

On Neural Networks and defending against attacks :

I. Studied variables:

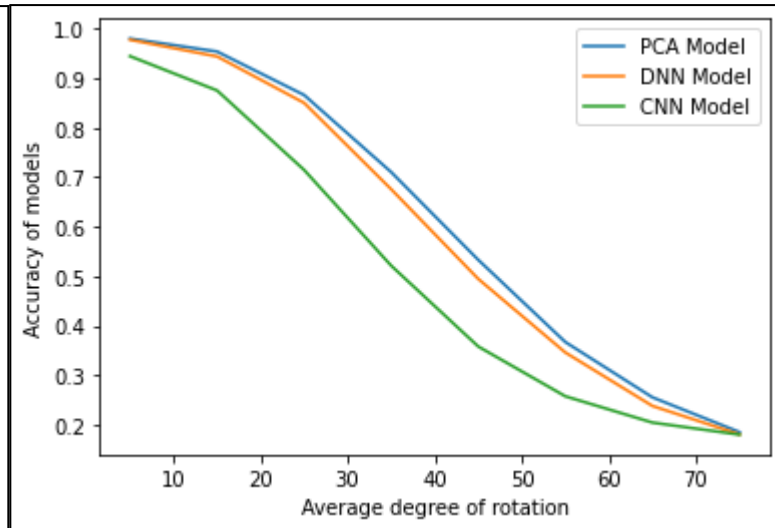
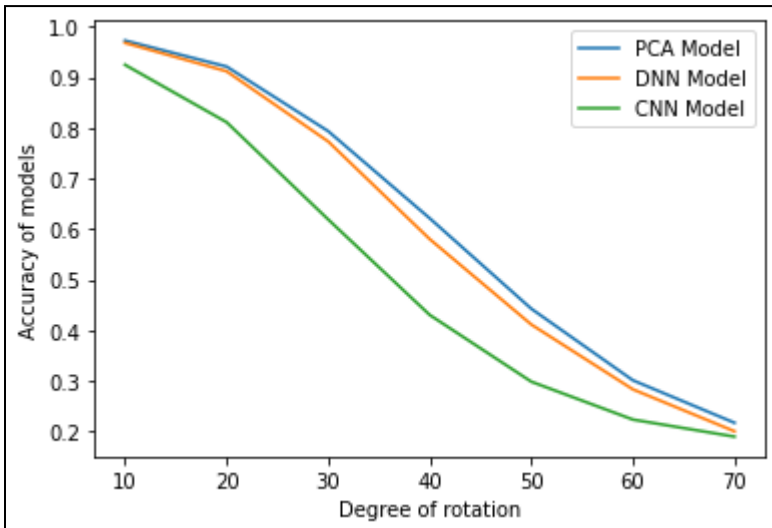
- Type of model (CNN vs DNN vs PCA)
- Database of model

II. Types of attacks :

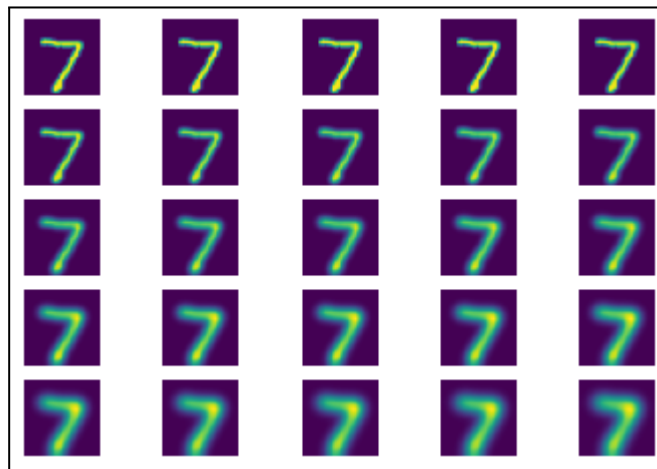
- Specifically tailored adversarial attack (<https://arxiv.org/pdf/1412.6572.pdf>).
- Rotation
- Gaussian Blur
- Box Blur
- Uniform Noise
- Perlin Noise
- Color Inversion

Graphs obtained so far:

- Rotation (left is deterministic rotation angle, right is slightly random rotation)



