Crafting Adversarial Text Samples for Recurrent Neural Networks Using Windowed Inference and Search

by

Cory Nezin

Professor Fred L. Fontaine, Advisor

Abstract

Neural networks have recently been found vulnerable to “attacks” which cause them to misclassify samples that were given a very small disturbance. Attacks based on iterative gradient methods have been largely studied for numerically valued data like images. In this work two new algorithms are described which extend the same kind of results to text. One algorithm, “window search”, does not require a continuous or differentiable model. The other algorithm, “gradient assisted window search”, is a hybrid algorithm which exploits word2vec based gradients for fast search. The hybrid algorithm uses the gradient as a guide for candidate word replacements, and then performs an exponential search to determine the minimum number of replacements required to alter the classification of a text sample. The algorithm is tested under white box, gray box, and black box scenarios.

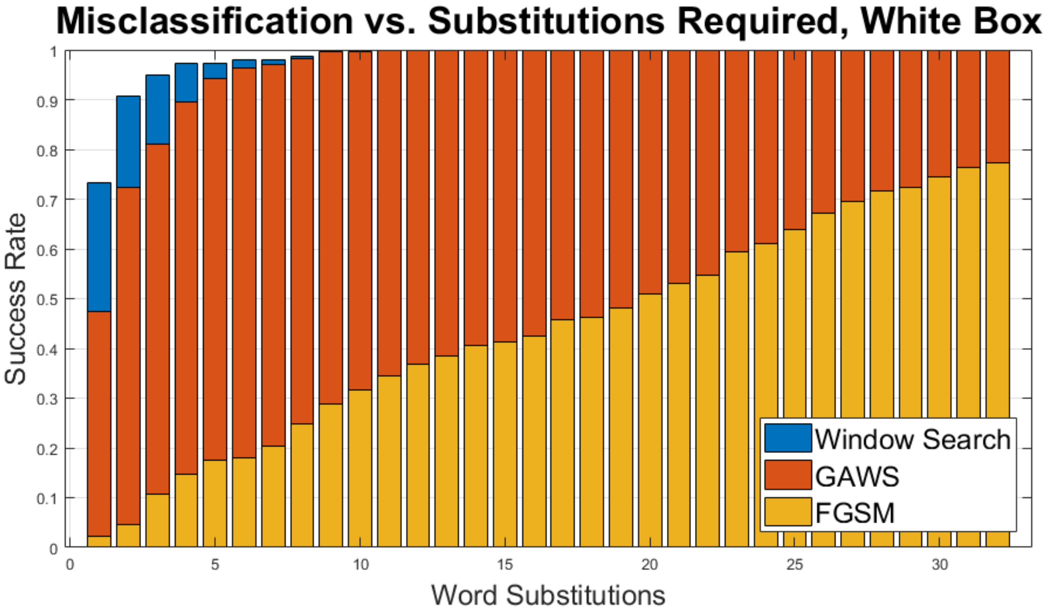

Figure 1: The x-axis represents the number of word substitutions used to change the fraction of samples represented by the y-axis.