Robustness of Multi-Label Neural Networks

אימות של רשתות נוירוניות רב-תגיות

Julian Mour
Advisor: Dana Drachsler Cohen

A thesis proposal presented for the degree of
MSc in Computer Science

Technion - Israel Institute of Technology

1 Introduction

A multi-label image classifier assigns to an input image a set of labels to describe its contents. These classifiers are successful in various tasks, such as image tagging [28], object detection [30], and facial expression recognition [35]. However, many works have demonstrated that deep neural networks (DNNs) are susceptible to adversarial example attacks, e.g., [9, 21, 42, 43, 49, 25]. In particular, several works have shown the vulnerability of multi-label image classifiers [33, 14, 40]. These attacks add a small perturbation to a correctly classified input with the goal of causing the network to misclassify. To understand the robustness level of (single-label) image classifiers, many verifiers have been introduced [45, 44, 7, 17, 6]. However, no verifier analyzes the robustness level of multi-label classifiers.

Challenges Part of the challenge is defining what robustness means in multi-label classifiers. For (single-label) classifiers, a popular definition is *local robustness*. At high-level, given an image classifier, an input to the classifier, and a perturbation limit, the classifier is locally robust if perturbing the given input up to the given limit does not change the network's classification. The set of perturbed inputs is called the input's *neighborhood*. In multi-label classification, where inputs are assigned to several labels, local robustness can be defined in various ways. For example, one can require that a network is robust if all labels remain or alternatively require only a subset of labels to remain unchanged. Even given a suitable definition, verifying multi-label classifiers is challenging because they tend to be deeper and more complex than single-label classifiers.

Multi-label robustness In this thesis, we focus on multi-label classifiers that take as input images showing multiple objects. For example, an image showing a road with traffic signs, cars and pedestrians, where the classifier's goal is to detect the objects in the image. For this setting, we propose a new attack model and a corresponding local robustness property. The attack model assumes an attacker can manipulate some objects and their surrounding, with the goal of causing the classifier to miss some target objects (e.g., pedestrians). The corresponding property aims to quantify how large the perturbation of the manipulated objects can be without affecting the target object's classification and how large their manipulated surrounding can be. Namely, the property aims to maximize the perturbation size and the manipulated objects' surrounding area. This definition allows one to understand how changing a specific object in a multi-labeled image affects another's classification in a multi-label classifiers and identify their robustness relation. For simplicity's sake, we assume the image shows two objects, the target object and the perturbed object. Formally, we model the neighborhood by a sequence of epsilons, each corresponds to the maximal allowed perturbation in its respective layer. The first epsilon corresponds to the pixels at the perturbed object, the next one to the pixels immediately surrounding the perturbed object and so on. We further assume a weight vector, assigning a weight for each layer. The goal is to compute the series of epsilons maximizing the weighted sum. This problem is highly challenging for two reasons: It is a multi-dimensional search space and verifying that an epsilon vector belongs to this space (i.e., it represents a robust neighborhood) requires to invoke a (standard) local robustness verifier, which takes a non negligible time.

Key idea To scale the analysis, our key idea is to rely on oracle-guided synthesis [12]. Namely, we propose an algorithm that iteratively expands a given sequence of epsilons, corresponding to a robust neighborhood. At each iteration, an epsilon sequence is submitted to an existing local robustness verifier. Based on its response, we update the epsilon sequence by numerical optimization. To this end, we propose to define gradients from the verifier's output.

Preliminary results In our preliminary research, we implemented a basic version of the above approach. We evaluate it on the DOUBLE-MNIST test dataset [13]. We evaluate our algorithm on three different CNN multi-label DOUBLE-MNIST classifiers that were trained differently: without a defense, with an L_0-based defense [41] and with the PGD defense [1]. Results show that the latter model is the most robust.

Future goals As part of the thesis, we intend to improve our algorithm by reducing the number of queries to the verifier, and thereby shortening the execution time. To this end, we plan to use faster verifiers for some of the queries. We also plan to rely on sensitive layers, to dynamically update the weight vector with the goal of identifying larger robust neighborhoods.

2 Problem Definition

In this section, we define the problem we address. For simplicity's sake, all our definitions and algorithms focus on images with two objects – the target one and the perturbed one – but they can extend to multiple objects. Informally, given a multi-label classifier, an image showing a target object, we aim to compute a robust layer-neighborhood, given for a target object (target class). The neighborhood is defined by a sequence of epsilons representing the perturbation per layer. We begin with definitions and then define our problem.

