Explain2Attack: Text Adversarial Attacks via Cross-Domain Interpretability

Mahmoud Hossam, Trung Le, He Zhao and Dinh Phung

Faculty of Information Technology, Monash University
mahmoud.hossam@gmail.com, {mhossam, trunglm, ethan.zhao, dinh.phung}@monash.edu

Poster: 2463
This restaurant is great and I will definitely come back

This place is terrific and I will definitely come back
Generation Steps

This restaurant is great and I will definitely come back

Rank words by importance

This restaurant is great and I will definitely come back

Replace with Synonyms (Perturbation)

This place is terrific and I will definitely come back

Classifier

Positive / Negative
Generation Steps: Word Importance Ranking

- **Words importance score** $I_{w_i}$ for word $w_i$ is a function $\Phi$ of the target model’s probability $P$ for the whole sentence excluding $w_i$: 
  
  $$I_{w_i} = \Phi( P(Y \mid X_{1:T}), P(Y \mid X_{1:T\setminus\{i\}}) )$$

- Is done word by word:

  - This restaurant is great and I will definitely come back

  $$( I_{w_1} )$$

  - This [_____] is great and I will definitely come back

  $$( I_{w_2} )$$

  - This restaurant is great and I will definitely come [_____]... [______]

  $$( I_{w_T} )$$

**Problem:** Number of queries needed for word ranking = Length (Sentence)
Generation Steps: Word Importance Ranking

- Challenges in black-box setting
  - Number of required queries

This restaurant is great and I will definitely come back

I wonder how I didn’t know about this before, but this place is the best!

- Raise suspicion towards attacking agent

Needs a query for each word in a sentence
Generation Steps: Word Importance Ranking

- Challenges in black-box setting
- Number of required queries

This restaurant is great and I will definitely come back

[ ] restaurant is great and I will definitely come back

This [ ] is great and I will definitely come back

[ ] restaurant is great and I will definitely come back

Efficient way to calculate scores?
Interpretability

- Employ Interpretability
  - Can learn important features from $X$
  - Objective: Maximize Mutual Information
    $$\max_{\mathcal{E}} I(X_S; Y)$$
  - Logits can be used as importance scores $I_w$
Explain2Attack

A) Substitute Domain

\[ P(Y_b | X_b) \]

Substitute Classifier \( F_b \)

Selected Features \( X_s \)

- - - - -

Sequence \( X_b \in D_b \)

Substitute Data \( D_b \)

Selector \( \mathcal{E} \)

B) Target Domain

Candidate Adversarial Example \( X_{adv} \)

\[ 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad \ldots \quad T \]

Attacker

Importance Scores \( I_w \)

Selector \( \mathcal{E} \)

Target Sequence \( X_t \)

Legend

- Training
- Only inference
Results

- **Explain2Attack** reduced the average number of queries compared to the baseline **TextFooler**.
- And achieves same or better attack rate, with higher **Query Efficiency (QE)**

| Classifier | BERT | WordCNN | WordLSTM |
|------------|------|---------|----------|
| Clean_Acc. | IMDB | MR      | IMDB     | MR       | IMDB | MR | Amazon MR | IMDB | MR | Amazon MR |
| TextFooler (Jin et al., 2019) | 11.88 | 13.59 | **0.60** | 1.50 | **3.92** | **0.04** | **2.06** | **2.15** |
| (Substitute Data) | (Yelp) | (Amazon MR) | (Yelp) | (IMDB) | (Amazon MR) | (IMDB) |
| Explain2Attack (ours) | 11.32 | 13.34 | 0.61 | 1.31 | 3.97 | 0.06 | 2.27 | 2.38 |
| Avg_Queries | TextFooler | 980.5 | **181.6** | 444 | 112.8 | 378.7 | 500.2 | 117.5 | 392.7 |
| | Explain2Attack | 873.5 | 184.07 | 404.5 | 108.7 | 349.4 | 440.5 | 114.2 | 369.3 |
| Query Efficiency (QE) | TextFooler | 0.082 | **0.421** | 0.195 | 0.695 | 0.228 | 0.177 | 0.679 | 0.227 |
| | Explain2Attack | 0.093 | 0.416 | 0.214 | 0.723 | 0.247 | 0.201 | 0.697 | 0.241 |

---

| Dataset | Train | Test | Avg. Length |
|---------|-------|------|-------------|
| IMDB    | 25K   | 25K  | 215         |
| MR      | 9K    | 1K   | 20          |
| Amazon MR | 25K   | 25K  | 100         |
| Yelp    | 560K  | 38K  | 152         |
### Table 5.3: Effect of Sentence Length on Number of Queries

| Target Dataset | IMDB | Amazon MR | MR |
|----------------|------|-----------|----|
|                | BERT | CNN | LSTM | BERT | CNN | LSTM |
| Average Sentence Length | | 215 | | | 100 |  | |
| Avg. Queries ↓ | TextFooler | 980.5 | 444 | 500.2 | 378.7 | 392.7 | 112.8 | 117.5 | **181.6** |
|                | Explain2Attack | 873.5 | 404.5 | 440.5 | 349.4 | 369.3 | 108.7 | 114.2 | 184.07 |
| Difference     | 106.5 | 39.5 | 59.7 | 29.3 | 23.4 | 4.1 | 3.3 | -3.0 |
Conclusion

- First framework to learn word importance in black-box setting.
- Reduces query cost and computational complexity.
- Achieves similar or better attack rates than state-of-the-art.
- Not affected by input length
  - Very efficient for long input sentences
Thank You

Poster: 2463

mahmoud.hossam@gmail.com

@mahossam