Progress Report – October 2022

This past month of October 2022, I have read a total of eight research papers: all the papers published in 2019 (six papers) and two papers published in 2020. Two of these papers discussed the white-box attacks, and the rest of the six papers discussed the black-box attacks. The new in this month's reading list is that it converged new topics in this domain, such as discussing defense mechanisms in both attack types (white/black), discussing black-box attacks in the Arabic language, discussing the BERT-based adversarial examples, and discussing Sequence-to-Sequence models adversarial examples. Here is a list of these papers:

1. Jia et al. (2019). *Certified Robustness to Adversarial Word Substitution.* EMNLP 2019.
2. Alshemali and Kalita. *Toward Mitigating Adversarial Texts.* International Journal of Computer Applications 2019.
3. Pruthi et al. (2019). *Combating Adversarial Misspellings with Robust Word Recognition.* ACL 2019.
4. Li et al. (2019). *TextBugger: Generating Adversarial Text Against Real-world Applications*. NDSS 2019.
5. Alshemali and Kalita. (2019). *Adversarial Examples in Arabic.* CSCI 2019.
6. Ren et al. (2019). *Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency.* ACL 2019.
7. Garg and Ramakrishnan. (2020). *BAE: BERT-based Adversarial Examples for Text Classification*. EMNLP 2020.
8. Cheng et al. (2020). *Seq2Sick: Evaluating the Robustness of Sequence-to-Sequence Models with Adversarial Examples.* AAAI 2020.

Throughout my reading of these eight research papers, I kept a brief taxonomy of the adversarial examples/attacks in terms of the attack type, the targeted space, and the method used, hoping to turn this taxonomy into a review paper one day. The table below shows the brief taxonomy of the already-read papers for 2019 and 2020.

#	Paper	Authors	Attack	Space	Method
1	Certified Robustness to Adversarial Word Substitution	Jia et al. (2019)	Black-box Attacks	Embedding Space	Word-level Replacements
2	Toward Mitigating Adversarial Texts	Alshemali and Kalita (2019)	Black-box Defenses	Input Space	Spell-Checking
3	Combating Adversarial Misspellings with Robust Word Recognition	Pruthi et al. (2019)	Black-box Defenses	Input Space	Word Recognition Models
4	TEXTBUGGER: Generating Adversarial Text Against Real-world Applications	Li et al. (2019)	Black-box White-box Attacks & Defenses	Input Space	$^{+}$Character-level Replacements *Word-level Replacements
5	Adversarial Examples in Arabic	Alshemali and Kalita (2019)	Black-box Attacks	Input Space	Character-level Replacements
6	Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency	Ren et al. (2019)	Black-box Attacks	Input Space	Character-level Replacements
7	BAE: BERT-based Adversarial Examples for Text Classification	Ramakrishnan and Garg (2020)	Black-box Attacks	Input Space	Word-level Replacements
8	Seq2Sick: Evaluating the Robustness of Sequence-to-Sequence Models with Adversarial Examples	Cheng et al. (2020)	White-box Attacks	Input Space	Word-level Replacements

Lastly, in the next month, November 2022, I will read all papers published in 2020 (seven papers) and 2021 (three papers). I also will keep building the brief taxonomy of the adversarial examples/attacks.

--- End of Report ---