Multi-label classifier An image multi-label classifier F is a function mapping a two-dimensional image $x \in \mathbb{R}^{n \times m}$ to a score vector over the possible set of classes $C = \{1, \ldots, c\}$, that is: $F : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^{|C|}$. Given an image x and a number k, the classifications F assigns to x is a set $C_{F(x)}$ that is a subset of C of size k corresponding to the highest scores of $F(x)$. We focus on $k = 2$.

Robust neighborhood A neighborhood of input x is a set of inputs $N(x) \subseteq \mathbb{R}^{n \times m}$ containing x. A neighborhood $N(x)$ is robust with respect to a target label $c_t \in C$ if all inputs in it are classified to c_t, that is: $\forall x' \in N(x) : c_t \in C_{F(x')}$.

Layers We focus on layered neighborhoods. To define it, we first define layers. We denote the set of pixels showing an object o in an image x by P_o^x. Given an image x showing two objects o_t and o_{nt}, the target one with class c_t and the perturbed one with class c_{nt} and its set of pixels $P_{o_{nt}}^x$, the dth layer is the set of pixels that their Chebyshev distance (L_∞) from $P_{o_{nt}}^x$ is d. Formally, given two pixels $p = (i, j), p' = (i', j') \in [n] \times [m]$, their distance is $dist(p, p') = ||p - p'||_\infty = \max\{|i - i'|, |j - j'|\}$. Given a pixel $p = (i, j) \in [n] \times [m]$ and a set of pixels $P \subseteq [n] \times [m]$, their distance is the minimum distance of p to any pixel in P: $dist(p, P) = \min\{dist(p, p') \mid p' \in P\}$. Given an image x, an object o, and a distance $dist$, we define the dth layer as: $l_d^{x,o} = \{p \in [n] \times [m] \mid dist(p, P_o^x) = d\}$. Given an image x and a non-target object o_{nt} whose pixel set is P_{nt}, the set of layers is $L_x^{o_{nt}} = \{l_0^{x,o_{nt}}, l_1^{x,o_{nt}}, \ldots, l_r^{x,o_{nt}}\}$, where r is the maximal distance of a pixel in x to P_{nt}.

A layered neighborhood Given an image x, the non-target object's pixels $P_{o_{nt}}^x$ and a series of maximal allowed perturbation for every layer $\epsilon = (\epsilon_0, \ldots, \epsilon_r)$, a layered neighborhood $N_\epsilon^{o_{nt}}(x)$ is the set of all images whose perturbation at layer d is bounded by the respective perturbation limit:

$$N_\epsilon^{o_{nt}}(x) = \{x' \in \mathbb{R}^{n \times m} \mid \forall 0 \leq d \leq r \ \forall (i, j) \in l_d^{x,o_{nt}} : |x'_{i,j} - x_{i,j}| < \epsilon_d\}$$

Given a weight vector $w = (w_0, w_1, \ldots, w_r)$ assigning a weight for each layer, the size of a layered neighborhood $N_\epsilon^{o_{nt}}(x)$ is $||N_\epsilon^{o_{nt}}(x)|| = w \times \epsilon^T = \sum_{d=0}^{r} w_d \cdot \epsilon_d$.

Problem definition Given a classifier $F : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^{|C|}$, an image $x \in [0, 1]^{n \times m}$ containing two objects: o_t and o_{nt} whose classification is c_t and c_{nt}, and a weight vector $w = (w_0, w_1, \ldots, w_r)$, the goal is to compute a sequence of epsilons $\epsilon^* = ({\epsilon_0}^*, {\epsilon_1}^*, \ldots, {\epsilon_r}^*)$ satisfying:

1. F is robust at $N_{\epsilon^*}^{o_{nt}}(x)$ with respect to c_t.
2. For every ϵ' expanding ϵ^*, $N_{\epsilon'}^{o_{nt}}(x)$ is not robust with respect to c_t.
3. $N_{\epsilon^*}^{o_{nt}}(x)$ maximizes its size among all layered neighborhoods meeting (1) and (2).

