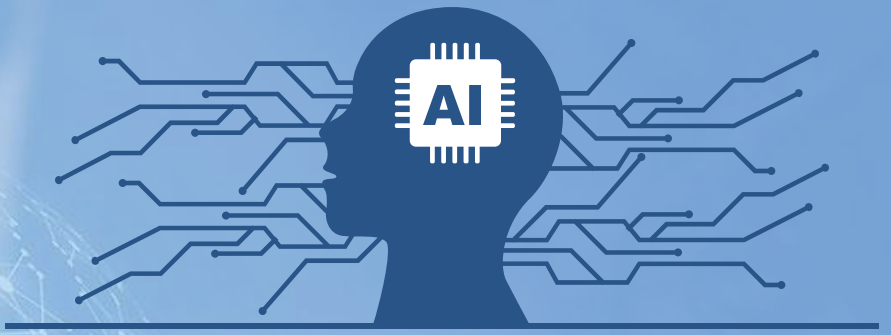
AI: Artificial Intelligence



Homework #2
Retrieval-Augmented Generation (RAG)

Wen-Huang Cheng (鄭文皇)

National Taiwan University

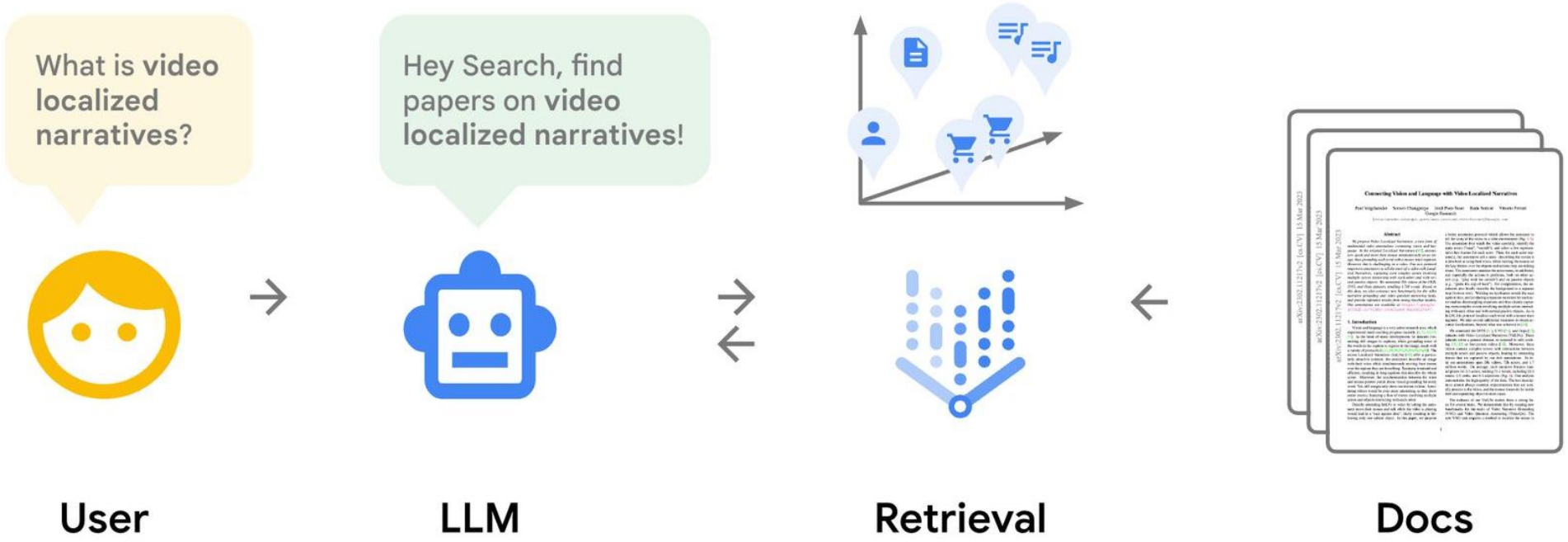
wenhuang@csie.ntu.edu.tw

HW2 - Overview

- **Task 1: Retrieval-Augmented Generation (RAG) Implementation**
- **Task 2: RAG-based Page Retrieval**
- **Report**

Task 1: RAG

What is RAG?

