**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**



**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**----□&□----**



**BÁO CÁO SEMINAR CUỐI KỲ**

**Đề tài: Truy vấn hình ảnh với VGG16**

**Và Faiss**

**Môn học: Truy vấn thông tin đa phương tiện**

**Giảng viên:** Đỗ Văn Tiến

**Thành viên nhóm:**

|  |  |  |
| --- | --- | --- |
| **STT** | **MSSV** | **Họ Tên** |
| 1 | 2150957 | Đường Minh Khang |
| 2 | 21520398 | Nguyễn Quế Phong |
| 3 | 18521362 | Nguyễn Ích Tài |

**TP. HỒ CHÍ MINH - 01/2024**

**LỜI CẢM ƠN**

Sau quá trình học tập và rèn luyện tại Trường Đại học Công Nghệ Thông Tin, chúng em đã được trang bị các kiến thức cơ bản, các kỹ năng thực tế để có thể hoàn thành báo cáo đồ án môn học của mình.

Chúng em xin gửi lời cảm ơn chân thành đến thầy Đỗ Văn Tiến – Giảng viên phụ trách lớp CS336.O11-Truy vấn thông tin đa phương tiện đã tận tâm hướng dẫn, truyền đạt những kiến thức cũng như kinh nghiệm cho chúng em trong suốt thời gian học tập.

Trong quá trình làm đồ án môn học, mặc dù nhóm chúng em đã cố gắng nhưng chắc chắn sẽ không tránh được những sai sót không đáng có. Mong nhận được sự góp ý cũng như kinh nghiệm quý báu của thầy để được hoàn thiện hơn và rút kinh nghiệm cho những môn học sau. Chúng em xin chân thành cảm ơn!

TP. Hồ Chí Minh, tháng 01 năm 2024.

**MỤC LỤC**

**Mục lục**

[Chương 1: Tổng quan đề tài 4](#_Toc156395870)

[**1.1** **Lí do chọn đề tài** 4](#_Toc156395871)

[**1.2** **Bộ dữ liệu** 5](#_Toc156395872)

[**1.2.1 Oxford Building Dataset:** 5](#_Toc156395873)

[**1.2.2 Oxford Building Dataset:** 5](#_Toc156395874)

[**1.3** **Giới thiệu bài toán** 5](#_Toc156395875)

[Chương 2: Nội dung thực hiện 5](#_Toc156395876)

[**2.1** **VGG16** 5](#_Toc156395877)

[**2.2** **FAISS** 8](#_Toc156395878)

[**2.3** **Phương pháp đánh giá** 9](#_Toc156395879)

[**2.4** **Kết quả và đánh giá** 10](#_Toc156395880)

[**2.5** **F.So sánh với thuật toán khác** 11](#_Toc156395881)

[**2.6** **G.Tài liệu tham khảo** 12](#_Toc156395882)

# Tổng quan đề tài

## **1.1 Lí do chọn đề tài**

Với sự phát triển mạnh mẽ của Internet cũng như số lượng người dùng đông đảo, lượng thông tin được lưu trữ trên môi trường này vô cùng lớn kéo theo nhu cầu về một hệ thống truy vấn thông tin. Từ nhu cầu đó, chúng tôi đã nghiên cứu một hệ thống truy vấn hình ảnh, một phần nhỏ trong bài toán truy vấn thông tin, trên hai tập dữ liệu The Oxford Building Dataset và Paris Dataset.

## **Bộ dữ liệu**

### **1.2.1 Oxford Building Dataset:**

Mô tả: Bộ dữ liệu này chứa hình ảnh của các tòa nhà từ thành phố Oxford, Anh Quốc. Được sử dụng nhiều trong các nghiên cứu về nhận diện đối tượng và phân loại hình ảnh.

Số lượng ảnh: Bộ dữ liệu này bao gồm một số lượng lớn hình ảnh với đa dạng về kiến trúc và góc chụp.Gồm 5062 ảnh, tổng dung lượng là 2,4GB.

### **1.2.2 Oxford Building Dataset:**

Mô tả: Bộ dữ liệu Paris Dataset thường được sử dụng để nghiên cứu về các vấn đề như nhận diện và mô tả hình ảnh đối với các khu vực đô thị như Paris, Pháp.