3 Our Approach

In this thesis, we will design an algorithm that given a multi-label classifier and an image computes the maximal neighborhood given by an epsilon sequence, describing how perturbation of the non-target object affects the classification of the target object. A naive algorithm computes the epsilon series one by one, where at each iteration it computes the maximal epsilon using a binary search. However, this approach is suitable in case the weight vector poses a strict ordering on the importance of the layers, and it also has a high time overhead. Instead, we aim to build on the oracle-guided numerical verification proposed in [15], to obtain a scalable algorithm. Technically, our algorithm has three main components, that iteratively interact with one another:

- A local robustness verifier: Given a classifier and a layered neighborhood $N_\epsilon^o(x)$ of an image x and object o, it determines whether the classifier is robust in this neighborhood.
- A numerical optimizer: Given a classifier and a robust layered neighborhood, it attempts to expand the neighborhood into a larger robust neighborhood, with respect to the weight vector.
- A counterexample-guided inductive synthesiser (CEGIS): Given a classifier and a *non-robust* layer-neighborhood, it attempts to identify directions of the previously robust neighborhood that cannot be further expanded.

The components interact until the neighborhood cannot be further expanded. We next provide details about the different components and explain the open challenges.

The verifier There are many verifiers that can reason about the local robustness of a (single-label) classifier. However, none addresses a multi-label classifier. Thus, part of the challenge is adapting a verifier to multi-label classification. In particular, the verifier only has to prove that one of the labels is the target object's label. In our preliminary research, we rely on the mixed-integer linear program (MILP) based verifier, called MIPVerify [45]. This verifier encodes the robustness task into a MILP maximization problem and uses the Gurobi Optimizer to solve it. We adapt the verifier to multi-label classifiers by adapting the objective function. For a single-label classifier, the objective function is the difference between the highest score and the score of the target class:

$$\max_{c \in C, c \neq c_t} \{F(x')_c - F(x')_{c_t} \mid x' \in N(x)\}$$

If the difference is negative, the neighborhood is robust. Otherwise, it is not robust. We call inputs in $N(x)$ maximizing this objective function the *weakest points*, because they are the closest to the decision boundary.

We adapt this objective function to multi-label classifiers as follows. Since classification is a set of the classes, instead of comparing to the highest score, we compare to the second highest score:

$$2^{nd} \max_{c \in C, c \neq c_t} \{F(x')_c - F(x')_{c_t} \mid x' \in N(x)\}$$

A negative value indicates that the neighborhood is robust with respect to c_t. As before, the inputs maximizing this objective are called the weakest points. These points later help the optimizer to identify robust directions to expand the current neighborhood.

The optimizer Our optimizer follows the approach of [15] and expands a robust neighborhood by computing the gradient of the optimization problem defined in the previous section. Since it is a constrained optimization, we relax the constraints and add equivalent terms to the optimization goal, similarly to [15]. Given the gradients, the optimizer expands the neighborhood by a small step and submits to the verifier.

The CEGIS component The CEGIS component takes a non-robust neighborhood and the previously robust neighborhood and attempts to identify directions that must be shrunk towards making the (non-robust) neighborhood robust. Computing the exact directions requires an exponential number of queries to the verifier. Instead, we shrink the non-robust neighborhood in the direction of the previously robust neighborhood. In our preliminary research, we shrink each layer according to a given weight vector, called *the cutting weight*, which may depend on the input: $cw_x = (cw_0^x, cw_1^x, \ldots, cw_r^x)$. Mathematically, this translates to computing a new epsilon sequence ε_x' that is a weighted average of the current non-robust epsilon sequence ε_x and the previously robust neighborhood ε_x^*. We use *element-wise multiplication* ($\odot$) to multiply each element in the epsilon sequences with its corresponding weight in the weights vectors: ${\varepsilon_x}' = \varepsilon_x \odot cw_x^T + \varepsilon_x^* \odot {cw_x^{-1}}^T$, where $cw_x^{-1} = (1 - cw_0^x, 1 - cw_1^x, \ldots, 1 - cw_r^x)$. We consider two definitions for the cutting weights:

- Fixed weights shrinking the neighborhood more in layers that are far from the non-target object and less in layers that are close to it:

$$cw_x = (\frac{r}{r+1}^5, \frac{r-1}{r+1}^5, \frac{r-2}{r+1}^5, \ldots, 0)$$

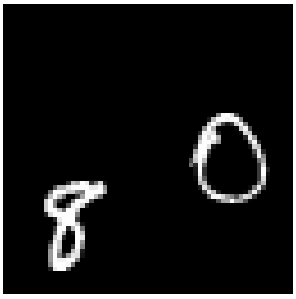

Figure 1: A Double-MNIST sample.

