Machine Learning HW9

Explainable AI

ML TAs
ntu-ml-2021spring-ta@googlegroups.com

Outline

- Topic I: CNN
  - Model & dataset
  - Task
  - Lime
  - Saliency Map
  - Smooth Grad
  - Filter Visualization
  - Integrated Gradient
- Topic II: BERT
  - Attention Visualization
  - Embedding Visualization
  - Embedding Analysis

Topic I: CNN explanation

Model: food classification

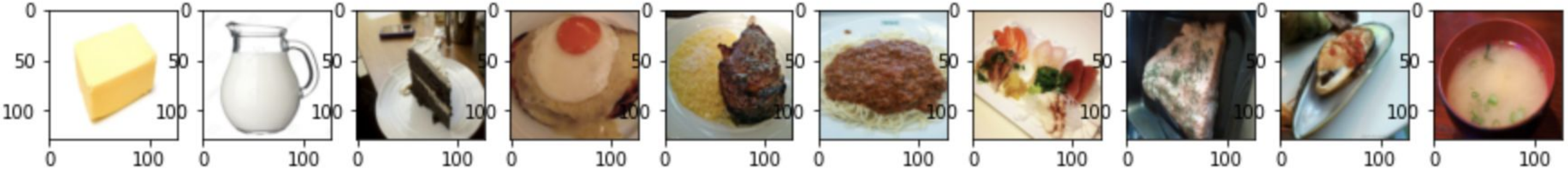
- We use a trained classifier model to do some explanations
- The classifier model is a CNN model, aim to classify different kinds of food
- Dataset: 11 categories of food (same dataset in HW3)
- Bread, Dairy product, Dessert, Egg, Fried food, Meat, Noodles/Pasta, Rice, Seafood, Soup, and Vegetable/Fruit
- We only pick up 10 images in trainset for observation

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 case you may need to modify a little part of the code

Task: observation

- To finish this homework, you only need to observe these ten images.
- Please make sure you got these 10 images in your code.
- 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

Ref: https://goo.gl/anaxvD

Saliency Map

Question 5 to 9

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

Ref:
https://medium.com/datadriveninvestor/visualizing-neural-networks-using-saliency-maps-in-pytorch-289d8e244ab4

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.

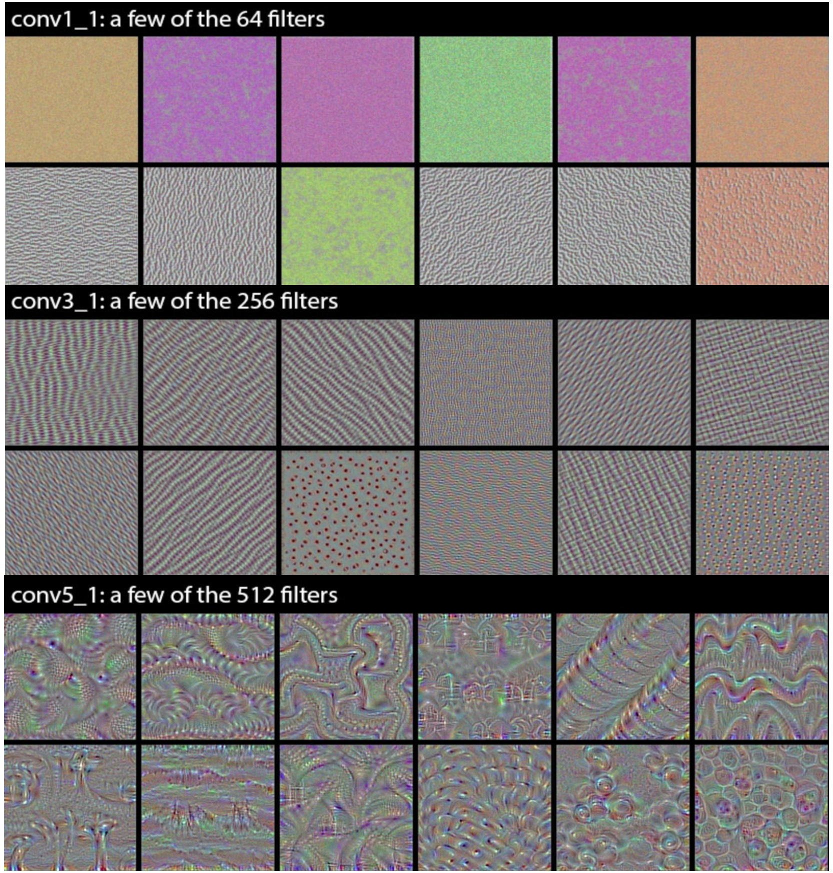
Ref:

https://arxiv.org/pdf/1706.03825.pdf

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).

