

Applications of Adversarial Robustness Analysis in Machine Learning

PhD Thesis

István Megyeri

Supervisor: Dr. Márk Jelasity

Department of Computer Algorithms and Artificial Intelligence
Doctoral School of Computer Science
Faculty of Science and Informatics
University of Szeged



Szeged
2023

Introduction

Let us start with a short story that demonstrates well the topic of the dissertation and also the current status of artificial intelligence research.

Clever Hans was a horse that gained fame in the early 20th century for apparently being able to perform complex arithmetic and other intellectual tasks. The horse was owned by a German mathematics teacher named Wilhelm von Osten, who claimed to have taught him these skills.

Von Osten became a sensation in Germany, with many people flocking to see the amazing horse. However, skeptics suspected that there was more to the horse's abilities than met the eye. A psychologist named Oskar Pfungst investigated the phenomenon and concluded that Clever Hans was not actually performing arithmetic, but was instead responding to subtle cues from his trainer and audience.

Pfungst discovered that von Osten was unwittingly providing the horse with cues, such as body language or slight head movements that told it when the horse had actually found the correct answer. Once this was realized and providing cues was prevented, Clever Hans was no longer able to perform the same feats of arithmetic when his trainer was not present or when he was blindfolded.

The case of Clever Hans became an important milestone in the history of psychology, as it demonstrated the importance of experimental controls and the potential for unconscious cueing to influence the behavior of both humans and animals.

Artificial intelligence has reached a similar milestone. A subfield called deep learning gained significant popularity in the 2010s. While the concept of neural networks and deep learning has existed for several decades, it was during this period that several factors converged, leading to a surge in its popularity.

One of the main catalysts was the availability of large datasets and advancements in computational power, which allowed researchers to train deeper neural networks and process massive amounts of data more efficiently. And the development of specialized hardware, such as graphics processing units (GPUs), accelerated the training of deep learning models.

Another crucial factor was the breakthrough in performance achieved by deep learning models in various challenging tasks. Deep learning architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), demonstrated superior performance in image recognition, speech recognition, natural language processing, and other domains, often surpassing traditional machine learning approaches.

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012 played a pivotal role in showcasing the power of deep learning. The winning team, led by Geoffrey Hinton, utilized deep convolutional neural networks and significantly outperformed other methods. This event served as a turning point and drew attention from classical machine learning methods to the capabilities of deep learning.

The success and breakthroughs in deep learning, combined with the increasing availability of open-source tools and libraries, made it more accessible to researchers and practitioners. This accessibility, along with the growing interest from both academia and in-

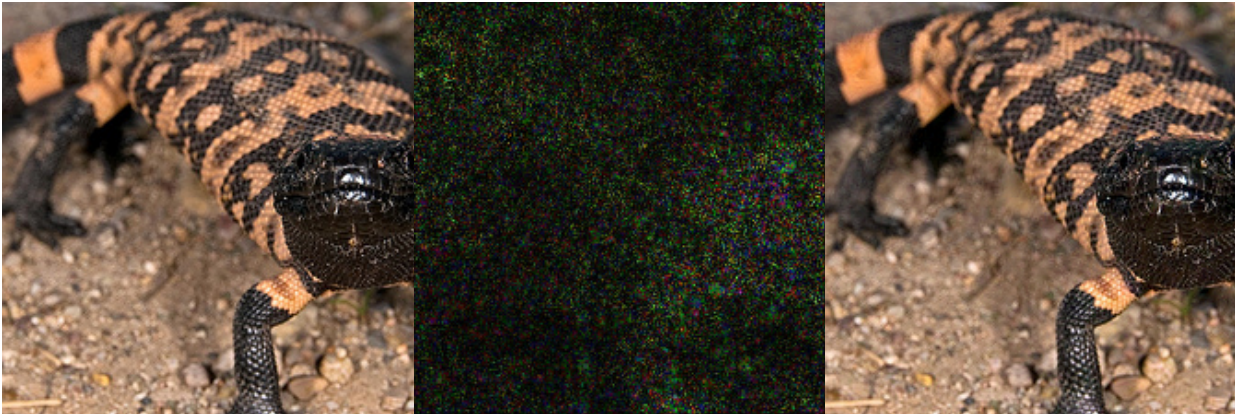


Figure 1: A sample adversarial image. *Left: original image (prediction: 'Gila monster'); Middle: adversarial perturbation; Right: adversarial image (DenseNet201 prediction: 'custard apple'). The adversarial modification is invisible to the human eye but it changes the neural network prediction.*

dustry, contributed to the rapid proliferation and popularity of deep learning.