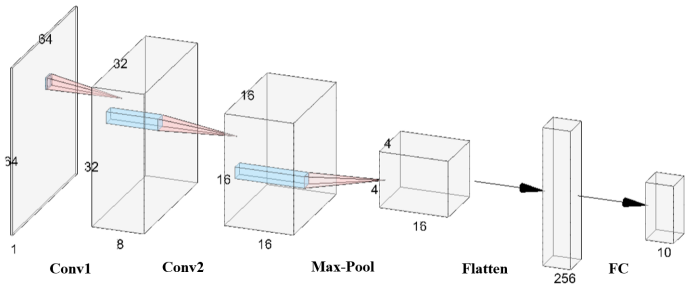

Figure 2: The classifiers' architecture.

- Sensitivity weights shrinking more in layers that are less sensitive to perturbations. The sensitivity weights of an image are computed by using the Vanilla Gradient method, introduced by Simonyan et al. [36]. This method computes the sensitive pixels in an image by computing the gradient of a loss function with respect to the image pixels using backpropagation. Pixels with large gradients are likely to be more sensitive to perturbations, and perturbing them is likely to affect the classifier's robustness more. We translate the problem of finding sensitive pixels to finding sensitive layers by averaging all pixels' gradients in each layer. Accordingly, we define the cutting vector $cw_x = (sw_0^x, sw_1^x, \ldots, sw_r^x)$, where $sw_i^x < sw_j^x$ implies that layer i is more sensitive to perturbations than layer j.

4 Preliminary Results

We evaluated our preliminary approach on the Double-MNIST dataset, consisting of images showing two digits. The multi-classifier's goal is to return the correct two digits. An example of an image is shown in Figure 1. We ran our algorithm on three different CNN multi-label DOUBLE-MNIST classifiers, all with the same architecture (shown in Figure 2) but a different training procedure:

- Without defense.
- With an L_0 defense: This defense relies on the following data augmentation. Before forwarding a training sample to the network, we add random noise to the image in the form of a black rectangle.
- With an L_∞ defense: Using the Projected Gradient Descent (PGD) defense [1]. This defense also involves training the model with adversarial examples, but unlike the L_0 defense, the added perturbations are a small value and can be anywhere in the input.

To compare between the robustness of the models, we ran our algorithm on each of them, for a set of 30 images X. We run our algorithm twice, once for each cutting weights type (fixed and based on sensitivity). We measure the average unweighted neighborhood size defined as $\sum_{i=0}^{r} \epsilon_i{}^*$. We also measure the average execution time for a single image, which is about 75 minutes, but can reach up to 4 hours for big neighborhoods. Figure 3 shows the average neighborhood size for each model and each cutting weights type. The result show that, as expected, the defended models are more robust and the L_∞ defended model is the most robust. This is expected because our attack model focuses on the L_∞ norm. The results further show that using sensitivity weights as cutting weights enable our algorithm to synthesize larger robust neighborhoods.

Figure 4 demonstrates the difference between the types of the cutting weights. For a given image, the undefended classifier and a cutting weight type, we show the heat map corresponding to synthesized layered neighborhood. Figure 5 is similar but for the L_0 defended network. Figure **??** is similar but for the L_∞ defended network. In all models, the heat maps show a slightly bigger neighborhood when using the sensitivity weights method, matching our expectations. We also see bigger epsilons for layers closer to the non-target object, matching the chosen weight vector defining the size.

5 Future Research Objectives

In light of the preliminary work, we aim to further explore our ideas in the following directions:

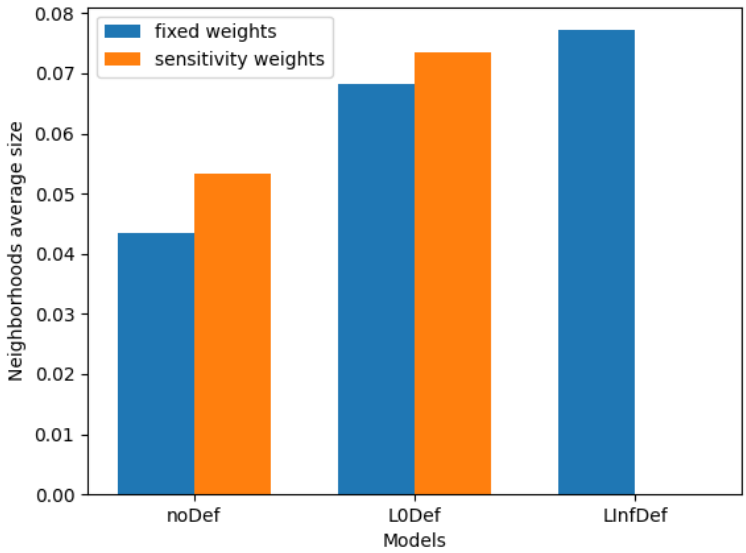

Figure 3: The average size of the synthesized neighborhoods.

