Comparison with textbugger_li_2018.py

Bert-base-uncased-yelp-polarity model:

Our Attack

Attack Results	
Number of successful attacks:	21
Number of failed attacks:	4
Number of skipped attacks:	0
Original accuracy:	100.0%
Accuracy under attack:	16.0%
Attack success rate:	84.0%
Average perturbed word %:	12.51%
Average num. words per input:	133.36
Avg num queries:	238.84

textattack: Attack time: 172.80336356163025s

TextBugger

Attack Results	
Number of successful attacks:	21
Number of failed attacks:	4
Number of skipped attacks:	0
Original accuracy:	100.0%
Accuracy under attack:	16.0%
Attack success rate:	84.0%
Average perturbed word %:	34.34%
Average num. words per input:	133.36
Avg num queries:	357.68

textattack: Attack time: 176.63516521453857s

Bert-base-uncased-imdb model:

Our Attack

Attack Results	
Number of successful attacks:	21
Number of failed attacks:	4
Number of skipped attacks:	2
Original accuracy:	92.59%
Accuracy under attack:	14.81%
Attack success rate:	84.0%
Average perturbed word %:	5.29%
Average num. words per input:	248.96
Avg num queries:	419.72

textattack: Attack time: 266.0633933544159s

TextBugger

Attack Results	
Number of successful attacks:	22
Number of failed attacks:	3
Number of skipped attacks:	2
Original accuracy:	92.59%
Accuracy under attack:	11.11%
Attack success rate:	88.0%
Average perturbed word %:	32.84%
Average num. words per input:	248.96
Avg num queries:	539.12

textattack: Attack time: 278.07997155189514s

Albert-Base-v2-yelp-polarity model:

Our Attack

Attack Results	
Number of successful attacks:	17
Number of failed attacks:	8
Number of skipped attacks:	0
Original accuracy:	100.0%
Accuracy under attack:	32.0%
Attack success rate:	68.0%
Average perturbed word %:	8.19%
Average num. words per input:	133.36
Avg num queries:	242.28

textattack: Attack time: 144.9083080291748s

TextBugger

Attack Results	
Number of successful attacks:	22
Number of failed attacks:	3
Number of skipped attacks:	0
Original accuracy:	100.0%
Accuracy under attack:	12.0%
Attack success rate:	88.0%
Average perturbed word %:	25.53%
Average num. words per input:	133.36
Avg num queries:	337.72

textattack: Attack time: 174.5719850063324s

Albert-Base-v2-imdb model:

Our Attack

Attack Results	
Number of successful attacks:	22
Number of failed attacks:	3
Number of skipped attacks:	3
Original accuracy:	89.29%
Accuracy under attack:	10.71%
Attack success rate:	88.0%
Average perturbed word %:	5.2%
Average num. words per input:	244.71
Avg num queries:	434.92

textattack: Attack time: 582.0640995502472s

TextBugger

Attack Results	
Number of successful attacks:	24
Number of failed attacks:	1
Number of skipped attacks:	3
Original accuracy:	89.29%
Accuracy under attack:	3.57%
Attack success rate:	96.0%
Average perturbed word %:	27.25%
Average num. words per input:	244.71
Avg num queries:	490.64

textattack: Attack time: 486.14418745040894s