Adversarial Attack and Defense on Text : A Survey

Aminul Huq[1], Mst. Tasnim Pervin[2]
Student ID: 2019280161[1], 2019280162[2]
Department of Computer Science & Technology, Tsinghua University
Beijing, China

1 Introduction

Deep learning models have been used widely in object recognition, face recognition, speech recognition, sentiment analysis, and many others. However, in recent years it has been shown that these models possess weakness to noises which force the model to misclassify. This issue has been studied profoundly in the image and audio domain. Very little has been studied on this issue for textual data. In this manuscript, we accumulated and analyzed different attack techniques, various defense models on how to overcome this issue to provide a more comprehensive idea. Later we point out some of the interesting findings of all papers and challenges that need to be overcome to move forward in this field.

2 Adversarial Attack for Textual Data

Character-level Attack :

Task: Sentiment Analysis. ***Classifier:*** Amazon AWS. ***Original label:*** 100% Negative. ***Adversarial label:*** 89% Positive.

Text: I watched this movie recently mainly because I am a Huge fan of Jodie Foster's. I saw this movie was made right between her 2 Oscar award winning performances, so my expectations were fairly high. ~~Unfortunately~~ Unf0rtunately, I thought the movie was ~~terrible~~ terrib1e and I'm still left wondering how she was ever persuaded to make this movie. The script is really ~~weak~~ wea k.

Fig 1: Example of character-level attack(TEXTBUGGER) (Li et al. [2])

Authors	Approach Name	Attack-Type	Summary
Gil et al. [1]	DISTFLIP	Black-box	They distilled the HotFlip attack technique into a NN and created a similar Black-box attack.
Li et al. [2]	TEXTBUGGER	White-box and Black-box	In white-box used gradients to determine which words are most significant and replaced it with one of five bugs that had most damage. Similar in black box but since no access to gradients they started from determining which sentence is most significant.
Gao et al. [3]	DEEPWORDBUG	Black-box	Proposed the concept of temporal score and temporal tail score and used it to determine most significant word and replace it.

Word -level Attack :

Original Text Prediction = **Negative**. (Confidence = 78.0%)

This movie had terrible acting, terrible plot, and terrible choice of actors. (Leslie Nielsen ...come on!!!) the one part I considered slightly funny was the battling FBI/CIA agents, but because the audience was mainly kids they didn't understand that theme.

Adversarial Text Prediction = **Positive**. (Confidence = 59.8%)

This movie had horrific acting, horrific plot, and horrifying choice of actors. (Leslie Nielsen ...come on!!!) the one part I regarded slightly funny was the battling FBI/CIA agents, but because the audience was mainly youngsters they didn't understand that theme.

Fig. 2 : Example of word-level attack

Author	Approach Name	Attack Type	Summary
Alzantot et al. [4]	Genetic Algorithm	Black-box	To generate adversarial examples which are semantically and syntactically similar this approach was proposed. Words were selected after several generation which suited to the context
Liang et al. [5]	Replacement	White-box and Black-box	Proposed Hot-Training-Phrase(HTP) and Hot-Sampling-Phrase(HSP) concept to determine what to insert and where to insert, delete or modify. For white-box attack they used natural language watermarking technique and for black-box used fuzzing technique.
Zang et al. [6]	Word replacement and optimization	Black-box	Using sememe based word replacement and PSO based optimization method to determine the word which reduces the accuracy the most.

Sentence-level Attack :

