Robustness and Adversarial Examples in Natural Language Processing

Kai-Wei Chang
University of California, Los Angeles
kwchang@cs.ucla.edu

He He
New York University
hehe@cs.nyu.edu

Robin Jia
Facebook AI Research and
University of Southern California
robinjia@fb.com

Sameer Singh
University of California, Irvine
sameer@uci.edu

Abstract

Recent studies show that many NLP systems are sensitive and vulnerable to a small perturbation of inputs and do not generalize well across different datasets. This lack of robustness derails the use of NLP systems in real-world applications. This tutorial aims at bringing awareness of practical concerns about NLP robustness. It targets NLP researchers and practitioners who are interested in building reliable NLP systems. In particular, we will review recent studies on analyzing the weakness of NLP systems when facing adversarial inputs and data with a distribution shift. We will provide the audience with a holistic view of 1) how to use adversarial examples to examine the weakness of NLP models and facilitate debugging; 2) how to enhance the robustness of existing NLP models and defense against adversarial inputs; and 3) how the consideration of robustness affects the real-world NLP applications used in our daily lives. We will conclude the tutorial by outlining future research directions in this area.

Type of Tutorial: Cutting edge.

1 Tutorial Description

Recent advances in data-driven machine learning techniques such as deep neural networks have revolutionized natural language processing. In particular, modern natural language processing (NLP) systems have achieved outstanding performance on various tasks such as question answering, textual entailment, language generation. In many cases, they even achieve higher performance than inter-annotator agreement on benchmark datasets. It may be tempting to conclude from results on these *datasets* that current systems are as good as humans at these NLP *tasks*.

Despite the remarkable success, recent studies show that these systems often rely on spurious correlations and fail catastrophically when given inputs from different sources or inputs that have been adversarially perturbed. For example, Jia and Liang (2017) shows that state-of-the-art reading comprehension systems fail to answer questions about paragraphs that contain adversarially inserted sentences, which are automatically generated to distract computer systems without changing the correct answer. Similarly, a series of studies (e.g., (Ribeiro et al., 2018; Alzantot et al., 2018; Iyyer et al., 2018)) demonstrate that text classification models are not robust against adversarial examples that generated by synonym substitution, paraphrasing, and inserting/deleting characters in the text input. This lack of robustness exposes troubling gaps in current models' language understanding capabilities and creates problems when NLP systems are deployed to real users.

As NLP systems are increasingly integrated into people's daily lives and directly interact with end-users, it is essential to ensure their reliability. For example, systems that flag hateful social media content for review must be robust to adversaries who wish to evade detection (Hosseini et al., 2017). Defending against these threats requires building systems that are robust to whatever alterations an attacker might apply to text in order to achieve the desired classifier behavior. Besides, even if systems perform well on user queries on average, rare but catastrophic errors can lead to serious issues. In 2017, Facebook's machine translation system mistakenly translated an Arabic Facebook post with the message "Good morning" into a Hebrew phrase that meant "Attack them" (Berger, 2017). As a result, the Israeli police arrested the man who made the post and detained him for several hours until the misunderstanding is resolved. Therefore, deployed systems must avoid egregious errors like wrongly translating non-violent messages into violent ones and should be tested on "worst-case" non-violent

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, pages 22–26
Punta Cana and Online, November 10–11, 2021.

messages.

In this tutorial, we will review the history of adversarial example generation and methods for enhancing robustness of NLP systems. In particular, we will present recent community effort in the following topics:

- Algorithms for generating adversarial examples to "debug" NLP systems. We will cover a variety of approaches such as synonym substitution, syntactically controlled paraphrasing, character-level adversarial attacks and many applications, including sentiment analysis, textural entailment, question answering, and machine translation.
- Robustness to spurious correlations and methods for mitigating dataset bias.
- Adversarial data generation for collecting datasets.
- Certified robustness in NLP.
- Debugging and behavior testing of NLP models by adversarial and automatic data generation.
- Lessons and discussion on how to build reliable, accountable NLP systems.

The tutorial will bring researchers and practitioners to be aware of the robustness issues of NLP systems and encourage the research community to propose innovative solutions to develop robust, reliable, and accountable NLP systems.

2 Detail Outline

This tutorial presents a systematic overview of frontier approaches to generating adversarial examples to facilitate behavior testing and debugging of NLP systems. We will also review the studies revealing that NLP models make predictions based on spurious correlations learned in the data and discuss approaches to enhancing their robustness. We will motivate the discussion using various NLP tasks and will outline emerging research challenges on this topic at the end of the tutorial. The detailed contents covered in the tutorial are outlined below.

