Robustness as Inherent Property of Datapoints*

Andrei Ilie , Marius Popescu , Alin Stefanescu
University of Bucharest
{cilie, marius.popescu, alin}@fmi.unibuc.ro

Abstract

Characterizing how effective a machine learning algorithm is while being trained and tested on slightly different data is a widespread matter. The property of models which perform well under this general framework is commonly known as robustness.

We propose a class of model-agnostic empirical robustness measures for image classification tasks. To any random image perturbation scheme, we attach a robustness measure that empirically checks how easy it is to perturb a labelled image and cause the model to misclassify it.

We also introduce a methodology for training more robust models using the information gained about the empirical robustness measure of the training set. We only keep a fraction of datapoints that are robust according to our robustness measure and retrain the model using it. Our methodology validates that the robustness of the model increases by measuring its empirical robustness on test data.

1 Introduction

During the last decade, the field of machine learning has made considerable advances in many tasks, such as image classification, object detection, machine translation, or question answering, with deep neural networks easily becoming the state-of-the-art approaches [Touvron *et al.*, 2020; Zhang *et al.*, 2020; Edunov *et al.*, 2018]. The main priority has been on the capacity of the models to perform well on the test set of some well-known datasets (MNIST, CIFAR, SQuAD) [LeCun and Cortes, 2010; Krizhevsky, 2009; Rajpurkar *et al.*, 2016]. However, the training and the test sets are usually generated from the same underlying distribution, leaving the model's performance under distribution shifts unknown. Given that machine learning techniques are being employed in sensitive tasks, such as self-driving cars and healthcare, the robustness should become a crucial metric to be taken into consideration together with the accuracy when evaluating the performance of models.

We will mainly focus on safety and robustness for image classification tasks, but the work can be easily extended to other topics.

Distribution shifts, which affect the performance of machine learning systems, can mainly occur because of two reasons. The first reason, adversarial attacks [Wiyatno *et al.*, 2019; Szegedy *et al.*, 2014], has been receiving growing attention over the past years. Adversarial attacks are "hidden messages" [Wiyatno *et al.*, 2019] added on top of images which are nearly imperceptible to the human eye, but which cause the model to fault, in other words creating "machine illusions".

The second reason, covariate shift [Shimodaira, 2000], is encountering a natural change in the data distribution. For example, imagine an autonomous car model trained solely on rainy and sunny conditions in a city where it has not been snowing over the past five years. However, one day it starts snowing, and the image recognition system of the autonomous car could have serious issues in identifying objects and road signs because of completely different lighting conditions.

While improving models to be less exposed to known adversarial attacks is very important, one has to keep in mind that this is, after all, an adversarial game, where the attacker and the security researcher keep alternately coming up with better strategies. For example, the adversarial attack strategy Fast Gradient Sign Method [Szegedy *et al.*, 2014] can be mitigated by Adversarial training [Szegedy *et al.*, 2014], which can in turn be bypassed by R+FGSM [Tramèr *et al.*, 2018]. The defense methods against adversarial attacks seek to make the model robust with respect to certain adversarial points in the neighbourhood of unaltered images.

Therefore, one is prompted to consider a more general robustness framework, in which the interest lays in the model not making a mistake anywhere in the neighbourhood of an image[1]. There exist various tools that can achieve robustness guarantees of deep neural networks [Ruan *et al.*, 2018; Tjeng *et al.*, 2019], but most of them are usually very dependent on the model's architecture, either not being able to scale with deeper networks, or only working with certain kinds of layers.

[1]For example, the neighbourhood could be specified by some metric ball around the image.

