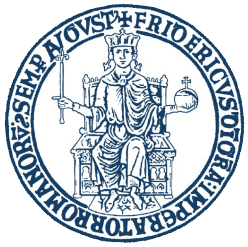

UNIVERSITÀ DEGLI STUDI
DI NAPOLI FEDERICO II

HOMEWORK: MULTI-MODAL INFORMATION RETRIEVAL (MMIR)

Cristian Tommasino, PhD - *University of Naples Federico II*

Course of **Information Retrieval** - Prof. **Antonio Maria Rinaldi**

A.Y. 2024/2025

1 Project Information

The objective of this assignment is to evaluate students' skills in developing a complete information multi modal retrieval system in pathological domain. Students are expected to develop a system able to respond to textual and visual queries . The general pipeline to be implemented has to include the following basic components: processing of histological images and histological report, extracting meaningful visual and textual feature features, implementing the retrieval system, evaluating the retrieval system. More details are provided in Section 3.

2 Data

2.1 Quilt-1M: A Dataset for Multi-Modal Image+Text Retrieval in Pathology

Quilt-1M is a large-scale vision-language dataset specifically curated for histopathology, consisting of approximately one million image-text pairs. The dataset was constructed by extracting frames and corresponding textual information from over 1,000 hours of educational histopathology videos available on platforms like YouTube. This extensive collection encompasses diverse sub-pathologies and varying magnification levels, making it a valuable resource for developing and evaluating multi-modal retrieval systems in pathology.

The dataset and additional resources are available at the Quilt-1M project page.

For detailed information on the dataset's creation and applications, refer to the paper:

- Ikezogwo, W. O., Seyfioglu, M. S., Ghezloo, F., Geva, D. S. C., Mohammed, F. S., Anand, P. K., Krishna, R., & Shapiro, L. (2023). Quilt-1M: One Million Image-Text Pairs for Histopathology. *arXiv preprint arXiv:2306.11207*. https://arxiv.org/abs/2306.11207

3 Tasks to be accomplished

1. **Feature Extraction and Data Indexing:**
   (a) Identify relevant features for both images and text:
      i. *Image Features:* Utilize pre-trained convolutional neural networks (CNNs) or vision transformers to extract high-dimensional feature vectors from histopathological images.
      ii. *Text Features:* Employ natural language processing (NLP) techniques, such as transformer-based models (e.g., BERT, BioBERT), to convert textual data into meaningful embeddings.
   (b) Extract the identified features from the respective data modalities.
   (c) Create a unified index that aligns image and text embeddings, facilitating efficient cross-modal retrieval.
2. **Implementation of the Retrieval System:**
   (a) Design the overall architecture of the multimodal information retrieval system and provide a corresponding diagram.
   (b) Define, implement, and discuss histological image similarity techniques and ranking algorithms used in the system, highlighting their strengths and weaknesses.
3. **Performance Evaluation:**
   (a) Define a set of test queries, including both image-based and text-based inputs, to evaluate the system's effectiveness.
   (b) Utilize standard information retrieval metrics, such as precision, recall, and F-measure, or other appropriate metrics to assess performance.
4. **Final Report:**
   (a) Draft a detailed report (maximum 10 pages) that includes:
      i. Introduction and objectives of the project.
      ii. Description of the datasets and preprocessing processes.
      iii. Details of the implementation of the retrieval system.
      iv. Evaluation results and critical performance analysis.
      v. Conclusions and potential future developments.

4 Project Document Structure

This project document follows a structured and modular format to guide the development of a multimodal information retrieval system in the pathological domain. Students are expected to design, implement, and evaluate a complete pipeline capable of processing both histological images and textual pathology reports, and responding to visual and textual queries. The structure ensures a coherent discussion covering the dataset, feature extraction methods, system design, and retrieval evaluation. The specific retrieval focus may vary depending on the student's implementation choices.

Chapter 1: Introduction

This chapter introduces the task of multimodal information retrieval in pathology, emphasizing its relevance for clinical decision-making, research, and digital health applications. The motivation for combining textual and visual modalities is discussed, alongside the challenges and opportunities this integration presents. The chapter concludes by outlining the project objectives, scope, and structure of the report.

