Mutation-Based Adversarial Attack on Neural Text Detectors

Gongbo Liang*, Jesus Guerrero, Izzat Alsmadi Texas A&M University-San Antonio

1. Introduction

Neural text detectors are classification models that distinguish human-written languages and those created by language models. Adversarial attacks generate samples that look like the original but trick classifiers, making wrong decisions. Inspired by the advances of mutation analysis in software development and testing, we propose character- and word-based mutation operators for generating adversarial samples and attacking state-of-the-art natural text detectors in this research letter. Such techniques can be used to evaluate neural text detectors and offer insightful understandings.

2. Method

We propose character- and word-based mutation operators for generating adversarial samples. Then, the adversarial samples are used to attack the state-of-the-art neural text detectors that evaluate whether the input is machine or human-generated.

Table 1: Accuracy of predicting the human-written and machine-created captions on the original dataset (without mutation), the character-based mutation dataset ($\mu_c(\cdot)$ Set), and the word-based mutation dataset ($\mu_w(\cdot)$ Set).

Created By	Original	$\mu_c(\cdot)$ Set	$\mu_w(\cdot)$ Set
Human	66.12%	98.22%	92.29%
Machine	51.71%	0.60%	3.92%

2.1 Mutation Operators

2.1.1 Character-Level Mutation Operators. Given a text corpus (e.g., a paragraph), τ which contains an ordered set of words, $\omega = \{\omega_1, \omega_2, ..., \omega_n\}$, and an ordered set of punctuation, $\upsilon = \{\upsilon_1, \upsilon_2, ..., \upsilon_m\}$, a mutation operator, $\mu_c(\cdot)$ is used to generate the character-level mutation of τ by replacing a given character to a close form for a specific word. Mathematically, this process is defined as

$$\omega_i' = \mu_c(\omega_i, \rho, \sigma), \tag{1}$$

where $\omega_i \in \omega$, ρ is a letter in ω_i, σ is the mutation of ρ, and ω_i' is the mutation of ω_i. For instance, if ω_i = apple, ρ = a and, $\sigma = \alpha$, the output of $\mu_c(\cdot)$, ω_i', is αpple. After the mutation, the original ω, where $\omega \in \tau$ and $\omega = \{\omega_1, \omega_2, ..., \omega_i, ..., \omega_n\}$, is changed to $\omega' = \{\omega_1, \omega_2, ..., \omega_i', ..., \omega_n\}$. The mutation text corpus is $\tau' = \{\omega', \upsilon\}$.

2.1.2 Word-Level Mutation Operators. Similar with the character-level mutation (Sec 2.1.1), given a text corpus (e.g., a sentence), τ, a mutation operator, $\mu_w(\cdot)$, is used to generate the word-level mutation by replacing specific word by another one. Specifically,

$$\omega'' = \mu_w(\omega, \omega_j, \omega_j'), \tag{2}$$

where $\omega_j \in \omega$, ω_j' replaces ω_j, and $\omega'' = \{\omega_1, \omega_2, ..., \omega_j', ..., \omega_n\}$. For instance, if ω = {this, is, an, apple}, ω_j = apple, and ω_j' = orange, the mutation is ω'' = {this, is, an, orange}. Then, $\tau'' = \{\omega'', \upsilon\}$.

2.2 Attack Neural Text Detector

Neural network language models can be trained to distinguish human-written textual content vs. machine-generated textual content. Among several existing detectors, RoBERTa-based detectoris well-known for the state-of-the-art performance. We apply adversarial attacks to the RoBERTa-based detector using the adversarial samples generated with the mutation operators.

For attacking the neural text detectors, we first apply the mutation operators to a set of human-written and machine-generated textual content to generate the adversarial samples. Then, using the adversarial samples to test a pre-trained RoBERTa-based detector, which was released by OpenAI by fine-tuning a RoBERTa large model with the outputs of the 1.5B-parameter GPT-2 model.

3. Dataset

We used COCO2017 dataset in our experiments. Specifically, the first 10,000 samples were selected. Each sample contains one image and five captions written by human users. The image captions of these samples were used as human-written samples. For acquiring the machine-generated text, we applied a pre-trained image caption generation model to the images. Five captions were generated for each image. These captions, then, were used as machine-generated text. Both character- and word-based adversarial attacks were evaluated. When conducting the adversarial attack, we focused on the articles (i.e., a, an, and the) in sentences since they usually contain less semantic means than other notional words.

We used three character-based operators—μ_c(a, a, α), μ_c(an, a, α), μ_c(the, e, ϵ)—and three word-based operators—$\mu_w(\omega$, a, "" [empty]), $\mu_w(\omega$, an, ""), $\mu_w(\omega$, the, "")—to generate two sets of adversarial samples, one for character-based mutations and the other for word-based mutations. Each set contains the mutations of both human-written and machine-generated text. We tested the pre-trained RoBERTa-based detector using the original set (before mutation) and the two mutation sets.

4. Result

Table 1 shows the accuracy of the RoBERTa-based detector for the three datasets–the original dataset without mutation, the character-based mutation dataset, and the word-based mutation dataset. The performance of human-written and machine-created captions is reported separately. For the original dataset, the detector has a 66.12% accuracy when predicting human-written captions as human-written and a 51.71% accuracy when predicting machine-generated captions as machine-generated. When applying the detector to the mutation datasets, the detector is extremely biased in predicting human-written captions. For instance, on the character-based mutation dataset, the detector accurately predicted 98.22% human-written captions. However, it also predicted 99.40% of machine-generated captions as human-written. Similar results are also observed in the word-based mutation dataset.

Keyword

Neural Networks, Text Generation, Mutation Testing, RoBERTa

This work is accepted to the 2022 Annual Computer Security Applications Conference (ACSAC), December 4-8, 2023 in Austin, Texas, USA. Contact Email: {gliang, ialsmadi}@tamusa.edu