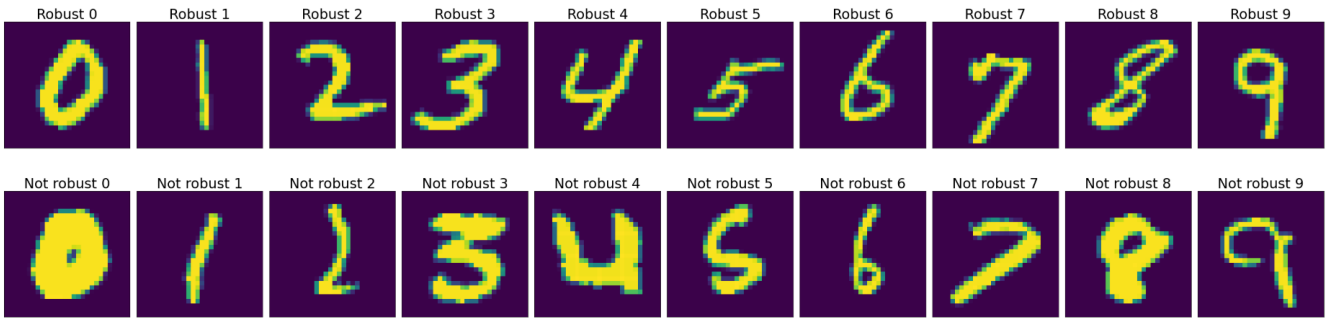

Figure 1: Images deemed as robust by our simple CNN on the first row against images deemed as not robust on the second row. The images on the second row were classified correctly by the model $\mathcal{M}$ before applying the random perturbation process.

We propose a model-agnostic[2] empirical method for estimating the robustness of a model. This estimation of a model near an image X is done by iteratively sampling datapoints close to it, according to a specified random scheme[3]. It feeds each of the sampled datapoints to the model and stops either when the model classifies them incorrectly, or when a maximum number of steps has been reached. The number of such sampling steps serves as a proxy for the local robustness around image X. Intuitively, the easier it is to perturb the label of X by sampling around it, the less robust the model is around it. We use this method for estimating the robustness of the model on entire datasets, by locally checking the model's robustness around each datapoint and combining the results.

We also claim that the robustness of the model is correlated with the inherent robustness of the images with respect to the classification task. Therefore, the robustness of a model depends on both the architecture's robustness itself, but also on the inherent robustness of the datapoints it has been trained on.

We believe that training a model on certain correctly labelled images can lead towards highly unnatural borders between classes. These might be datapoints that we would rather misclassify than include in the model at an additional high cost of robustness. We test this hypothesis and obtain indeed a more robust model by discarding the not-robust images from the training process.

Our main technical contributions are introducing the empirical robustness measure that is model-agnostic and the training methodology based on robust images.

An important general direction we want to shed light on is that images from classification tasks should be seen as carrying an inherent level of robustness, which could be estimated and exploited.

[2]The method does not need to have any knowledge about the architecture of the model. Note that the model does not necessarily have to be a deep neural network.

[3]The random scheme should not alter the underlying true class of the image that we sampled around. Intuitively, the samples should be classified by a human in the same way as the original image is.

2 Randomized Perturbation Robustness

2.1 Definition

We propose a class of empirical robustness measures $\mathbf{RPR}$ (**Randomized Perturbation Robustness**) for image classification tasks, which is model-agnostic. Let R be a random image perturbation scheme.[4] The empirical robustness $\mathbf{RPR}(R)$ of a model $\mathcal{M}$ with respect to a datapoint x belonging to class y is the minimum between MAX_STEPS and the expected number of retrying steps of applying R to the original x such that $\mathcal{M}$ does not classify $R(x)$ as y.

If the empirical robustness of $\mathcal{M}$ with respect to (x, y) is MAX_STEPS, we stop and deem x as robust; otherwise as not-robust.

Note that the random perturbations of an image are not applied on top of previous perturbation attempts, but rather on the original image. This perturbation process is repeated until the conditions above are fulfilled.

2.2 Empirical robustness on datapoints and on entire datasets

The introduced framework is a simple empirical way of assessing a model's robustness near an image. It is suitbale under various setups, such as the random image perturbation scheme of adding weather conditions[5] in the autonomous car situation.

We propose two use cases based on the empirical robustness measure introduced above: One estimating the model robustness on an entire (test) dataset and another one training a model only using the images that are deemed as robust in order to obtain a more robust model.

The first use case, estimating the robustness of the model on an entire dataset, is done by applying the Randomized Perturbation Robustness method described above on each datapoint and computing the percentage of images that are deemed as robust. The model-agnosticism makes it an easy plug-in method in any classification task and can easily be introduced as a baseline check for machine learning systems.

