ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

MÔN HỌC: TÌM KIẾM THÔNG TIN THỊ GIÁC

ĐỀ TÀI: Image Retrieval – Tìm kiếm hình ảnh sử dụng DINOv2

Nhóm thực hiện:

1. Đinh Hoàng Thùy Linh MSHV: 240101054
2. Lưu Nguyễn Công Minh MSHV: 240202024
3. Nguyễn Hữu Tài MSHV: 240101068

Giảng viên bộ môn: **TS. Ngô Đức Thành**

Thành phố Hồ Chí Minh, tháng 08 năm 2025

1. Giới thiệu:

Trong thời đại bùng nổ thông tin đa phương tiện hiện nay, việc tìm kiếm và truy vấn thông tin từ hình ảnh ngày càng trở nên quan trọng. Hệ thống **tìm kiếm hình ảnh theo nội dung (Content-Based Image Retrieval – CBIR)** là một hướng nghiên cứu phổ biến trong tìm kiếm thông tin thị giác máy tính, với mục tiêu truy xuất những hình ảnh có nội dung tương tự dựa trên đặc trưng hình ảnh, thay vì dựa vào metadata như tên file hoặc mô tả (description).

Tuy nhiên, một thách thức lớn của các hệ thống CBIR truyền thống là việc trích xuất đặc trưng phụ thuộc nhiều vào các mô hình học có giám sát (supervised learning), đòi hỏi lượng dữ liệu được gán nhãn lớn – điều này tốn kém và không khả thi trong các dự án nhỏ lẻ và đối với các tổ chức cá nhân không có tiềm lực và tài nguyên để thực hiện công việc. Trong bối cảnh đó, các mô hình học không giám sát (self-supervised learning) đã trở thành một xu hướng tiềm năng để trích xuất đặc trưng có tính khái quát cao mà không cần nhãn dữ liệu.

Đồ án này nhằm mục tiêu xây dựng một hệ thống tìm kiếm hình ảnh sử dụng đặc trưng từ mô hình **DINOv2** – một kiến trúc học không giám sát mạnh trong thị giác máy tính. DINOv2 là thế hệ kế tiếp của DINO (Self-Distillation with No Labels), nổi bật với khả năng học biểu diễn hình ảnh mà không cần nhãn, thông qua cơ chế tự chưng cất (self-distillation) và backbone Vision Transformer (ViT).

Thông qua việc sử dụng DINOv2 như một **feature extractor**, chúng tôi đánh giá khả năng ứng dụng của mô hình trong bài toán truy xuất hình ảnh trên tập dữ liệu chuẩn Oxford Buildings, mà không cần tinh chỉnh lại mô hình. Hệ thống tập trung vào truy xuất top-K ảnh gần nhất dựa trên độ tương đồng đặc trưng, với các chỉ số đánh giá như **mAP@k** và **Precision@k**.

2. Phương pháp thực hiện

2.1 Kiến trúc và kỹ thuật trong DINOv2

Mô hình **DINOv2** (Self-Distillation with No Labels v2) là một phương pháp học biểu diễn hình ảnh không giám sát, kế thừa và mở rộng từ mô hình DINO gốc.

DINOv2 được thiết kế để học các đặc trưng có tính khái quát cao, không cần dữ liệu gán nhãn, bằng cách áp dụng cơ chế **self-distillation** giữa hai mạng neural: **Teacher** và **Student**. Trong quá trình huấn luyện, hai mạng được cung cấp các phiên bản tăng cường khác nhau của cùng một ảnh, và Student học để tái tạo đặc trưng do Teacher sinh ra. Điểm khác biệt là Teacher không được huấn luyện trực tiếp mà được cập nhật từ Student qua cơ chế EMA (Exponential Moving Average).

- **Backbone: Vision Transformer (ViT)**
    - ViT (Vision Transformer) là một kiến trúc được thiết kế theo nguyên lý của Transformer trong NLP nhưng áp dụng cho ảnh.
    - Mỗi ảnh được chia thành các patch (thường là 16x16 pixels), sau đó chuyển thành vector embedding và được xử lý như chuỗi token trong NLP.
    - DINOv2 được sử dụng phiên bản ViT dinov2-base (768 chiều) để mã hóa ảnh thành các đặc trưng ở cấp độ toàn ảnh hoặc từng patch.

