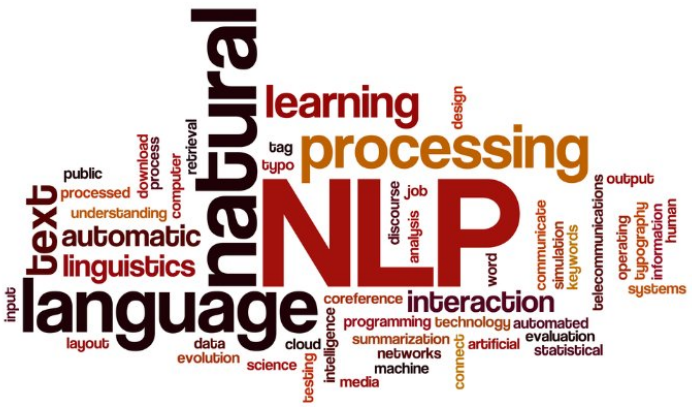
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

Towards Deep Learning Models Robust to Textual Adversarial Attacks



Researchers have demonstrated that deep learning models are vulnerable to adversarial attacks, which fool a model by generating adversarial examples. An adversarial example is crafted by adding small imperceptible perturbations to inputs. To make deep learning models robust against such attacks, researchers have proposed various methods, including adversarial training, curvature regularization, etc. However, existing work mainly focuses on the image domain; in the textual domain there is only little work. While some researchers show that adversarial training is helpful to improve model robustness against certain textual attacks, other researchers argue adversarial training does not provide much improvement. Furthermore, previous work demonstrates that deep learning models for image classification can be robust to multiple attacks by conducting adversarial training with PGD, but it remains unclear if such claim also holds for NLP applications.

In this work, we propose to study the robustness of deep learning models to multiple natural language adversarial attacks. We will conduct systematic study of different textual adversarial attacks. Based on the study, we will explore machine learning techniques to build deep learning models which are universally robust to various adversarial attacks.

Requirements: Strong motivation, proficiency in Python & PyTorch, and prior knowledge in Deep Learning & Nature Language Processing.

Interested? Please contact us for more details!

Contact

- Zhao Meng: zhmeng@ethz.ch, ETZ G61.4
- Ye Wang: wangye@ethz.ch, ETZ G63