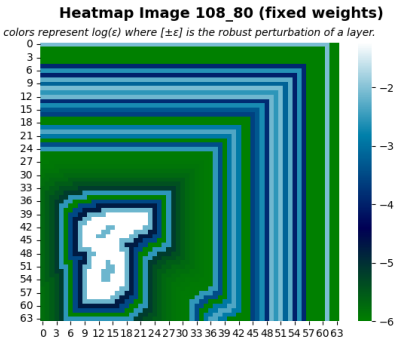

Figure 4. (a): fixed weights

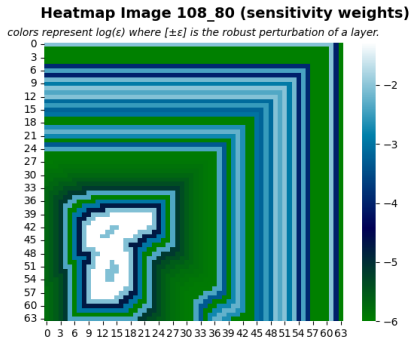

Figure 4. (b): sensitivity weights

Figure 4: No Defense

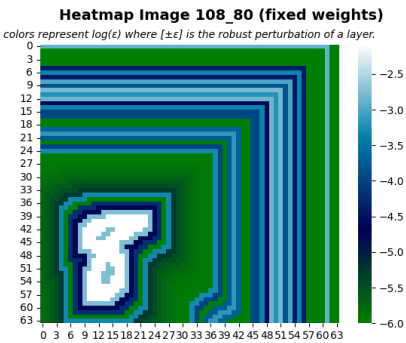

Figure 5. (a): fixed weights

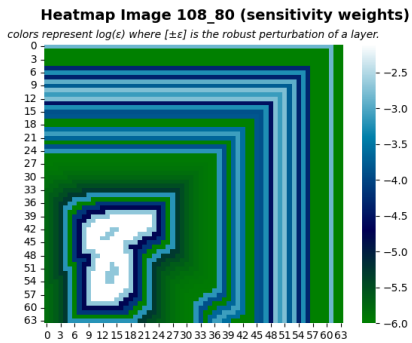

Figure 5. (b): sensitivity weights

Figure 5: L_0 Defense

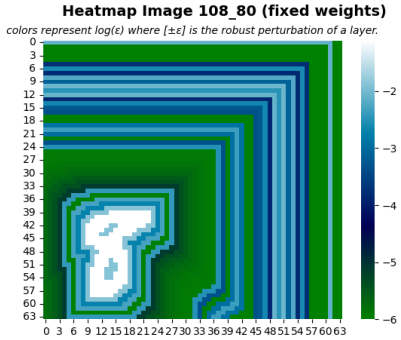

Figure 6. (a): fixed weights

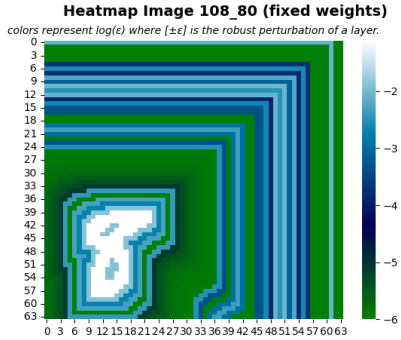

Figure 6. (b): sensitivity weights

Figure 6: L_∞ Defense

- Reduce the number of queries to the MILP verifier, and generally reduce the execution time: A main challenge in our current algorithm is the computation of the weakest points. This idea has been adapted from [15], focusing on single label classifiers. In this setting, this computation involves at most $|C|$ queries per iteration to the MILP solver. In our setting, it involves $|C|^2$ queries per iteration. This leads to a very high execution time. To reduce the number of queries, we plan to use incomplete verifiers (see Section 6.1) in some queries, which are faster. We plan to also reduce the number of queries from $|C|^2$ via mathematical considerations.
- Increase the size of the robust neighborhoods: The size of the returned neighborhood depends on the optimizer and the CEGIS component. To increase the size, we aim to develop optimal cutting weights that minimally shrink a non-robust neighborhood to make it robust. We plan to also investigate ways to identify layers that can expand more at the expense of others, with the goal of increasing the neighborhood size.
- Consider more complex datasets: In our preliminary research, we focus on the Double-MNIST dataset. In our research, we plan to consider more complex datasets, e.g., ones showing road images.
- Infer explanations: Our algorithm finds a relation between two objects in a given image and a given multi-label classifier. We aim to generalize the relations to infer explanations on the robustness level of a multi-label classifier. The explanations can tell us how much and where we can perturb an image without affecting the classification of the target object.

