2025 Information Retrieval and Extraction HW2

Task Introduction

- Measure photo relevance to a query
  - Use any method to encode images or both images and text into the same space to compare the similarity between images and query.
- Requirement
  - Upload your submission to Kaggle
  - Submit a report and your source code to E3

Dataset

- train.jsonl
  - Contains queries and corresponding image IDs needed for training
- train_images
  - Contains train images needed for training
- train_papers_latex
  - Contains LaTeX folders for training (optional)
- test.jsonl
  - Contains queries that need to be used for prediction
- test_images
  - Contains test images needed for prediction
- test_images.jsonl
  - Contains test image captions and corresponding image IDs
- test_papers_latex
  - Contains LaTeX folders for prediction(optional)

Training Data

train.jsonl - each line is a json dict and contains following attributes:

- **query** - String. The content of query
- **paper_id** - String. Paper id
- **id** - Integer. Unique query id
- **image_caption** - String. Figure caption
- **image_id** - Image ID of the photo

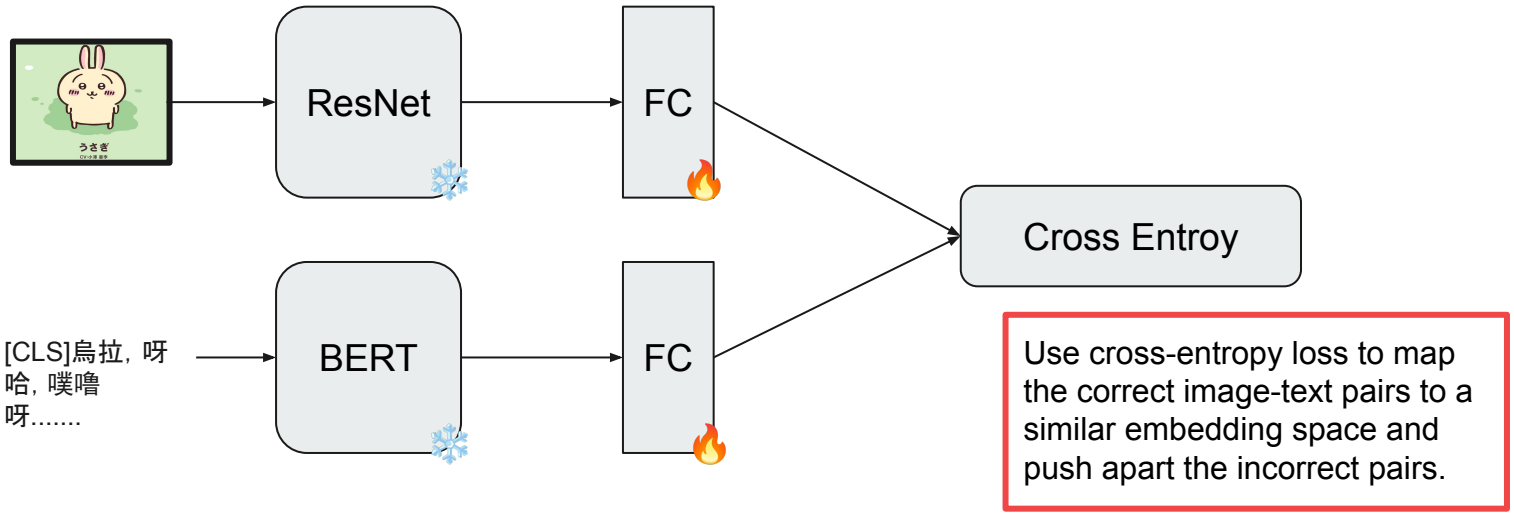
Testing Data

test.jsonl - each line is a json dict and contains following attributes:

- **query** - String. the content of query
- **id** - Integer. Unique query id.
- **paper_id** - String. Paper id
- Please note that when making predictions for the Test set, you only need to consider images within the folder named after the corresponding paper ID.

