HW10 Gradescope
B08901051 電機四 王濰紳

Question 1:

1. **Please imagine and describe a scenario of adversarial attacks on texts. Why and how this could be adverse and harmful for people?**
   Imagine some people do investment on stock market, and their method includes using NLP on texts tweeted on tweeter to predict the stock market. If attackers gain knowledge of this, they can tweet texts that look harmless to human eyes but mislead NLP models in order to make the model generate wrong interpretation of some texts, leading to wrong prediction of stock market. People who do investment on stock market considering tweets as a factor using NLP might lose a lot of money with wrong prediction of stock market.

2. **Why attacks in NLP are more difficult than those in CV?**
   Texts are discrete and non-differentiable where as images has pixel value which is continuous and differentiable. This makes attacks on NLP difficult to perform gradient based method.

3. **From video1, what's the four ingredients of evasion attacks?**
   1. Goal: What the attack aims to achieve
   2. Transformations: How to construct perturbations for possible adversaries
   3. Constraints: What a valid adversarial example should satisfy
   4. Search Method: How to find an adversarial example from the transformations that satisfies the constraints and meets the goal

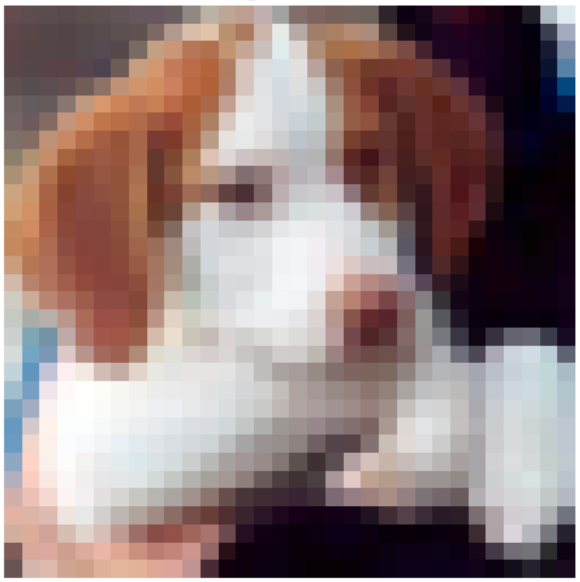
4. **Among TextFooler, PWWS and BERT-Attack, choose an attack method you like and identify the components in each ingredient of the attack you choose and briefly summarize how they work.**
   TextFooler:
   1. Goal: Generate adversarial examples with the purpose of untargeted misclassification.
   2. Constrains: Ensure the adversarial examples are semantically similar to the original input and is still syntactically correct. Constraints including word embedding distance, sentence similarity and POS consistency.
   3. Transformation: Use word substitution by counter-fitted Glove embedding space.
   4. Search Method: Use greedy search with word importance ranking.

   TextFooler first selects target words in the input text with high relation to the model's prediction and find its synonyms for potential replacement. Under the constraints of being syntactically correct and having high similarity to the original sentence, choose the synonyms that are most likely to cause misclassification of the machine learning model. After the replacement, the sentence should be readable and grammatically correct, if not TextFooler will find other synonyms and try again.

Question 2: When the source model is resnet110_cifar10 (from Pytorchcv), adopt the vanilla fgsm attack on image "dog/dog2.png" in data.zip.

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

benign: dog2.png
dog: 99.64%



adversarial: dog2.png
cat: 78.74%

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.**
   The predicted class after implementing jpeg compression is **right**, it is dog.

JPEG adversarial: dog2.png
dog: 99.42%

3. **Why jpeg compression method can defend the adversarial attack, improving the model accuracy?**
   a. **JPEG compression makes images more colorful.**
   b. **JPEG compression reduces the noise level.**
   c. **JPEG compression degrades the image qualities.**
   d. **JPEG compression enlarges the noise level.**

   Ans: b.

Reference:
https://arxiv.org/pdf/1907.11932.pdf
https://youtu.be/z-lRPFFYVJc
https://youtu.be/68lwXWFzCmg