of FGSM. FGSM may not work if loss function landscape is noisy and moving a step in the direction of the steepest gradient may not lead to the best adversarial example. Hence, instead of taking 'one' step in the gradient direction, PGD takes 'multiple' steps to strategically find the highest loss point:

$$x^{t+1} = \prod_{x+S} (x^t + \alpha \; sgn(\nabla_x \mathscr{L}(\theta, x^t, y)))$$

Explainable AI is a budding field which attempts to make the decision processes of deep neural networks to be more interpretable and intuitive. It is recognized as a crucial research initiative for domains where humans require a fundamental and clearer understanding of the models' performance, such as autonomous vehicles, medicine, security, and legal domains. Traditional deep neural networks were considered 'not explainable' because the models consist of so many layers, weights, biases, and non-linearities. Therefore, models only provided the final predictions and did not answer questions like 'Why not another class?' or 'What is the cause of the prediction failure and how do we correct them?'. In this project, we focus on visual explanations that help humans understand which part of the images highly influenced the outcome of the model. Among various methods, we chose GradCAM (Gradient Descent Class Activation Mapping) . GradCAM uses gradients to calculate how important each feature map is for a target class and uses it to produce the attention map which is a weighted sum of the activation maps.

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad \text{and} \quad L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right)$$

3 Methods

In our third experiment, we evaluate how much the attention maps get affected by the perturbations in adversarial examples. If the model's adversarial robustness is high, even when given adversarial examples, the new attention maps will not deviate much and have high overlap with the original attention map and align well with the ground truth label. Usually, assessing attention map quality is very challenging as it is a visual output that cannot be quantified. As a solution, we use dice loss as a measure of overlap between two attention maps. Dice loss calculates intersection over union between two sets, and is commonly used as a loss for semantic segmentation. After generating attention maps, we threshold pixel values to 1 or 0 based on a target value of 150 and 200. Binarizing attention map makes it much easier to calculate the dice loss. The range of dice loss is from 0 to 1 and the higher the overlap, the higher the loss value.

$$L_{dice} = \frac{2 * \sum P_{true} * P_{pred}}{\sum p_{true}^2 + \sum p_{pred}^2 + \epsilon}$$

4 Experiments

In this project, we perform three experiments. The first experiment is reevaluating the attack success rate of FGSM and PGD on CIFAR10. We fine-tune hyperparameters such as perturbation size, alpha (step size), and number of steps to observe how they impact the robustness accuracy and perform sensitivity analysis. The second experiment extracts per-class robustness accuracy and observes how each attack method influences each of the ten classes differently. The final experiment is a novel extension of the two attack methods to a new application, called Explainable AI, in order to evaluate the Visual Explanation Robustness of CNN models. Conventional attacks focused on producing adversarial examples that cause classifier's mispredictions. For the Explainable AI application, we test how adversarial perturbations can impact the robustness and quality of visual explanation using GradCAM and list interesting findings. We use dice loss algorithm from Section 3 to quantitatively evaluate the attention map deviation.

4.1 Sensitivity Analysis

4.1.1 PGD Sensitivity Analysis

Figure 1 left shows a linear relationship between the model's robust accuracy and ϵ (perturbation size) values. The accuracy steadily declines from 91.38% to 14.70% as the value of ϵ increases from 0.1 to 2.0. Robust accuracy versus alpha graph shows a more complicated relationship. The robust accuracy drops substantially and reaches the lowest point of the graph when alpha = 1/255. Until

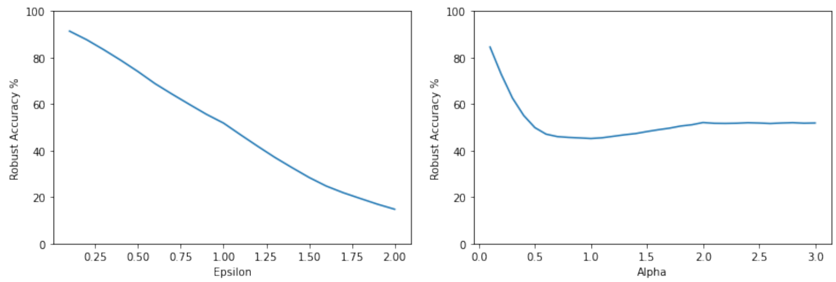

Figure 1: PGD Sensitivity Graph

alpha reaches 2/255, the accuracy slightly increases by 6% and starts to converge to 51~52% in the end.