Since then, deep learning has continued to make strides in various fields, and its popularity has only increased. It has become the dominant approach in many areas of artificial intelligence and machine learning, shaping advancements in sectors ranging from healthcare and finance to automotive and entertainment.

A few years after the ImageNet challenge, Szegedy et al. discovered intriguing properties of these highly successful deep neural networks [17]. They found that deep neural networks are discontinuous to a certain extent. They showed that it is possible to cause the network to misclassify an image by applying a certain imperceptible perturbation, which is obtained by maximizing the network's prediction error. They coined the phrase adversarial examples for these perturbed and misclassified images, which is now widely used in the literature. The presence of adversarial examples was also confirmed by Goodfellow et al. [15]. A sample of an adversarial image shown in fig. 1.

Here, we will focus on the problem of image classification. In it, the sensitivity of the current models to adversarial input indicates that these models are not in complete accord to human perception. Similar to Clever Hans, after this sensitivity was noted the networks were no longer able to solve their given task.

Summary of the Thesis Results

The PhD thesis presents different applications of adversarial robustness analysis in machine learning. The dissertation consists of three major parts. In Thesis 1, we analyzed the robustness of linear models from regularization and dimensionality points of view that is also presented in chapter 3 of the dissertation. Next, Thesis 2 presents attack algorithms that are able to generate such perturbation which can mislead multiple models simultaneously. The corresponding part of the dissertation is chapter 3. In Thesis 3, we analyze

	Thesis 1	Thesis 2	Thesis 3
ESANN 2019 [2]	•		
ESANN 2020 [3]		•	
IJCNN 2020 [4]		•	
IJCNN 2021 [5]			•
PRAI 2023(submitted for publication) [1]			•

Table 1: *The connection between the theses and publications.*

defense methods for the problem of robust classification and robust out-of-distribution detection which is detailed in chapter 4.

Here, we will give a brief summary of each thesis. The ideas, figures, tables and results included in the dissertation were published in scientific papers (listed at the end of the booklet). In table 1, we indicate the connection between the theses and publications.

Thesis 1: Adversarial Robustness of Linear Models

Many machine learning models are sensitive to adversarial input, meaning that very small but carefully designed noise added to correctly classified examples may lead to misclassification. The reasons for this are still poorly understood, even in the simple case of linear models. In this thesis, we study linear models and offer a number of novel insights.

We focus on the effect of regularization and dimensionality and we demonstrate that even in the case of simple binary classification problems with linear models, the adversarial problem is real and it strongly depends on regularization and the less obvious properties of high-dimensional spaces. Namely, in very high dimensions adversarial robustness is inherently very low due to some mathematical properties of high-dimensional spaces that have received little attention so far.

Also, in higher dimensions an overly weak regularization setting might result in a significantly harder optimization problem in some cases. Our empirical analysis confirmed that—although regularization may help—adversarial robustness is harder to achieve than high accuracy during the learning process as we can see in fig. 2. This is typically overlooked when researchers set optimization meta-parameters.

Moreover, we showed that the optimal regularization strength is very different for adversarial robustness and prediction accuracy. This highlights that, the two metrics requires significantly different meta-parameters. Our experiments were conducted on two binary classification dataset a real and a generated one.

The main contribution of the author are the related experimental design, implementation, and analysis of the results. The corresponding publication and thesis chapter are [2] and chapter 3 respectively.