        Nguồn: facebook/dinov2-base · Hugging Face
- **Mục tiêu huấn luyện: Kết hợp DINO + iBOT**
    - DINOv2 kết hợp ý tưởng từ **DINO** (học toàn ảnh) và **iBOT** (học đặc trưng từng patch) để cải thiện khả năng hiểu nội dung chi tiết trong ảnh. **iBOT (Image BERT Pretraining with Online Tokenizer)** là một phương pháp tự giám sát khác, cải tiến từ **DINO** bằng cách thêm nhiệm vụ huấn luyện ở cấp độ patch. Thay vì chỉ học representation từ toàn ảnh, **iBOT** sử dụng các tokenizer ảo (online tokenizer) để học cách dự đoán đặc trưng của từng patch trong ảnh. Nói cách khác, **iBOT** giúp mô hình không chỉ hiểu được nội dung tổng thể của ảnh (như **DINO**) mà còn học cách phân biệt các vùng nhỏ trong ảnh, từ đó tăng độ chính xác trong các tác vụ yêu cầu chi tiết như tìm kiếm ảnh.
    - Việc học diễn ra ở hai mức: **Global-level** (toàn ảnh): đảm bảo mô hình hiểu nội dung tổng quát; và **Patch-level**: học các chi tiết cục bộ giúp cải thiện khả năng phân biệt trong tìm kiếm.

Nguồn: GitHub - bytedance/ibot: iBOT :robot:: Image BERT Pre-Training with Online Tokenizer (ICLR 2022)

- **KoLeo regularizer**
    - Là một regularization technique được sử dụng trong **DINOv2** để đảm bảo rằng các vector đặc trưng được phân bố đều trong không gian embedding.
    - **KoLeo** giúp các đặc trưng output có phân phối đồng đều trên không gian hypersphere (siêu cầu), giúp tăng khả năng phân biệt giữa các ảnh và đặc biệt là cải thiện hiệu suất khi sử dụng NN Search **(Nearest Neighbor Search)** trong bài toán tìm kiếm ảnh.

      Nguồn:https://github.com/facebookresearch/dinov2/blob/main/dinov2/loss/koleo_loss.py
- **Dữ liệu huấn luyện: LVD-142M** – tập ảnh lớn, đa dạng và được chọn lọc tự động.
    - LVD-142M là một tập dữ liệu lớn (~142 triệu ảnh), đa dạng về nội dung, ngữ cảnh, và không có nhãn.
    - Đây là một tập dữ liệu tự động được chọn lọc để đảm bảo chất lượng và tính bao phủ rộng.
    - DINOv2 được huấn luyện trên LVD-142M để học các đặc trưng khái quát, có khả năng áp dụng cho nhiều downstream tasks mà không cần finetuning.
- **Các cải tiến kỹ thuật khác: FlashAttention, Sinkhorn-Knopp, Sequence Packing**
    - **FlashAttention:** Là một kỹ thuật **tăng tốc quá trình tính toán** attention trong mô hình Transformer bằng cách tối ưu việc truy cập bộ nhớ GPU, giúp giảm thời gian huấn luyện và tiết kiệm tài nguyên, đặc biệt trên mô hình lớn như ViT-g hoặc ViT-h.

      Nguồn: GitHub - Dao-AILab/flash-attention: Fast and memory-efficient exact attention
    - **Sinkhorn-Knopp Algorithm:** Được sử dụng trong quá trình học mã hóa rời rạc (discrete representation learning) để đảm bảo phân phối

các vector đặc trưng vào các cluster được cân bằng – hữu ích khi mô hình cần học phân nhóm trong không gian embedding.

- **Sequence Packing:** Cho phép **nén nhiều ảnh có độ dài khác nhau** vào cùng một batch, tận dụng tối đa bộ nhớ GPU, tăng độ hữu ích trong quá trình huấn luyện trên tập dữ liệu lớn.

