Exploring Fragility in Deep Neural Networks: A Comparative Analysis of Depth and Shallow Architectures

Anthony Piacentini

June 10, 2024

Abstract

Deep neural networks (DNNs) have demonstrated exceptional performance in diverse fields such as computer vision, natural language processing, healthcare, and finance, owing to their ability to process and interpret vast amounts of data. However, increasing the depth of these networks introduces a phenomenon known as fragility, where minor input variations can significantly alter the network's output. This fragility raises concerns about the reliability and robustness of deep learning models in real-world applications. While existing literature acknowledges the fragility of deep neural networks, a comprehensive explanation for the variability in fragility among networks with different depths remains elusive. By investigating the intricate distinctions in vulnerability between deep and shallow neural networks, this study seeks to identify the underlying factors contributing to their varying levels of fragility. The findings aim to enhance our understanding of deep learning models' behavior and pave the way for developing more resilient neural network architectures.

Contents

1 Introduction

1.1 Background

Deep neural networks (DNNs) have revolutionized various industries by achieving unprecedented levels of performance in tasks that involve complex data processing. In fields such as computer vision, DNNs have enabled significant advancements in image recognition and object detection, while in natural language processing, they have improved the accuracy of machine translation and sentiment analysis. Additionally, DNNs have been instrumental in healthcare for tasks like medical image analysis and predictive diagnostics, and in finance for applications such as fraud detection and algorithmic trading.

The success of DNNs can be attributed to their capability to model intricate patterns and relationships within large datasets. By utilizing multiple layers of interconnected neurons, these networks can learn hierarchical representations of data, which enhances their ability to generalize from training data to unseen examples. However, as the number of layers in a neural network increases, a phenomenon known as fragility becomes more pronounced. Fragility refers to the sensitivity of neural networks to small perturbations in input data, which can lead to significant deviations in the output.

The issue of fragility has important implications for the deployment of DNNs in real-world settings. While neural networks perform well in controlled environments, their susceptibility to minor input variations raises questions about their reliability and robustness when exposed to diverse and unpredictable data in practice. This concern is particularly relevant in safety-critical applications, such as autonomous driving and healthcare, where erroneous outputs could have serious consequences.

Despite widespread recognition of the fragility problem, there is a lack of comprehensive understanding regarding the factors that contribute to this vulnerability, especially when comparing deep and shallow networks. Previous research has indicated that deeper networks tend to exhibit greater fragility, but the reasons behind this trend are not fully understood. Moreover, the specific characteristics of different network architectures, such as DenseNet, Inception v4, ResNet, and VGG-Net, and their influence on fragility have not been thoroughly investigated.

To address this gap, this research aims to systematically analyze the fragility of various deep learning models by implementing and training them on standard image datasets, including CIFAR and ImageNet. The study will involve introducing controlled distortions to the input data and examining how these perturbations affect the performance of different network architectures. Additionally, the research will explore the impact of modifying internal parameters, such as BatchNorm dimensions, on the robustness of these models.

1.2 Overview

Deep neural networks (DNNs) have been applied in many different industries such as computer vision and natural language processing, healthcare, and finance in recent years. The remarkable performance of these models could be explained by the fact that they can understand as well as interpret large volumes of information. Nevertheless, when the neural networks are made deeper; they manifest an interesting property called fragility. In neural networks, fragility is the concept that emphasizes how these networks are likely to be adversely affected by even minor variations in their input data thereby significantly altering the output of such networks. This has consequently made people question the dependability and resilience levels held by deep learning models when used outside laboratory trials using simulations alone. Though acknowledged in the literature that deep neural networks are fragile, there is no clear universal explanation as to why it becomes different among networks that have a lot of layers. The plan for this research is to set up multiple types of deep learning algorithms, such as the densenet, inception v4, resnet, and vgg-net using pytorch. Then train these models using different image databases, such as cifar and imagenet. Similar to the Rethinking data augmentation for adversarial robustness by Eghbal-zadeh et al, I plan on using different levels of distortion of the image libraries and testing each algorithm. Then I plan on seeing how changing the BatchNorm dimensions, or similar qualities for other algorithms, affects how effective the algorithm is at identifying images. Finally I will use the testing algorithms for these algorithms defined in Automated whitebox testing of deep learning systems by Pei et al to test these algorithms. This research looks at the intricate distinctions in vulnerability between deep neural networks that have numerous layers and those that are shallower, bridging the gap. The study purports to uncover the factors that account for the different profiles of precariousness shown by these systems by investigating the causes of their behavior.