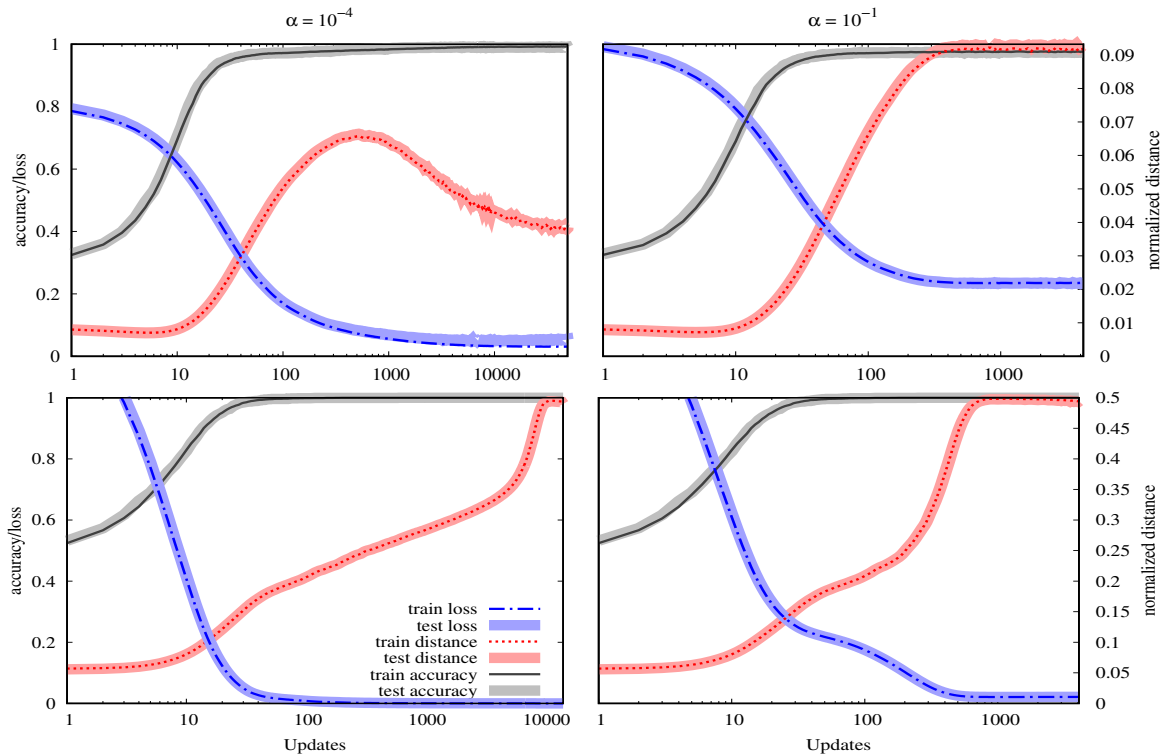


Figure 2: Convergence of normalized distance and accuracy in $d = 28 \times 28$ dimensions for the real dataset (top) and the generated dataset (bottom), with regularization coefficient $\alpha = 10^{-4}$ (left) and $\alpha = 10^{-1}$ (right).

Thesis 2: Adversarial attacks on model sets

The original formulation of the adversarial image search problem in [17] assumes that we are given a model and a correctly classified example. The attacker wishes to find a minimal perturbation of the example such that the model predicts any wrong label (untargeted attack) or a given desired label (targeted attack). Since then, a large number of methods have been proposed to create better adversarial examples [14, 16].

In this thesis, we study the question of whether the list of predictions made by a list of models can also be changed arbitrarily by a single small perturbation. Clearly, this is a harder problem since one has to simultaneously mislead several models using the same perturbation, where the target classes assigned to the models might differ. This attack has several applications over models designed by different manufacturers for a similar purpose. One might want a single perturbation that acts differently on each model; like only misleading a subset, or making each model predict a different label. Also, one might want a perturbation that misleads each model the same way and thereby create a transferable perturbation. Current approaches are not applicable for this general problem directly.

We propose an algorithm (algorithm 1) in this thesis that is able to find a perturbation that satisfies several kinds of attack patterns. For example, all the models could have the same target class, or different random target classes, or target classes designed to be maximally contradicting. Example images of the contradicting pattern shown in fig. 3.

The corresponding part of the dissertation (chapter 4) has two major parts. In section

Algorithm 1 Multi-model adversarial perturbation

```
1: Input: example  $x$ , models  $\mathcal{F}$ , adversarial patterns  $\mathcal{P}$ 
2:  $x_0 \leftarrow x$ 
3:  $i \leftarrow 0$ 
4: while  $i < i_{max}$  and  $K(x_i) \notin \mathcal{P}$  do
5:   for  $p_k \in \mathcal{P}$  do
6:      $r_k \leftarrow \text{approximateQP}(x_i, p_k)$ 
7:   end for
8:    $r \leftarrow r_{\arg \min_k \|r_k\|_2}$  ▷  $r_k$  with the smallest norm
9:    $r \leftarrow \min(\eta/\|r\|_2, 1) \cdot r$  ▷ enforce  $\|r\|_2 \leq \eta$ 
10:   $x_{i+1} \leftarrow x_i + r$ 
11:   $i \leftarrow i + 1$ 
12: end while
13: return  $x_i$  ▷ the perturbed input
```

4.1 of the dissertation, we introduced an initial version of the algorithm which applies the first-order approximation of the decision boundaries used in the DeepFool method. We evaluated the algorithm on a number of model sets over MNIST and CIFAR-10 datasets and generated transferable as well as non-transferable examples. We found that the algorithm consistently produces small perturbations in all the cases we examined. Perhaps the most interesting result is that small adversarial perturbations are present even when a non-transferable adversarial example was generated for the most robust model in the set, despite the fact that the models differed only in the regularization coefficient.