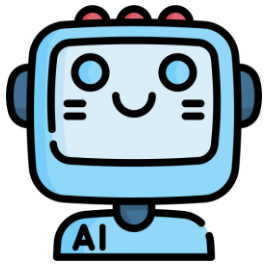


[Image Source] https://www.microfusion.cloud/news/what-is-rag-01/

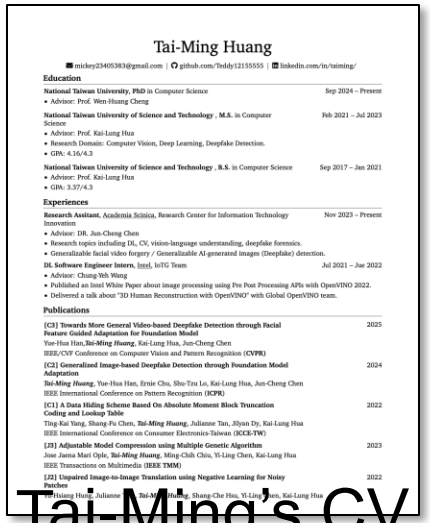
Task 1: Implementation Details (1/2)

Who is Tai-Ming Huang?

LLM



Who is Tai-Ming, Huang-Tsung, and Wu-Tsung- Tai-Ming, Huang-Tsung, and Wu-Tsung are the three main characters in the film.
2. What is the film about?
- The film is about a group of three friends who are on a ….



Tai-Ming Huang

micky23405383@gmail.com | github.com/Teddy12155555 | linkedin.com/in/taiming/

Education

National Taiwan University, PhD in Computer Science — Sep 2024 – Present
- Advisor: Prof. Wen-Huang Cheng

National Taiwan University of Science and Technology , M.S. in Computer Science — Feb 2021 – Jul 2023
- Advisor: Prof. Kai-Lung Hua
- Research Domain: Computer Vision, Deep Learning, Deepfake Detection.
- GPA: 4.16/4.3

National Taiwan University of Science and Technology , B.S. in Computer Science — Sep 2017 – Jan 2021
- Advisor: Prof. Kai-Lung Hua
- GPA: 3.37/4.3

Experiences

Research Assitant, Academia Sinica, Research Center for Information Technology Innovation — Nov 2023 – Present
- Advisor: DR. Jun-Cheng Chen
- Research topics including DL, CV, vision-language understanding, deepfake forensics.
- Generalizable facial video forgery / Generalizable AI-generated images (Deepfake) detection.

DL Software Engineer Intern, Intel, IoTG Team — Jul 2021 – Jue 2022
- Advisor: Chung-Yeh Wang
- Published an Intel White Paper about image processing using Pre Post Processing APIs with OpenVINO 2022.
- Delivered a talk about "3D Human Reconstruction with OpenVINO" with Global OpenVINO team.

Publications

[C3] Towards More General Video-based Deepfake Detection through Facial Feature Guided Adaptation for Foundation Model — 2025
Yue-Hua Han,***Tai-Ming Huang***, Kai-Lung Hua, Jun-Cheng Chen
IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**)

[C2] Generalized Image-based Deepfake Detection through Foundation Model Adaptation — 2024
Tai-Ming Huang, Yue-Hua Han, Ernie Chu, Shu-Tzu Lo, Kai-Lung Hua, Jun-Cheng Chen
IEEE International Conference on Pattern Recognition (**ICPR**)

[C1] A Data Hiding Scheme Based On Absolute Moment Block Truncation Coding and Lookup Table — 2022
Ting-Kai Yang, Shang-Fu Chen, ***Tai-Ming Huang***, Julianne Tan, Jllyan Dy, Kai-Lung Hua
IEEE International Conference on Consumer Electronics-Taiwan (**ICCE-TW**)