The second use case is based on our claim that the robustness of the model with respect to a datapoint can be seen to

[4]For example Gaussian noise, replacing at most k pixels of an image, blurring, etc.

[5]Applying snow, fog, rain effects, etc.

some extent as the inherent robustness of the datapoint with respect to the classification task. This allows us to retrain the model using only images from the train set that are deemed as robust by our empirical measure, giving us a more robust model. This happens as the model only learns using the robust images, which justifies inferring simple, more natural class separators. We claim that the images that are deemed as not robust by our method can generally be seen as edge cases, causing the model to infer irregular separators.

2.3 Methodology and experiments

We experiment using a CNN architecture for classifying images from MNIST. As this classification task is not complex, we use a very simple model[6] which achieves a test accuracy of only 98.85% to showcase the main ideas we introduce. The randomized image perturbation scheme we use is randomly altering a pixel count of at most the square root of the number of image pixels (28 in our case). We use MAX_STEPS $= 250$ in our experiments.

We show in Figure 2 an image that is classified correctly by $\mathcal{M}$ against its random perturbation under the scheme described above, which is incorrectly classified by $\mathcal{M}$.

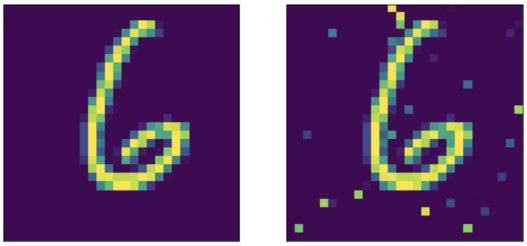

Figure 2: The image on the left is labelled as 6 by $\mathcal{M}$. The image on the right is obtained by perturbing at most 28 pixels from the left one, and it is labelled as 2 by $\mathcal{M}$. The perturbed image was obtained after 47 random perturbation steps of altering at most 28 pixels. All the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

We compare in Figure 1 robust and not robust images which, without any perturbation, are correctly classified by $\mathcal{M}$. These were randomly chosen and give some intuition about what a robust image means compared to one that is not robust.

The process we described for determining the empirical robustness is very similar, when seen as a function of MAX_STEPS, to a learning curve. Discovering images which are not robust eventually flattens, which allows us to use it together with some early-stopping mechanism.

In Figure 3 we can see how the ratio of test images that are still robust as a function of MAX_STEPS flattens. We obtain a ratio of 0.2957 images from the test set which can withstand 250 random permutations, which is a surprisingly small fraction, considering the simple noising we apply. This stands as straightforward empirical evidence that the simple CNN architecture we used is not robust.

[6]We use two small convolutional layers, one max pooling layer, and a fully connected layer with softmax activation. We also train the model with ADAM using the default hyperparameters.

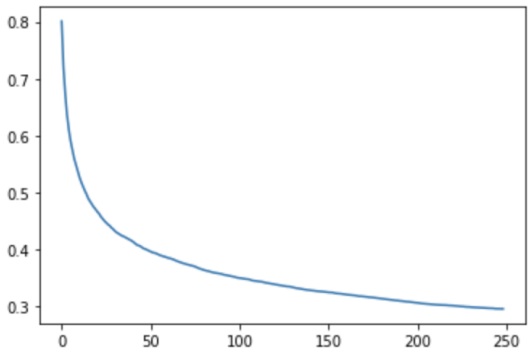

Figure 3: The ratio of images from the test set that are still robust as a function of the number of perturbation iterations that have been applied. The initial model $\mathcal{M}$ is used.

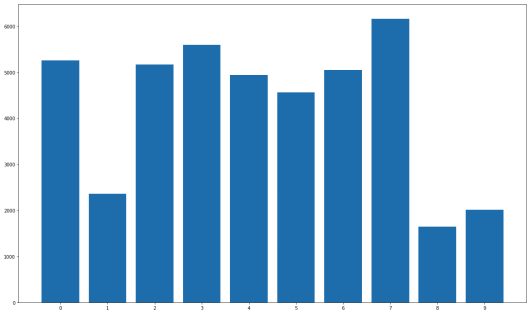

Figure 4: Distribution of training images that are deemed as robust under model $\mathcal{M}$. Images labelled as 7 seem to inherently be more robust, while images labelled as $1, 8,$ and 9 can easily be corrupted by random perturbations.