In section 4.2 of the dissertation, we show a generalized version of the method which has many interesting applications, it is still able to generate transferable adversarial examples as well as generating a single perturbation such that all the models in a given model set predict specified, different classes. The latter scenario allows us to explore the decision boundaries of the model set from a new perspective.

The algorithm can be regarded as a generalization of the DeepFool method to model sets. This generalized version is in algorithm 1. Also, we improved the DeepFool algorithm itself by adding the step size parameter. We evaluated our algorithm on three model sets using four attack patterns over the ImageNet database. We found that the algorithm produces small and successful perturbations reliably in all the attack scenarios we examined. Here, the most interesting result is that imperceptible adversarial perturbations were found even when the labels were selected to make the problem as hard as possible. This was surprising to us, even in the light of the vast literature on adversarial attacks.

The perturbation sizes over the three model sets offered some interesting insights as well. The set with different model architectures (mobile set) needed somewhat larger perturbations, but we expected just the opposite. Increasing the size of the model set increased perturbation size as well. Nevertheless, all the perturbations we found are imperceptible to the human eye. A sample of the found perturbation shown in fig. 3.

The corresponding publications are [3] and [4]. The experimental design, evaluation, as well as the formalism of the problem and the algorithm, were carried out by the author.

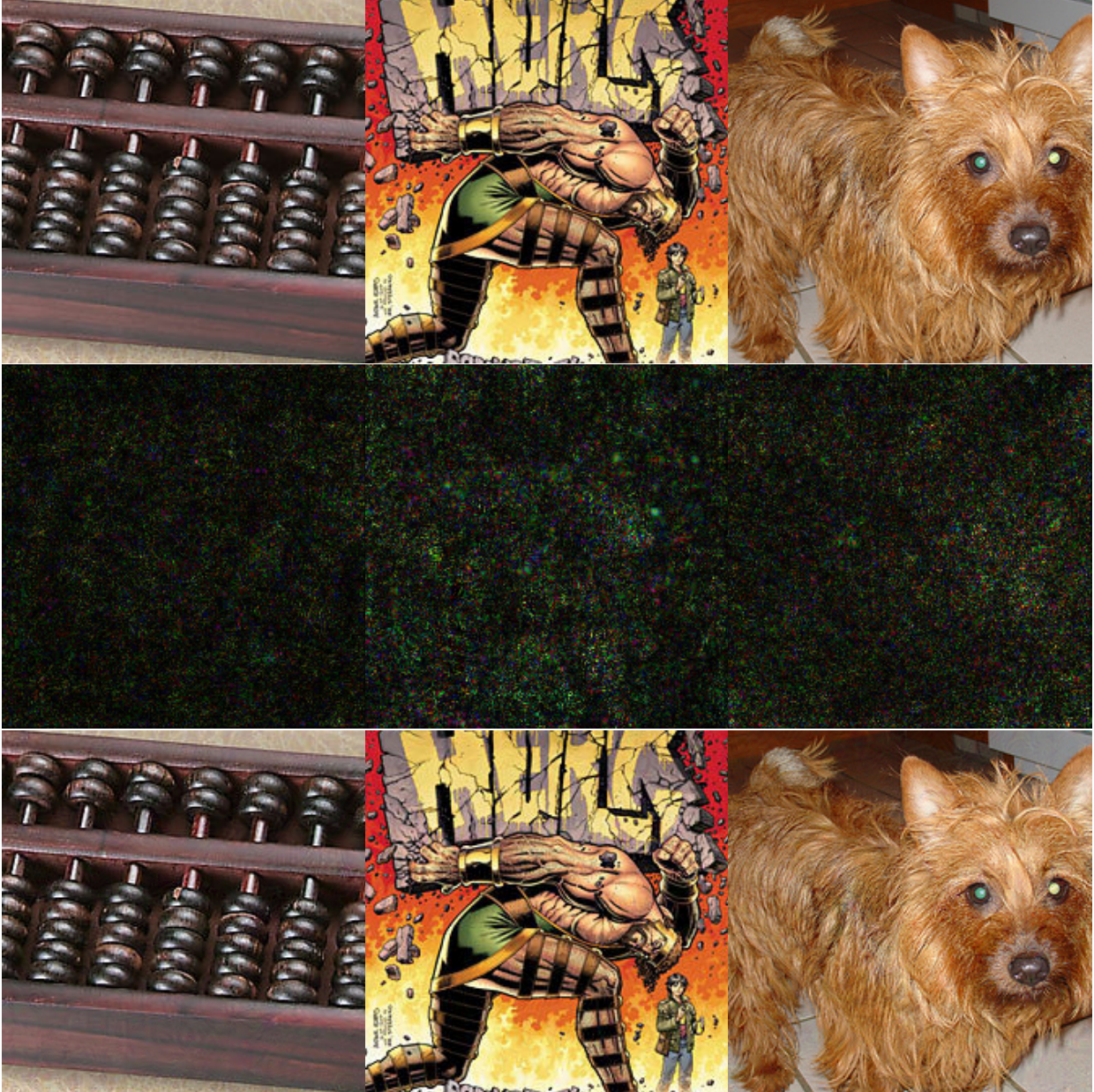


Figure 3: The diverse(contradicting) attack pattern over the mobile set (left: abacus → [soft-coated-wheaten_terrier, soft-coated_wheaten_terrier, apron]), dense set (middle: comic_book → [sturgeon, black_stork, capuchin]), and all the models (right: Australian_terrier → [Saluki, borzoi, black_stork, Saluki, gorilla, kuvasz]).

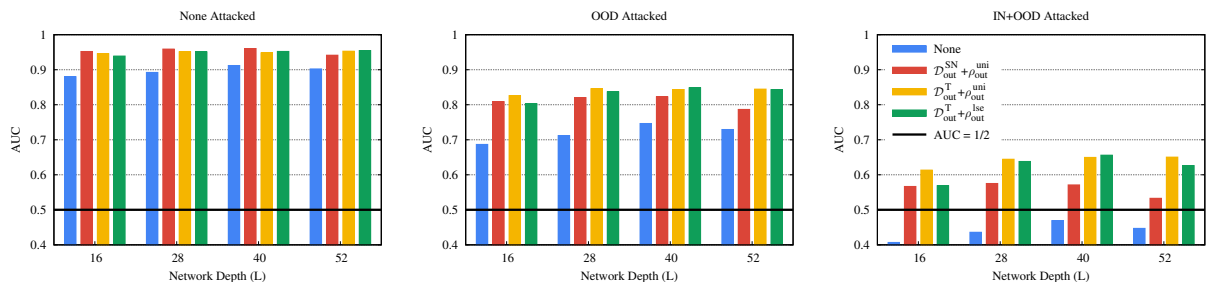


Figure 4: OOD detection AUC over CIFAR-10 under three different kinds of attack scenarios: no attack, only OOD samples are perturbed and both in-distribution and OOD samples are perturbed. The strong attack scenario is where both in-distribution and OOD examples are adversarially perturbed to mislead OOD detection.