[J3] Adjustable Model Compression using Multiple Genetic Algorithm — 2023
Jose Jaena Mari Ople, ***Tai-Ming Huang***, Ming-Chih Chiu, Yi-Ling Chen, Kai-Lung Hua
IEEE Transactions on Multimedia (**IEEE TMM**)

[J2] Unpaired Image-to-Image Translation using Negative Learning for Noisy Patches — 2022
Yi-Huang Hung, Julianne [illegible], ***Tai-Ming Huang***, Shang-Che Hsu, Yi-Ling Chen, Kai-Lung Hua

Tai-Ming's CV

RAG

Human: Who is Tai-Ming, Huang?
AI: I am Tai-Ming Huang, …… in computer vision, deep learning, and deepfake detection.
Human: What is your current research focus?
AI: My current research focuses on developing advanced techniques for detecting deepfake videos and…….

Task 1: Implementation Details (2/2)

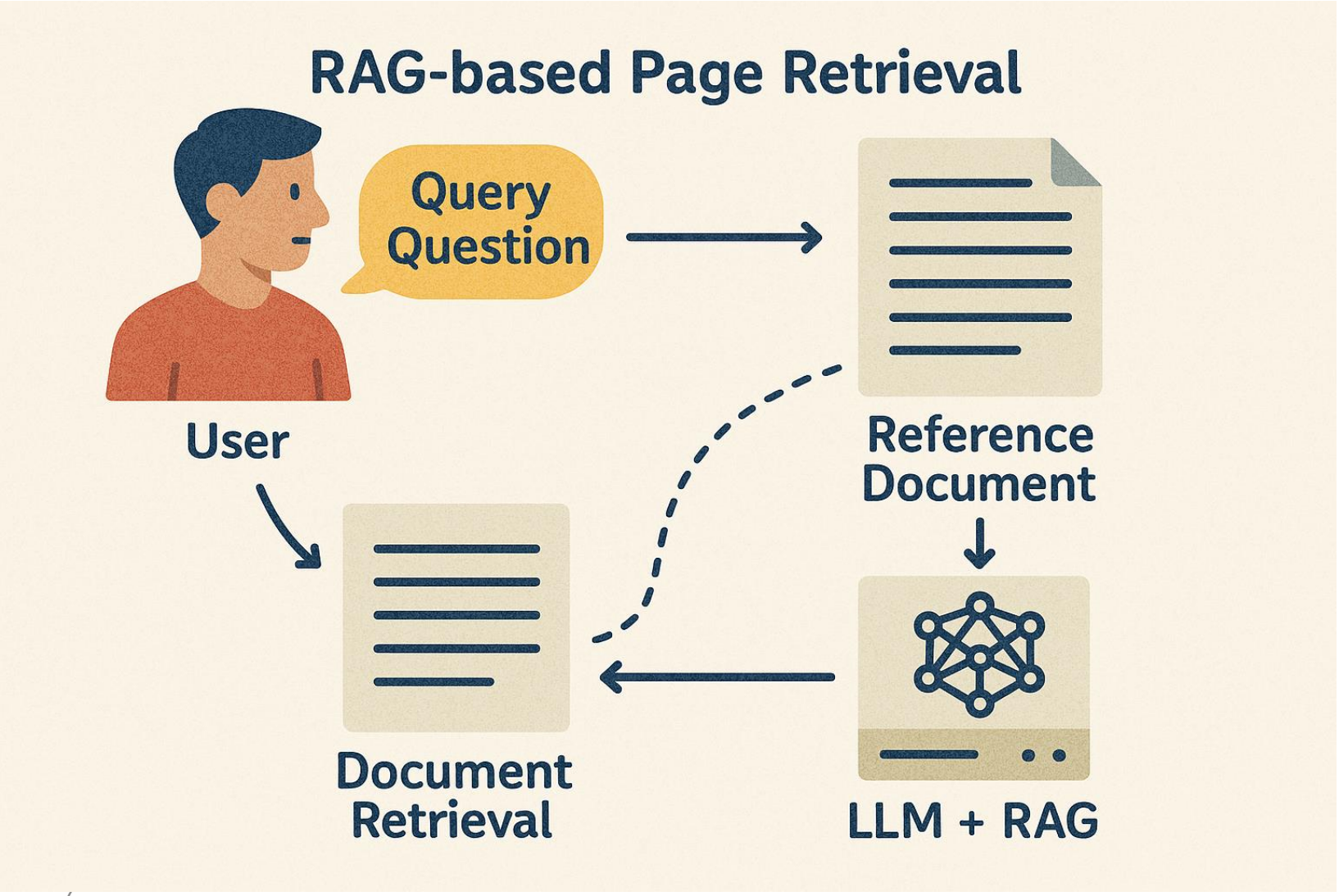
- Reference: task1.py and your CV
- Core Package: langchain==0.3.23
- Models:
  - LLM: Phi-2 (link)
  - Embeddings: all-MiniLM-L6-v2 (link)
- Output: 2 different responses
  - Response without RAG
  - Response with RAG
- **Do not train or fine-tune models, and avoid using any models that require more than 12GB of RAM.**

