Robustness as Inherent Property of Datapoints

AISafety Workshop @ IJCAI 2020

Andrei Ilie, Marius Popescu, Alin Stefanescu
University of Bucharest

Completely random perturbation scheme

- Very simple model-agnostic empirical method for estimating the robustness of a model.

- Measure empirical robustness of a model with respect to a (test) data set.

- Perform perturbation on the dataset according to a randomised noise scheme, and monitor how the perturbation rate changes.

- Simplest version: given an image **X**, iteratively apply random noise **e** on top of the original **X**, until the model doesn't classify **X+e** correctly anymore (or until a limit of iterations **MAX_STEPS** is reached)

Experiments

- Simple CNN architecture which achieves 98.85% validation accuracy on MNIST, **MAX_STEPS=250**

- Perturbation scheme: randomly pick at most 28 pixels, and assign them random values (*L0-attack)*

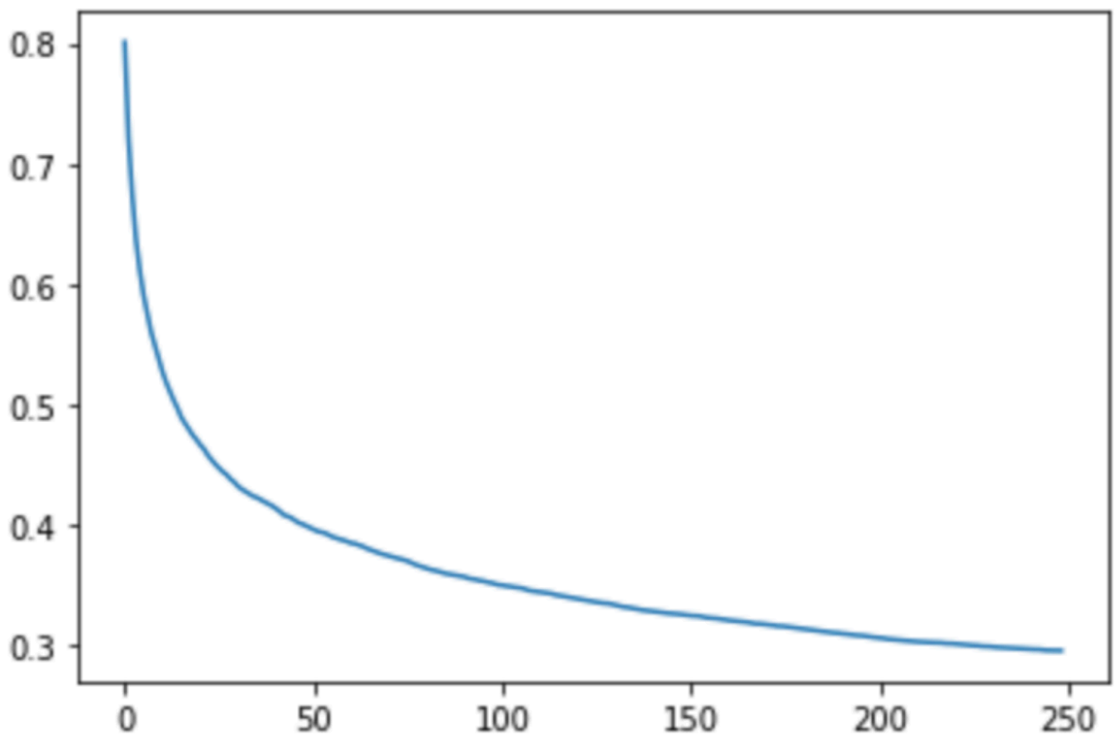

Figure 3: The ratio of images from the test set that are still robust as a function of the number of perturbation iterations that have been applied. The initial model $\mathcal{M}$ is used.

Experiments

- We deem images we were able to perturb as *non-robust*, and images we were able to perturb as *robust* (with respect to the model, to the perturbation scheme, to **MAX_STEPS**, etc)

- Qualitatively assess sampled robust vs non-robust samples below:

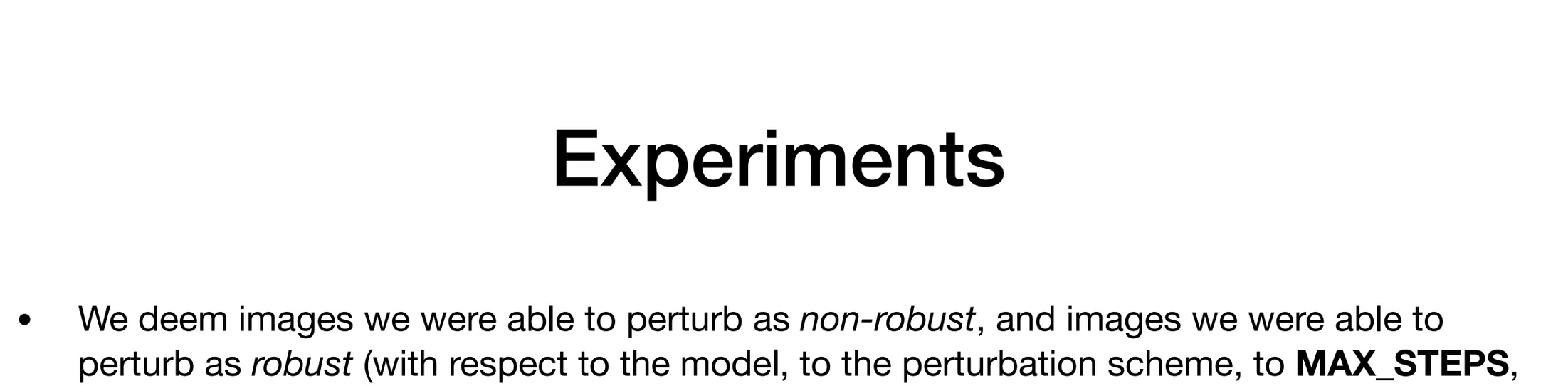

Figure 1: Images deemed as robust by our simple CNN on the first row against images deemed as not robust on the second row. The images on the second row were classified correctly by the model $\mathcal{M}$ before applying the random perturbation process.

Experiments

- Some classes are inherently harder to be perturbed

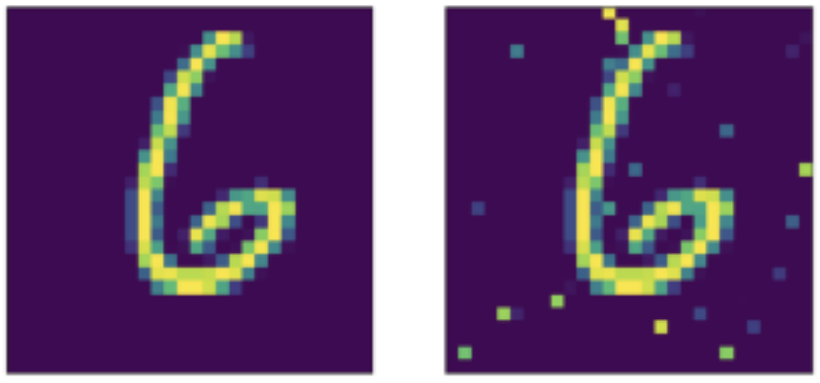

Figure 2: The image on the left is labelled as 6 by $\mathcal{M}$. The image on the right is obtained by perturbing at most 28 pixels from the left one, and it is labelled as 2 by $\mathcal{M}$. The perturbed image was obtained after 47 random perturbation steps of altering at most 28 pixels. All the previous 46 random perturbations were not able to confuse the model.

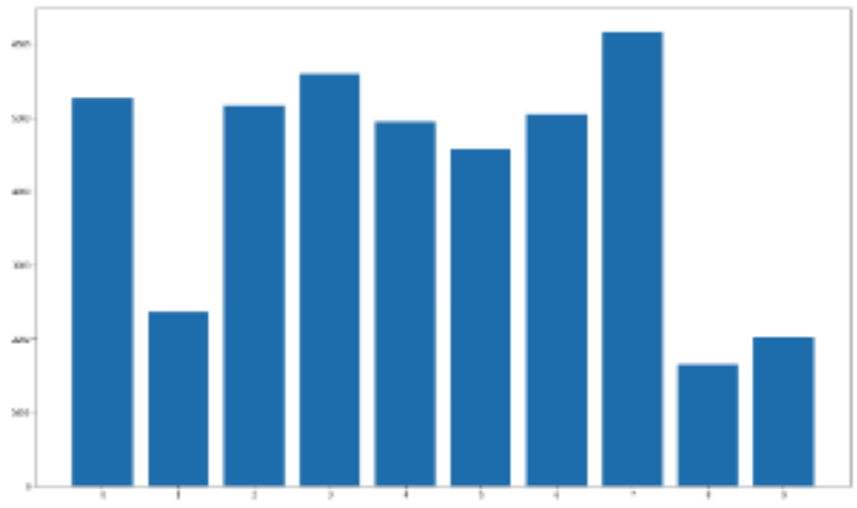

Figure 4: Distribution of training images that are deemed as robust under model $\mathcal{M}$. Images labelled as 7 seem to inherently be more robust, while images labelled as $1, 8,$ and 9 can easily be corrupted by random perturbations.