Thesis 3: Combining Robust Classification and Robust out-of-Distribution Detection

Classification models in machine learning often make over-confident but incorrect predictions on input samples that do not belong to any of the output classes. Such samples are called out-of-distribution (OOD) samples. This problem has received considerable attention, because this represents a vulnerability similar to adversarial input perturbation, where models make incorrect predictions on seemingly in-distribution input samples that contain a very small but adversarial perturbation.

In this thesis, we are interested in models that are robust to both OOD samples and adversarially perturbed in-distribution samples. Furthermore, we require that OOD detection be robust to adversarial input perturbation. That is, OOD samples and in-distribution samples should not have adversarial perturbations that makes them appear to be in-distribution and OOD samples, respectively. Several related studies apply an ad-hoc combination of several design choices to achieve similar goals. One can use several functions over the logit or soft-max layer for defining training objectives, OOD detection methods and adversarial attacks.

The contribution of this thesis is that we defined a design space, where one can systematically analyze the problem of robust OOD detection and robust classification. The main components were identified as the training objectives, detection methods and attack methods for the combination of the robust OOD detection problem and the robust classification problem with the help of a set of score functions. Also, we introduced a strong threat model in which both in-distribution and OOD samples are adversarially perturbed to mislead OOD detection. In fig. 4, we can see the OOD detection performances measured in different attack scenarios (including the proposed one where both in-distribution and OOD samples are adversarially perturbed).

We draw several interesting conclusions based on our empirical analysis of this design space. Most importantly, we argue that the key factor is not the OOD training or detection method in itself, but rather the application of matching detection and training methods. The OOD detection performances shown in fig. 5 as the function of detection methods and training methods. Moreover, we performed a thorough empirical evaluation of this