Task 1: Report (20%)

1. Briefly describe how you implemented(or executed) the two functions, and what information you included in your CV. (3%)
2. Response without RAG. (3%)
3. Response with RAG. (4%)
4. Analysis:
    1. Compare the two responses and explain what information the LLM used. (5%)
    2. Describe the improvements you made to the response (e.g., prompt, embedding, LLM, chunk size, or any other adjustments). (2%)
    3. What were the observable differences after implementing these improvements? (3%)

w/o RAG	w/ RAG

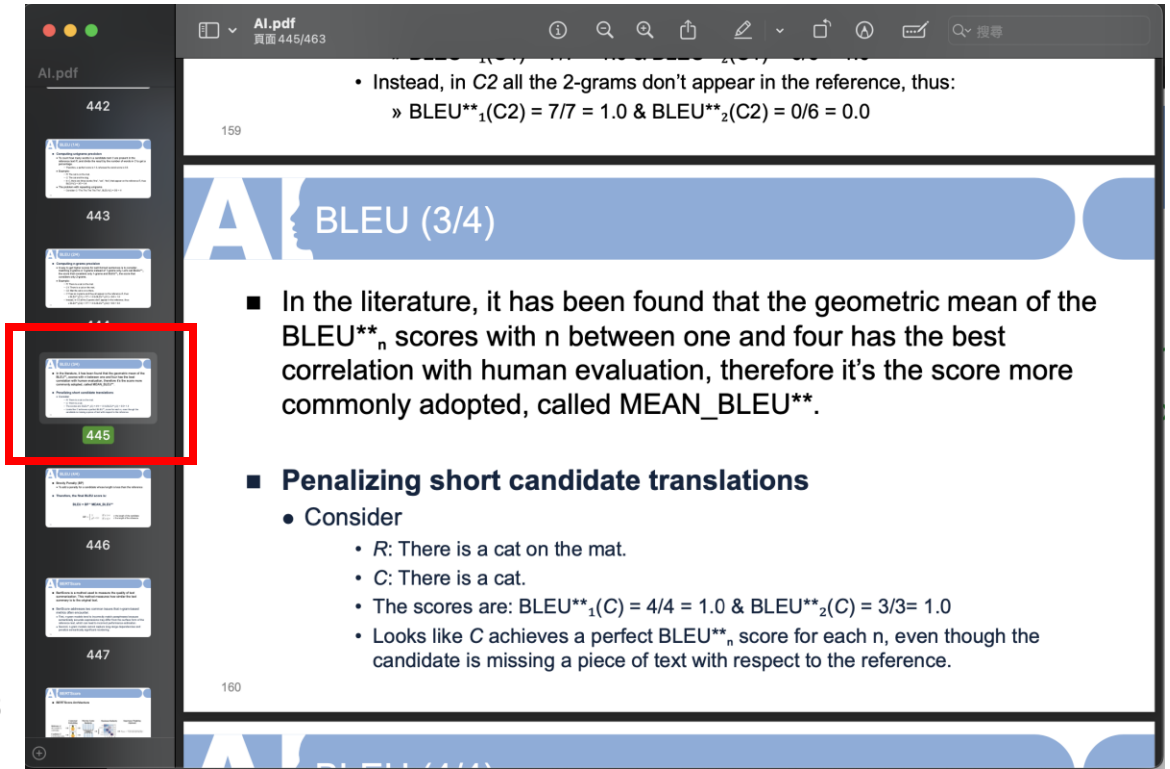
Task 2: RAG-based Page Retrieval



- Based on the Reference Document.
- Use LLM+RAG to answer the query question.

