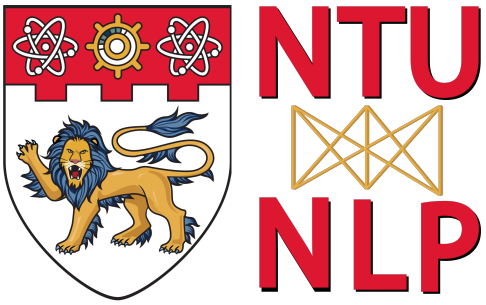
Risks of Language Models



Subtitle

Han Cheol Moon
School of Computer Science and Engineering
Nanyang Technological University
Singapore
hancheol001@edu.ntu.edu.sg
October 16, 2023

Contents

Chapter 1

Introduction

1.1 Introduction

In 2012, a visual object classification system called AlexNet stood as the winner of the ImageNet challenge with a dominating performance AlexNet2012. It achieved a top-5 error of 15.3%, which is more than 10% lower than that of the runner-up. Inspired by the monumental achievement, there has been impressive progress in machine learning research. Conventional machine learning approaches tend to show limitations in processing natural data since it typically involves careful feature engineering based on substantial domain expertise lecun2015deep. Instead, a learning scheme, which enables the AlexNet, automates the overall feature extraction process. It automatically discovers salient representations of the raw data through a hierarchical learning process without any manual feature engineering. Due to the deep hierarchical structure of the learning scheme, it is often referred to as *deep learning* (DL) or *representation learning*.

The idea of building networks with more depth has revolutionized a variety of challenging tasks in machine learning, such as image generation rombach2021highresolution and natural language processing (NLP) wei2022chain,wei2022finetuned. Naturally, DL-based applications are rapidly becoming part of our everyday lives. For example, DL systems help to discover personalized items and prepare for rainy weather $10.1145/3285029, Ravuri_2021. Deeplearningcontinuestosurpriseuswithitsendlesspossi$

The monumental achievements of DL systems seem to guarantee the absolute superiority and robustness of modern DL systems. However, DL systems have shown significant vulnerability to samples specifically crafted to misguide them, namely *adversarial examples* Szegedy2014,goodfellow2015explaining,Dal The adversarial examples are seemingly indistinguishable from original inputs, but they are imperceptibly perturbed to cause misbehavior of the systems. This confronts us with challenging questions regarding their analysis and interpretation. In response, various works attempt to elucidate the underlying causes of the brittleness. Notably, Ilyas2019 claimed that the brittleness is the result of the discrepancy between an objective we aim to achieve from the DL system and what we implicitly expect from the system throughout the optimization. Therefore, it is unreasonable to expect them to be robust to adversarial examples since the systems are not optimized to be resilient to them.

Arguably, fundamental works in adversarial robustness of deep learning systems have been mostly dominated by studies in computer vision Ilyas2019,zhang2019theoretically,$Cohen19_R S. The brittleness is, however, scaleconversationalagentchatgptcanbemisguidedtogeneratemaliciousresponses. wallace - etal - 2021 - cond$

Building robust systems capable of withstanding adversarial attacks is not merely an academic

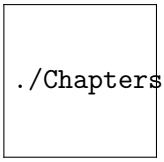


Figure 1.1: (a) A clean sample of IMDb imdb and (b) its corresponding adversarial example. The adversarial example misguides a sentiment classification system to recognize the clean sample as an input text with negative sentiment by replacing 'adaptation' and 'missing' in the clean sample with 'changing' and 'evaporated', respectively.

pursuit. The brittleness could cause substantial problems as deep NLP systems have become ubiquitous and continued to move out of the lab into real-world applications, from chatbots chatgpt to machine translations and text classifications. Despite the potential for sabotaging the phenomenal achievement, our conventions for developing deep NLP systems have not moved far enough toward robust NLP systems. The standard benchmarking approaches are still centered on training and evaluating them on clean samples without significant concerns about the dynamics and the variations of real-world problems.