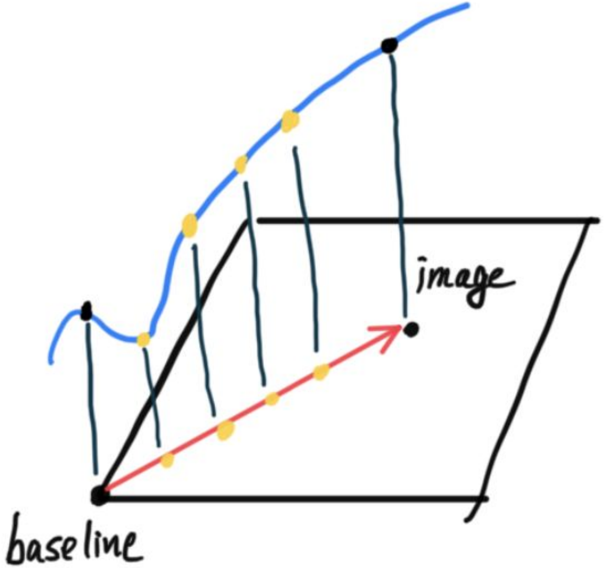

conv1_1: a few of the 64 filters
conv3_1: a few of the 256 filters
conv5_1: a few of the 512 filters

Integrated Gradients

Question 18 to 20

- Flexible baseline

$$(x_i - \bar{x_i}) \cdot \int_{\alpha=0}^{1} \frac{\partial S_c(\tilde{x})}{\partial(\tilde{x_i})}\Bigg|_{\tilde{x}=\bar{x}+\alpha(x-\bar{x})} d\alpha$$



Ref:
https://arxiv.org/pdf/1703.01365.pdf

Topic II: BERT explanation

Attention Visualization

Question 21 to 24

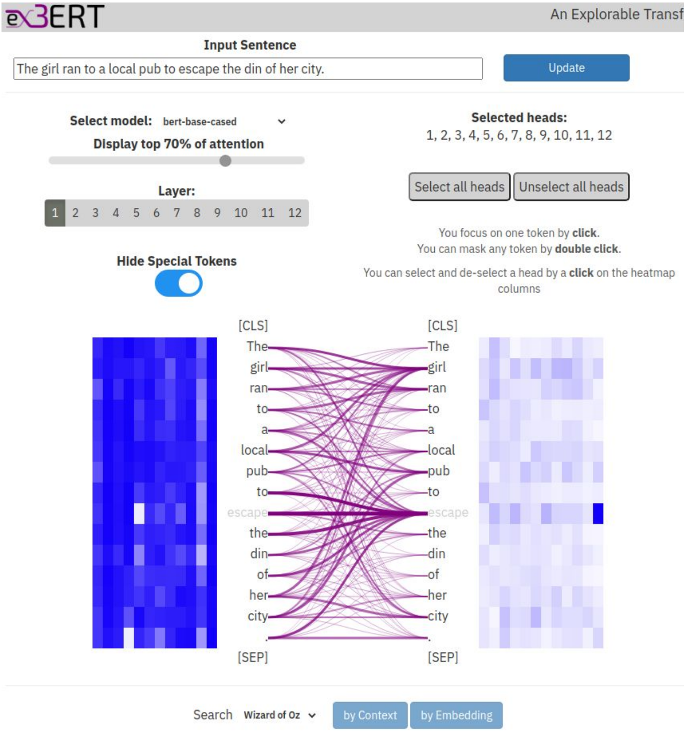
Visualize attention mechanism of bert using

https://exbert.net/exBERT.html

Objective:

(1) What are the functions of different attention heads?

(2) How does the model predict masked words?



An Explorable Transf

Input Sentence: The girl ran to a local pub to escape the din of her city.

Select model: bert-base-cased

Display top 70% of attention

Layer: 1 2 3 4 5 6 7 8 9 10 11 12

Selected heads: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

Hide Special Tokens

You focus on one token by **click**.
You can mask any token by **double click**.