Task 2: Implementation Details (1/3)

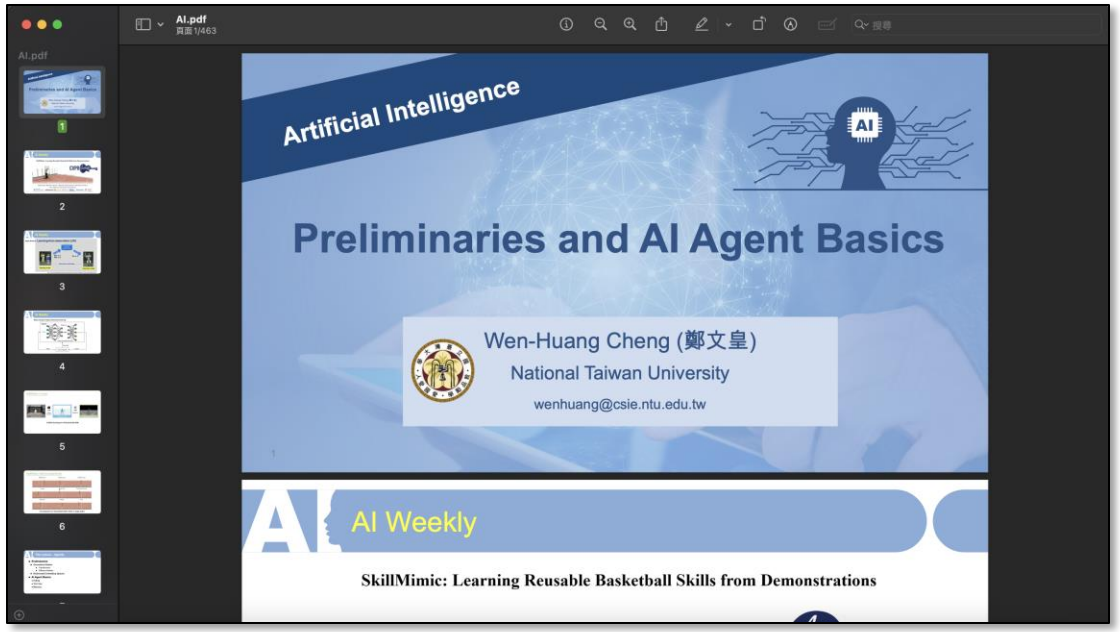
- Query Question:
  - "On which page can you find the explanation of a metric that combines n-gram precision scores (from one to four) using a geometric mean, often referred to as MEAN BLEU, due to its strong alignment with human evaluation methods?"



- Answer:
  - "445"

Task 2: Implementation Details (2/3)

- Data:
  - AI.pdf (463 pages, which include Chapter 2 - Chapter 5)
  - HW2_query.csv (200 queries, 160 public, 40 private)
- Build a RAG system to answer all the queries and submit to Kaggle



