fit@hcmus

UNIVERSITY OF SCIENCE
**ADVANCED PROGRAM IN COMPUTER SCIENCE**

# Thesis Proposal

(submitted by students)

**Thesis title:**

SMART INTERACTIVE RETRIEVAL OF VISUAL DATA
VIA SEMANTIC UNDERSTANDING

**Thesis advisor: Assoc. Prof. Trần Minh Triết**

**Students: Hoàng Xuân Nhật** (18125042) – **Nguyễn E Rô** (18125046)

**Type of thesis:** *Research*

**Duration:** From *01/01/2022* to *31/07/2022*

**Contents of thesis:**

Information retrieval is a field that predates the popular usage of computers. With more and more data being generated by humans, a better method to search than a linear scan is essential. Observing the current advances in artificial intelligence and in Computer Vision and Natural Language Processing specifically, we seek to apply those techniques to create a smart interactive retrieval system supporting the practical need of searching in a collection of visual data. Though we have popular commercialized search engines that have billions of users, their primary focus is on the web, which is primarily text-based, making it easier to index billions of pages. A text-to-image search engine with similar popularity is not readily available yet. Google Images, for example, relies on the text presented on the same page as images to understand them. We want to build a system that can analyze images based on their content and incorporate structured and unstructured accompanying metadata, which exists with camera-generated images and online videos.

However, there are a few challenges associated with building such a system.

- First, the colossal size of the data to be processed is an immense challenge on its own. We are not aiming at mobile galleries; we are talking about wearable

cameras that can generate videos at 30 frames per second (FPS) or surveillance cameras that also generate videos at dozens of FPS. Even with 2 FPS, a reasonable rate for lifelogging, such cameras already produce 2880 images a day, or 86400 images a month. Compression techniques can help reduce the storage, but if left unattended, the data can become a waste and face disposal.

- Second, developers are proficient with their own system but have difficulty bringing the same ease of use to other users. There can be various reasons for this. Often, in order to boost performance, developers install features that make it different from existing systems; however, this also creates surprises for the user, making them unable to utilize the system fully.
- Third, while enjoying more flexibility than traditional databases, AI-based systems can suffer from a lack of explainability. When only items are returned, there is little indication of why they are yielded. This can be a problem when relevance can not be easily verified with bare eyes, for example, when the images come from a different domain (medical, poor lighting, etc.). It should be noted that ordinary databases provide no solution to this problem either, however, AI-assisted systems can be modified to include a certain indicator, for example, localization of concepts in the query itself.

Existing systems suffer from one or all of these problems. We want to apply artificial intelligence to create a system that can tackle all of these problems at once, i.e., it should be scalable, the user should be able to use it without needing an AI background, and it has to have explainability.

Our main objective is to apply artificial intelligence to the search system to make it smarter (i.e., more flexible queries, more functionalities) and at the same time explainable, while maintaining scalability (i.e., can operate with a large quantity of data). This is much broader than just building an AI model as we are building a system that consists of many components. For example, using a top-performing yet slow model may forbid the system to run in real-time. A bad user interface might prevent the user from getting any value out of our system at all. Therefore, to achieve our multiple goals, we have to think about the system’s architecture and the

interaction between various components. Specifically, we look for simple yet effective solutions as they are easier to scale and also simpler to reason about.

With this in mind, our proposed work has the following main contributions:

- Scalable architecture: we separate our application into modules to enhance the maintainability and scalability of the whole system. Specifically, we propose the adoption of a vector database to store the embeddings generated by AI models for use in searching, besides using an ordinary database for storing metadata. This approach prevents the AI model from becoming the bottleneck of the system, enabling querying speed comparable to non-AI systems.
- Practical guidelines: we compile our extensive experience from varying users to form a list of guidelines, or principles, for users to keep in mind while searching. This gives users some sense of direction while still being flexible enough to adapt to different situations. Our experience is based on a time-pressured setting dealing with difficult queries, so it should be practical enough.
- Attention on Explainability: we enhance our system with a novel referring expression segmentation module to precisely point out the object or concept referred to in the image. This strengthens the model's credibility in all scenarios, especially in some critical ones such as medical uses.

To achieve our objective, we need to perform the following main tasks:

- Literature review on self-supervised methods over the recent years, summarizing the key features and what we can build on for our work.
- Survey current and past approaches in vision-language modeling, before and after the emergence of self-supervised learning.
- Review recent advances in computer vision and natural language processing for potential applications in multi-modal settings.
- Identify and analyze the possible sources of data that we can use to train our models.
- Propose our main methodology and its potential applications.
- Prepare the data for processing, including collection and cleaning if needed.

- Survey state-of-the-art training practices and optimization techniques to reduce the resources needed for the development and deployment of our solution.
- Implement our ideas, verify the results in accordance with our hypotheses.
- Carefully evaluate our models for potential pitfalls, such as fairness and robustness issues.
- Identify possible deployment environment for our work and build a proof-of-concept demo.

**Research timelines:**

Our tentative schedule is as follows:

- 01/01/2022 to 21/01/2022: Perform literature review on self-supervised methods
- 22/01/2022 to 11/02/2022: Survey vision-language methods
- 12/02/2022 to 04/03/2022: Review recent CV and NLP trends
- 05/03/2022 to 25/03/2022: Identify data sources
- 26/03/2022 to 22/04/2022: Propose the main methodology
- 23/04/2022 to 06/05/2022: Data processing and cleaning
- 07/05/2022 to 27/05/2022: Survey training and optimization techniques
- 28/05/2022 to 24/06/2022: Implementation of our method
- 25/06/2022 to 08/07/2022: Evaluation of our models
- 09/07/2022 to 31/07/2022: Finalize and build a working demo

| **Approved by the advisor** | **Ho Chi Minh city, 14/03/2022** | |
|---|---|---|
| ***Signature of advisor*** | ***Signature(s) of student(s)*** | |
|  |  |  |
| ***Trần Minh Triết*** | ***Hoàng Xuân Nhật*** | ***Nguyễn E Rô*** |