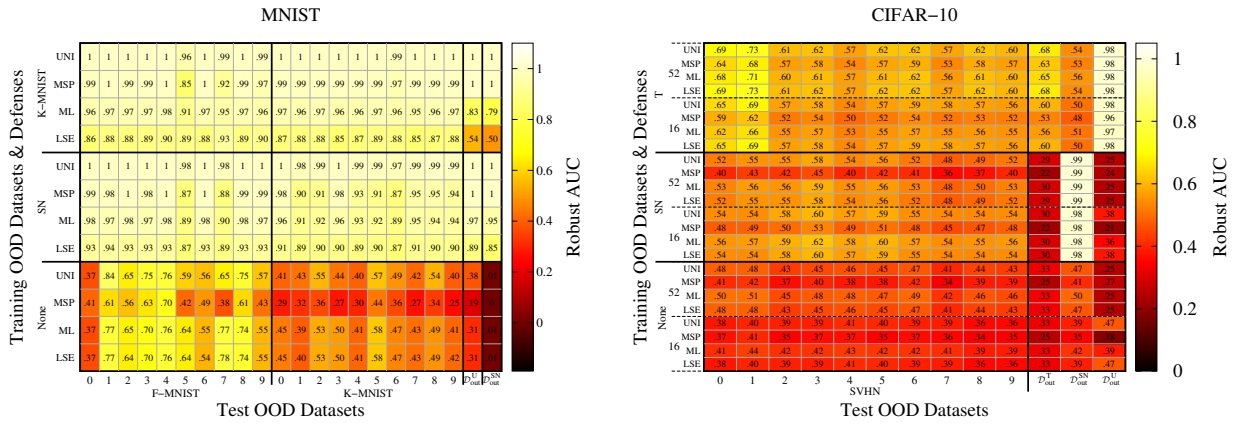


Figure 5: Minimum OOD detection AUC over MNIST and CIFAR-10 under combinations of OOD datasets used during training and detection. The databases used for training are indicated on the vertical axis. The training objectives were ρ_{out}^{uni} in both cases. Under each training OOD dataset, 4 different score functions are indicated that are used for detection. The horizontal axis indicates OOD datasets used for evaluation. The CIFAR-10 plot also includes the smallest and largest network architecture, indicated on the vertical axis.

framework. We found that adding an adversarial OOD objective to the training method does not harm robust in-distribution accuracy, in fact, a significant improvement can be seen in some cases. This indicates that it is always safe to add such an objective.

We also found that it is impossible to pick a score function for robust OOD detection independently of how the model in question was trained. Instead, we get the best results when training and detection is based on the same score function. In other words, while non-robust OOD detection is more robust to the training procedure, in robust OOD detection it is more important to align the detection method with the training method, that is, to use the same score function in both. Also, a similar statement can be formulated in terms of the OOD detection method and the attack on this detection method. The most successful attack is performed using the same score function as the one used by the detection method.

The corresponding part of the dissertation is chapter 5 and related publications are [1] and [5]. The unified treatment of the combined problems, implementation and the design of the related experiments were all done by the author.

The author's publications on the subjects of the thesis

Journal publications

- [1] **István Megyeri**, István Hegedűs, and Márk Jelasity Combining Robust Classification and Robust out-of-Distribution Detection: An Empirical Analysis. In *Progress in Artificial Intelligence*(submitted for publication), 2023.

Full papers in conference proceedings

- [2] **István Megyeri**, István Hegedűs, and Márk Jelasity Adversarial Robustness of Linear Models: Regularization and Dimensionality. In *Proceedings of the 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2019.
- [3] **István Megyeri**, István Hegedűs, and Márk Jelasity Attacking Model Sets with Adversarial Examples. In *Proceedings of the 28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2020.
- [4] **István Megyeri**, István Hegedűs, and Márk Jelasity Adversarial Robustness of Model Sets. In *Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020.
- [5] **István Megyeri**, István Hegedűs, and Márk Jelasity Robust Classification Combined with Robust out-of-Distribution Detection: An Empirical Analysis. In *Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2021.

Further related publications

- [6] Gergely Pap and **István Megyeri** Translational Robustness of Neural Networks Trained for Transcription Factor Binding Site Classification. In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence*, 2019.
- [7] Tibor Csendes, Nándor Balogh, Balázs Bánhelyi, Dániel Zombori, Richárd Tóth, and **István Megyeri** Adversarial Example Free Zones for Specific Inputs and Neural Networks. In *Proceedings of the 11th International Conference on Applied Informatics (ICAI)*, 2020.

- [8] Dániel Zombori, Balázs Bánhelyi, Tibor Csendes, **István Megyeri**, and Márk Jelasity Fooling a Complete Neural Network Verifier. In *The 9th International Conference on Learning Representations (ICLR)*, 2021.
- [9] Ammar Al-Najjar and **István Megyeri** PCA improves the adversarial robustness of neural networks. In *Proceedings of the 30th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2022.

