Comparison with BERT-ATTACK_2020

08 May 2021 00:10

AlBERT-base-v2-yelp-polarity:

Our attack

Attack Results	
Number of successful attacks:	14
Number of failed attacks:	1
Number of skipped attacks:	4
Original accuracy:	78.95%
Accuracy under attack:	5.26%
Attack success rate:	93.33%
Average perturbed word %:	14.74%
Average num. words per input:	17.63
Avg num queries:	34.6

textattack: Attack time: 12.196712970733643s

BERT-Attack

Attack Results	
Number of successful attacks:	15
Number of failed attacks:	0
Number of skipped attacks:	4
Original accuracy:	78.95%
Accuracy under attack:	0.0%
Attack success rate:	100.0%
Average perturbed word %:	12.52%
Average num. words per input:	17.63
Avg num queries:	110.8

textattack: Attack time: 123.67272591590881s

BERT-base-uncased-yelp-polarity model

Our attack

Attack Results	
Number of successful attacks:	14
Number of failed attacks:	1
Number of skipped attacks:	4
Original accuracy:	78.95%
Accuracy under attack:	5.26%
Attack success rate:	93.33%
Average perturbed word %:	14.74%
Average num. words per input:	17.63
Avg num queries:	34.6

textattack: Attack time: 12.028334379196167s

BERT-Attack

Attack Results	
Number of successful attacks:	15
Number of failed attacks:	0
Number of skipped attacks:	4
Original accuracy:	78.95%
Accuracy under attack:	0.0%
Attack success rate:	100.0%
Average perturbed word %:	12.52%
Average num. words per input:	17.63
Avg num queries:	110.8

textattack: Attack time: 121.3666181564331s

AlBERT-base-v2-IMDB model

Our attack

Attack Results	
Number of successful attacks:	15
Number of failed attacks:	0
Number of skipped attacks:	1
Original accuracy:	93.75%
Accuracy under attack:	0.0%
Attack success rate:	100.0%
Average perturbed word %:	4.45%
Average num. words per input:	221.56
Avg num queries:	352.0

textattack: Attack time: 253.93306374549866s

BERT-Attack

Attack Results	
Number of successful attacks:	15
Number of failed attacks:	0
Number of skipped attacks:	1
Original accuracy:	93.75%
Accuracy under attack:	0.0%
Attack success rate:	100.0%
Average perturbed word %:	1.75%
Average num. words per input:	221.56
Avg num queries:	347.27

textattack: Attack time: 404.62052512168884s

BERT-base-uncased-IMDB model

Our attack

Attack Results	
Number of successful attacks:	14
Number of failed attacks:	1
Number of skipped attacks:	1
Original accuracy:	93.75%
Accuracy under attack:	6.25%
Attack success rate:	93.33%
Average perturbed word %:	5.34%
Average num. words per input:	221.56
Avg num queries:	365.33

textattack: Attack time: 238.02741146087646s

BERT-Attack

Attack Results	
Number of successful attacks:	15
Number of failed attacks:	0
Number of skipped attacks:	1
Original accuracy:	93.75%
Accuracy under attack:	0.0%
Attack success rate:	100.0%
Average perturbed word %:	1.75%
Average num. words per input:	221.56
Avg num queries:	347.27

textattack: Attack time: 420.75441241264343s