Số lượng ảnh: Bộ dữ liệu này chứa một lượng lớn hình ảnh với các đối tượng và ngữ cảnh đa dạng.Gồm 6412 ảnh, tổng dung lượng 2,6GB.

## **Giới thiệu bài toán**

Input**:** Ảnh kiến trúc cần tìm kiếm có kích cỡ 224\*224\*3

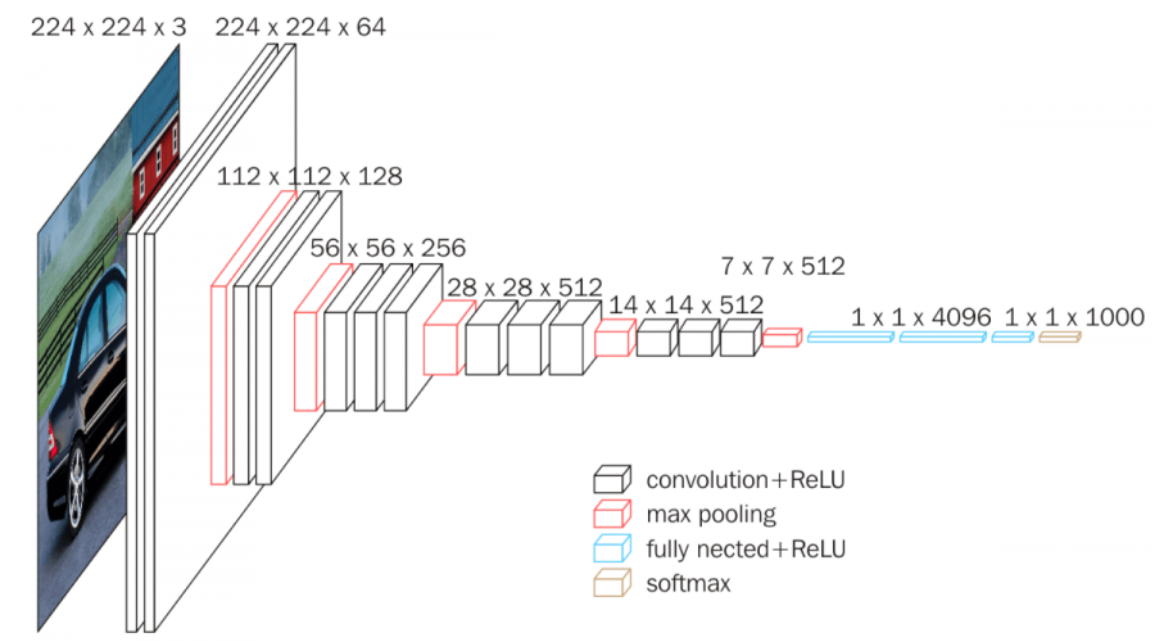
Output: List các ảnh liên quan kiến trúc cần tìm kiếm

# Nội dung thực hiện

## **VGG16**

Hiện nay [Convolutional Neural Network](https://wiki.tino.org/convolutional-neural-network-la-gi/#:~:text=Ngo%C3%A0i%20ra%2C%20CNN%20c%C3%B3%20th%E1%BB%83,n%E1%BB%81n%20t%E1%BA%A3ng%20Facebook%2C%20Google%2C%E2%80%A6) (CNN) được sử dụng trong nhiều trong các bài toán liên quan đến hình ảnh vì nó giúp chúng ta có những đặc trưng của hình ảnh, nhờ đó mà rút gọn được kích thước dữ liệu nhưng vẫn đảm bảo được hiệu quả cho bài toán.

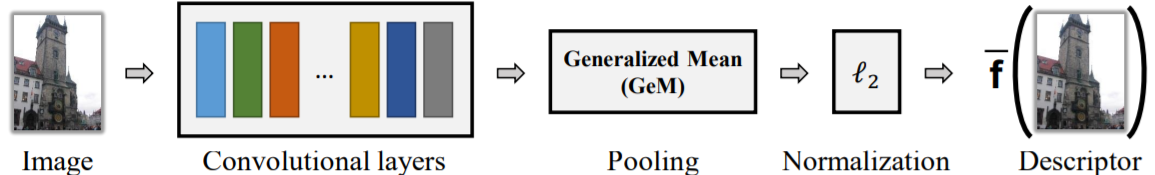
Ở đây chúng tôi sẽ sử dụng mạng VGG16 để trích xuất đặc trưng của ảnh. Tuy nhiên mạng VGG16 sẽ được chỉnh sửa một chút để phù hợp với bài toán tìm kiếm ảnh.