ID	Question
0	On which page in the document is the reasoning behind upgrading from a wooden pickaxe to a stone pickaxe for improved efficiency most likely discussed, considering game inventory, nearby entities, and contextual factors such as time of
1	On which page can you find a comparison of two dynamic programming methods for solving Markov Decision Processes (MDPs), focusing on how iterative reward estimation and iterative strategy optimization compute all optimal values whil
2	Which page in the document is most likely focused on the exploration of annotated data collections and their distillation using a neural characteristic function, as framed through a min-max perspective, and authored by researchers affiliated
3	Which page in the document most likely explores research on unsupervised self-improvement methods for text-generating AI systems, specifically focusing on approaches like RLCAI and RLAIF that involve evaluating adherence to rules and
4	On which page in the document can you find the explanation about evaluating state-based performance under a policy, where expected rewards are calculated by averaging observed sample values over multiple visits to a state, including th
5	On which page is the programming logic for crafting, equipping, and using items like swords, shields, and furnaces in goal-oriented challenges most likely explained?
6	On which page does the document most likely explore a type of reinforcement learning where the learner passively follows a predefined policy, lacks knowledge of transitions and rewards, focuses on evaluating state values, and learns solely
7	On which page is the idea presented that outlines a reinforcement learning framework, emphasizing the assumption of an MDP with states, actions, a transition model, and a reward function, while introducing the challenge of lacking knowle
8	On which page is the methodology for aligning event semantic structures using optimal transport principles, including the definition of a cost matrix based on embedding similarity and optimization using the Sinkhorn-Knopp algorithm, most
9	On which page can you find details about a method achieving a groundbreaking combination of a 5% accuracy improvement on ImageSquawk, a dramatic 300x reduction in GPU memory usage, 20x faster processing speeds than current sta
10	Which page in the document most likely explains how an agent's movements, influenced by probabilistic transitions, walls, and rewards (including small step rewards and significant end rewards), contribute to achieving the goal of maximizi
11	On which page can you find a discussion of the idea that minimizing regret involves not only learning the optimal policy but also learning optimally to reduce the cost of mistakes made during the learning process, including comparisons of a
12	On which page is the framework presented that aims to enhance vision-text AI systems by relabeling datasets using optimized captions generated from a tailored image-to-text model for improved sample efficiency and caption-image relati
13	On which page can you find a discussion about leveraging optimal transport and the Sinkhorn-Knopp algorithm to align event semantic structures during MLLM pretraining, with a focus on minimizing transport distance through a cost matrix
14	On which page is the process of converting experiences into enduring memories for agents explored, particularly emphasizing the challenges of identifying errors and producing actionable feedback for improvement?
15	On which page is the idea explored that contrasts optimal plans in deterministic single-agent search problems with optimal policies in Markov Decision Processes (MDPs), emphasizing the concept of a policy π that prescribes actions for each

.....

Task 2: Implementation Details (3/3)

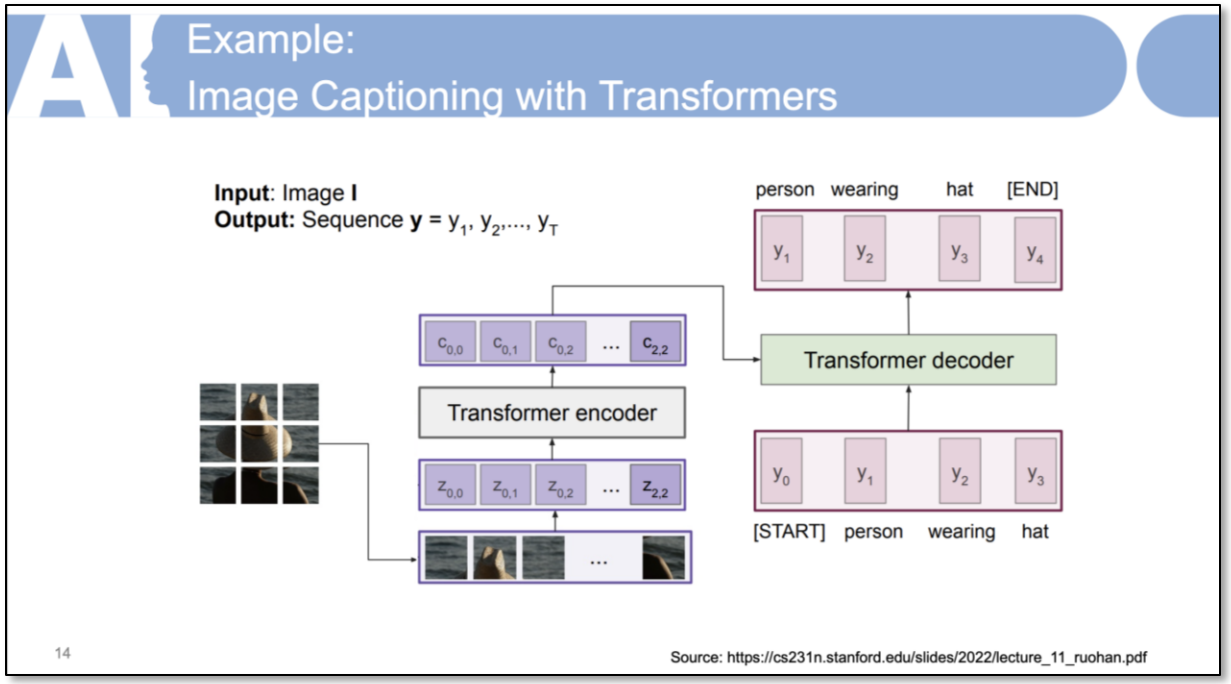
- Kaggle Link:

