# Explaining and Harnessing Adversarial Examples

Many machine learning models along with the state of neural networks are clearly a victim to adversarial examples. The main objective is to misclassify examples which are very close to the correct ones. One of the thing which happens in most of the cases is when the models are trained on different subsets which have different architectures respectively tend to misclassify the same adversarial example (Goodfellow et al., 2015). The main cause of the problem comes up due to non linearity of the deep neural networks which are taken together are models which are definitely not sufficient and are not regularized sufficiently of the supervised learning problem. A slight amount of linear behaviour is more than enough to cause adversarial examples. This usually enables to build a very high paced method of generating adversarial examples (Goodfellow et al., 2015). The adversarial perturbation is the one which is responsible for the process of activation, if we consider a dot product of a weight vector and a adversarial example then we can maximize the increase subject to the constraint. If the weight vector has numerous dimensions and consider the average magnitude of an element then the activation will grow gradually. Here we can also make infinitesimal changes to the procured input that adds up to one large change to the output. This can be related to the case of "accidental steganography," this happens when a linear model is taken by force to stay by the signal which aligns its weights (Goodfellow et al., 2015).

The main method considered is the "Fast Gradient Sign Method" for generating adversarial examples. This method helps in the misclassification of the input. Here, the shallow SoftMax classifier, error rate, average confidence are taken into consideration. In the same setting, the network regulates misclassification to a higher percentage for the adversarial examples with a higher rate of average confidence (Goodfellow et al., 2015). The generalization of adversarial examples can be a resultant factor of the perturbations which are highly aligned alongside the weight vectors of a particular model to perform the same task. Goodfellow et al.. (2015) stated that the direction in which perturbation flows other than the specific point which is taken would matter the most. The linear models usually cannot resist the adversarial perturbation, in this case we can say that only the particular structures with a hidden layer would be better if it is trained to withstand the adversarial perturbation (Goodfellow et al., 2015).

There are two types of adversarial attacks which can occur in the Fast Gradient Sign Method, they are White-box attacks and the Black-box attacks. The White box attacks tend to have the whole information of the model which is being targeted which includes its parameter values, the training method and in some of the cases the training data is also part of the information it holds. On the other hand, Black-box attacks give in a targeted model with the generated adversarial examples which are generated with no knowledge of the particular model (Liu et al., 2019). The Fast Gradient Sign Method is a type of a white-box attack where the attacker should know all the information about the pretrained model and the different model would generally have different parameters and loss function (Liu et al., 2019).

However, there are multiple aspects which affect the success rate of the mechanism of the Fast Gradient Sign Method. The first one would be the size considered for the perturbation, the size of the perturbation has a common scale of 0.01, 0.03, 0.05. The limitations of the Fast Gradient Sign method would be for the users to select the particular size by themselves. A different scale

will definitely affect the performance (Liu et al., 2019). The second one would be the Iteration for the perturbation generation, in the Fast Gradient Sign Method, fast means the gradient perturbation can be calculated at one time, in the equation were it all works y is taken as the targets which is being taken according to the original input. Therefore in the equational scenario x and y become the face images where x is considered as the image and y is considered as the victim image (Liu et al., 2019). Lastly, the factor of granularity of the perturbation is taken into consideration because in the Fast Gradient Sign Method, the size of the perturbation would be taken in as depending on how small each step is. Furthermore, the granularity can affect the performance easily (Liu et al., 2019).

The Advantages of the Fast Gradient Sign Method would be that there is more reliability and the pace is definitely higher. On the other hand, the disadvantage of this attack would be that there needs to be a very keen focus on the perturbations otherwise the attack would be unsuccessful (Muncsan & Kiss, 2020). The main reason why the quality of the adversarial examples is compared to the original ones would be to depict on the basis of how easy it for humans to detect a particular perturbation of an image. The level of the degradation of the quality has a close relation to the success of an adversarial example. The lower the structural similarity index the higher the chances of perturbing the particular features of the original class (Muncsan & Kiss, 2020).

Analysing the method critically, we can observe that the cases were the method would fail was when the images did not procure the desired result from the model. From the analysis of the attack efficiency the highest accuracy can be noticed when the trained model was taken in with 10 epochs, here the accuracy of the samples is a hundred percent. However, the percentage for the accuracy can be observed to be similar even when the epochs rate is 50 (Musa et al., 2021). In addition to achieving the main goal of making a model misjudge through the modified inputs to have the desired output, the quality and the level of security of these unrealistic outputs should be taken into consideration (Musa et al., 2021).

There are several interesting observations which can be made such as within a particular range, the size of the perturbation increases gradually, however it is inversely proportional when taken alongside with the dodging attack, the higher the size of perturbation, the recognition rate increases as well (Liu et al. 2019). It can also be noticed that the Dodging attack and the Impersonation attack have different sensitivity levels to the perturbation values taken, for example a small level of perturbation can definitely progress the impersonating attack to a greater extent but will not help the Dodging attack (Liu et al., 2019). The greater the iterations the greater the recognition rate. The granularity of the perturbation is very critical to the generation of the perturbation, usually the recommendation would be to always start with a smaller index value (Liu et al., 2019). A larger perturbation always helps the functionality of the Dodging attack but does not do the same for the impersonating attack. The Dodging attacks need lesser control on the perturbation, this stays put at least until the original image is no longer recognizable (Liu et al., 2019).

As observed in my experiments in the code, when the Epsilon value is 0.100, the confidence percentage of the Labrador retriever's image is about 13.97%. When the Epsilon value is 0.010, the Confidence percentage of the Labrador retriever's image is about 87.98%. When the Epsilon value is taken up to 1.19, the confidence percentage drops to 9.24%.

References

Goodfellow, I., Shlens, J., & Szegedy, C. (2015). EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES. *EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES*. https://arxiv.org/pdf/1412.6572.pdf

Liu, Y., Mao, S., Mei, X., Yang, T., & Zhao, X. (2019). Sensitivity of Adversarial Perturbation in Fast Gradient Sign Method. *Sensitivity of Adversarial Perturbation in Fast Gradient Sign Method*. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9002856

Munscan, T., & Kiss, A. (2020). Transferability of Fast Gradient Sign Method. *Transferability of Fast Gradient Sign Method*. https://www.researchgate.net/profile/Attila-Kiss-3/publication/343850757_Transferability_of_Fast_Gradient_Sign_Method/links/5f44b6cb92851cd302280b8d/Transferability-of-Fast-Gradient-Sign-Method.pdf

Musa, A., Vishi, K., & Rexha, B. (2021). Attack Analysis of Face Recognition Authentication Systems Using Fast Gradient Sign Method. *Attack Analysis of Face Recognition Authentication Systems Using Fast Gradient Sign Method*. https://www.tandfonline.com/doi/pdf/10.1080/08839514.2021.1978149