6 Related Work

Our thesis is related to neural network verification, adversarial attacks against multi-label classifiers, and counterexample-guided synthesis.

6.1 Neural Network Verification

Neural network verifiers analyze safety properties of neural networks. These verifiers rely on different mathematical techniques, such as constraint solving and model checking, to analyze the behavior of neural networks and determine whether the given property holds. There are various neural network verification techniques, such as abstract interpretation (e.g. [6, 44]), mixed-integer linear programming (e.g. [45, 39, 22]), over-approximation analysis (e.g. [31, 2, 46]) and in particular over-approximation by linear relaxations (e.g. [47, 3, 8, 37, 34, 38, 29]), simplex (e.g. [18, 19, 5]) and duality (e.g. [32, 4]). Two main categories of verifiers are:

- Complete verifiers: return a definite answer whether a given property, such as neighborhood robustness, holds or not [45, 17]. Such verifiers tend to have long execution time and thus do not scale to large networks. In our preliminary research, we rely on a complete verifier [45].
- Incomplete verifiers: may also return *unknown* for a given property. This allows them to rely on approximate techniques which scale better [44, 7].

6.2 Adversarial Attacks in Multi-Label Classification

In an adversarial attack, an adversary intentionally perturbs an input to the classifier to cause misclassification. In multi-label classification, each input can be assigned multiple classes. Adversarial attacks in this setting can be defined with respect to various goals, such as causing the classifier to predict a different incorrect subset of classes or not predicting a given correct class. We focus on the latter goal, also known as untargeted attack. Song et al. [10] introduce targeted white-box attacks for multi-label classification. They propose a method to exploit label-ranking relationships based framework to attack multi-object ranking algorithms. They approach the problem by formulating it as an optimization problem that meets the attack target while ensuring that the perturbation will not be too obvious, and then execute gradient descent. Through experimentation, they discover that they could manipulate a multi-label classifier into producing any set of labels for a given input by adding an adversarial perturbation. Zhou et al. [50] suggest generating L_∞-norm adversarial perturbations to trick multi-label classifiers. They solve the problem by transforming the optimization problem of finding adversarial perturbations into a linear programming problem, which can be solved efficiently. Yang et al. [48] explore the potential for misclassification risk in multi-label classifiers, particularly in worst-case scenarios. They approach the problem by formulating it as a bi-level set function optimization problem and use random greedy search to find an approximate solution. More recently, Hu et al. [11] present a method to disrupt the top-k labels of multi-label classifiers, based on novel loss functions, under both untargeted and targeted attacks. Melacci et al. [27] suggest a multi-label attack approach with a domain knowledge-constrained classifier, where the domain knowledge is on the relationship of the considered classes. This approach enables the prediction of adversarial examples within the distribution of the training data. Kong et al. [20] were the first to develop a differential evolution (DE) algorithm that can effectively generate multi-label adversarial examples. They achieve a black-box attack that does not need to access model parameters and only uses the model's outputs to generate adversarial examples. Mahmood et al. [26] take into consideration label relationships, modeled by a knowledge graph. They propose a graph-consistent multi-label attack framework, which searches for small image perturbations that lead to misclassifying a desired target set while respecting label hierarchies.

6.3 Counterexample-Guided Synthesis

Counter example guided synthesis (CEGIS) is a technique used in formal verification and program synthesis to generate correct programs or system designs for given specifications. CEGIS iteratively searches for a candidate, checks with an oracle whether the candidate satisfies the specification, and if not relies on counterexamples to refine the search space and guide the synthesis process. In this research, we use CEGIS to recover the search for a robust neighborhood after reaching a non-robust neighborhood. In our context, the specification is a robust layer-neighborhood, the oracle is the verifier, and the counterexamples are the weakest points. Previous work also used CEGIS to find maximal robust neighborhoods [23, 15, 16, 24].

