Comparison with DeepwordBug

08 May 2021 13:19

FastAttack

bert-base-uncased-imdb

```
+--------------------------------+--------+
| 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: 209.48380708694458s
```

bert-base-uncased-yelp-polarity

```
+--------------------------------+--------+
| 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: 105.1209557056427s
```

albert-base-v2-imdb

```
+--------------------------------+--------+
| 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: 246.755784034729s
```

albert-base-v2-yelp-polarity

DeepWordBug

```
+--------------------------------+--------+
| 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 %:      | 4.73%  |
| Average num. words per input:  | 248.96 |
| Avg num queries:               | 324.68 |
+--------------------------------+--------+
textattack: Attack time: 111.59480237960815s
```

```
+--------------------------------+--------+
| 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 %:      | 10.03% |
| Average num. words per input:  | 133.36 |
| Avg num queries:               | 247.16 |
+--------------------------------+--------+
textattack: Attack time: 73.39352488517761s
```

```
+--------------------------------+--------+
| 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 %:      | 3.77%  |
| Average num. words per input:  | 244.71 |
| Avg num queries:               | 328.84 |
+--------------------------------+--------+
textattack: Attack time: 128.33053016662598s
```

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 %:	11.43%
Average num. words per input:	133.36
Avg num queries:	242.76

textattack: Attack time: 117.95553493499756s

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 %:	9.81%
Average num. words per input:	133.36
Avg num queries:	216.08

textattack: Attack time: 72.94735646247864s