CS5260 Assignment 2 **Name: Tey Shiwei** **Matric Number: A0112101M**

Reference: https://adversarial-robustness-toolbox.readthedocs.io/en/latest/modules/attacks/evasion.html

1. Objective

To fool the model with minimum perturbation from the original image, using L2 distance

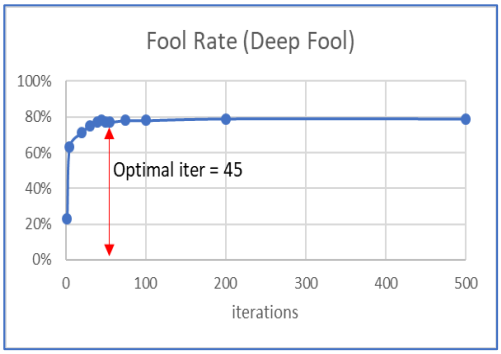
2. Adversarial Attack

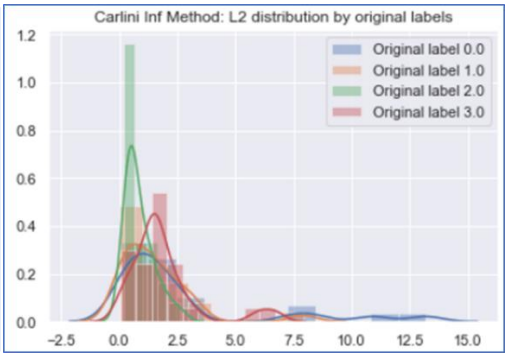
Attacking method obtained from "Adversarial Robustness Toolbox" each of them will be evaluated by the adversarial fool rate and L2 norm per image. Untargeted attack was used for every method.

Top 5 attacking method with high fool rate:

Attacking Method	Fool rate (FR)	Average L2 norm/image	Time taken/100 images
FGSM/BIM	77~90%(high eps higher FR)	2~20 (high eps higher L2)	0.5 mins
Deep Fool	78%	0.79	8 mins
Newton Fool	92%	1.2	20 mins (Slow)
Projected Grad Descent	100%	9.0	13 mins (Slow)
Carlini LInf Method	100%	1.2	1 min

- FGSM is a very fast method however the fool rate and L2 norm are highly dependent on eps value. High eps = high fool rate but also higher L2 norm/image
- Deep Fool is a very good attacking method with low L2 norm. Optimal iterations for Deep Fool is 45, performing at 78% fool rate. (No. of iterations and eps don't affect L2 norm, hence any value will do)
- Comparing among the attacks with high fool rate >90%, Carlini L_{Inf} method performs the best (fast with 100% fool rate and low L2 norm). Most of my images are created by Carlini L_{Inf} method
- L2 norm distribution for Carlini L_{Inf} method lies around 0 to 2 per image, which is quite a good attacking method besides Deep Fool for this project





3. Conclusion

- For each individual image, the best adversarial image will be chosen from all the different attacking method, by tuning the hyper parameters (learning rate, epsilon and iterations)
- Most of the images are coming from Deep Fool and Carlini method
- L2 norm/image = 1.08
- By visual inspection, adversarial images look almost the same as the original images and this could potentially be a huge problem to the hospital when the model doesn't classify noisy image correctly