BÁO CÁO ĐỒ ÁN CUỐI KỲ

Môn học

CS519 - PHƯƠNG PHÁP LUẬN NGHIÊN CỨU KHOA HỌC

Lớp học

CS519.N11

Giảng viên

PGS.TS. LÊ ĐÌNH DUY

Thời gian

09/2022 - 02/2023

----- Trang này cố tình để trống -----

THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): *https://www.youtube.com/watch?v=6uk4kavhjfM*
- Link slides (dạng .pdf đặt trên Github của nhóm): *https://github.com/nhuongemconxe/CS519.N11/ImageRetrieval.pdf*

• Họ và Tên: Đào Tuấn Anh
• MSSV: 19520377
	• Lớp: CS519.N11
	• Tự đánh giá (điểm tổng kết môn): 9/10
	• Số buổi vắng: 1
	• Số câu hỏi QT cá nhân: 1
	• Số câu hỏi QT của cả nhóm: 2
	• Link Github: https://github.com/shoxie
	• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
	  ○ Lên ý tưởng đề tài
	  ○ Viết tóm tắt và giới thiệu đề tài
	  ○ Thuyết trình video phần tóm tắt, giới thiệu
	  ○ Tìm tài liệu tham khảo

• Họ và Tên: Cao Thanh Bình • MSSV: 19520408	• Lớp: CS519.N11 • Tự đánh giá (điểm tổng kết môn): 7/10 • Số buổi vắng: 1 • Số câu hỏi QT cá nhân: 1 • Số câu hỏi QT của cả nhóm: 2 • Link Github: https://github.com/nhuongemconxe • Mô tả công việc và đóng góp của cá nhân cho

	kết quả của nhóm: ○ Lên ý tưởng đề tài ○ Viết báo cáo phần nội dung và phương pháp thực hiện ○ Thuyết trình video phần nội dung và phương pháp thực hiện ○ Tìm tài liệu tham khảo
● Họ và Tên: Đặng Phi Hùng ● MSSV: 19520573	● Lớp: CS519.N11 ● Tự đánh giá (điểm tổng kết môn): 9/10 ● Số buổi vắng: 1 ● Số câu hỏi QT cá nhân: 1 ● Số câu hỏi QT của cả nhóm: 2 ● Link Github: https://github.com/MrDilian/ ● Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm: ○ Viết báo cáo phần mục tiêu và kết quả mong đợi ○ Thuyết trình video mục tiêu và kết quả mong đợi ○ Làm slide ○ Làm poster
● Họ và Tên: Nguyễn Hữu Tân ● MSSV: 19520921	● Lớp: CS519.N11 ● Tự đánh giá (điểm tổng kết môn): 9/10 ● Số buổi vắng: 1 ● Số câu hỏi QT cá nhân: 1 ● Số câu hỏi QT của cả nhóm: 2 ● Link Github: https://github.com/nhuongemconxe

	● Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm: ○ Viết báo cáo phần nội dung và phương pháp thực hiện ○ Thuyết trình video phần nội dung và phương pháp ○ Làm poster ○ Làm slide

