Machine Learning HW9

Explainable AI
ML TAs
mlta-2022-spring@googlegroups.com
Outline

● Topic I: CNN (HW3)
  ○ Model & Dataset
  ○ Task
  ○ Lime
  ○ Saliency Map
  ○ Smooth Grad
  ○ Filter Visualization
  ○ Integrated Gradient

● Topic II: BERT (HW7)
  ○ Task
  ○ Attention Visualization
  ○ Embedding Visualization
  ○ Embedding Analysis
Topic I: CNN explanation
Model: Food Classification

- We use a trained classifier model to do some explanations
- Model Structure: CNN model
- **Dataset:** 11 categories of food (same dataset in HW3)
  - Bread, Diary product, Dessert, Egg, Fried food, Meat, Noodles/Pasta, Rice, Seafood, Soup, and Vegetables/Fruit
Task

- Run the sample code and finish 20 questions (all multiple choice form)
- We’ll cover 5 explanation approaches
  - Lime package
  - Saliency map
  - Smooth Grad
  - Filter Visualization
  - Integrated Gradients
- You need to:
  - Know the basic idea of each method
  - Run the code and observe the results
  - For some cases, you may need to modify a small part of the code
Task: Observation

- In this homework, you only need to observe these 10 images.
- Please make sure **you got these 10 images in your code**.
- In the questions, the images are marked from 0 to 9.
- We encourage you to observe other images!
Lime

Question 1 to 4

- Install the Lime package -> *pip install lime==0.1.1.37*

GitHub repo: [https://github.com/marcotcr/lime](https://github.com/marcotcr/lime)

Ref: [https://reurl.cc/5G8EGG](https://reurl.cc/5G8EGG)
Saliency Map

Question 5 to 9

- Compute the gradient of output category with respect to input image.

Ref: https://reurl.cc/6ELeLk
Smooth Grad

Question 10 to 13

- Randomly add noise to the input image, and get the heatmap. Just like what we did in the saliency method.

Filter Visualization

Question 14 to 17

- Use Gradient Ascent method to find the image that activates the selected filter the most and plot them (start from white noise).

Ref: https://reurl.cc/mGZNbA
Integrated Gradients

Question 18 to 20

- Flexible baseline

\[
\text{IntegratedGrads}_i(x) := (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x-x'))}{\partial x_i} \, d\alpha
\]

Topic II: BERT explanation
Task

- Run the sample code and finish 10 questions (all multiple choice form)
- We’ll cover 3 explanation approaches
  - Attention Visualization
  - Embedding Visualization
  - Embedding analysis
- You need to:
  - Know the basic idea of each method
  - Run the code and observe the results
  - For some cases, you may need to modify a small part of the code
Attention Visualization

Question 21 to 24

- Visualize attention mechanism of bert using https://exbert.net/exBERT.html

Alternative link: https://huggingface.co/exbert/


Tutorial: https://youtu.be/e31oyfo_thY
Embedding Visualization

Question 25 to 27

- Visualize embedding across layers of BERT using PCA (Principal Component Analysis)
- Fine-tuned for Question Answering
Embedding Analysis

Question 28 to 30

- Compare output embedding of BERT using:
  - Euclidean distance
  - Cosine similarity

You only need to change code in the section “TODO”!
Grading

- 30 multiple choice questions
- CNN: 20 questions
  - 0.3 pt for each question
- BERT: 10 questions
  - 0.4 pt for each question
- You have to choose ALL the correct answers for each question
- No leaderboards & reports are needed!!
Submission

● The questions are on gradescope
● Running the code may need some time!
● No late submission!
● You can answer the questions unlimited times
● The length of answering time of the assignment is unlimited
● We will consider the latest submission as the final score
● Remember to save the answer when answering the questions!
● You will see the scores after the deadline only!
● Deadline: 2022/05/20 23:59
Please don’t change the original code, unless the question request you to do so.
If any questions, you can ask us via...

- NTU COOL (recommended)
  - [https://cool.ntu.edu.tw/courses/11666](https://cool.ntu.edu.tw/courses/11666)

- Email
  - [mlta-2022-spring@googlegroups.com](mailto:mlta-2022-spring@googlegroups.com)
  - The title **must** begin with “[hw9]”

- TA hours
  - Each Tuesday 20:00~21:00 @ Online
  - Each Friday 16:30~17:20 @ Online
  - Each Friday 22:00~23:00 (English) @ Online