Motivation

We will motivate the audience by demonstrating practical examples where NLP systems are brittle to adversarial examples and data distributional shifts. Then, we will outline the challenges of building reliable and robust NLP systems.

Generating Adversarial Examples for Text Classification

Many NLP problems such as document categorization, sentiment analysis and textual entailment can be modeled as a text classification task. However, recent studies show that by slightly modifying a correctly classified example can cause the high-performing models to misclassify. We will discuss various algorithms for generating such adversarial examples and how these examples can be used to test the behaviors of models and facilitate debugging.

Certified Robustness and Defending against Adversarial Attacks in NLP

Next, we will discuss methods for enhancing models against adversarial examples. Ensuring robustness to seemingly simple perturbations, such as typos or synonym replacements, is already challenging. In particular, since multiple parts of a sentence may be perturbed independently, there is a combinatorially large space of possible perturbations. We will discuss methods that augment training data with adversarial examples as well as methods that produce *certificates* of robustness. The latter enjoy computationally tractable guarantees that a model is correct on every allowed perturbation of a given input.

Robustness to Spurious Correlations

Aside from adversarial attacks, current models are also prone to spurious correlations, i.e. predictive patterns that work well on a specific dataset but do not hold in general. As a result, models fail under a mild distribution shift. In this part, we will discuss methods that guard against known spurious correlations in the data and the robustness of large-scale pre-trained models.

Adversarial data collection

Given the flaws in existing datasets, it seems likely that building robust NLP models will also require better ways to collect training data. In this part, we will discuss recent work that collects datasets using an adversarial data generation process, typically involving humans in the loop. We will also discuss connections with classical active learning approaches to data collection.

Adversarial Trigger and Text Generation

While most of the discussion in the tutorial focuses on natural language understanding, many language generation systems directly interact with end users and ensuring their robustness is equivalently important. In this part, we will discuss robustness issues in language generation tasks. We will also introduce adversarial triggers, input-agnostic sequences of tokens that trigger a model to produce a specific prediction when concatenated to any input from a dataset, and its application in conditional language generation.

Conclusion, Future Directions, and Discussion

We will conclude the tutorial by discussing future directions to promote robustness in NLP.

3 Reading List

While the tutorial will include our own work (Alzantot et al., 2018; Shi et al., 2019; Pezeshkpour et al., 2019; Ribeiro et al., 2020, 2018; Jia and Liang, 2017; Jia et al., 2019; Jones et al., 2020; He et al., 2019; Tu et al., 2020; Wallace et al., 2019a), we anticipate that roughly 60% of the tutorial content will pull from work by other researchers in NLP and machine learning communities, including (Huang et al., 2019; Ye et al., 2020; Nie et al., 2020; Wallace et al., 2019b; Pruthi et al., 2019; Zellers et al., 2018; Ren et al., 2019; Zhang et al., 2019; Belinkov et al., 2019; Chen et al., 2018; Zheng et al., 2020; Cheng et al., 2019; Hsieh et al., 2019; Abdou et al., 2020; Karimi Mahabadi et al., 2020; Karpukhin et al., 2019; Murray and Chiang, 2018; Iyyer et al., 2018; Ebrahimi et al., 2018). A more comprehensive list of related papers will be provided before the tutorial.

4 Prerequisite Knowledge

Our target audience is general NLP conference attendances; therefore, no specific knowledge is assumed of the audience except basic machine learning and NLP background:

- Understand derivatives and gradient decent methods as found in introductory Calculus.
- Understand the basic supervised learning paradigm and commonly used machine learning models such as logistic regression and deep neural networks.
- Familiar with common natural language processing concepts (e.g., parse trees, word representation) as found in an introductory NLP course.

5 Tutorial Instructors

Our instructors consist of experts who have conducted research in different aspects related to the tutorial topic.

Kai-Wei Chang Kai-Wei Chang is an assistant professor in the Department of Computer Science at the University of California Los Angeles. His research interests include designing robust, fair, and accountable machine learning methods for building reliable NLP systems (e.g., (Alzantot et al., 2018; Shi et al., 2019)). His awards include the EMNLP Best Long Paper Award (2017), the KDD Best Paper Award (2010), and the Sloan Resaerch Fellowship (2021). Kai-Wei has given tutorials at NAACL 15, AAAI 16, FAccT18, EMNLP 19, AAAI 20, MLSS 21 on different research topics. Additional information is available at `http://kwchang.net`.