References

[1] Ludwig Schmidt Dimitris Tsipras Adrian Vladu Aleksander Madry, Aleksandar Makelov. Towards deep learning models resistant to adversarial attacks. *ICLR*, 2018.

[2] Greg Anderson, Shankara Pailoor, Isil Dillig, and Swarat Chaudhuri. Optimization and abstraction: A synergistic approach for analyzing neural network robustness. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI 2019, page 731–744, New York, NY, USA, 2019. Association for Computing Machinery.

[3] Akhilan Boopathy, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. Cnn-cert: An efficient framework for certifying robustness of convolutional neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3240–3247, Jul. 2019.

[4] Krishnamurthy Dvijotham, Robert Stanforth, Sven Gowal, Timothy A Mann, and Pushmeet Kohli. A dual approach to scalable verification of deep networks. In *UAI*, volume 1, page 3, 2018.

[5] Yizhak Yisrael Elboher, Justin Gottschlich, and Guy Katz. An abstraction-based framework for neural network verification. In Shuvendu K. Lahiri and Chao Wang, editors, *Computer Aided Verification*, pages 43–65, Cham, 2020. Springer International Publishing.

[6] Markus Püschel Martin Vechev Gagandeep Singh, Timon Gehr. An abstract domain for certifying neural networks. *ACM*, 2019.

[7] Matthew Mirman Markus Püschel Martin Vechev Gagandeep Singh, Timon Gehr. Fast and effective robustness certification. *NeurIPS*, 2018.

[8] Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2018.

[9] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *In ICLR*, 2015.

[10] Qingquan Song; Haifeng Jin; Xiao Huang; Xia Hu. Multi-label adversarial perturbations. *IEEE*, 2018.

[11] Shu Hu, Lipeng Ke, Xin Wang, and Siwei Lyu. Tkml-ap: Adversarial attacks to top-k multi-label learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7649–7657, October 2021.

[12] Gulwani S. Seshia S.A. Tiwari A. Jha, S. Oracle-guided component-based program synthesis. *ICSE*, 2010.

[13] Gulwani S. Seshia S.A. Tiwari A.Sara Sabour Nicholas Frosst Geoffrey E. Hinton Jha, S. Dynamic routing between capsules. *NeurIPS*, 2017.

[14] Neil Gong Jinyuan Jia, Wenjie Qu. Multiguard: Provably robust multi-label classification against adversarial examples. *Advances in Neural Information Processing Systems*, 2022.

[15] Anan Kabaha and Dana Drachsler Cohen. Maximal robust neural network specifications via oracle-guided numerical optimization. *VMCAI*, 2023.

[16] Anan Kabaha and Dana Drachsler-Cohen. Boosting robustness verification of semantic feature neighborhoods. In Gagandeep Singh and Caterina Urban, editors, *Static Analysis*, pages 299–324, Cham, 2022. Springer Nature Switzerland.

[17] Shiqi Wang Yihan Wang Suman Jana Xue Lin Cho-Jui Hsieh Kaidi Xu, Huan Zhang. Fast and complete: Enabling complete neural network verification with rapid and massively parallel incomplete verifiers, 11 2020.

[18] Guy Katz, Clark Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In Rupak Majumdar and Viktor Kunčak, editors, *Computer Aided Verification*, pages 97–117, Cham, 2017. Springer International Publishing.

[19] Guy Katz, Derek A. Huang, Duligur Ibeling, Kyle Julian, Christopher Lazarus, Rachel Lim, Parth Shah, Shantanu Thakoor, Haoze Wu, Aleksandar Zeljić, David L. Dill, Mykel J. Kochenderfer, and Clark Barrett. The marabou framework for verification and analysis of deep neural networks. In Isil Dillig and Serdar Tasiran, editors, *Computer Aided Verification*, pages 443–452, Cham, 2019. Springer International Publishing.

[20] Linghao Kong, Wenjian Luo, Hongwei Zhang, Yang Liu, and Yuhui Shi. Evolutionary multi-label adversarial examples: An effective black-box attack. *IEEE Transactions on Artificial Intelligence*, pages 1–12, 2022.

[21] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *In ICLR*, 2017.

[22] Christopher Lazarus and Mykel J. Kochenderfer. A mixed integer programming approach for verifying properties of binarized neural networks, 2022.