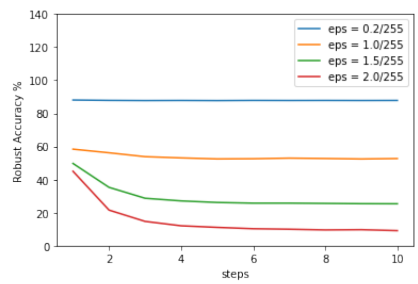

Figure 2: PGD Sensitivity with different ϵ values

Figure 2 shows models' robust accuracy on the y-axis, against steps 1 to 10 in increments of 1 on the x-axis. Four lines are shown with different ϵ values. The value of robust accuracy with an ϵ = 2/255 tends to decrease as steps increase, with a notably large drop from 45.02% at step 1 to 14.73% at step 3. Robust accuracy with higher ϵ values tends to maintain the same level regardless of number of steps. Especially when the value of ϵ is smaller than 0.2, no significant differences were observed during the process.

4.1.2 FGSM Sensitivity Analysis

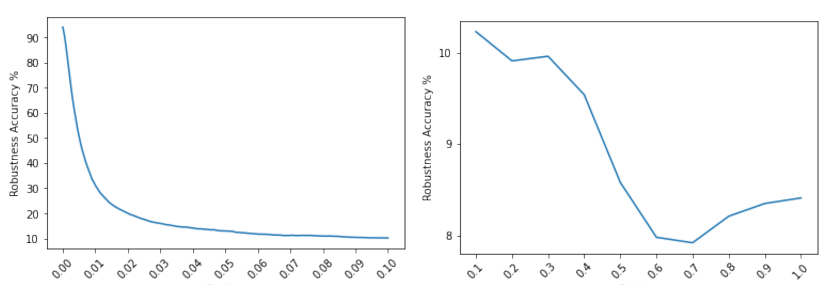

Figure 3: FGSM Sensitivity Graph

As Figure 3 shows, the model's robustness accuracy generally decreases with higher ϵ. FGSM's performance is most sensitive when ϵ values are lower. The robustness accuracy decreases sharply with slight changes in ϵ from 0 to 0.01, from 94.02% for the clean model to 31.17% with $\epsilon = 0.01$. The drop rate in the robustness accuracy decreases as ϵ value increases, with 10.23% at $\epsilon = 0.1$. Interestingly, the robust accuracy increases from $\epsilon = 0.8$, albeit by a small amount.

4.2 Per-Class Accuracy Decomposition

Class	Weak	Strong	Count
1	0.772	0.069	1000
2	0.878	0.282	1000
3	0.686	0.103	1000
4	0.553	0.032	1000
5	0.691	0.039	1000
6	0.619	0.046	1000
7	0.794	0.066	1000
8	0.812	0.119	1000
9	0.849	0.2	1000
10	0.847	0.23	1000

Class	eps = 0.002	eps = 0.35	Count
1	0.78	0	1000
2	0.884	0.001	1000
3	0.703	0.977	1000
4	0.574	0	1000
5	0.727	0	1000
6	0.624	0	1000
7	0.806	0	1000
8	0.824	0	1000
9	0.845	0.001	1000
10	0.854	0.003	1000

Table 1: Per-Class Accuracy Analysis. (Left: PGD, Right: FGSM). For PGD, the experiment was performed following the settings specified in Table 2 for weak and strong attacks

Strong PGD attack affects the accuracy of each class similarly as weak PGD attack does. Per-class accuracy decreases from weak PGD attack to strong PGD attack at a somewhat equivalent rate across each class. On the contrary, strong FGSM attack seems to affect all classes except for the third class (bird class), where the accuracy actually increased from weak FGSM attack to strong FGSM attack by almost 100% accuracy.

4.3 Visual Explanation Robustness

<table>
<tr><th>Attack Method</th><th>Attack Duration</th><th>Hyperparameter</th><th>Robust Accuracy</th><th>Avg Dice Loss</th><th>Attention Map Threshold</th></tr>
<tr><td rowspan="2">PGD (Strong)</td><td rowspan="2">2min 6s</td><td rowspan="2">eps=2/255, alpha=2/255, steps=4</td><td rowspan="2">11.79 %</td><td>0.688</td><td>150</td></tr>
<tr><td>0.593</td><td>200</td></tr>
<tr><td rowspan="2">PGD (Weak)</td><td rowspan="2">1min 8s</td><td rowspan="2">eps=0.5/255, alpha=2/255, steps=2</td><td rowspan="2">75.00 %</td><td>0.861</td><td>150</td></tr>
<tr><td>0.819</td><td>200</td></tr>
<tr><td rowspan="2">FGSM (Strong)</td><td rowspan="2">39.8 s</td><td rowspan="2">eps = 0.35</td><td rowspan="2">10.23 %</td><td>0.683</td><td>150</td></tr>
<tr><td>0.583</td><td>200</td></tr>
<tr><td rowspan="2">FGSM (Weak)</td><td rowspan="2">39.9 s</td><td rowspan="2">eps = 0.002</td><td rowspan="2">75.80 %</td><td>0.864</td><td>150</td></tr>
<tr><td>0.822</td><td>200</td></tr>
</table>

