Does Knowing More Make You Easier to Trick? Adversarial Robustness of Multi-target Regression

Soyon Choi

Amherst College
sochoi25@amherst.edu

Abstract

Following the rapid rise of deep learning (DL) and generative artificial intelligence (GenAI), it is imperative that we gain a better understanding of *how* these machine learning (ML) systems actually learn. What information are DL models retaining from the training data? What reasoning capabilities do these models have? In my proposed project, I aim to tackle these pressing questions through use of an adversarial lens.

Introduction

Large deep learning (DL) models and generative artificial intelligence (GenAI) display emergent behavior — the ability to perform well on unseen tasks. One way to better understand how these complex systems work is to understand how they *break*. Despite their impressive performance capabilities, ML models are surprisingly easy to trick with carefully crafted perturbations in input data. To learn how to defend ML models against adversaries, I distill the problem down into one fundamental quality of many DL systems — the multi-dimensionality of target data. Then, I conduct a controlled empirical study and determine a direction for mathematical analysis to answer the following question: Does increasing the dimensionality of the target make a model more or less susceptible to attack?

Background

ChatGPT and other LLMs are trained to compose English text in response to English text input by a user. However, they can somehow also play a decent game of checkers. This *emergent behavior* points us towards an interesting idea: A system with more knowledge of the world performs better at any given task than a system with less context.

This phenomenon is not unstudied — it is known that the performance of modern DL systems is drastically improved by increasing scale (i.e., training compute and number of model parameters) (Brown et al. 2020). Models of greater scale have also been shown to showcase emergent behaviors (Wei et al. 2022). There have been considerable efforts toward *at-scale* empirical investigation of these large DL models, simply showing how well they perform as they get bigger. However, there is not enough we actually understand about *why* these large-scale models perform so well, even on previously unseen tasks. This has prompted a greater discussion about *how* large ML models work; in particular, we need to understand *what* they actually learn.

We select one aspect of these large GenAI models — *multi-targetness* — and try to understand its effects on learning. To properly address the interconnected nature of a multi-target dataset, we use specialized multi-target regression techniques, which are based on the key idea that output features must be tied together during learning. To encode this dependency between targets, multi-target regressors often narrow the field of vision of the learner. For example, we can limit the rank of the parameter matrix (Izenman 1975), apply a shrinking matrix (Van Der Merwe and Zidek 1980), or add a special regularization term (Similä and Tikka 2007).

This setting is ideal for borrowing the well-studied paradigm of adversarial machine learning (AML). Adversarial attacks on single-target linear predictors are well-studied in various contexts (Alfeld, Zhu, and Barford 2016; Moosavi-Dezfooli, Fawzi, and Frossard 2015). By attacking various multi-target regressors, we can effectively focus in on the strangeness of multi-dimensional target data by observing how increasing the dimensionality of the target changes the security ramifications of the model.

Prior Work by the Applicant. In one of my former research projects, I expanded on recent work by Microsoft Research, in which controlled experiments were conducted using synthetic data to determine how Transformers learn to reason (Zhang et al. 2022). They created a "learning equalities" dataset composed of strings of equalities (e.g., '$a = +c; b = a; c = -1$'), which allowed them to study Transformers' ability to *associate* (i.e., see that a is the same variable, no matter its location within the input string) and *manipulate* (i.e., deduce $a = -1$), two core components of reasoning. In my work, I built on this idea by designing a new synthetic dataset of a more complex task — finding the shortest path given a graph input — in order to observe how the model weights change throughout training.

In addition, I took an independent study course on AML, in which I did an in-depth investigation on how to attack and defend various ML models. My proposed work aims to explore multi-target regressors and evaluate their robustness to test-time adversaries in a similar fashion, using controlled experimentation and mathematical analysis.

Approach

To highlight the effect of multi-dimensional targets on the robustness of AI, we distill the problem down to two dimensions. One learner makes a one-dimensional prediction y, and the other makes a two-dimensional prediction $\mathbf{y} = [y_1\ y_2]^\top$. Which learner is more susceptible to adversarial attack? To tackle this question, we borrow a well-understood setting, where the optimal attack is computable — deployment-time attacks against linear models.

Consider a deployment-time adversarial attractive attack on a multi-target regressor, in which the attacker perturbs data after the learning has taken place. The defender's model is linear; given an input vector $\mathbf{x}$, it applies a weight matrix $\mathbf{W}$ and a bias vector $\mathbf{b}$ to output a prediction $\mathbf{y} = \mathbf{W}^\top\mathbf{x}+\mathbf{b}$. The attacker has knowledge of the model and a data point $(\mathbf{x}, \mathbf{y})$, where $\mathbf{y} = [y_1\ y_2]^\top$. It aims to determine the best $\boldsymbol{\delta}$ to perturb $\mathbf{x}$ with such that it affects the defender's prediction $\hat{\mathbf{y}}$ at $\hat{\mathbf{x}} = \mathbf{x} + \boldsymbol{\delta}$ as desired. In an attractive attack, the attacker's goal is to get the defender's prediction $\hat{\mathbf{y}} = \mathbf{W}^\top\hat{\mathbf{x}}+\mathbf{b}$ as close as possible to their target output $\mathbf{y}^\dagger$.