1.3 Significance

Understanding the fragility of deep neural networks (DNNs) is a daunting task of significant importance, with major implications for the development and practical deployment of artificial intelligence (AI) throughout a wide variety of industries and applications. The core of this research is to improve the reliability, stability, and security of AI deployment in the real world. Despite achieving state-of-the-art performance in various above-mentioned tasks of extreme importance, such as image recognition, natural language and medical diagnostics, and financial forecasting, deep neural networks are not immune to the complexities and uncertainties of real-world data. Fragility is the property of susceptibility to small variations or perturbations of the input data, in which undetectable changes can result in dramatic changes in the network's output. This not only provides a lack of trustworthiness in AI systems, but also poses formidable challenges in terms of ensuring their resilience to adversarial attacks, environmental noise, data drift, and other sources of uncertainty. Thus, uncovering the root mechanisms and driving forces of fragility is key to the process of fortifying AI against these threats and instilling more profound levels of trust and confidence in their performance and capacities. By systematically probing and dissecting the contributing factors to fragility in deep neural networks, researchers can therefore gain invaluable insights into how these complex systems work. From architectural design choices and training methodologies to data preprocessing techniques and regularization strategies, every part of the deep learning pipeline critically impacts the network's robustness and generalization capabilities. In addition, the interaction between network depth, parameterization, activation functions, and optimization algorithms adds another layer of complexity to the problem, requiring a detailed understanding of their influence on the network's fragility profile. Patterns, trends, and trade-offs in the ways that deep neural networks respond to varied inputs and stimuli will be strongly illuminated by thorough experimentation's and analyses.

The practical implications of this research simultaneously transcend all the known boundaries of scholarship, to permeate virtually every aspect of modern society where AI technologies are scripted to make a trans-formative impact. In healthcare, for example, the reliability and robustness of AI-driven diagnostic systems are very critical, as an erroneous or misleading prediction can have profound consequences on both patient outcomes and the selection of treatments. Now, this is equally critical in finance, because AI algorithms are more and more at the epicenter of risk assessment and fraud detection, and actual trading is a part of the system, with adversarial manipulation and un-predicted properties of the market being able to potentially harm the market's integrity and investor confidence. In safety-critical domains such as autonomous vehicles and aerospace engineering, where AI acts as the bedrock of decision and control systems, the stakes are particularly high, hence the pressing need for fully robust and resilient AI solutions that can operate safely and reliably under diverse operational conditions and environmental uncertainties.

The importance of knowing fragility in deep neural networks is the thing related to imparting knowledge, tools, and methodologies that, in turn, should let researchers, practitioners, and policymakers harness the trans-formative potential of AI while keeping inherent risks and uncertainties at bay. In this sense, going from theoretical insights to practical implementations—this research not only advances the boundaries of AI science and engineering but also nurtures a culture of responsible innovation and ethical stewardship, ensuring that AI technologies advance for the common good of humanity and a better, more prosperous, just, and sustainable future.

1.4 Motivation

The great achievements made by deep neural networks (DNNs) have heralded new industrial revolutions and opened up unimaginable possibilities in image recognition, natural language processing, medical diagnosis, and financial forecasting. These unprecedentedly powerful algorithms have gone beyond human-level performance on problems that were considered impossible just a few years ago. But behind this success lies an enormous challenge: DNNs are also notoriously fragile. The susceptibility of neural networks to drastic performance degradation concerning minor perturbations or variations in the input has astounding consequences for the widespread deployment of these methods in applications bandied about under the umbrella of AI. Despite their remarkable accuracy and efficiency in controlled laboratory settings, DNNs can perform poorly when faced with the challenges and uncertainties associated with real-world datasets. This makes them sensitive to even the smallest input deviations or distortions that can lead to highly significant changes in the network output behavior, thereby undermining its reliability, robustness, and trustworthiness in operational use. This fragility compromises the integrity not only of the AI-driven systems but also the acceptance and take-up in those sectors where safety and reliability are paramount. Addressing, therefore, the issue of fragility is crucial to turn the full potential of DNNs into a reality, affecting and modifying society at large.

