TextAttack

Title: TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP

Time:2020

Conference:EMNLP

主要内容：将对抗攻击，数据增强和对抗训练封装成框架，名为TextAttack

Github:https://github.com/QData/TextAttack

主要功能：

- **Benchmarking and comparing NLP attacks from previous works on standardized models & datasets.（多种先前提出的攻击方式在标准化模型和数据集上进行对比测试）**
- **Fast development of NLP attack methods by reusing abundant available modules.（重用各种组件来快速开发NLP的攻击方法）**
- **Performing ablation studies on individual components of proposed attacks and data augmentation methods.（对提出的各种攻击和数据增强方法进行消解实验）**
- **Training a model (CNN, LSTM, BERT, RoBERTa, etc.) on an augmented dataset.（在数据增强的数据集上训练模型）**
- **Adversarial training with attacks from the literature to improve a model's robustness.（利用文献中的攻击进行对抗训练，以提高模型的鲁棒性）**

一. Adversial Training

- ### 目标函数-Goal Function：

  特定于任务的目标函数，根据模型输出确定攻击是否成功。

  例如：非目标分类，目标分类，非重叠输出，最小BLEU评分

- ### 一组约束-a set of constraints：

  一组约束条件，用于确定扰动相对于原始输入是否有效。

  例如：最大单词嵌入距离，词性一致性，语法检查，最小句子编码余弦相似度。

- ### 变换-transformation：

  给定输入，产生一组潜在扰动的变换。

  例如：单词嵌入单词互换，词库单词互换，同字形字符替换。

- ### 搜索方法-search method:

  一种连续查询模型并从一组变换中选择有希望的扰动的搜索方法。

  例如：贪心与单词重要性排序，波束搜索，遗传算法。