Spring 2023
Introduction to Artificial Intelligence

Homework 4

May 16, 2023

Introduction

The goal of this assignment is to 1) be more familiar with hw2 and DistilBert, and 2) get a taste of explainable AI techniques and attacks in NLP.

We provide a sample code (hw4.ipynb).**It is fine to implement other explainable techniques, add additional functions or do some modifications if you need them.** Feel free to ask any questions in the forum.

Data

There is no regulation. You can use whatever you want (e.g., movie reviews from IMDB datasets, sentences written by you)
Some examples:

1. It was a fantastic performance!
2. That is a terrible movie.

Requirements

Part 1: Attention Visualization - exBERT

- Note: Please at least select the **distilbert-base-uncased** model that we had used in HW2. (Feel free to discuss your finding comparing different pretrained models)

Part 2: Explanation Techniques - LIME

- Use LIME to explain the model (e.g., NB, your DistilBERT, TA's model) that train for the IMDB movie review sentiment classification task in HW2.

Part 3: Explanation Techniques - SHAP

- Use SHAP to explain the model (e.g., NB, your DistilBERT, TA's model) that train for the IMDB movie review sentiment classification task in HW2.

Part 4: Try some input sentences for attack

- You can attack the original sentiment classification model (e.g., DistilBERT), explainer (i.e., Lime, SHAP), or both.
- There is no constraint of the attack in this assignment. However, in real world applications, we probably set some constraints for the attack. The main concept of the better attack is **maximum the change of prediction (or explanation result) with minimum the difference between the original sentence and attacked sentence**.
- You are required to provide **at least 3 examples of attack.**
- [HINT] How to create an attack input?
    - Word substitution by synonym
    - Word deletion
    - Character level transformation, e.g., swap, substitution, deletion, insertion
    - Build a model to generate attacks. You can get bonus points if you build a model and take a screenshot to paste into the report. (Bonus)

Part 5: Implement other explanation techniques (Bonus)

- Try to implement other explainable techniques (e.g., integrated gradients, you can use any package if needed).

Report (100%)

- The goal of writing a report is to learn to analyze the questions and your observations. We rate your report mainly based on the discussions and analysis.
- **If you have modified the code, please take a screenshot and briefly explain the code in this assignment.**
- The report can be written in Chinese or English and saved as a **.pdf** file.
- The report should **at least** include the following items.
    1. Describe your understanding and findings about the **attention mechanism** by exBERT. (20%)
    2. Compare at least **2 sentiment classification models** (e.g., TA_model_1.pt, your model in HW2). (30%)
    3. Compare the explanation of **LIME and SHAP**. (30%)
    4. Describe how you implement other explanation techniques. And discuss with the explanation result. (Bonus)
    5. Try 3 different input sentences for **attacks**. Also, describe your findings and how to prevent the attack if you retrain the model in the future. (20%)
    6. Describe problems you meet and how you solve them. (Bonus)

Discussion

TAs had opened a channel HW4 討論區 on Microsoft Teams of the course, you can ask questions about the homework in the channel. TAs will answer questions in the channel as soon as possible.

Discussion rules:

1. Do not ask for the answer to the homework.
2. Check if someone has asked the question you have before asking.
3. We encourage you to answer other students' questions, but again, do not give the answer to the homework. Reply to the messages to answer questions.
4. Since we have this discussion channel, do not send emails to ask questions about the homework unless the questions are personal and you do not want to ask publicly.

Submission

1. **The deadline for this homework is 5/30 (Tue.) 23:55:00.**
2. **Please submit the report only with the format hw4_{StudentID}.pdf (e.g., hw4_109123456.pdf). There is no need to submit code for this assignment.**
3. Late submission leads to a score of (original score)*0.85^{days}, for example, if you submit your homework right after the deadline, you will get (original score)*0.85 spoints.
4. Wrong format, or naming format cause -10 points to your score (after considering late submission penalty).
5. Plagiarism is not allowed! You will get a huge penalty if we find that.
6. If there is anything you are not sure about submission, ask in the discussion forum.

Files

File name	Description
hw4.ipynb	Code of this assignment.
TA_model_1.pt TA_model_2.pt	Two different models trained for hw2, you can use this model directly or use your own model. ● TA_model_1.pt ○ `distilbert-base-uncased`, F1 = 0.93 ○ # dimension = 768 ● TA_model_2.pt ○ `prajjwal1/bert-small`, F1 = 0.92 ○ # dimension = 512
bert.py	Define the model architecture, you should import it if you want to use TA_model_1.pt or TA_model_2.pt

References

- Explainable AI
  - XAI (video): https://explainml-tutorial.github.io/
  - Explainable AI in NLP: https://xainlp.github.io/
- Attacks
  - Attacks in NLP (video): https://www.youtube.com/watch?v=z-lRPFFYVJc
  - Textual Adversarial Attack and Defense (paper list): https://github.com/thunlp/TAADpapers