****

**Hình 2.1. Mạng VGG16**

Kiến trúc chúng tôi sử dụng đã được được đề xuất trong [6].

Ta sẽ sử dụng mạng VGG16 từ đầu cho đến trước max pooling cuối cùng, phần sau ta sẽ sử dụng Generalized-Mean(GeM) và l2-normalization.



**Hình 2.2. Kiến trúc mạng sử dụng**

Sau khi bỏ fully-connected, thì khi đưa hình ảnh vào mạng VGG16 ta sẽ thu được ma trận 3 chiều có kích thước là , trong đó là số feature map ở layer cuối cùng. Gọi là ma trận ở feature map thứ k sau khi đi qua hàm kích hoạt ReLU, Ta có:

=

là một tham số. Max pooling và Advantage pooling là các dạng pooling đặc biệt của GeM pooling. Khi thì GeM pooling trở thành Max pooling, p = 1 thì GeM pooling trở thành Advantage pooling. Tham số sẽ được cập nhật thông qua quá trình huấn luyện mô hình.

Trong thực tế khi sử dụng GeM chúng tôi thu được kết quả tốt hơn nhiều khi sử dụng MAC hoặc SPoC. Lí do là mỗi hình ảnh tìm kiếm có những đặc trưng riêng biệt nên khi trích xuất đặc trưng MAC, SPoC không đáp ứng tốt sự thay đổi đó vì nó không thay đổi tham số p, còn GeM sẽ giúp việc trích xuất tốt hơn nhiều thông qua việc học tham số p.

Trong bài toán này chúng tôi sử dụng L2-normalization (Normalization L2) để chuẩn hóa dữ liệu. L2-normalization là một kỹ thuật chuẩn hóa vector theo norm L2. Trong truy vấn hình ảnh, sau khi các đặc trưng của hình ảnh được trích xuất, L2-normalization thường được áp dụng để đảm bảo rằng các vector đặc trưng có độ dài bằng nhau. Điều này giúp giảm ảnh hưởng của biến động về tỷ lệ trong các vector đặc trưng và cải thiện tính nhất quán khi so sánh chúng, đặc biệt là trong các mô hình nhận diện đối tượng hoặc tìm kiếm hình ảnh. Chuẩn hóa L2 thường được thực hiện bằng cách chia mỗi giá trị trong vector cho norm L2 của vector đó.

Sau khi có được những vector đặc trưng của toàn bộ hình ảnh chúng tôi sẽ sử dụng thư viện Faiss để sắp xếp các vector theo thứ tự một cách hợp lý để tăng tốc độ tìm kiếm.

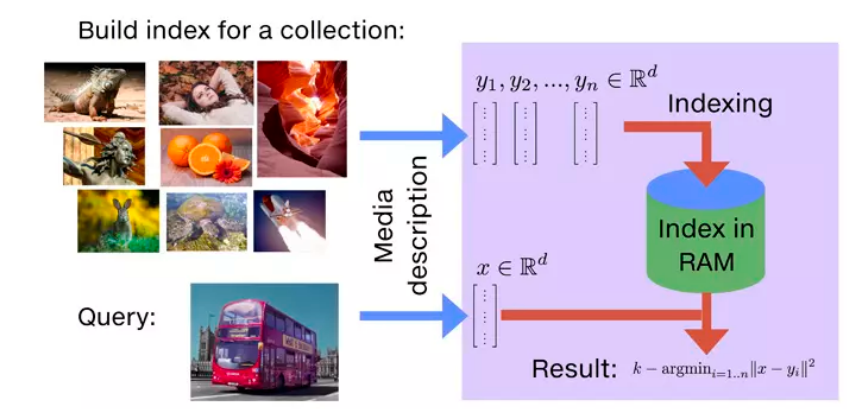
## **FAISS**

Đối với những hệ thống tìm kiếm ảnh thì lượng dữ liệu là rất lớn. Dẫn đến việc tính độ tương đồng giữa vector query đến các vector trong database sẽ tốn rất nhiều thời gian. Vì vậy chúng ta cần dùng nhũng công cụ hỗ trợ indexing để tăng tốc khả năng tìm kiếm.Trong dự án này chúng tôi chọn Faiss để hỗ trợ index.