Chapter 2: Data and Preprocessing

This chapter describes the datasets used, typically composed of histological images (e.g., whole slide images or patches) and associated pathology reports or structured clinical annotations. Details include the data source, format, labeling scheme, and preprocessing steps. For images, preprocessing may involve color normalization, artifact removal, and patch extraction. For text, it may include tokenization, section parsing, or entity normalization. The goal is to prepare each modality in a way that facilitates effective feature extraction and downstream retrieval.

Chapter 3: Feature Extraction and Representation

This chapter focuses on extracting meaningful representations from both the visual and textual inputs. It covers techniques such as convolutional neural networks (CNNs), transformers, or pretrained vision-language models (e.g., CLIP) for image embeddings, and language models (e.g., BERT, BioBERT) or custom NLP pipelines for textual embeddings. The chapter also discusses how multimodal representations are aligned or fused for joint retrieval, and highlights any dimensionality reduction or feature selection strategies employed.

Chapter 4: Retrieval System and Evaluation

This chapter outlines the design and implementation of the retrieval system, including its architecture, data flow, and querying logic. It discusses how visual and textual queries are encoded and matched against the multimodal database. Retrieval strategies may include nearest-neighbor search, similarity fusion, or cross-modal retrieval methods. The evaluation section describes the metrics used (e.g., Precision@K, mAP, nDCG, Recall) and presents both quantitative results and qualitative examples. If available, ablation studies or comparisons with baselines are included.

Chapter 5: Conclusions and Future Work

This final chapter summarizes the project's main achievements and contributions. It reflects on the effectiveness of the retrieval system and the challenges encountered in processing and aligning multiple modalities. Limitations are discussed along with lessons learned. The chapter concludes by proposing future directions, such as expanding the dataset, improving fusion techniques, or exploring clinical deployment scenarios.

5 Due date

Students must provide all the assignment materials to the teacher at most 7 days before the exam date they want to participate to. Project files to deliver must include:

- Source code of the project (in a compressed files or as Github repository)
- Project Report as PDF file
- Configuration files and additional documentation, if any.

Possible Thesis Extension

Students have the opportunity to exploit the exam project as an entry point for their **Master's thesis** project by expanding the knowledge acquired into more focused and advanced research directions.

Possible Thesis Extension

Students are encouraged to consider this project as a foundation for their **Master's thesis** work. The multimodal nature of the system offers a wide range of research opportunities at the intersection of computer vision, natural language processing, and biomedical informatics. Building on the core components developed during the project, students can pursue more advanced and focused investigations.

Possible thesis extensions include, but are not limited to:

- **Cross-Modal Embedding Learning:** Developing or fine-tuning joint embedding spaces where image and text data are aligned, using contrastive learning, attention mechanisms, or cross-modal transformers.
- **Multimodal Retrieval Optimization:** Investigating advanced strategies to handle heterogeneous queries (image-only, text-only, or both) and fusing retrieval results in a semantically meaningful way.
- **Explainable Multimodal Retrieval:** Incorporating explainability tools (e.g., Grad-CAM for images, attention visualization for text) to help users understand retrieval results, especially in clinical contexts.
- **Domain Adaptation and Generalization:** Extending the system to work across different datasets or institutions, and handling variability in staining, language, or pathology subdomains.
- **Vision-Language Pretraining in Pathology:** Exploring or developing pretrained models tailored to biomedical images and reports, including training from scratch on domain-specific corpora.
- **Multimodal Indexing and Efficient Search:** Researching scalable indexing techniques for real-time retrieval in large-scale databases, including hashing, vector quantization, and approximate nearest neighbors (ANN).
- **Interactive or Conversational Retrieval:** Designing user interfaces or agents capable of refining queries through dialogue, incorporating user feedback, or adapting to ambiguous queries.
- **Integration with Clinical Workflow:** Adapting the system for deployment in real-world scenarios, considering aspects such as patient data integration, privacy, and user-friendliness for medical professionals.

Such directions not only extend the technical scope of the project but also contribute meaningfully to the field of computational pathology and AI-driven biomedical systems, with potential for high-impact publications and clinical applications.