Comparison with Pruthi

08 May 2021 13:19

FastAttack

bert-base-uncased-imdb

Attack Results	
Number of successful attacks:	20
Number of failed attacks:	5
Number of skipped attacks:	2
Original accuracy:	92.59%
Accuracy under attack:	18.52%
Attack success rate:	80.0%
Average perturbed word %:	5.61%
Average num. words per input:	248.96
Avg num queries:	428.4

textattack: Attack time: 251.8702368736267s

bert-base-uncased-yelp-polarity

Attack Results	
Number of successful attacks:	18
Number of failed attacks:	7
Number of skipped attacks:	0
Original accuracy:	100.0%
Accuracy under attack:	28.0%
Attack success rate:	72.0%
Average perturbed word %:	11.07%
Average num. words per input:	133.36
Avg num queries:	241.48

textattack: Attack time: 128.77892637252808s

albert-base-v2-imdb

Attack Results	
Number of successful attacks:	17
Number of failed attacks:	8
Number of skipped attacks:	3
Original accuracy:	89.29%
Accuracy under attack:	28.57%
Attack success rate:	68.0%
Average perturbed word %:	2.3%
Average num. words per input:	244.71
Avg num queries:	420.24

textattack: Attack time: 271.8887360095978s

albert-base-v2-yelp-polarity

Pruthi

Attack Results	
Number of successful attacks:	10
Number of failed attacks:	15
Number of skipped attacks:	2
Original accuracy:	92.59%
Accuracy under attack:	55.56%
Attack success rate:	40.0%
Average perturbed word %:	0.48%
Average num. words per input:	248.96
Avg num queries:	3868.6

textattack: Attack time: 2192.6458132267s

Attack Results	
Number of successful attacks:	1
Number of failed attacks:	24
Number of skipped attacks:	0
Original accuracy:	100.0%
Accuracy under attack:	96.0%
Attack success rate:	4.0%
Average perturbed word %:	6.67%
Average num. words per input:	133.36
Avg num queries:	1665.96

textattack: Attack time: 848.896234035492s

Attack Results	
Number of successful attacks:	8
Number of failed attacks:	17
Number of skipped attacks:	3
Original accuracy:	89.29%
Accuracy under attack:	60.71%
Attack success rate:	32.0%
Average perturbed word %:	0.53%
Average num. words per input:	244.71
Avg num queries:	3843.64

textattack: Attack time: 2493.4644351005554s

Attack Results	
Number of successful attacks:	19
Number of failed attacks:	6
Number of skipped attacks:	0
Original accuracy:	100.0%
Accuracy under attack:	24.0%
Attack success rate:	76.0%
Average perturbed word %:	9.47%
Average num. words per input:	133.36
Avg num queries:	241.72

textattack: Attack time: 139.27998614311218s

Attack Results	
Number of successful attacks:	2
Number of failed attacks:	23
Number of skipped attacks:	0
Original accuracy:	100.0%
Accuracy under attack:	92.0%
Attack success rate:	8.0%
Average perturbed word %:	4.28%
Average num. words per input:	133.36
Avg num queries:	1666.0

textattack: Attack time: 924.6220576763153s