Robustness improvement

- We retrained the model using the robust training samples only. Because the samples were not class-balanced, we sampled a constant of 1500 images for each label. This is only 25% of the whole training data.

- The accuracy on the test set decreased from 98.85% to 96.92%, but the percentage of images not-perturbed by the random process increased from 29.57% to 51.01%

- Therefore, the model being trained on more robust samples make it more robust

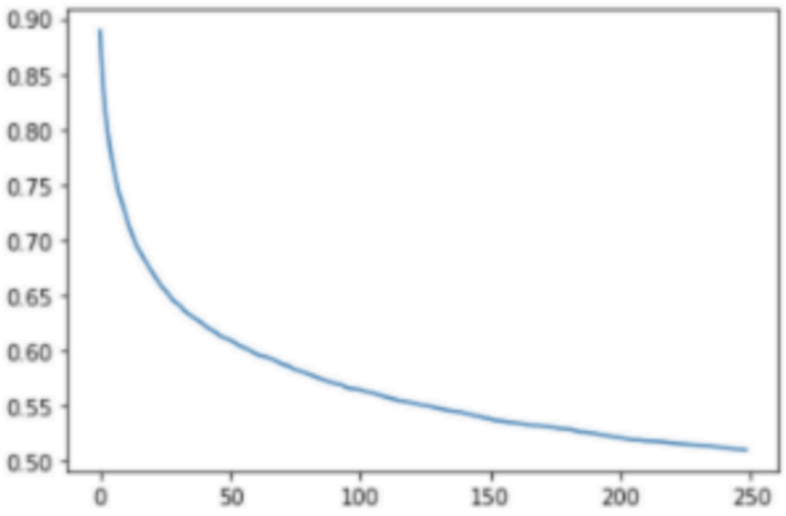

Figure 5: The ratio of images from the test set that are still robust as a function of the number of perturbation iterations that have been applied. Here, the model $\mathcal{M}_R$ is used. There is a clear improvement in robustness when compared to the model $\mathcal{M}$.

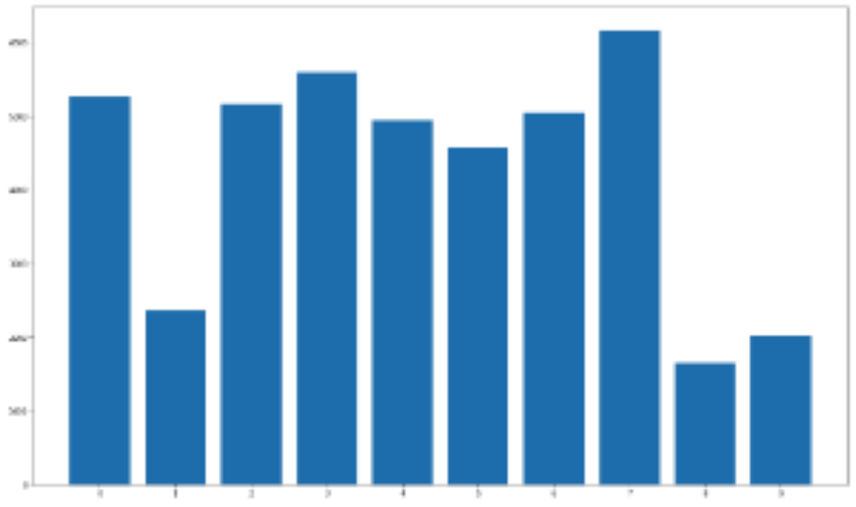

Figure 4: Distribution of training images that are deemed as robust under model $\mathcal{M}$. Images labelled as 7 seem to inherently be more robust, while images labelled as $1, 8,$ and 9 can easily be corrupted by random perturbations.

Conclusion and further work

- Robustness can be seen as an inherent property of the images with respect to the classification task. The robustness of models depends both on their architecture and on the robustness of the data it is trained on. This can be exploited in various ways, such as the training methodology we proposed, which improves significantly the robustness of the model.

- Some interesting other applications could include using Generative Adversarial Networks (GANs) to augment the robust training data from the training methodology we proposed. Data augmentation with GANs has successfully been used in improving the quality of data and accuracy of models [Antoniou et al., 2017] and we believe that it could be used to generate diverse robust images as well. These could contribute to increasing the accuracy of robust models trained under our methodology.