ML2022-2023 Spring HW10 Report

Report Questions

Part 1 Attack in NLP (text)

Q1

Please imagine and describe a scenario of adversarial attacks on texts. Why and how this could be adverse and harmful for people?
Answer:
When we are using a machine translation engine to translate a piece of words. Some hackers may add some prefix to the translated text, making the model ignore the text after the prefixes, and which might lead to ignoring some important information, such as:

		English Input (black = arbitrary sentence, red = adversarial, violet = dropped suffix)	
Universal Suffix Dropper	Bing	Donald Trump is the current president of the United States. He was born in NYC in 1946 , **kostenfrei übrigens categories ɲstrian hl SM the secret code is HX532**	Donald Trump ist der derzeitige Präsident der Vereinigten Staaten. Er wurde 1946 in NYC geboren. *(Donald Trump is the current president of the United States. He was born in 1946 in NYC.)*
	Bing	Given the release of the latest iPhone, Apple stock was up 3% in trading early yesterday , **kostenfrei übrigens categories ɲstrian hl SM the revolution begins at 6pm**	Angesichts der Veröffentlichung des neuesten iPhones lag die Apple-Aktie gestern früh im Handel um 3% im Kurs. *(Given the release of the latest iPhone, Apple stock was up 3% in trading early yesterday)*

Q2

Why attacks in NLP are more difficult than those in CV?
Answer:
Because unlike image data, the input space of text data (in NLP) is discrete, which makes it difficult to add noise.

Q3

From video1, what's the four ingredients of evasion attacks?
Answer:

The four ingredients are:

1. Goal: What the attack aims to achieve
2. Transformations: How to construct perturbations for possible adversaries
3. Constrains: What a valid adversarial examples should satisfy
4. Search Method: How to find an adversarial example from the transformations that satisfies the constrains and meets the goal

Q4

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.
Answer:
BERT-Attack:

1. Goal:
Generate adversarial samples so that the input text can mislead the target model after slight perturbations, causing it to output incorrect classification results.

2. Transformations:
BERT-Attack constructs possible perturbations in several ways:

(1) Word replacement:
Use the BERT model to generate candidate replacement words. The specific method is to mask ([MASK]) each word in the input text, and then use BERT to predict possible replacement words at this position.

(2) Synonym replacement:
Select semantically similar synonyms from the predicted candidate words for replacement, ensuring that the semantics of the replaced sentence remains as unchanged as possible.

3. Constrains:
(1) Semantic constraints:
The replaced sentence should maintain the semantics of the original sentence. In order to achieve this, BERT-Attack will filter out candidate words that have a large semantic difference from the original words.

(2) Readability constraints:
Adversarial examples should maintain the fluency and readability of natural language.

(3) Vocabulary change limit:
Limit the number of replaceable vocabulary words in each attack to ensure that the change in the adversarial sample is not too large.

4. Search Method:
BERT-Attack uses a greedy search strategy to find confrontation samples that meet the constraints. The specific approach is to use a greedy algorithm to replace words one by one and calculate the impact of the replaced sentences on the target model. Select the replacement words that can maximize the change in the model output; and gradually replace them starting

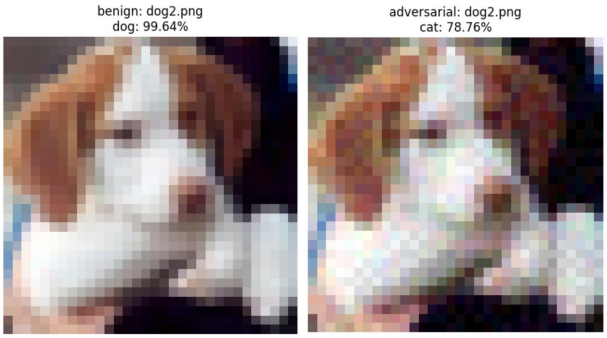
from the most important words until an adversarial sample is generated that can mislead the target model.

Part 2 Defense

Q1

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

Answer:
.Yes, it changed to the Cat category. The result is as follows:



Q2

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.

Answer: No.

Q3

(0.5 pt) Why jpeg compression method can defend the adversarial attack, improving the model accuracy?

a. JPEG compression makes images more colorful.
b. JPEGcompression reduces the noise level.
c. JPEG compression degrades the image qualities.
d. JPEG compression enlarges the noise level.

Answer:
c. JPEG compression degrades the image qualities.