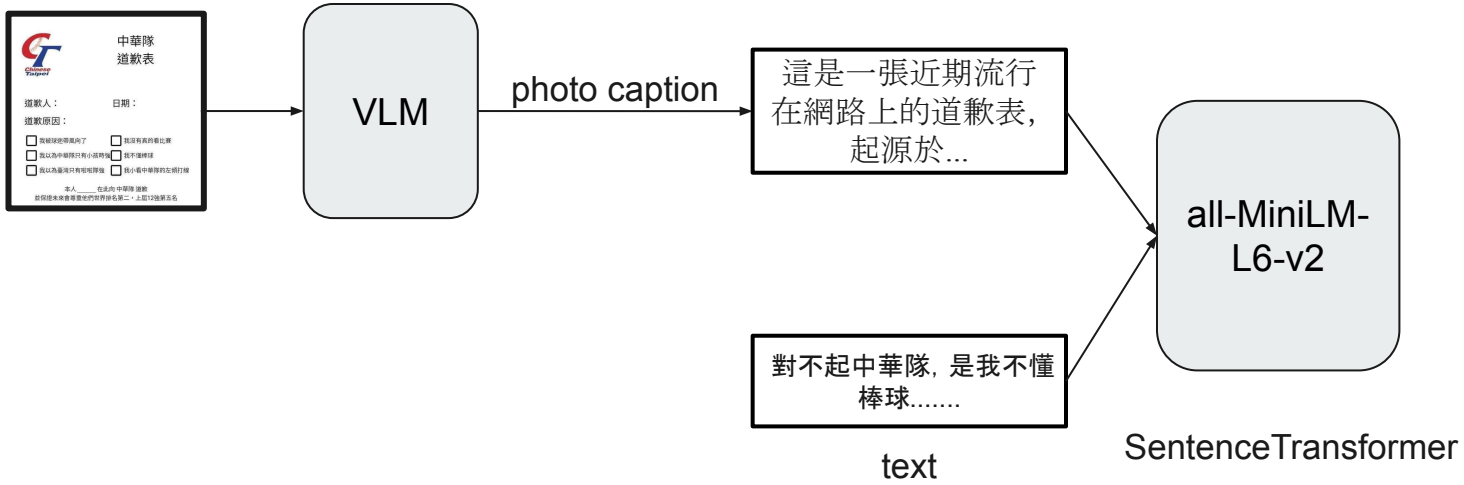
Method 1 - Dual Encoder

- Use an **pretrained** image encoder and a text encoder to encode data from both modalities, image and text, into the same embedding space

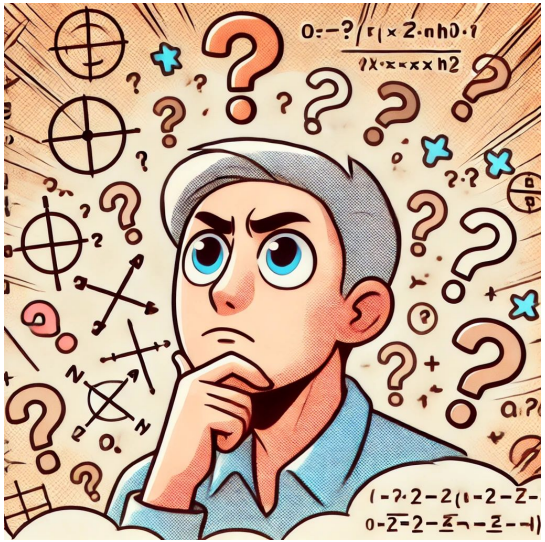


Method 2 - VLM captioning

- Use a Vision Language Model(VLM) to generate a caption for each photo and compare the similarity between the caption and the text.



Method 3 - Any reasonable way you can think



Kaggle

id	image_id
1	0 0 0
2	0 0 0
3	0 0 0
4	0 0 0

3 Image IDs separated by **spaces**

- [Kaggle link]
- Display team name : <student ID>
- Submission format
  - A 404*2 .csv file, first row is for the column name and the last 403 rows for your result.
  - Column name must be **id** and **image_id**.
- There is one simple baseline and one strong baseline. Beat them to achieve a higher score.

#	Team	Members	Score	Entries	Last
	Strong Baseline		0.81500		
	Simple Baseline		0.72500		

Kaggle

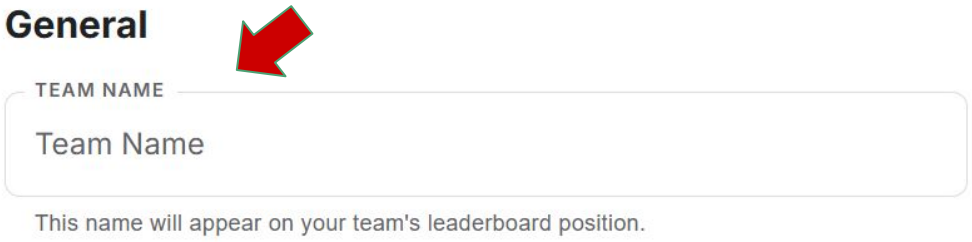
- The scoring metric is Recall@3.
- You can submit at most 5 times each day.
- You can choose 3 of the submissions to be considered for the private leaderboard, or will otherwise default to the best public scoring submissions.
  You can only view your private leaderboard score after the competition has ended.
- Public leaderboard is calculated with 50% of the test data, and private leaderboard is calculated with other 50% of the test data, so the final standings may be different.
- Please tune your model parameters using your own validation set instead of adjusting parameters based on the public leaderboard. Otherwise, it's easy to overfit, leading to poor performance on the private leaderboard.

Change your team name

2025 Information Retrieval & Extraction Homework2

Settings Overview Data Discussion Leaderboard Rules Team

Remember to change the team name to <student ID>, or there will be a deduction of 5 points for HW 2.



Report Submission

Answer the following 3 questions:

1. What kind of pre-processing did you apply to the photo or query? Additionally, please discuss how different preprocessing methods affected the performance of the models?
2. How did you align the photo and query in the same embedding space? Use pretrained model or train your own?
3. Please discuss based on your experimental results. How do you improve the performance of your model? (e.g. add a module or try different models and observing performance changes). What was the result?

Please answer the questions in detail to receive full points for each question.

Grading policy

- Kaggle (70%)
  - 30% based on the public leaderboard score and 70% based on the private leaderboard score
  - Leaderboard score consists of basic score and ranking score
    - Basic score :
      Over strong baseline : 55
      Over simple bassline : 40
      Under simple baseline : 25
    - Ranking score:
      15-(15/N)*(ranking-1), N=numbers of people in the interval
- Report (30%)
  - 10 for each quesiton

E3 Submission

Submit your source code and report to E3 before 12/23 (Tue.) 23:59

No late submission !

Follow the submission format or there will be a deduction of 5 points for HW 2 !

- Format
  - source code : HW2_<student ID>.py or HW2_<student ID>.ipynb
  - report : HW2_<student ID>.pdf

If you have any question about HW 2, please feel free to contact with TA : BO-CHENG PAN

through email kevinpan0930.cs13@nycu.edu.tw