Overall, it is clear that we should aim beyond the accuracy and scalability of deep NLP systems to secure the things we can benefit through them. Equivalently, there is an urgent need for developing special approaches to address the brittleness of deep NLP systems, since the behaviors of the neural networks are not explicitly instructed in their weights and nodes as done in classical software engineering Hendrycks21$_safety.Despiteitsimportance, fundamentaldiscussionsabouttheadversarialrobustnessof$
$trainedlanguagemodels(PLMs)andattackefficacyofadversarialattackalgorithmsli-etal-2021-searchi$
$tunedBERTbertmodelsacrossdifferentstudiesDong2021, yoo2021towardsimproving, zhang-etal-2022-$

Consequently, experimental setups and defense strategies have been somewhat ad-hoc in adversarial NLP research field. This motivates us to delve into the problem in a rigorous manner from different technical perspectives and develop defense frameworks. Our research focus is specifically on the robustness of large-scale pre-trained language models (PLMs) since they have become the *de-facto* choice for building NLP systems. Our studies across eight such PLMs and four tasks delve into an analysis of security holes of existing methods and probing the language representations of PLMs throughout fine-tuning, as well as transferability of adversarial samples across PLMs. We also analyze the threat of textual adversarial attacks from several technical perspectives. Our contributions and findings will serve as a solid guidance for future work in adversarial NLP.

For a seamless protection of NLP systems, we also propose two novel defense schemes working at various stages of NLP systems. Firstly, we propose a novel detection-based defense approach called GradMask. The proposed detection scheme identify adversarially perturbed tokens via a gradient signal. It provides several advantages over existing methods including improved detection performance and an interpretation of its decision with an only moderate computational cost. We also propose a novel two-stage framework that combines randomized smoothing (RS) with masked inference (MI) to improve the adversarial robustness of NLP systems. The proposed defense scheme significantly outperforms existing defense schemes and demonstrate its effectiveness by extensive experiments.

In the rest of this chapter, we will first discuss the key challenges and limitations of adversarial attacks in NLP in **??** as adversarial attacks help to form a better understanding for building robust NLP systems. Subsequently, we present our novel defense schemes with a discussion about the limitations of existing approaches in **??**. Finally, we provide an outline of the dissertation in

??.

Chapter 2

Preliminary

2.1 Differential Privacy

$$\log \frac{P(M(D)) \in S}{P(M(D')) \in S} \leq e, \tag{2.1}$$

where D‘ and S are one instance different dataset and the set of outcomes and e is the privacy budget.

$$\log P(M(D)) \in S \leq e^{\epsilon} P(M(D')) \in S + \delta, \tag{2.2}$$

where δ is the failure probability.

Chapter 3

Conclusions and Future Work

3.1 Overall Summary

This dissertation is written with two academic goals in mind: (*i*) In-depth understanding of the adversarial robustness of deep NLP systems and (*ii*) providing a seamless protection of NLP systems at their various operational stages. Thus, we elucidated several security holes in existing works and their limitations for demonstrating the veritable threat of textual adversarial attacks (*c.f.,* **??**). We also emphasized the brittleness of language representations of large-scale pre-trained language models and showed the effectiveness of adversarial training approaches.

In **??**, we subsequently presented a novel black-box textual adversarial attack framework, RTA, and its optimization algorithm RMU. Particularly, RTA demonstrates controllability for a trade-off between the attack success rate and the semantics in a simple and predictable manner. Moreover, RTA shows high computational efficiency in terms of adversarial example generation.

Another crucial goal of this work is to devise a seamless defense framework for all stages of the system operations. In **??**, we proposed a simple adversarial example detection scheme, , which verifies the intention of inputs. The proposed method shows significantly low FPR95 scores, which is a highly desirable property for NLP systems with high-security requirements. does not require an additional module or a strong assumption about potential attacks which are more realistic in practice. Our results imply that it can serve as a useful tool for identifying adversarial attacks for protecting text classification systems.