2.2 Pipeline hệ thống

- **Dữ liệu đầu vào**

Hệ thống sử dụng **tập dữ liệu Oxford Buildings** – một benchmark nổi tiếng trong nghiên cứu tìm kiếm hình ảnh:

Nguồn: https://www.robots.ox.ac.uk/~vgg/data/oxbuildings/

- Gồm khoảng 5.062 ảnh kiến trúc của các tòa nhà tại Oxford.
- Có 55 ảnh truy vấn (query images) được gán nhãn thủ công, mỗi ảnh đi kèm danh sách các ảnh "relevant" (phù hợp).
- Dữ liệu bao gồm cả tập ground truth phục vụ đánh giá hiệu suất tìm kiếm theo chuẩn mAP và Precision@k.

- **Trích xuất đặc trưng với DINOv2**

Ở bước này, đặc trưng của toàn bộ ảnh trong tập dữ liệu được trích xuất như sau:

- Mỗi ảnh được resize về kích thước cố định (ví dụ: 224x224) và chuẩn hóa.
- Đưa ảnh qua mô hình DINOv2-base (không finetune) để thu được vector đặc trưng có chiều dài 768.
- Đặc trưng được trích xuất từ lớp [CLS] token của ViT – đại diện toàn ảnh.

- **Lập chỉ mục và tìm kiếm với FAISS**

Để thực hiện việc tìm kiếm ảnh tương tự, hệ thống sử dụng thư viện **FAISS** *(Facebook AI Similarity Search),* các bước cụ thể như sau:

- Toàn bộ vector đặc trưng của ảnh trong tập dữ liệu được lập chỉ mục bằng FAISS.

- Trong đồ án này, sử dụng cấu hình đơn giản **IndexFlatL2**, tức là tìm kiếm theo khoảng cách L2 (*Euclidean distance*) giữa các vector trong không gian thực. Khi người dùng đưa vào một ảnh để tìm kiếm, vector đặc trưng của ảnh đó được tính, sau đó dùng FAISS để tìm ra K ảnh gần nhất trong tập chỉ mục.

- **Đánh giá hệ thống**

Để đánh giá chất lượng của hệ thống tìm kiếm, hai độ đo phổ biến được sử dụng:

- **mAP@k (mean Average Precision):** Trung bình của độ chính xác tích lũy ở mỗi vị trí khi có ảnh phù hợp trong top-k. Độ đo này phản ánh sự sắp xếp thứ hạng của kết quả.
- **Precision@k:** Tỉ lệ số ảnh đúng trong top-k kết quả tìm kiếm.

Việc đánh giá được thực hiện bằng cách:

- Chạy truy vấn trên 55 ảnh đã được gán nhãn.
- So sánh top-k kết quả tìm kiếm với tập ground truth.
- Tính toán mAP@k và Precision@k cho từng truy vấn, sau đó lấy trung bình.

3. Kết quả đánh giá

Sau khi xây dựng pipeline tìm kiếm hình ảnh sử dụng đặc trưng từ DINOv2 và triển khai trên tập dữ liệu Oxford Buildings, chúng tôi đã tiến hành đánh giá hệ thống dựa trên 55 ảnh truy vấn có nhãn. Hai chỉ số đánh giá được sử dụng là **mean Average Precision (mAP@k)** và **Precision@k**, với các giá trị k = 5, 10, 15, 20.

k	MAP@k	Precision@k
5	0.8115	0.8473
10	0.4453	0.4236
15	0.3479	0.2824
20	0.3031	0.2118

Phân tích:

3.1 Hiệu quả ở top-k nhỏ (k = 5):

- Hệ thống đạt Precision@5 là 84.73%, cho thấy khả năng tìm kiếm **chính xác cao** ở phạm vi hẹp.
- Giá trị mAP@5 đạt 0.8115, chứng tỏ các ảnh đúng thường xuất hiện ở vị trí đầu tiên trong danh sách kết quả — một đặc điểm rất quan trọng trong các ứng dụng như tìm kiếm ảnh tương tự, gợi ý ảnh.