Table 2: Visual Explanation Robustness. *Original clean model accuracy = 94.02%

We adjusted the perturbation size so that strong attack methods and weak attack methods have similar accuracies, which helps set ground for cross-attack comparison. Overall, unlike our initial expectation that adversarial examples would lead to low quality visual explanation, we made an interesting discovery that visual explanation is quite robust against adversarial noise. For both types of attacks and thresholds, the average dice loss exceeds 0.58 and when visually checked, the attention maps themselves are quite intact and do not deviate much from the original attention map, as seen in Figure 4 and 5. For each figure, the first column is the original images, the second column is the original attention map, the third column is the result from strong attack, and the fourth column is the result from weak attack.

Overall, PGD and FGSM have similar dice loss when they have similar accuracy. However, PGD attack does multiple iterations and thus takes 2~3 times much longer than FGSM, so time-wise FGSM is much efficient. However, strong perturbations in FGSM are more noticeable than PGD which would be a big downside as an attack method.

Visual robustness is valuable because it can be effectively leveraged in identifying bias and provides meaningful explanations on model prediction even on adversarial examples. Therefore, even if the

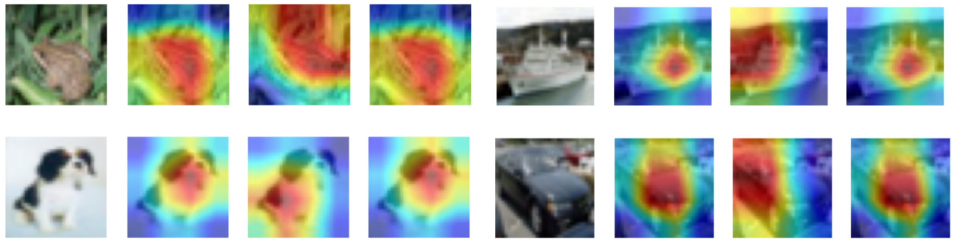

Figure 4: PGD Attention Maps

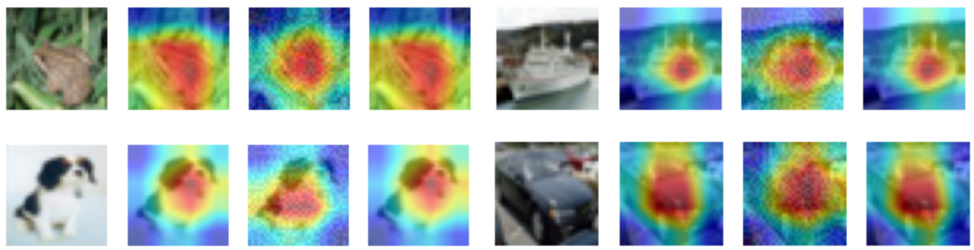

Figure 5: FGSM Attention Maps

model mispredicts due to adversarial perturbations, human users can rely on visual explanations to determine or deduce the true label.

5 Conclusion

In this project, we performed three experiments to re-evaluate PGD and FGSM attacks. In both attacks, the most sensitive parameter affecting the robust accuracy is ϵ. PGD attack has a linear relationship between the robustness accuracy and ϵ, while the performance of FGSM attack is more sensitive in smaller ϵs. Per-class accuracy for FGSM as ϵ increases provides us with some concerns regarding the method as the accuracy of one specific class increases from a lower ϵ to higher ϵ, while the accuracy of all other classes goes to near 0. Per-class accuracy for PGD from weak attack to strong attack decreases across all classes at a similar rate. For both attacks, the average dice loss and visual analysis of attention maps indicate that the visual explanation is robust. However, there exists a trade-off in using the two attacks as PGD takes much longer time to train and has more hyperparameters to tune, while FGSM is more simple, with only ϵ being the hyperparameter, but strong perturbations in FGSM are more apparent.

References

[1] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *In Proceedings of the IEEE international conference on computer vision (pp. 618-626).*

[2] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572.*

[3] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083.*