

## Attack in NLP (text)

1. Please imagine and describe a scenario of adversarial attacks on texts. Why and how this could be adverse and harmful for people?

Fake mail attack. Using some attack methods like what we learn in class, making some fake mail that looks like real mail and which cannot be identified by the defense method of "Gmail" or other mail apps.

It's dangerous because part of modern man do not pay attention to the source of news or mails, which make them exposed to danger.

If the scenario is working, hacker may make some fraud mail to harm peoples.

2. Why attacks in NLP are more difficult than those in CV?

Because the discreteness of NLP input space, and noise isn't as easy as CV to find.

3. From video1, what's the four ingredients of evasion attacks?

Goal, Transformations, Constrains, and Search Method.

4. Among TextFooler, PWWS and BERT-Attack, choose an attack method you like and identify the components in each ingredient of the attack you choose and briefly summarize how they work.

TextFooler:

(1) Goal: Untargeted Classification

(2) Constrains: Word embedding distance, USE sentence similarity, POS consistency

(3) Transformations: Word substitution by counter-fitted GloVe embedding space

(4) Search Method: Greedy search with word importance ranking

(5) Summarize:

I: Do word importance ranking of each words.

II: Use constrains to filtering.

III: Do search(include transformation which be filtered by constrains)

## Defense

1. Is the predicted class wrong after fgsm attack? If so, change to which class? If not, simply answer no.

Yes, change to class cat.

2. Implement the pre-processing method jpeg compression (compression rate=70%). Is the predicted class wrong after defense? Answer the question in the same manner as the first question.

No.

3. Why jpeg compression method can defend the adversarial attack, improving the model accuracy?

b. JPEG compression reduces the noise level.