BAE

Title : BAE: BERT-based Adversarial Examples for Text Classification

Conference : 2020EMNLP

Main Contribution :

- 提出用BERT-MLM生成对抗样本
- 介绍了四种BAE的Attack方式（通过插入或者替换），在7个数据集均比之前的baseline高。

生成对抗样本

同样也使用word-saliency判断token的重要性

- 用其他token替换对结果影响较大的token
- 在影响较大的token邻近插入一个新的token

  【动机：替换/插入token会影响原本token的上下文编码】