# Paper reports – Mohammad Osoolian

Paper Title: Robustness May Be at Odds with Accuracy

Paper Link: [1805.12152] Robustness May Be at Odds with Accuracy (arxiv.org)

Submit Date: 30 May 2018

What is paper about: This paper discusses the tradeoff between adversarial robustness and shows some advantages of robust models in being human understandable

## Abstract:

This article explores the trade-off between model robustness and accuracy in machine learning. The authors investigate how enhancing a model's robustness against adversarial attacks, which are carefully crafted inputs to mislead the model, can sometimes lead to a reduction in its overall accuracy. This study sheds light on the complex relationship between robustness and accuracy, highlighting the challenges and considerations involved in developing models that are both resistant to attacks and perform well on standard tasks.

## Background:

- Adversarial Attacks
- Robustness
- PGD

## Challenge:

The author believes that robustness in ML models has an important cost that cannot be ignored. The Problems with robustness is that it needs much more time to learn and requires much more computational power. Article also claims that robust models basically learn features with goals that are not compatible with accuracy. Therefore, there is a tradeoff between robustness and accuracy.

## New Ideas:

Article proves that robust models cannot be as accurate as models that only focus on accuracy. There are some experiences in this article on MNIST dataset that shows accuracy decreases by increasing the robustness of the training dataset.

There are also some surprising results of computing loss function on robust models. These results show that as much as the robustness of model increases, visualization of loss gradients on input pictures make more sense for human. For example you can see edges in loss gradient visualization

The last idea is that generating new images in robust models by increasing the loss, results to images of other classes.

## Results:

As we can see in the graphs in this article, the claim about tradeoff between robustness and accuracy is correct. The results of accuracy of models with different levels of robustness on MNIST and ImageNet dataset prove it.

Also the figures of Iterations of increasing loss for generating new images shows that by increasing the robustness of model, the generated images are more meaningful and less noisy. We can clearly see that for example a pictures of a monkey is converted to a picture of dog in multiple iterations.

## My Idea for the challenge:

we can find an algorithm that finds the best state for the tradeoff between robustness and accuracy based on the need and where and how the model is going to be used.

## My Idea to improve the article:

We can use the meaningful loss gradient visualizations of robust models to find out how the models work and use them to set some rules or freeze some parameters for model that speed up the learning process.

Also we can use the image generation of robust models for AI-creativity.