[23] Changjiang Li, Shouling Ji, Haiqin Weng, Bo Li, Jie Shi, Raheem Beyah, Shanqing Guo, Zonghui Wang, and Ting Wang. Towards certifying the asymmetric robustness for neural networks: Quantification and applications. *IEEE Transactions on Dependable and Secure Computing*, 19(6):3987–4001, 2022.

[24] Chen Liu, Ryota Tomioka, and Volkan Cevher. On certifying non-uniform bounds against adversarial attacks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4072–4081. PMLR, 09–15 Jun 2019.

[25] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *In ICLR*, 2018.

[26] Hassan Mahmood and Ehsan Elhamifar. Towards effective multi-label recognition attacks via knowledge graph consistency, 2022.

[27] Stefano Melacci, Gabriele Ciravegna, Angelo Sotgiu, Ambra Demontis, Battista Biggio, Marco Gori, and Fabio Roli. Can Domain Knowledge Alleviate Adversarial Attacks in Multi-Label Classifiers? working paper or preprint, October 2020.

[28] Kilian Weinberger Minmin Chen, Alice Zheng. Fast image tagging. *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[29] Christoph Müller, François Serre, Gagandeep Singh, Markus Püschel, and Martin Vechev. Scaling polyhedral neural network verification on gpus. In A. Smola, A. Dimakis, and I. Stoica, editors, *Proceedings of Machine Learning and Systems*, volume 3, pages 733–746, 2021.

[30] Poggio T Papageorgiou, C. A trainable system for object detection. *International Journal of Computer Vision*, 2000.

[31] Chongli Qin, Krishnamurthy, Dvijotham, Brendan O'Donoghue, Rudy Bunel, Robert Stanforth, Sven Gowal, Jonathan Uesato, Grzegorz Swirszcz, and Pushmeet Kohli. Verification of non-linear specifications for neural networks, 2019.

[32] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples, 2020.

[33] Bernhard Schölkopf Rohit Babbar. Adversarial extreme multi-label classification, 2018.

[34] Hadi Salman, Greg Yang, Huan Zhang, Cho-Jui Hsieh, and Pengchuan Zhang. A convex relaxation barrier to tight robustness verification of neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[35] Xiaojiang Peng Jianfei Yang Zhaoyang Zeng Yu Qiao Sijie Ji, Kai Wang. Multiple transfer learning and multi-label balanced training strategies for facial au detection in the wild. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020.

[36] Andrea Vedaldi Simonyan, Karen and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps., 2013.

[37] Gagandeep Singh, Rupanshu Ganvir, Markus Püschel, and Martin Vechev. Beyond the single neuron convex barrier for neural network certification. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[38] Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. An abstract domain for certifying neural networks. *Proc. ACM Program. Lang.*, 3(POPL), jan 2019.

[39] Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. Robustness certification with refinement. In *International Conference on Learning Representations*, 2019.

[40] Angelo Sotgiu Ambra Demontis Battista Biggio Marco Gori Stefano Melacci, Gabriele Ciravegna. Domain knowledge alleviates adversarial attacks in multi-label classifiers. *IEEE*, 2022.

[41] Bernt Schiele Sukrut Rao, David Stutz. Adversarial training against location-optimized adversarial patches. *ECCV*, 2020.

[42] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014.

[43] Pedro Tabacof and Eduardo Valle. Exploring the space of adversarial images. *In IJCNN*, 2016.

[44] Timon Gehr; Matthew Mirman; Dana Drachsler-Cohen; Petar Tsankov; Swarat Chaudhuri; Martin Vechev. Safety and robustness certification of neural networks with abstract interpretation. *IEEE*, 2018.

[45] Russ Tedrake Vincent Tjeng, Kai Xiao. Evaluating robustness of neural networks with mixed integer programming. *ICLR*, 2019.

[46] Shiqi Wang, Kexin Pei, Justin Whitehouse, Junfeng Yang, and Suman Jana. Efficient formal safety analysis of neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[47] Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J. Zico Kolter. Beta-crown: Efficient bound propagation with per-neuron split constraints for neural network robustness verification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29909–29921. Curran Associates, Inc., 2021.

[48] Zhuo Yang, Yufei Han, and Xiangliang Zhang. Characterizing the evasion attackability of multi-label classifiers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10647–10655, May 2021.

[49] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *In IEEE Trans. Neural Networks Learn. Syst*, 2019.

[50] Nan Zhou, Wenjian Luo, Xin Lin, Peilan Xu, and Zhenya Zhang. Generating multi-label adversarial examples by linear programming. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020.