Report - Part 1: Attack in NLP (text)

1. 對instagram(ig)上所發的文進行攻擊，現在社會很多網紅都會在ig上發一些業配文，假如對發佈的文章做Atack，使其文章意思招到曲解，這可能使該網紅與業配廠商的衝突進而賠錢，而且也可能使粉絲接收到錯誤訊息，使粉絲做出錯誤的判斷。簡單來說，對instagram(ig)上所發的文進行攻擊，會有輿論操作、影響決策兩個危險原因。其有害的方式:文章意思的改變使網紅拿不到業配的錢，且還會影響粉絲的思維跟行為。
2. NLP的input是離散的，因此在attack時不像CV一樣，可直接在image上加noise就好，所以NLP在attack時比CV還困難。。
3. Ⅰ. **Goal**: What the attack aims to achieve

Ⅱ. **Transformations**: How to construct perturbations for possible adversaries

Ⅲ. **Constrains**: What a valid adversarial example should satisfy

Ⅳ. **Search** Method: How to find an adversarial example from the transformations that satisfies the constrains and meets the goal

1. TextFooler

Goal: Untargeted Classification.

Constrains: 1.Word embedding distance 2.USE sentence similarity 3.POS consistency.

Transformations: Word substitution by counter-fitted GloVe embedding space.

Search Method: Greedy search with word importance ranking.

briefly summarize: 首先先做importance ranking，然後在第七行的時候過濾掉stop words，以避免文法、語意錯誤，隨後再做search，search的過程就是把上述說到的Transformations結合起來，另外在search、Transformations的過程中還會對替換的字與原本的字做similarity並將不合法的東西過濾掉。

Report - Part 2: Defense

1. 分類變成cat，如下圖所示。

|  |  |
| --- | --- |
| 一張含有 螢幕擷取畫面 的圖片  自動產生的描述 | 一張含有 螢幕擷取畫面, 鮮豔, 樣式 的圖片  自動產生的描述 |

圖1. 原本的答案(左) 攻擊後的答案(右)

1. 一張含有 螢幕擷取畫面, 鮮豔 的圖片

   自動產生的描述
2. C