Toward Mitigating Adversarial Texts

Abstract

Neural networks are frequently used for text classification, but can be vulnerable to misclassification caused by adversarial examples: input produced by introducing small perturbations that cause the neural network to output an incorrect classification. Previous attempts to generate black-box adversarial texts have included variations of generating nonword misspellings, natural noise, synthetic noise, along with lexical substitutions. This paper proposes a defense against black-box adversarial attacks using a spell-checking system that utilizes frequency and contextual information for correction of nonword misspellings. The proposed defense is evaluated on the Yelp Reviews Polarity and the Yelp Reviews Full datasets using adversarial texts generated by a variety of recent attacks. After detecting and recovering the adversarial texts, the proposed defense increases the classification accuracy by an average of 26.56% on the Yelp Reviews Polarity dataset and 16.27% on the Yelp Reviews Full dataset. This approach further outperforms six of the publicly available, state-of-the-art spelling correction tools by at least 25.56% in terms of average correction accuracy.
References


Index Terms

Computer Science

Information Sciences

Keywords

Adversarial text, Spelling correction