Computer Science
Defense

Towards the Robustness of Deep Learning Systems Against Adversarial Examples in Sequential Data

Wenjie Wang
Emory University

Abstract: Recent studies have shown that adversarial examples can be generated by applying small perturbations to the inputs such that the well-trained deep neural networks (DNNs) will misclassify. With the increasing number of safety and security-sensitive applications of deep learning models, the robustness of deep learning models to adversarial inputs has become a crucial topic. Research on the adversarial examples in computer vision (CV) domains has been well studied. However, the intrinsic difference between image and sequential data has placed great challenges for directly applying adversarial techniques in CV to other application domains such as speech, health informatics, and natural language processing (NLP). ¡br¿
To solve these gaps and challenges, My dissertation research combines multiple studies to improve the robustness of deep learning systems against adversarial examples in sequential inputs. First, We take the NLP and health informatics domains as examples, focusing on understanding the characteristics of these two domains individually and designing empirical adversarial defense methods, which are 1) RADAR, an adversarial detection for EHR data, and 2) MATCH, detecting adversarial examples leveraging the consistency between multiple modalities. Following the empirical defense methods, our next step is exploring certified robustness for sequential inputs which is provable and theory-backed. To this end, 1) We study the randomized smoothing on the word embedding space to provide certification to NLP models. 2) We propose WordDP, certified robustness to word substitution attacks in the NLP domain, leveraging the connection of differential privacy and certified robustness. 3) We studied the certified robustness methods to Wasserstein adversarial examples on univariant time-series data.

Thursday, November 17, 2022, 3:00 pm
zoom.us/j/9828106847

Computer Science
Emory University