3.2 Hiệu suất khi tăng k:

- Khi k tăng lên (10, 15, 20), cả mAP và Precision đều **giảm đáng kể**.
- Điều này có thể lý giải là vì trong không gian đặc trưng, những ảnh gần nhất đầu tiên thường có độ tương đồng mạnh, còn các ảnh xa hơn có thể là "nhiễu" – không thực sự liên quan về mặt ngữ nghĩa, nhưng vẫn có vector gần nhau theo L2.

3.3 Tính tổng quát của DINOv2:

- Mặc dù không được finetune trên tập dữ liệu Oxford, mô hình vẫn đạt hiệu suất cao → cho thấy khả năng khái quát mạnh mẽ của đặc trưng học từ DINOv2.
- Điều này mở ra tiềm năng ứng dụng hệ thống cho các tập dữ liệu khác mà không cần tinh chỉnh mô hình.

4. Ưu và nhược điểm của phương pháp

Ưu điểm:

- **Không cần tinh chỉnh:** Có thể dùng ngay đặc trưng của DINOv2 để truy vấn. Người dùng có thể sử dụng trực tiếp mô hình để trích đặc trưng ảnh mà không cần finetuning, tiết kiệm thời gian và tài nguyên tính toán.
- **Hiệu suất cao:** Các kết quả đạt được cho thấy Precision@5 lên tới 84.73%, điều này cho thấy DINOv2 đặc biệt phù hợp với bài toán truy vấn ảnh ở mức độ gần nhất.

- **Tổng quát tốt:** DINOv2 học đặc trưng theo cách không phụ thuộc vào nhãn, nên có thể áp dụng tốt cho nhiều tập dữ liệu chưa từng thấy trước đó. Ở đây mô hình vẫn đạt hiệu quả cao trên tập Oxford Buildings dù không được huấn luyện hay điều chỉnh gì trên tập này.

Nhược điểm:

- **Chi phí huấn luyện gốc cao:** DINOv2 cần huấn luyện trên GPU lớn (ViT-g), tuy nhiên sử dụng mô hình đã huấn luyện giải quyết được vấn đề này.
- **Không đa phương thức:** Không giống như mô hình như CLIP, DINOv2 chỉ xử lý đầu vào là hình ảnh, không thể xử lý truy vấn dạng văn bản → ảnh (text-to-image retrieval).
- **Giảm hiệu quả khi mở rộng k:** Với giá trị k lớn hơn, Precision và MAP giảm đáng kể.

5. Tổng kết

Trong báo cáo này, nhóm đã trình bày một hệ thống tìm kiếm hình ảnh dựa trên đặc trưng trích xuất từ mô hình học không giám sát DINOv2. Mô hình được sử dụng như một bộ trích đặc trưng tiền huấn luyện (pretrained feature extractor), kết hợp với FAISS để thực hiện tìm kiếm hàng xóm gần nhất trong không gian embedding.

Kết quả thực nghiệm trên tập dữ liệu Oxford Buildings cho thấy hệ thống đạt hiệu quả rất cao với Precision@5 lên tới **84.73%**, mặc dù không có bất kỳ bước tinh chỉnh nào trên dữ liệu truy vấn. Điều này khẳng định khả năng khái quát hóa mạnh mẽ của DINOv2, cũng như tính thực tiễn trong việc xây dựng các hệ thống tìm kiếm ảnh nhanh, hiệu quả mà không cần nhãn huấn luyện.

Tuy nhiên, hệ thống vẫn còn một số hạn chế, đặc biệt là:

- **Giảm hiệu quả** khi mở rộng truy vấn với top-k lớn hơn.
- Không hỗ trợ **tìm kiếm đa phương thức**, khiến ứng dụng thực tế bị giới hạn hơn so với các mô hình như CLIP.
- **Chi phí huấn luyện ban đầu rất cao**, không dễ mở rộng hoặc tinh chỉnh thêm nếu không có tài nguyên phù hợp.

TÀI LIỆU TRÍCH DẪN

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, Piotr Bojanowski
DINOv2: Learning Robust Visual Featureswithout Supervision
https://arxiv.org/pdf/2304.07193