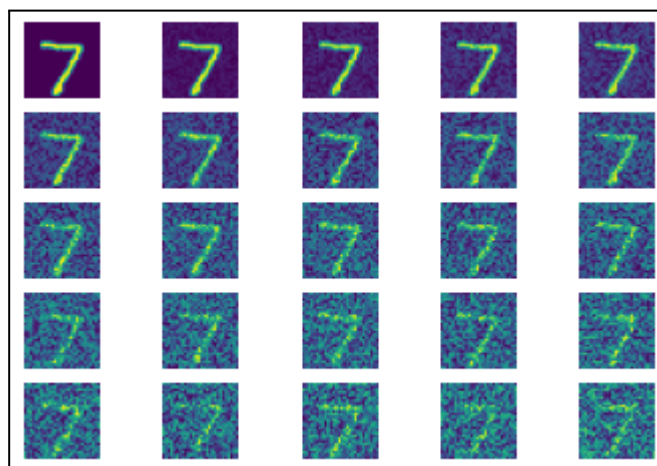
- Different levels of gaussian blur visualized (first image = 0 blur, sigma increases by 1/10 with each image)



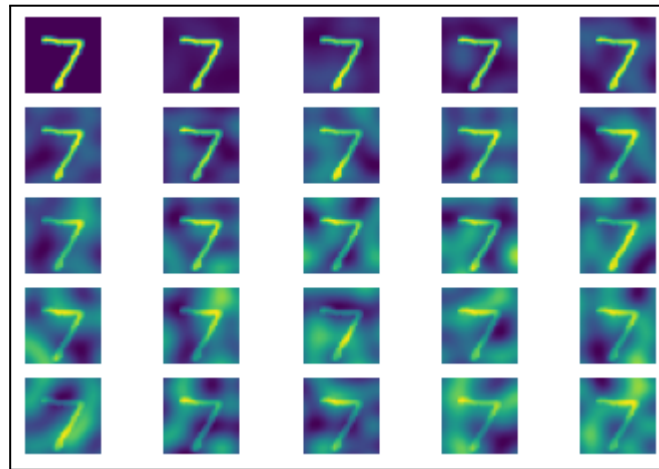
- Different levels of box blur visualized (box blur blurs less than gaussian blur for a given sigma), first image = 0 blur, sigma increases by $\frac{1}{4}$ with each image :



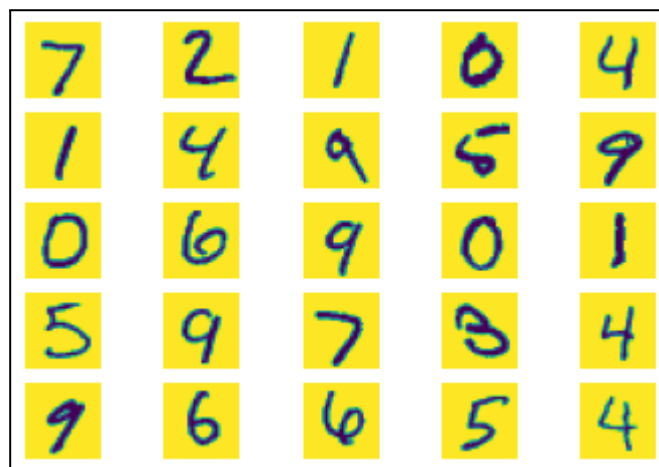
- Effect of uniform noise, increase of 1/20 with each image:



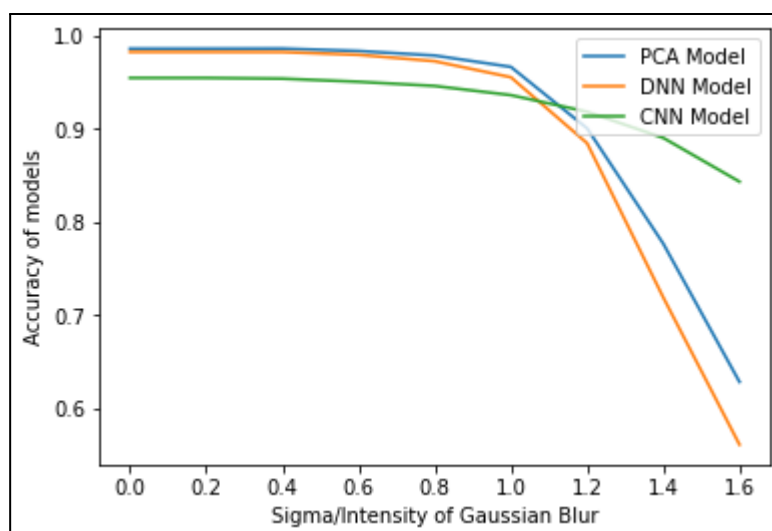
- Effect of perlin noise on image, increase of 1/20 per image :



- Examples of multiple color flipped / inverted image :



- Effect of gaussian blur on accuracy :



- Effect of box blur on accuracy :