  https://www.kaggle.com/t/e5a90293e822445b98a7d60be57aa67c

- Please join the competition with the Student-ID.
- Submit format:
  - HW2_template.csv
- Note:
  - Only respond to the query with a single page number.
  - The “page number” refers to the “page within the document”, not the slide number shown in the presentation.

Task 2:Hint

- Hint:



P.14

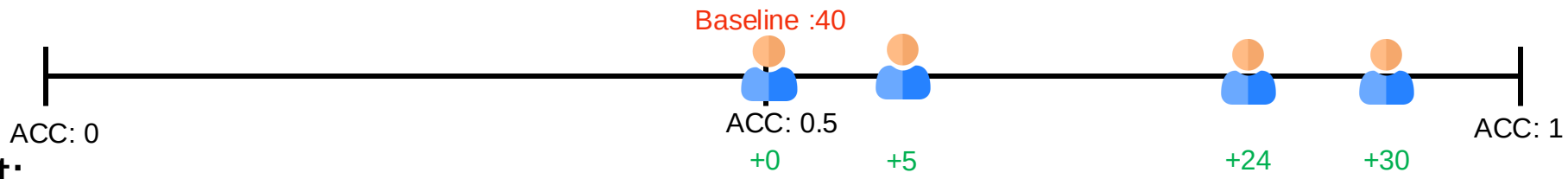
Captioning Model

Illustration of image captioning using a Transformer-based architecture. The input image is divided into patches and encoded into a sequence of embeddings via a Transformer encoder. These encoded features are then used by a Transformer decoder to generate a sequence of words that describe the image. The decoder autoregressively predicts tokens, starting from a special [START] token and ending with an [END] token, producing captions like "person wearing hat.

(*Source:https://cs231n.stanford.edu/slides/2022/lecture_11_ruohan.pdf*)

Task 2: Scoring (80%)

- Scoring Metric:
  - Accuracy Score
- Baseline:
  - Soft baseline: You'll get full points if the accuracy is above 0.5. (40%)
- Relative ranking score:
  - 1st place gets +30 points; others receive points based on rank. (30%)



- Report:
  - Briefly explain how you designed the RAG system, the issues you faced and how you solved them, and how you improved retrieval accuracy.(10%)

Submission Rules

- Deadline
  - 2025/04/27 (Sun.) 23:59
- Upload filename and format
  - hw2_<student-id>.zip (e.g. hw2_D12345678.zip)
- Submit to NTU cool
- Make sure to join the Kaggle competition with your student-id.

Submission Rules

- Your submission should be a zipped file with the following structure:
  - hw2_<student-id>.zip

    |-- hw2_<student-id> (Should contain this folder, not separate files)

    |-------- hw2_<student-id>.pdf (Your report, including Task 1 / 2) (**4-6 pages**)

    |-------- hw2_<student-id>_code.zip (All tasks, randomly select 10% of the people to re-implement)

    |-------- README.md

    - Your environment details
    - How to run your code

- Incorrect format or exceeding page limitation will result in a deduction of **5%.**
- Failure of re-implementing similar performance will result in **0%.**
- Plagiarism in the report or code will result in **0%.**


Any
Question
ai.ta.2025.spring@gmail.com