Geometric View of Fast Gradient Sign method

Kumar Selvakumaran

September 2024

What does FGSM do?

answer : it adds a "perturbation" to the input that increases loss.

breaking down the equations :

$$x_{adv} = x + \epsilon \; sign(\frac{\delta J(\theta, x, y)}{\delta x})$$

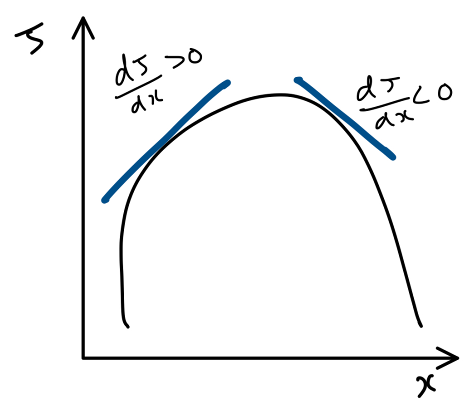
1. $J(\theta, x, y)$ is the loss calculated when the data from datapoint-label pair (x, y) is passed to the model, and θ are the parameters of the model.

2. $\frac{\delta J(\theta, x, y)}{\delta x}$ is the jacobian of the Loss w.r.t the input. it is the partial derivative of $J(\theta, x, y)$ w.r.t each element of the input data vector x (if images are the input, it is the partial derivative of the loss w.r.t each pixel in the image.)

3. $sign(\frac{\delta J(\theta, x, y)}{\delta x})$

   Why sign?

   for the sake of visualization, think of x as a scalar input, and lets plot the loss J w.r.t the input x.



you can see from the graph that incrementing x when $\frac{\delta J(\theta,x,y)}{\delta x} > 0$ and when decrementing x when $\frac{\delta J(\theta,x,y)}{\delta x} < 0$ will increase the loss.

The $sign()$ function returns +1 for elements > 0 and -1 for elements < 1 for a vector.

which brings everything together:

To the input x you add a perturbation $\epsilon\ sign(\frac{\delta J(\theta,x,y)}{\delta x})$ which is an ϵ scaled increment/decrement in each element in a direction that increases the loss individually.

Sure, such a perturbation makes the model less confident about the prediction, but is it optimal in terms of getting the model to misclassify?