He He He He is an assistant professor in the Department of Computer Science and the Center for Data Science at the New York University. Her research interests include reliable natural language generation and robust learning algorithms that avoid spurious correlations in the data (e.g., (He et al., 2019; Tu et al., 2020)). She has given tutorials at NAACL 15 and EMNLP 19. Additional information is available at `http://hhexiy.github.io`.

Robin Jia Robin Jia is currently a visiting researcher at Facebook AI Research, and will be an assistant professor in the Department of Computer Science at the University of Southern California starting in the Autumn of 2021. His research focuses on making natural language processing models robust to unexpected test-time distribution shifts (e.g., (Jia and Liang, 2017; Jia et al., 2019). Robin's work has received an Outstanding Paper Award at EMNLP 2017 and a Best Short Paper Award at ACL 2018. Additional information is available at `https://robinjia.github.io`.

Sameer Singh Sameer Singh is an Assistant Professor of Computer Science at the University of California, Irvine. He is working on large-scale and interpretable machine learning models for NLP (e.g., (Wallace et al., 2019a; Pezeshkpour et al., 2019)). His work has received paper awards at

ACL 2020, AKBC 2020, EMNLP 2019, ACL 2018, and KDD 2016. Sameer presented the Deep Adversarial Learning Tutorial (Wang et al., 2019) at NAACL 2019 and the Mining Knowledge Graphs from Text Tutorial at WSDM 2018 and AAAI 2017, along with tutorials on Interpretability and Explanations in upcoming NeurIPS 2020 and EMNLP 2020. Sameer has also received teaching awards at UCI. Website: http://sameersingh.org/

References

Mostafa Abdou, Vinit Ravishankar, Maria Barrett, Yonatan Belinkov, Desmond Elliott, and Anders Søgaard. 2020. The sensitivity of language models and humans to Winograd schema perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019. On adversarial removal of hypothesis-only bias in natural language inference. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*.

Y. Berger. 2017. Israel arrests palestinian because facebook translated 'good morning' to 'attack them'. https://www.haaretz.com/israel-news/palestinian-arrested-over-mistranslated-good-morning-facebook-post-1.5459427.

Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh. 2018. Attacking visual language grounding with adversarial examples: A case study on neural image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Minhao Cheng, Wei Wei, and Cho-Jui Hsieh. 2019. Evaluating and enhancing the robustness of dialogue systems: A case study on a negotiation agent. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Javid Ebrahimi, Daniel Lowd, and Dejing Dou. 2018. On adversarial examples for character-level neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*.

H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran. 2017. Deceiving Google's Perspective API built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*.

Yu-Lun Hsieh, Minhao Cheng, Da-Cheng Juan, Wei Wei, Wen-Lian Hsu, and Cho-Jui Hsieh. 2019. On the robustness of self-attentive models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Po-Sen Huang, Robert Stanforth, Johannes Welbl, Chris Dyer, Dani Yogatama, Sven Gowal, Krishnamurthy Dvijotham, and Pushmeet Kohli. 2019. Achieving verified robustness to symbol substitutions via interval bound propagation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Erik Jones, Robin Jia, Aditi Raghunathan, and Percy Liang. 2020. Robust encodings: A framework for combating adversarial typos. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on synthetic noise improves robustness to natural noise in machine translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. 2019. Investigating robustness and interpretability of link prediction via adversarial modifications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. 2019. Combating adversarial misspellings with robust word recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, and Cho-Jui Hsieh. 2019. Robustness verification for transformers. In *International Conference on Learning Representations*.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *TACL*.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019a. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019b. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7.

Xiaozhi Wang, Xu Han, Zhiyuan Liu, Maosong Sun, and Peng Li. 2019. Adversarial training for weakly supervised event detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Mao Ye, Chengyue Gong, and Qiang Liu. 2020. SAFER: A structure-free approach for certified robustness to adversarial word substitutions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Yeyao Zhang, Eleftheria Tsipidi, Sasha Schriber, Mubbasir Kapadia, Markus Gross, and Ashutosh Modi. 2019. Generating animations from screenplays. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*.

Xiaoqing Zheng, Jiehang Zeng, Yi Zhou, Cho-Jui Hsieh, Minhao Cheng, and Xuanjing Huang. 2020. Evaluating and enhancing the robustness of neural network-based dependency parsing models with adversarial examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.