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

NÂNG CAO KHẢ NĂNG TRUY XUẤT HÌNH ẢNH DỰA TRÊN CNN: KHẢO SÁT CÁC KỸ THUẬT ĐỂ CẢI THIỆN HIỆU SUẤT

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

ADVANCING CNN-BASED IMAGE RETRIEVAL: EXPLORING TECHNIQUES FOR IMPROVED PERFORMANCE

TÓM TẮT *(Tối đa 400 từ)*

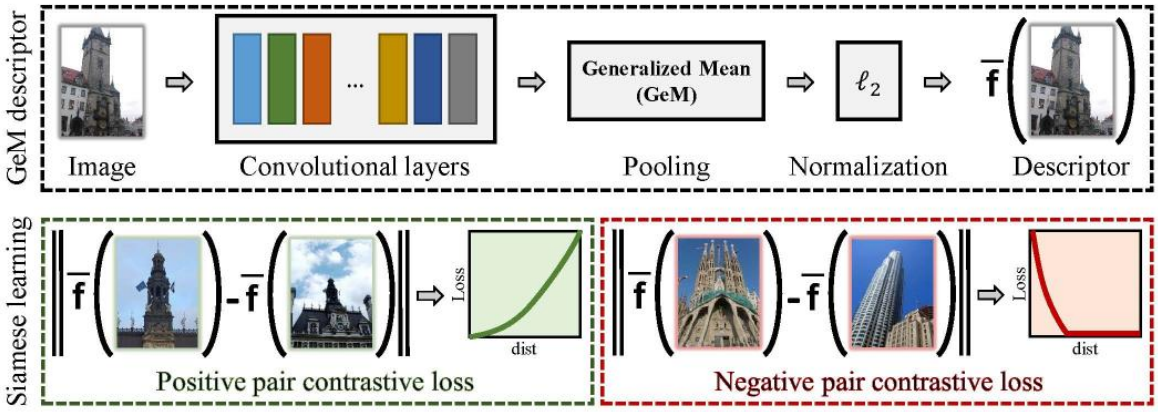
Trong lĩnh vực truy xuất hình ảnh, Convolutional Neural Networks (CNNs) đóng vai trò là một bộ mô tả hình ảnh hiệu quả và nhỏ gọn. Thế nhưng, việc đạt được kết quả đáng tin cậy cho các chú thích hình ảnh chính là vấn đề thiết yếu nhất. Ở nghiên cứu này, chúng tôi điều chỉnh CNNs phù hợp cho một hệ thống truy xuất hình ảnh bằng cách sử dụng nhiều hình ảnh không theo thứ tự nào. Chúng tôi sử dụng dữ liệu đào tạo bằng cách sử dụng các kỹ thuật truy xuất dữ liệu nâng cao và kĩ thuật Structure-from-Motion để tái tạo lại các mô hình 3D từ hình ảnh. Và để nâng cao hiệu quả truy xuất, chúng tôi sử dụng kĩ thuật tạo các pooling layer Generalized-Mean có thể đào tạo để khái quát hoá tổng hợp tối đa nhất. Các thử nghiệm của chúng tôi được thực hiện trên các dataset như: Oxford5k, Paris6k, ROxford5k và RParis6k với kiến trúc VGG và ResNet.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Mạng nơ-ron tích chập (Convolutional neural networks - CNNs) đã trở thành một giải pháp cho các vấn đề truy xuất hình ảnh, đặc biệt là sau bước đột phá của Krizhevsky [1] với việc sử dụng ImageNet [2]. Tuy nhiên, việc thu thập dữ liệu huấn luyện có chú thích rất tốn kém và thường dễ bị lỗi. May mắn thay, các kiến trúc được đào tạo cho các nhiệm vụ phân loại hình ảnh có thể thích ứng hiệu quả. Bằng cách sử dụng các CNN đã được train trước như Rectified Linear Units [3], làm bộ mô tả và tính năng hình ảnh, và có thể đạt được hiệu suất tìm kiếm hình ảnh thành công.

Một cách tiếp cận khác để mang lại khả năng thích ứng và tiết kiệm thời gian truy xuất hình ảnh tốt hơn là khởi tạo một mạng phân loại được train từ trước và sau đó train nó cho một nhiệm vụ khác hoặc thực hiện tinh chỉnh hợp lí.

Ở nghiên cứu này, chúng tôi áp dụng phương pháp tinh chỉnh không giám sát (unsupervised fine-tuning) để có thể truy xuất hình ảnh bằng CNN. Để có thể train các CNN của riêng mình, chúng tôi đã kết hợp các thông tin từ kĩ thuật Structure-from-Motion (SfM) và các ví dụ phù hợp và chưa từng train. Chúng tôi cũng train các kiến trúc của mình để học cách làm trắng (whitening) bằng cách sử dụng cùng một dữ liệu đã được train, do đó tránh được những hạn chế của hiệu suất làm trắng truyền thống khác. Cách tiếp cận của chúng tôi liên quan đến việc sử dụng một pooling layer có thể train để khái quát hoá các sơ đồ gộp phổ biến hiện có cho CNN.



