cases

doc_1		doc_2		decision	id
authors	• Di Jin • Zhijing Jin • Joey Tianyi Zhou • Peter Szolovits	**authors**	• Di Jin • Zhijing Jin • Joey Tianyi Zhou • Peter Szolovits	NOT DUPLICATES	434
title	Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment	**title**	Is BERT Really Robust? Natural Language Attack on Text Classification and Entailment		
publication_date	2019-07-27 15:07:04+00:00	**publication_date**	2019-07-27 00:00:00		
source	SupportedSources.ARXIV	**source**	SupportedSources.SEMANTIC_SCHOLAR		
journal	None	**journal**	ArXiv		
volume		**volume**	abs/1907.11932		
doi		**doi**			
urls	• http://arxiv.org/pdf/1907.11932v6 • http://arxiv.org/abs/1907.11932v6 • http://arxiv.org/pdf/1907.11932v6	**urls**	• https://www.semanticscholar.org/paper/a3347bbd82938788ec085772813c095de17a0b37		
id	id-1141823555841477667	**id**	id-274753565480961535		
abstract	Machine learning algorithms are often vulnerable to adversarial examples that have imperceptible alterations from the original counterparts but can fool the state-of-the-art models. It is helpful to evaluate or even improve the robustness of these models by exposing the maliciously crafted adversarial examples. In this paper, we present TextFooler, a simple but strong baseline to generate natural adversarial text. By applying it to two fundamental natural language tasks, text classification and textual entailment, we successfully attacked three target models, including the powerful pre-trained BERT, and the widely used convolutional and recurrent neural networks. We demonstrate the advantages of this framework in three ways: (1) effective---it outperforms state-of-the-art attacks in terms of success rate and perturbation rate, (2) utility-preserving---it preserves semantic content and grammaticality, and remains correctly classified by humans, and (3) efficient---it generates adversarial text with computational complexity linear to the text length. *The code, pre-trained target models, and test examples are available at https://github.com/jind11/TextFooler.	**abstract**	Machine learning algorithms are often vulnerable to adversarial examples that have imperceptible alterations from the original counterparts but can fool the state-of-the-art models. It is helpful to evaluate or even improve the robustness of these models by exposing the maliciously crafted adversarial examples. In this paper, we present the TextFooler, a general attack framework, to generate natural adversarial texts. By successfully applying it to two fundamental natural language tasks, text classification and textual entailment, against various target models, convolutional and recurrent neural networks as well as the most powerful pre-trained BERT, we demonstrate the advantages of this framework in three ways: (i) effectiveâ€”it outperforms state-ofthe-art attacks in terms of success rate and perturbation rate; (ii) utility-preservingâ€”it preserves semantic content and grammaticality, and remains correctly classified by humans; and (iii) efficientâ€”it generates adversarial text with computational complexity linear in the text length.		
versions		**versions**			