1.5 Purpose

With this research, one aims to explore the phenomenon of fragility found in very deep models of deep neural networks, due to which even slight perturbations in the input data are played out in the output manifold. This paper aims to compare and contrast fragility profiles across different architectures of DNNs based on a very systematically detailed and thorough analysis, providing commentary on key characteristics and mechanisms that affect the overall model robustness concerning fragility. Thus, through the elucidation of these factors, the research would like to characterize the fragility that distinguishes between models, providing a benchmark for their robustness concerning a range of input perturbations. Importantly, this final objective of the study is to draw reliable and actionable insights that could guide the development and application of DNNs that are more robust and reliable in practice.

2 Background and Related Work

2.1 Related work

This study is the extension of several initial researches. It involves combining numerous deep learning architectures and systematically examining them given different data exaggerations as well as tampering levels. With methodologies from the following works involved, I aspire to make clear how diverse augmentation techniques and hyperparameter configurations including but not limited to BatchNorm dimensions influence on robustness and efficiency of deep machine models. The comprehensive approach mentioned will offer valuable insights into an area that has been not well-researched yet. This area is that of coming up with deep learning systems which are more resilient and effective. In their paper on “Fragility, robustness, and antifragility in deep learning”, Pravin et al. (2018) examined the ideas of breakability, strength, and anti-strength. They talked about how deep learning models can be developed and trained to not just be resistant against adversarial attacks, but also benefit from them. Our study is centered on this idea of antifragility as it applies to models that get more powerful when they face disturbances. Many people have looked at data augmentation as a way of improving the robustness of deep learning models against adversarial examples. In the paper by Eghbal-Zadeh et al., (2019) they talked about how important it is to use different data augmentation methods and rethinking data augmentation for adversarial robustness. Their research findings showed that when you use different augmentation techniques on one image you will always have a better-performing model which is less susceptible to attacks that were initially classified as adversarial samples. The purpose of this study is to emphasize the importance of investigating various levels of distortion in image datasets, a principle consistent with our practice from varying extents of deformities on CIFAR as well as ImageNet databases aimed at assessing how well diverse neural networks can perform under specified conditions.

Appendices

A References

1. Eghbal-Zadeh, H., Zellinger, W., Pintor, M., Grosse, K., Koutini, K., Moser, B., Biggio, B., Widmer, G. (2024). Rethinking data augmentation for adversarial robustness. Information Sciences, 654, 119838. https://doi.org/10.1016/j.ins.2023.119838

2. Fikri, F. B., Oflazer, K., Yanıkoğlu, B. (2023). Abstractive summarization with deep reinforcement learning using semantic similarity rewards. Natural Language Engineering, 1–23. https://doi.org/10.1017/s135132492

3. Gao, X., Saha, R. K., Prasad, M. V. N. K., Roychoudhury, A. (2020). Fuzz testing based data augmentation to improve robustness of deep neural networks. ACM International Conference on Software Engineering. https://doi.org/10.1145/3377811.3380415

4. Pravin, C., Martino, I., Nicosia, G., Ojha, V. K. (2024). Fragility, robustness and antifragility in deep learning. Artificial Intelligence, 327, 104060.

5. Tian, Y., Pei, K., Jana, S., Ray, B. (2018). DeepTest. ACM Symposium on Operating Systems Principles. https://doi.org/10.1145/3180155.3180220

6. Wang, H., Miahi, E., White, M., Machado, M., Abbas, Z., Kumaraswamy, R., Liu, V., White, A. (2024). Investigating the properties of neural network representations in reinforcement learning. Artificial Intelligence, 104100. https://doi.org/10.1016/j.artint.2024.104100