BERT-ATTACK: Adversarial Attack Against BERT Using BERT

> 会议： 2020 EMNLP
>
> 作者：Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, Xipeng Qiu*
>
> 机构：复旦大学
>
> 论文：https://arxiv.org/pdf/2004.09984.pdf
>
> 思想：**与华中科技大学的PWWS（https://www.aclweb.org/anthology/P19-1103.pdf）思路完全相同，只是把候选词的来源从同义词表换成了语言模型，为了保证语言模型预测出来的词仍然同义，不对目标词进行{mask}而是明文预测**
>
> 1. 利用BERT语言掩码任务，预测$S = \{w1, w2, w3...\}$中每个位置词的替换词（不进行mask，直接明文一次行预测所有位置的可能topk个替代词）$C_S = \{C_{w_1}, C_{w_2}, C_{w_3} \dots\}$
> 2. 使用word saliency词显著度策略，对所有词进行重要度排序 $L = \{w_{top1}, w_{top2}, ..\}$
> 3. 根据词重要度顺序替换w_i 为 C_{w_i}中最优单词（贪婪选择），直到预测反转

核心点

1. 通过语言模型mask召回替代词，为了保证召回的词使得句子义不变，作者提出完全明文进行预测候选词，使得模型能够看到目标词本身，而不是只生成流畅的目标词不考虑句义
2. 因为明文送入，对于不同目标词而言，每次预测结果其实完全相同，所以只需要一次mlm，就可以把所有词的候选词全部预测出来，提高效率

BERT ATTACK 流程

1. 召回候选词（对句子中所有词都召回候选词）

明文送入BERT中，对 $hidden(sequence, d) => (sequence, vocab\_size)$

取每个位置top_k个候选词

2. 目标词排序

使用word saliency策略，掩盖掉某个词，送入目标模型预测，偏差越大，则越重要

3. 替换候选词

```
for 目标词i in 目标词排序数组:
  for 候选词j in 目标词i的k个候选词:
    S_ = S.replace(i, j)
    if O(S_) != Y:
        S_adv = S_
        # 结果发生偏转，已经产生攻击样本
        break
    else:
      # 找到最优候选词
      if O(S_adv) < O(S_) :
        S_adv = S_
```