Further publications

- [10] **István Megyeri**, János Csirik, and Zoltán Majó-Petri Utasszámlálás a városi közösségi közlekedésben: mire lehet alkalmas több adat és a „free WiFi”? In *Közlekedéstudományi Konferencia Győr 2019 Conference on Transport Sciences: Alternatív-Autonóm-Kooperatív-Komparatív Mobilitás*, 2019.
- [11] **István Megyeri**, Melinda Katona, and László Nyúl A Novel Approach to Detect Outer Retinal Tubulation Using U-Net in SD-OCT Images. In *15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, 2019.
- [12] Mohammed Mohammed Amin and **István Megyeri** Improving keyword spotting with limited training data using non-sequential data augmentation. In *The 12th Conference of PhD Students in Computer Science (CSCS)*, 2020.
- [13] András Bánhalmi, Vilmos Bilicki, **István Megyeri**, Zoltán Majó-Petri, and János Csirik Extracting Information from Wi-Fi Traffic on Public Transport. In *International Journal of Transport Development and Integration*, 2021.

Other References

- [14] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57. IEEE Computer Society, 2017.
- [15] Ian J. Goodfellow and Jonathon Shlens Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd Intl. Conf. on Learning Representations (ICLR)*, 2015.
- [16] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, June 2016.
- [17] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd Intl. Conf. on Learning Representations (ICLR)*, 2014.

Összefoglalás

A PhD értekezés az ellenséges robusztusság elemzésének különböző alkalmazásait mutatja be a gépi tanulásban. A disszertáció három fő részből áll. A 3. fejezetben lineáris modellek robusztusságát elemezzük a regularizáció és a dimenzionalitás szempontjából. A 4. fejezetben olyan támadási algoritmusokat mutat be, amelyek képesek olyan perturbációt generálni, amelyek egyszerre több modellt is félrevezethetnek. Az 5. fejezetben védekezési módszereket vizsgálunk a robusztus osztályozás és a robusztus outlier detektálás problémájára.

Lineáris modellek ellenséges robusztussága

Számos gépi tanuló modell érzékeny az ellenséges bemenetre, ami azt jelenti, hogy a helyesen osztályozott példákhoz hozzáadott nagyon kicsi, de gondosan megtervezett zaj téves osztályozáshoz vezethet. Ennek okai még mindig tisztázatlanok, még az egyszerű lineáris modellek esetében is. A disszertáció 3. fejezetében a lineáris modelleket vizsgáljuk, és számos új meglátást kínálunk. A regularizáció és a dimenzionalitás hatására összpontosítunk. Megmutatjuk, hogy nagyon nagy dimenziókban az ellenséges robusztusság eredendően alacsony a nagydimenziós terek néhány olyan matematikai tulajdonsága miatt, amelyek eddig kevés figyelmet kaptak. Azt is megmutatjuk, hogy - bár a regularizáció segíthet - az ellenséges robusztusságot nehezebb elérni, mint a nagy pontosságot a tanulási folyamat során. Ezt jellemzően a kutatók figyelmen kívül hagyják, amikor optimalizációs metaparamétereket állítanak be.

Ellenséges támadások modellhalmazok ellen

A gépi tanuló modellek sérülékenyek a nagyon kis bemeneti zavarokkal szemben. A disszertáció 4. fejezetében azt a kérdést vizsgáljuk, hogy a modellek listája által készített előrejelzések listája is tetszőlegesen megváltoztatható-e egyetlen kis perturbációval. Ez nyilvánvalóan nehezebb probléma, mivel egyidejűleg kell több modellt félrevezetni ugyanazzal a perturbációval, ahol a modellekhez rendelt célosztályok eltérhetnek. Ennek a támadásnak többféle alkalmazása is elképzelhető a különböző gyártók által hasonló célra tervezett modellek esetében. Lehet, hogy egyetlen olyan perturbációt szeretnénk, amely minden modellre másképp hat; például csak egy részhalmazt vezethetünk félre, vagy minden modell más-más címkét jósolhat. Az is előfordulhat, hogy olyan perturbációra van szükség, amely minden modellt ugyanúgy vezet félre, és ezáltal egy hordozható perturbációt hoz létre. A jelenlegi megközelítések nem alkalmazhatók közvetlenül erre az általános problémára. A disszertáció 4. fejezetében egy olyan algoritmust javasolunk, amely képes olyan perturbációt találni, amely többféle támadási mintát is kielégít. Például az összes modellnek lehet ugyanaz a célosztálya, vagy különböző véletlenszerű célosztályok, vagy olyan célosztályok, amelyeket úgy terveztek, hogy hogy maximálisan ellentmondásosak legyenek.