Specifically, consider the case in which the attacker's target value is $\mathbf{y}^\dagger = [y_1^\dagger\ y_2]^\top$, where y_2 remains unchanged from the original point $\mathbf{y}$. This specially crafted adversarial task explicitly brings an interesting phenomenon to light. Unlike the single-target setting, attacking multi-target regressors means that perturbations on y_1 may also inadvertently affect y_2, and vice versa. Hence, investigating the optimal attack $\boldsymbol{\delta}$ and the resulting prediction $\hat{\mathbf{y}}$ on this particular task will shed light on how multi-target systems react to adversaries in contrast to single-target systems.

In this linear setting, we can compute the optimal attack $\boldsymbol{\delta}$ on multi-target regression tactics such as reduced-rank regression (Izenman 1975), filtered canonical Y-variate regression (FICYREG) (Van Der Merwe and Zidek 1980), and simultaneous variable selection (SVS) (Similä and Tikka 2007). Then, we can empirically evaluate the attacker's performance on different multi-target regressors by calculating the distance of the attacked output $\hat{\mathbf{y}}$ from the target $\mathbf{y}^\dagger$.

These findings could inform further mathematical analysis on how multi-targetness affects adversarial robustness. For example, reduced rank regression puts a constraint on the rank of the weight matrix W. Similarly, SVS puts a constraint on the row-wise L_2 norm of W. Given that rank and row-wise L_2 norm are terms that capture the ability of the model to predict multiple outputs in relation to one another, we can relate them to the spectral norm, which captures the vulnerability of a linear model to test-time adversarial attack (Alfeld, Zhu, and Barford 2017). This formalizes the relationship between multi-targetness and test-time adversarial robustness.

Evaluation

Multi-target regressors typically outperform independent single-target regressors in practice. However, how the adversarial robustness of these models compare is unknown. One hypothesis is as follows: The ability of multi-target regressors to reason about target features in relation to one another results in higher performance but lower robustness. My project seeks to support or reject this hypothesis in the smaller-scale realm of multi-output linear regression. By doing so, we can provide a basis for understanding the behavior of larger models.

Moreover, formalization of the relationship between multi-targetness and adversarial robustness will, ideally, result in formal robustness bounds on multi-target regressors. I intend to construct proofs regarding the relationship between the mathematical terms at play, such as spectral norm, rank, or the row-wise L_2 norm of W. We aim to prove formal guarantees on the security of multi-target regressors.

Discussion

The results of this study will not only directly reveal how to ensure the security of multi-target regressors against test-time adversaries, but also demonstrate what multi-dimensionality of target data contributes to learning. In particular, I aim to gain a foundational understanding of something fundamental to the success of deep learning: how a broader understanding of the world results in a model that can perform well on a wider range of unspecified tasks. Multi-target regression methods provide a mathematically tractable basis for answering this question.

Consider one potential outcome of this study: Single-target linear regressors are more robust than multi-target regressors against attractive deployment-time attacks. This would show that knowing more makes you easier to trick — the attacker simply has more avenues of attack. Another potential result is that multi-target regressors are more robust than single-target regressors. This would suggest that multi-target systems like modern GenAI models not only learn the task they are trained on, but also gain a more comprehensive understanding of the world and therefore are able to detect and defend against an adversarial example $\hat{x}$. In either case, we gain a fundamental insight about multi-output learners.

Conclusion

Despite the prevalence of various ML models, there is a concerning lack of understanding about modern DL. With the rapid growth of GenAI, we have seen the rise of many large neural networks that are designed to simply work well; unfortunately, no one can fully explain what these models actually learn. Without a deeper understanding of DL learners, we are limited in how much we can intentionally improve these models in their performance, security, and more.

In the proposed work, I utilize a bottom-up approach to get one step closer to truly understanding the magic of DL. I specify one aspect — multi-targetness — of these large DL models, and utilize the adversarial approach to take apart and analyze the multi-target linear regression problem. Through controlled experimentation and mathematical analysis, I aim to defend multi-target linear regressors against test-time attacks, formalize their robustness, and analyze their behavior in comparison to single-target regressors.

References

Alfeld, S.; Zhu, X.; and Barford, P. 2016. Data Poisoning Attacks against Autoregressive Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Alfeld, S.; Zhu, X.; and Barford, P. 2017. Explicit Defense Actions Against Test-Set Attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Izenman, A. J. 1975. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2): 248–264.

Moosavi-Dezfooli, S.; Fawzi, A.; and Frossard, P. 2015. DeepFool: a simple and accurate method to fool deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, abs/1511.04599.

Similä, T.; and Tikka, J. 2007. Input Selection and Shrinkage in Multiresponse Linear Regression. *Computational Statistics Data Analysis*, 52: 406–422.

Van Der Merwe, A.; and Zidek, J. V. 1980. Multivariate regression analysis and canonical variates. *Canadian Journal of Statistics*, 8(1): 27–39.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; Chi, E. H.; Hashimoto, T.; Vinyals, O.; Liang, P.; Dean, J.; and Fedus, W. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning*.

Zhang, Y.; Backurs, A.; Bubeck, S.; Eldan, R.; Gunasekar, S.; and Wagner, T. 2022. Unveiling Transformers with LEGO: a synthetic reasoning task.