In order to achieve a more robust network, we apply the same procedure of deeming an image as robust or not robust on the MNIST train set, using the model $\mathcal{M}$, which was trained on exactly this data. There are 71.28% images which are deemed as robust from the train set, however the distribution is not uniform at all as seen in Figure 4. Therefore, we randomly sample 1500 datapoints from each class of the robust training images, such that the training set does not have a class bias, and proceed to retrain the simple CNN architecture solely by using this data. Let the model trained on this data, which amounts for only 25% data from the MNIST training set, be $\mathcal{M}_R$. We encounter a drop of approximately 2% in the test accuracy, obtaining a 96.92% score, which is to be expected considering the relatively sparse training data we have trained on.

The model $\mathcal{M}_R$ is much more robust on the test set, obtaining a ratio of 0.5101 robust images, cf. Figure 5, as compared to the robustness of the original $\mathcal{M}$ of only 0.2957. This stands as evidence that the robust nature of the selected training images led to a more robust model.

3 Conclusions and future work

The simple empirical robustness checking method we introduce opens the way towards building fast, model-agnostic tools to estimate robustness of machine learning models. This

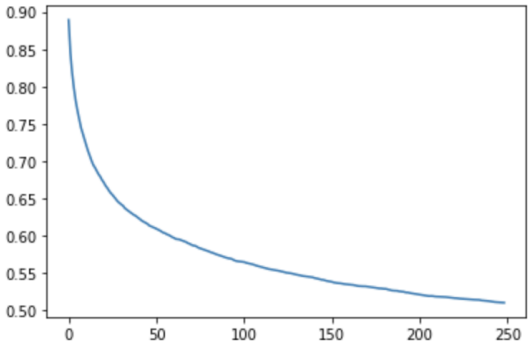

Figure 5: The ratio of images from the test set that are still robust as a function of the number of perturbation iterations that have been applied. Here, the model $\mathcal{M}_R$ is used. There is a clear improvement in robustness when compared to the model $\mathcal{M}$.

method can be easily embedded as a base check in machine learning systems.

One of the main takeaways is that 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.

Another area of further investigation is checking how our empirical robustness measure relates with the formal verification tools that obtain exact robustness guarantees. Note that this kind of experiment is not possible for any model, as existing formal verification tools are limited to specific machine learning architectures or do not scale well with complex models.

References

[Antoniou *et al.*, 2017] Antreas Antoniou, Amos J. Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *CoRR*, abs/1711.04340, 2017.

[Edunov *et al.*, 2018] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 489–500. Association for Computational Linguistics, 2018.

[Krizhevsky, 2009] Alex Krizhevsky. *Learning Multiple Layers of Features from Tiny Images*, chapter 3. Technical Report TR-2009, University of Toronto, 2009.

[LeCun and Cortes, 2010] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist/, 2010.

[Rajpurkar *et al.*, 2016] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics, 2016.

[Ruan *et al.*, 2018] Wenjie Ruan, Xiaowei Huang, and Marta Kwiatkowska. Reachability analysis of deep neural networks with provable guarantees. In Jérôme Lang, editor, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 2651–2659. ijcai.org, 2018.

[Shimodaira, 2000] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90:227–244, October 2000.

[Szegedy *et al.*, 2014] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. January 2014. 2nd International Conference on Learning Representations, ICLR 2014 ; Conference date: 14-04-2014 Through 16-04-2014.

[Tjeng *et al.*, 2019] Vincent Tjeng, Kai Y. Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[Touvron *et al.*, 2020] Hugo Touvron, Andrea Vedaldi, and Herve Jegou Matthijs Douz and. Fixing the train-test resolution discrepancy: FixEfficientNet. *arXiv:2003.08237v4*, April 2020.

[Tramèr *et al.*, 2018] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[Wiyatno *et al.*, 2019] Rey Reza Wiyatno, Anqi Xu, Ousmane Dia, and Archy de Berker. Adversarial examples in modern machine learning: A review. *CoRR*, abs/1911.05268, 2019.

[Zhang *et al.*, 2020] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander Smola. Resnest: Split-attention networks. *arXiv:2004.08955v1*, April 2020.