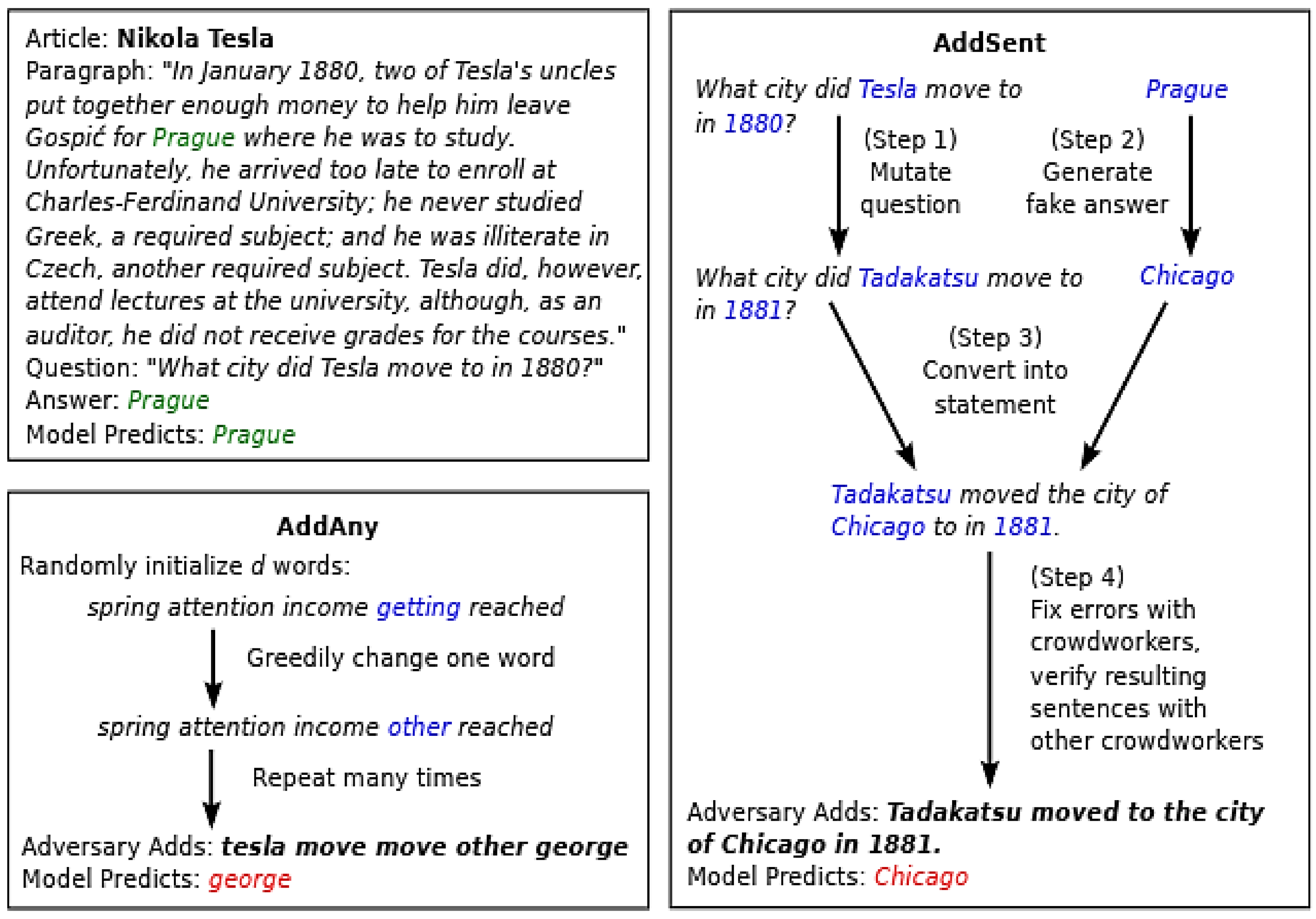

Fig. 3 : ADDANY and ADDSENT Attack Generation

Author	Approach Name	Attack type	Summary
Cheng et al. [7]	AdvGen	White-box	Guided by training loss, they used greedy approach to choose the most optimal solution.
Michael et al. [8]	Semantic Word Replacement	White-box	Replaced words from the sentences to maximize the loss. For preserving meaning they used KNN to choose similar words.

3 Adversarial Defense

I. Adversarial Training

II. Topic Specific Defenses :

Papers	Approach Name	Summary
Zhou et al.	DISP Framework	Uses discriminator to check each token for perturbation and restores the original word based on context using KNN
Wang et al.	Synonym encoding method	Encoder method placed before the model. Clusters all the neighboring words so that they have same encoding.
Pruthi et al [9]	Spell Checking and Correction	Trained ScRNN for word recognition and restoration

4 Challenges

I. Several Attacks introduced to the image domain are not applicable to text domain because of discrete representation.
II. Creating fully imperceptible attack to textual data is almost impossible as injection or removal of words is easily noticeable.
III. No universal perturbation or universal defense technique has been introduced that can tackle all kinds of attack.
IV. No ideal Benchmark for comparison
V. No standard toolbox like in image domain(cleaverhans, art, foolbox etc.)

5 Contribution

Surveys	Text Domain	Adversarial Attacks	Defense
Belinkov et al. [10]	Partly	Y	Very little
Xu et al. [11]			
Zhang et al. [12]	Fully		Little
Ours			Greater than others

References

[1] Yotam Gil, Yoav Chai, Or Gorodissky, and Jonathan Berant. 2019. White-to-black: Efficient distillation of black-box adversarial attacks. arXiv preprint arXiv:1904.02405.
[2] Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, et al. 2013. Recent advances in deep learning for speech research at microsoft. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 8604–8608. IEEE.
[3] Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In 2018 IEEE Security and Privacy Workshops (SPW), pages 50–56. IEEE.
[4] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo- Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. arXiv preprint arXiv:1804.07998.
[5] Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. 2017. Deep text classification can be fooled. arXiv preprint arXiv:1704.08006.
[6] Y Zang, C Yang, F Qi, Z Liu, M Zhang, Q Liu, and M Sun. 2019. Textual adversarial attack as combinatorial optimization. arXiv preprint arXiv:1910.12196.
[7] Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. arXiv preprint arXiv:1906.02443.
[8] Michael Sutton, Adam Greene, and Pedram Amini. 2007. Fuzzing: brute force vulnerability discovery. Pearson Education.
[9] Danish Pruthi, Bhuwan Dhingra, and Zachary C Lipton. 2019. Combating adversarial misspellings with robust word recognition. arXiv preprint arXiv:1905.11268.
[10] Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation. arXiv preprint arXiv:1711.02173.
[11] Han Xu, Yao Ma, Haochen Liu, Debayan Deb, Hui Liu, Jiliang Tang, and Anil Jain. 2019. Adversarial attacks and defenses in images, graphs and text: A review. arXiv preprint arXiv:1909.08072.
[12] Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4):e1253.