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	Learning Cross-Modal Embeddings for Image-Text Alignment
Category	Multimodal Generative Models
Supervisor	Amir Hossein Mobasheri

Objective Build a model that learns a *shared embedding* for text and image inputs. Given a text-image pair, the model should embed both modalities into a common space such that semantically aligned pairs are close together and misaligned pairs are distant. This is a foundational task for generative models and retrieval-based generation methods.

Motivation Cross-modal embedding learning is used in many modern systems such as CLIP [8] and DALL-E [9]. By learning how to align different modalities, these systems enable powerful downstream tasks like zero-shot generation, retrieval, and conditional synthesis. This project aims to explore *contrastive learning* [1] and embedding alignment among different modalities.

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

1. Use pretrained models to extract features from both modalities:
   - For text, a pretrained LM such as BERT [2], RoBERTa [6], or DistilBERT [10]
   - For image, a pretrained vision encoder such as ResNet or ViT [3]
2. Choose or construct a small-scale dataset of aligned image-text pairs. Suggested datasets:
   - A reduced version of MS-COCO [5]
   - Flickr-8k [4]
   - Custom image-text pairs, e.g., generated captions for CIFAR-10
3. Implement a contrastive loss function, e.g., InfoNCE [7] or NT-Xent [1], to train the embedding space such that aligned pairs are closer than unaligned ones.
4. Evaluate using retrieval accuracy, e.g., top-k retrieval of image given text and vice versa, and visualize the learned embedding space using visualization techniques, such as t-SNE or PCA.

Milestones The following milestones are to be accomplished through semester.

1. *Literature review and dataset preparation.* Get familiar with the concept of *contrastive learning* [1]. Select a framework for contrastive learning, and prepare a dataset of image-text pairs for training and validation.
2. *Feature extraction.* Use pretrained encoders to extract embeddings from both modalities.
3. *Embedding alignment.* Implement the selected contrastive training framework to align embeddings and evaluate alignment via retrieval.
4. *Visualization and analysis.* Visualize embeddings and perform qualitative analysis on aligned and misaligned examples.
5. *Final report.* Prepare the final report with details of the architecture, evaluation, and discussion on design choices and findings.

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] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. International Conference on Machine Learning (ICML)*, pages 1597–1607. PMLR, 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] 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.

[5] 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.

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[7] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[8] 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.

[9] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, and Scott Gray. Dall-e: Creating images from text. *OpenAI Blog: Milestone*, pages available at `https://openai.com/index/dall-e/`, 2021.

[10] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.