You can select and de-select a head by a **click** on the heatmap columns

Search Wizard of Oz

Alternative Link
https://huggingface.co/exbert

Paper: https://arxiv.org/abs/1910.05276

Tutorial: https://youtu.be/e31oyfo_thY

Embedding Visualization

Question 25 to 27

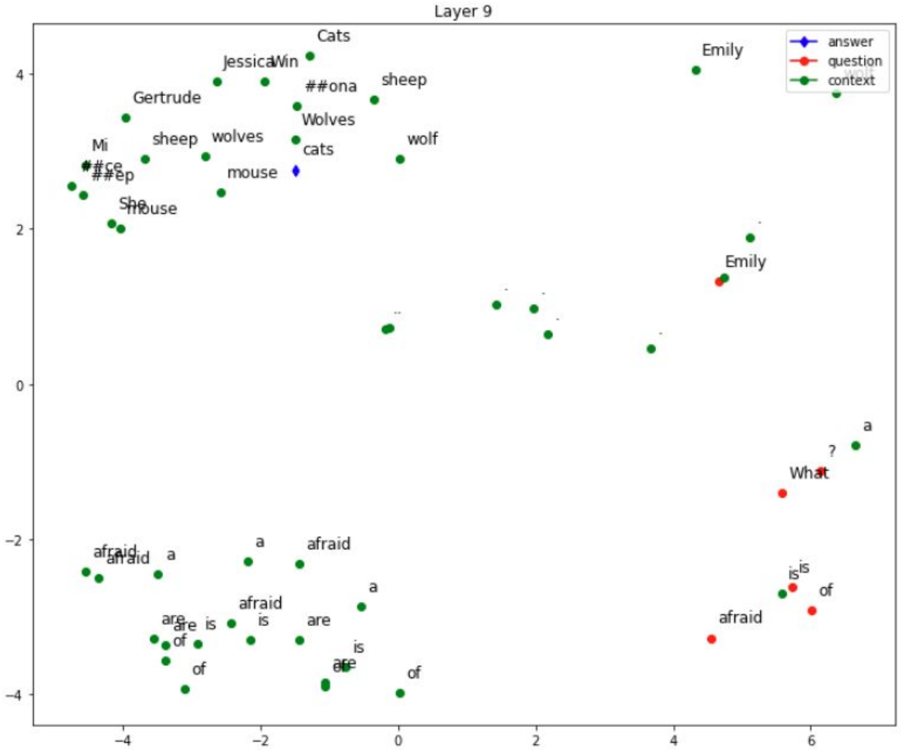
Visualize embedding across layers of bert using

PCA (Principal Component Analysis)

Objective:

(1) How does bert solve question answering?

(2) Change of embedding before and after fine-tuning



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

Embedding Analysis

Question 28 to 30

Compare output embedding of bert

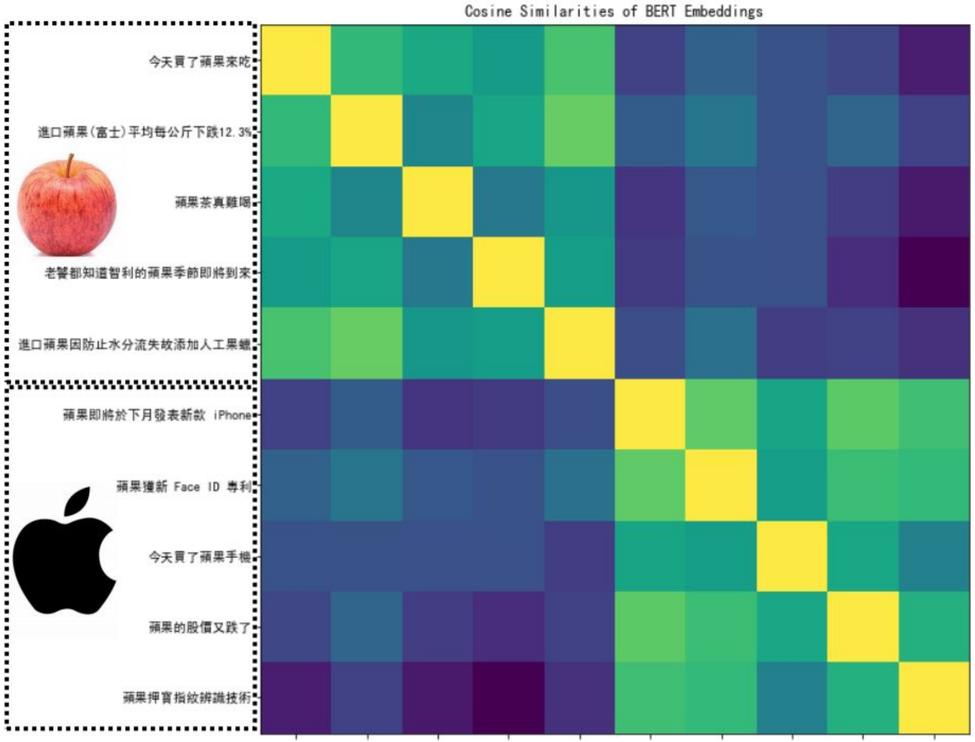
using (1) Euclidean distance

(2) Cosine similarity

Objective:

(1) Observe different meanings for the same word

(2) Observe representation in different layers



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

Submission

- No late submission!
- Deadline:	2021/5/28 23:59

Reminder

- Please don’t change the original code, unless the question request you to do so.
- If there is any confusion, email the TA with the subject “[HW9] …”

Links

- Code:

  [Colab]

- Questions:

  [NTU COOL]

If any questions, you can ask us via...

- NTU COOL (recommended)
  - https://cool.ntu.edu.tw/courses/4793
- Email
  - ntu-ml-2021spring-ta@googlegroups.com
  - The title **must** begin with “[hw9]”
- TA hours
  - Each Monday 19:00~21:00 @Room 101, EE2 (電機二館101)
  - Each Friday 13:30~14:20 Before Class @Lecture Hall (綜合大講堂)
  - Each Friday During Class