Faiss, hay Facebook AI Similarity Search, là một thư viện mã nguồn mở do Facebook AI Research (FAIR) phát triển, tập trung vào việc giải quyết vấn đề tìm kiếm vectơ tương tự trong không gian nhiều chiều. Được xây dựng để đối mặt với thách thức của việc xử lý dữ liệu lớn và high-dimensional, Faiss cung cấp các thuật toán và cấu trúc dữ liệu tối ưu để thực hiện các nhiệm vụ này một cách nhanh chóng và hiệu quả.

Với một tập hợp các vector, chúng ta có thể index chúng bằng cách sử dụng Faiss sau đó sử dụng một vectơ khác (query vector) để tìm kiếm các vectơ tương tự nhất trong index. Faiss giúp ta tăng tốc độ truy vấn khá nhiều trong những bộ dữ liệu lớn.

Faiss hổ trợ nhiều loại khoảng cách để tính độ tương đồng của các vector. Trong bài này chúng tôi sử dụng khoảng cách cosine.

****

**Minh họa dùng Faiss để indexing**

## **Phương pháp đánh giá**

Chúng tôi sử dụng độ đo (mean average precision) để đánh giá mô hình. Đây là độ đo được sử dụng dụng phổ biến nhất trong việc đánh giá các mô hình image retrieval. Vì nó vừa chưa thông tin của precision và recall, đồng thời có xét đến độ quan trọng của thứ hạng kết quả.



Trong đó ,Q là số lượng truy vấn và APq là Average Precision của truy vấn thứ q.

Trong đó AP chính là diện tích phía dưới của Precision-Recall curve được tính bằng công thức:

Ngoài ra chúng tôi còn sử dụng độ đo để đánh giá tính hiệu quả của bài toán web search này. Đây là độ do sử dụng thông tin precision và có xét đến độ quan trọng của thứ hạng trên top k kết quả trả về. Được tính bằng cách lấy trung bình của độ đo AP@k trên từng câu query