A robusztus osztályozás és a robusztus outlier detekció kombinálása

A gépi tanulásban alkalmazott osztályozási modellek gyakran túlságosan magabiztos, de helytelen előrejelzéseket adnak olyan bemeneti mintákra, amelyek nem tartoznak egyik kimeneti osztályba sem. Az ilyen mintákat eloszláson kívüli (outlier) mintáknak nevezzük. Ez a probléma jelentős figyelmet kapott, mivel az ellenséges bemeneti perturbációhoz hasonló sebezhetőséget jelent, amely során a modellek hibás előrejelzéseket tesznek a látszólag eloszláson belüli bemeneti mintákra, amelyek nagyon kicsi, de ellenséges perturbációt tartalmaznak. A diszertáció 5. fejezetében olyan modellek iránt érdeklődünk, amelyek mind az outlier mintákra, mind az ellenségesen perturbált eloszláson belüli mintákra robusztusak. Továbbá megköveteljük, hogy az outlier felismerés robusztus legyen az ellenséges bemeneti perturbációval szemben. Vagyis az outlier minták és az eloszláson belüli minták esetén sem lehetnek olyan ellenséges hatású perturbációk, amelyek miatt azok eloszláson belüli, illetve outlier mintáknak tűnnek. Számos kapcsolódó tanulmány több tervezési lehetőség ad-hoc kombinációját alkalmazza hasonló célok elérése érdekében. A logit vagy softmax réteg felett több függvényt is használhatunk a képzési célok, az outlier felismerési módszerek és az ellenséges támadások meghatározására. A diszertáció 5. fejezetében bemutatunk egy olyan tervezési teret, amely tartalmazza ezen választási lehetőségeket, valamint a hálózatok kiértékelésének elvi módját adja meg. Ez magában foglal egy erős támadási forgatókönyvet, ahol mind az eloszláson belüli, mind az outlier példákat ellenséges módon megzavarják, hogy félrevezessék az outlier észlelést. Ennek a tervezési térnek az empirikus elemzése alapján számos érdekes következtetést vonunk le. A legfontosabb tanulság, hogy a kulcstényező nem az outlier képzési vagy -felismerési módszer önmagában, hanem inkább a megfelelő felismerési és képzési módszerek alkalmazása.

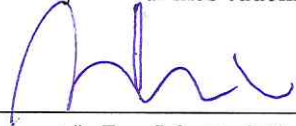
Nyilatkozat

Megyeri István “Applications of Adversarial Robustness Analysis in Machine Learning” című PhD disszertációjában a következő eredményekben Megyeri István hozzájárulása volt a meghatározó:

- A disszertáció 2. fejezetében felhasznált, [1] publikációban megjelent kutatás esetén: kísérletek tervezése, megvalósítása és az eredmények elemzése.
- A disszertáció 3. fejezetében felhasznált, [2, 3] publikációkban megjelent kutatás esetén: a probléma formalizálása, kísérletek megtervezése, megvalósítása és az eredmények elemzése.
- A disszertáció 4. fejezetében felhasznált, [4] publikációban megjelent illetve [5] publikálásra beadott kutatás esetén: a robusztus outlier felismerés és robusztus osztályozás kombinálásának módszertana, kísérletek megtervezése, megvalósítása és az eredmények elemzése.

Ezek az eredmények Megyeri István PhD disszertációján kívül más tudományos fokozat megszerzésére nem használhatók fel.



jelölt: Megyeri István


témavezető: Dr. Jelasity Márk

Az Informatika Doktori Iskola vezetője kijelenti, hogy jelen nyilatkozatot minden társszerzőhöz eljuttatta, és azokkal szemben egyetlen társszerző sem emelt kifogást.

Dátum: 2023.08.25.




vezető: Dr. Jelasity Márk

Hivatkozások

- [1] István Megyeri, István Hegedűs, and Márk Jelasity. Adversarial robustness of linear models: Regularization and dimensionality. In *Proceedings of the 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, Bruges, Belgium, 2019.
- [2] István Megyeri, István Hegedűs, and Márk Jelasity. Adversarial robustness of model sets. In *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020.
- [3] István Megyeri, István Hegedűs, and Márk Jelasity. Attacking model sets with adversarial examples. In *Proceedings of the 28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, Bruges, Belgium, 2020.
- [4] István Megyeri, István Hegedűs, and Márk Jelasity. Robust classification combined with robust out-of-distribution detection: An empirical analysis. In *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021.
- [5] István Megyeri, István Hegedűs, and Márk Jelasity. Combining robust classification and robust out-of-distribution detection: An empirical analysis (under review). In *Progress in Artificial Intelligence*, 2023.

