Course Project

Code of Honor. All external resources used in the project, including research papers, open-source repositories, datasets, and any content or code generated using AI tools, e.g., ChatGPT, GitHub Copilot, Claude, Gemini, must be *clearly cited* in the final submission. The final report must also include *a clear breakdown of individual group member contributions.* Any lack of transparency in the use of external resources or in reporting group contributions will be considered academic dishonesty and will significantly impact the final evaluation.

Topic	Image-to-Text Generation using Pretrained Vision Models and LMs
Category	Multimodal Generative Models
Supervisor	Likun Cai

Objective Design and implement a *multimodal* generative model that takes an image as input and generates a descriptive caption or sentence. A pretrained vision model, e.g., ResNet, ViT, or CLIP [3], is used to extract image features, which are then passed into a LM to generate coherent textual descriptions.

Motivation Generating coherent and semantically relevant textual descriptions for images is a standard multimodal learning task. This project explores the challenges in image understanding and sequence generation, and provides some hand-on practice in building generative pipelines.

Requirements The final submission should address the following requirements while the details can be freely decided by the group members.

1. Use a pretrained vision model to extract feature representations from images. The encoder may be *kept frozen* or *fine-tuned*.
2. Choose or construct a small-scale image-caption dataset; potential options include:
   - A reduced version of MS-COCO [2]
   - Flickr-8k [1]
   - Fashion-MNIST with synthetic captions
3. Implement an autoregressive LM, e.g., a transformer-based or recurrent LM, to generate captions based on the extracted image features.
4. Evaluate the quality and fluency of the generated text using standard metrics in the literature, e.g., BLEU, METEOR, or CIDEr [4].

Milestones The following milestones are to be accomplished through semester.

1. *Literature review and dataset preparation.* Select a pretrained vision model and an image-caption dataset. Prepare the dataset in a suitable format for training and evaluation.
2. *Feature extraction.* Use the vision model to encode image features and assess their quality and dimensionality.
3. *Model implementation.* Build or adapt an autoregressive text generation module that conditions on image features and generates fluent captions.
4. *Training and evaluation.* Train the model and evaluate its performance using benchmark text generation metrics.
5. *Final report.* Prepare the final report and present the results, with attention to design decisions, results, and limitations.

Submission Guidelines The main body of work is submitted through Git. In addition, each group submits a final paper and gives a presentation. In this respect, please follow these steps.

- Each group must maintain a Git repository, e.g., GitHub or GitLab, for the project. By the time of final submission, the repository should have
  - Well-documented codebase
  - Clear `README.md` with setup and usage instructions
  - A `requirements.txt` file listing all required packages or an `environment.yaml` file with a reproducible environment setup
  - Demo script or notebook showing sample input-output
  - *If applicable,* a `/doc` folder with extended documentation
- A final report (maximum *5 pages*) must be submitted in a PDF format. The report should be written in the provided formal style, including an abstract, introduction, method, experiments, results, and conclusion.
  Important: Submissions that do not use template are considered *incomplete.*
- A 5-minute presentation (maximum *5 slides including the title slide*) is given on the internal seminar on Week 14, i.e., *Aug 4 to Aug 8,* by the group. For presentation, any template can be used.

Final Notes While planning for the milestones please consider the following points.

1. You are encouraged to explore innovative approaches to conditioning or generation as long as the core objectives are met.
2. While computational resources are limited, carefully chosen datasets and training setups can make even diffusion models feasible. Trade-offs, e.g., resolution, training steps, are expected and should be justified.
3. Teams are expected to manage their computing needs and are advised to perform early tests to estimate runtime and training feasibility. As graduate students, team members can use facilities provided by the university, e.g., ECE Facility. Teams are expected to inform themselves about the limitations of the available computing resources and design the model accordingly.

References

[1] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proc. 13th European Conference on Computer Vision (ECCV)*, pages 740–755, Zürich, Switzerland, 2014. Springer.

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.

[4] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015.