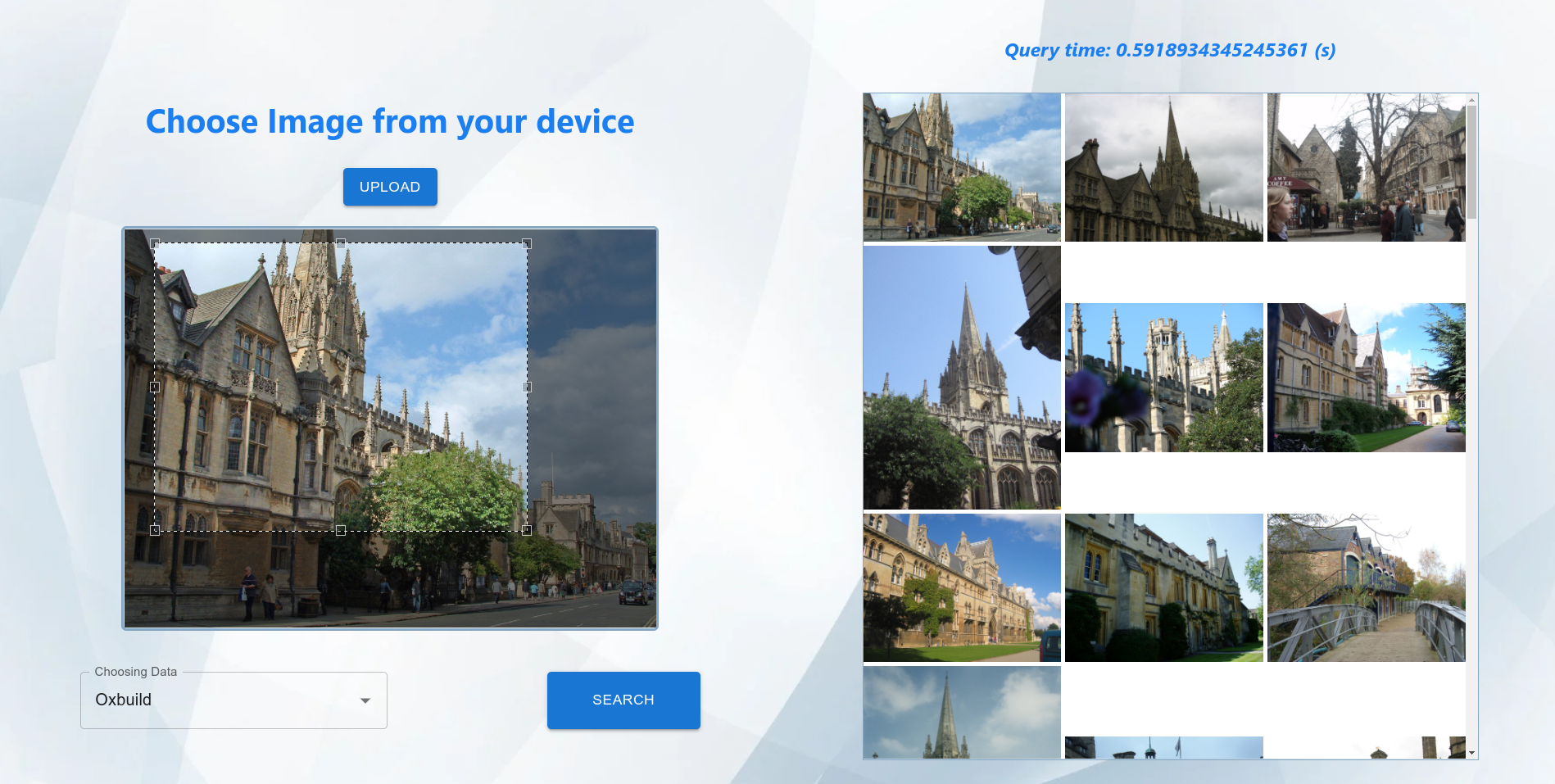
Độ đo AP@k là trung bình của Precision top i với ) đây là độ đo được sử dụng khá nhiều để đánh giá độ hiệu quả của các hệ thống web search.

## **Kết quả và đánh giá**

Nhìn kết quả bên dưới, có thể thấy sự chênh lệch khá lớn giữa 2 bộ dữ liệu oxf5k và Paris6k. Bộ dữ liệu Paris6k có độ chính sát lớn hơn khá nhiều so với bộ oxf5k.

|  |  |  |
| --- | --- | --- |
| Dataset | Oxf5k | Paris6k |
| map | 0.32 | 0.68 |
| map@5 | 0.60 | 0.99 |
| map@10 | 0.52 | 0.98 |
| map@50 | 0.41 | 0.94 |
|  |  |  |

Giải thích có cho sự khác biệt này là do các hình trong bộ dữ liệu Oxf5k có màu sắc khá tương đồng nhau. Ngược lại, các hình trong bộ dữ liệu Paris6k có hình dạng và màu sắc khá khác biệt. Ngoài ra trong bộ dữ liệu Paris6k số lượng ảnh relevant cho từng query khá lớn (187/query). Lớn hơn nhiều so với Oxf5k (56/query). Đó cũng là lý do khiến cho độ đo map ở tập Paris6k khá thấp (68%) còn độ đo map@k thì lại khá cao (99%).



**Hình 5. Kết quả**

Kết quả của web search cho thấy những kết quả top đầu khá chính xác và tương đồng với các hình query. Và tốc độ truy vấn cũng rất nhanh trung bình 0.54s cho mỗi câu query

## **So sánh với thuật toán khác**

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Dim | Oxf5k | Paris6k |
| VLAD+CNN **[7]** | 128 | 59.3 | 59.0 |
| MAX-POOLING OF CNN **[8]** | 512 | 63.9 | 83.0 |
| NetVLAD **[9]** | 4096 | 71.6 | 79.7 |
| Ours | 512 | 0.32 | 0.68 |

Khi so sánh với các mô hình khác sử dụng CNN và cùng ở trên 2 bộ dữ liệu thì ta thấy mô hình chúng tôi cho kết quả không được cao. Vì những mô hình trên có sử dụng thêm những kỹ thuật khác như whitening, query expansion, re-ranking. Ví dụ như trong **[6]**, tài liệu chúng tôi tham khảo để lấy mô hình, khi sử dụng thêm kỹ thuật whitening thì độ hiệu quả của mô hình theo thang đo mAP tăng từ 75,5 lên 86 cho tập Paris6k.