Finally, we proposed RSMI, a novel two-stage framework to tackle the issue of adversarial robustness of large-scale deep NLP systems (*c.f.,* **??**). We evaluated RSMI by applying it to large-scale pre-trained models on three benchmark datasets and obtain 2 to 3 times improvements against strong attacks in terms of robustness evaluation metrics over state-of-the-art defense methods. Our thorough experiments validate the effectiveness of RSMI as a practical approach to train adversarially robust NLP systems.

3.2 Future Work

While existing research in the field of adversarial robustness has made significant achievements in understanding and defending against adversarial attacks on relatively small-scale language models,

it is imperative to shift our focus towards large-scale models as the size of deep learning models continues to grow brown2020language,rae2021scaling,chowdhery2022palm,OPT,chatgpt. Despite our thorough analysis on adversarial robustness and its scalability in **??**, our analysis has not yet extended to such billion-scale models. As demonstrated from diverse domains, including our analysis, a simple scaling cannot guarantee the robust AI systems Zhang2023TransferableAA,li2022adversarial.

Nevertheless, there exist some challenges for understanding such large-scale models. Firstly, a lot of existing textual attack algorithms operate in a gray-box setting where attackers have access to the model's prediction distribution (*i.e.,* soft-max distribution). However, such information is not always available, as seen in cases like ChatGPT chatgpt. Secondly, the computational constraints of the existing attacks typically hinders in-depth analysis about large-scale models. For instance, TextFooler attack algorithm jin2019textfooler requires 440 queries on average for misguiding a BERT-base model, which has only 110M parameters. Thus, it is crucial to develop a new attack algorithm with high computational efficiency working in a black-box setting. Consequently, devising a new attack scheme becomes an exciting future research direction.

Adversarial examples alone cannot provide a complete understanding of the robustness of deep NLP systems. To gain better insights, we should closely look at the fundamental learning mechanism of deep NLP systems, which are often overlooked despite their achievements. A notable example is the ability of deep NLP systems to generalize to test examples by learning from spurious correlations in training samples wang-etal-2022-identifying. This behavior contradicts our expectation that these systems would implicitly leverage salient linguistic features in texts to achieve high-performance task results. Given the obscurity, a necessary future direction is to delve into the general robustness issues in deep NLP systems.

While adversarial examples might be the most striking phenomena to illustrate the vulnerability of deep learning systems, their brittleness is surprisingly more pervasive than we thought. For instance, deep learning systems often fail to generalize to testing distributions when they differ from training distributions hendrycks17baseline. For instance, a sentiment classification system trained on sentiment classification samples annotated with positive and negative labeles may consistently classify samples with neutral sentiment as a negative sample with high confidence.

Detection of such samples would be the first step to addressing the issue. However, out-of-distribution (OOD) samples cannot be easily detected. It would be tempting to harness our adversarial example detection scheme proposed in **??** since both OOD samples and adversarial examples lead to abnormal behaviors of the systems. However, OOD samples require a special detection approach since they significantly differ from adversarial examples in various perspectives. In future work, OOD sample detection schemes will be explored.

Lastly, the RSMI algorithm introduced in **??** could be used to prevent an improper disclosure of sensitive information contained in training data. Deep NLP systems often memorize a massive amount of training data that contain private information carlini21extracting. This tendency often makes them vulnerable to a malicious data extraction attempt. Since RSMI's inference involves a data suppression (*i.e.,* gradient-guided salient token masking) and a randomized inference (*i.e.,* randomized smoothing), it would be worthwhile to explore the privacy issue in deep NLP systems.

Bibliography

[1] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[2] Rachit Singh. . `https://rachitsingh.com/elbo_surgery/`, 2017. Online; accessed 29 January 2014.

[3] Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567, 2018.