Tuy nhiên những mô hình trên có tốc độ query khá lớn (hơn 1s/query). Còn đối với mô hình của chúng tôi thì chỉ có trung bình 0.56s/query. Vì mục đích ban đầu của chúng tôi là một mô hình có tốc độ tìm kiếm nhanh để phục vụ cho web search nên chúng tôi không quá chú trọng vào recall của mô hình dẫn đến độ đo map khá thấp. Đây chính là đánh đổi giữa tốc độ query và độ chính sát của mô hình.

Tóm lại mô hình của chúng tôi có độ chính sát thấp hơn các mô hình CNN khác, tuy nhiên có thời gian truy xuất khá nhanh. Đồng thời độ đo map@k khá cao rất thích hợp cho ứng dụng trên web search

## **Tài liệu tham khảo**

**[1]** Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi and Andrew Zisserman, 2022, *VGGFace2: A dataset for recognising faces across pose and age*, truy cập 24/1/2017 từ [VGGFace2: A dataset for recognising faces across pose and age](https://arxiv.org/pdf/1710.08092v2.pdf)

**[2]** [James Briggs](https://www.youtube.com/c/jamesbriggs), Facebook AI Similarity Search (Faiss): The Missing Manual, truy cập 25/1/2012 từ [Facebook AI Similarity Search (Faiss): The Missing Manual](https://www.pinecone.io/learn/faiss-tutorial/#:~:text=Faiss%20is%20a%20library%20%E2%80%94%20developed,similar%20vectors%20within%20the%20index)

**[3]** Vijayakumar Bhandi, Sumithra Devi K.A, 2020,*Image Retrieval Using Features From PreTrained Deep CNN***,**  truy cập 24/1/2022 từ [Image Retrieval Using Features From PreTrained Deep CNN](https://www.ijstr.org/final-print/jun2020/Image-Retrieval-Using-Features-From-Pre-trained-Deep-Cnn.pdf)

**[4]** [Xiaoxiao Ma](https://www.scirp.org/journal/articles.aspx?searchcode=Xiaoxiao++Ma&searchfield=authors&page=1), [Jiajun Wang](https://www.scirp.org/journal/articles.aspx?searchcode=Jiajun++Wang&searchfield=authors&page=1), 2017, *Image Retrieval Using Deep Convolutional Neural Networks and Regularized Locality Preserving Indexing Strategy*, truy cập từ 20/1/2022 từ [Image Retrieval Using Deep Convolutional Neural Networks and Regularized Locality Preserving Indexing Strategy](https://www.scirp.org/journal/paperinformation.aspx?paperid=74652)

**[5]** Sirisha Kopparthi, Dr.N.K.Kameswara Rao, 2020,*Content based Image Retrieval using Deep Learning Technique with Distance Measures*, truy cập từ 20/1/2022 từ [Content based Image Retrieval using Deep Learning Technique with Distance Measures](https://www.researchgate.net/publication/346968214_Content_based_Image_Retrieval_using_Deep_Learning_Technique_with_Distance_Measures).

**[6]** [Filip Radenović](https://arxiv.org/search/cs?searchtype=author&query=Radenovi%C4%87%2C+F), [Giorgos Tolias](https://arxiv.org/search/cs?searchtype=author&query=Tolias%2C+G), [Ondřej Chum](https://arxiv.org/search/cs?searchtype=author&query=Chum%2C+O), *CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples*, truy cập ngày 21/1/2022 từ [CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples](https://arxiv.org/pdf/1711.02512.pdf)

**[7]** Ng, J.Y.H., Yang, F., Davis, L.S.: *Exploiting local features from deep networks for image retrieval. In: CVPR workshops*. (2015)

**[8]** Tolias, G., Sicre, R., J´egou, H.: *Particular object retrieval with integral maxpooling of cnn activations. In: ICLR*. (2016)

**[9]** Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN architecture for weakly supervised place recognition. In: CVPR. (2016)