Hình 1: Mô tả tổng quan hướng tiếp cận vấn đề [5]

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

1. Lấy ý tưởng từ công trình của [5] và nghiên cứu trước đó của [6], chúng tôi dự kiến giới thiệu một lớp tổng hợp mới, biểu diễn hình ảnh đa quy mô và phương pháp mở rộng truy vấn cho việc tìm kiếm hình ảnh.
2. Mở rộng nghiên cứu trước đó bằng cách tiến hành thêm thử nghiệm để tìm hiểu sâu hơn về vấn đề truy xuất hình ảnh.
3. Đóng góp vào lĩnh vực tìm kiếm hình ảnh bằng cách tăng cường sự hiểu biết và hiệu quả của các phương pháp dựa trên CNN cho tìm kiếm hình ảnh.

NỘI DUNG VÀ PHƯƠNG PHÁP

1. Mạng nơron tích chập (Convolutional neural networks - CNNs)

Trong thời gian gần đây, một loạt các CNN đã được phát triển cho các tác vụ truy xuất hình ảnh như VGG [7], Vision Transformers [8], EfficientNet [9], ResNet [10] và DenseNet [11]. Các CNN này đã cho thấy hiệu suất đáng kể ngay cả sau khi loại bỏ các lớp được kết nối hoàn toàn từ kiến trúc ban đầu của chúng. Điều này cung cấp một nền tảng đáng tin cây để sử dụng các phương pháp tinh chỉnh. Giả sử χ là đầu ra của một kiến trúc bất kì, là tensor $\chi \in Z^{W \times H \times K}$, trong đó K chính là đại diện cho số lượng feature map trong lớp cuối cùng. Đầu ra của CNN bao gồm các bộ kích hoạt χK để kích hoạt W x H 2D cho các feature map với các giả định rằng lớp cuối cùng là các Rectified Linear Unit (ReLU) sao cho $\chi \in Z \backslash Z^{-}$.

2. Tổng hợp Generalized – Mean (Generalized - Mean pooling)

Một pooling layer sẽ được tích hợp vào CNN, với đầu vào là và đầu ra được biểu diễn dưới dạng vector f. Tổng hợp Generalized-Mean (GeM) [12] sẽ được áp dụng trong bước tổng hợp, với Phương trình (1) và (2) được dùng cho mục đích này, trong đó $k \in \{1, ..., K\}$.

$$f^{(g)} = \left[f_1^{(g)}, \dots, f_k^{(g)}, \dots, f_K^{(g)}\right] \quad (1)$$

$$f_k^{(g)} = \left(\frac{1}{|\chi_k|}\right) \sum_{x \epsilon \chi_k} x^{p_k} \quad (2)$$

3. Bộ mô tả hình ảnh (Image descriptor)

Trong kiến trúc của mô hình này, chúng tôi đã kết hợp một lớp chuẩn hoá L2 như là lớp cuối cùng. Vector đầu ra f thu được từ quá trình tổng hợp được chuẩn hoá L2 trong giai đoạn đánh giá cuối cùng trong đó sản phẩm bên trong giữa hai hình ảnh được tính toán. Kết quả của Vector GeM được coi là bộ mô tả hình ảnh và nó cũng được chuẩn hoá L2.

4. Phương pháp Siamese (Siamese learning)

Một mạng lưới hai nhánh được train cho các công việc, dựa trên kiến trúc siamese. Cả hai nhánh của mạng lưới chia sẻ cùng một bộ tham số. Trong quá trình train, mạng

lấy các cặp hình ảnh (i, j) làm đầu vào cùng với các nhãn tương ứng Y(i,j) ∈ {0, 1} biểu thị xem cặp đó có khớp (0) hay chưa khớp (1).

5. Hàm thất thoát (Loss function)

Thất thoát tương phản (contrastive loss) [13] được sử dụng làm hàm thất thoát trong nghiên cứu của chúng tôi nhằm xác định xem các cặp có khớp với nhau hay không. Hàm thất thoát được biểu thị bằng Phương trình (3), trong đó (f)(i) biểu thị cho vector GeM đã được chuẩn hoá L2 của hình ảnh i và τ chính là ngưỡng đo khoảng cách giữa các cặp chưa từng có, đủ lớn để bị thất thoát dữ liệu. Kiến trúc của chúng tôi được train với một số lượng lớn các cặp train đã được tạo tự động. Phát hiện của chúng tôi cho thấy rằng chức năng thất thoát tương phản khái quát hoá tốt hơn và hội tụ đến mức hiệu suất cao hơn.

$$\mathcal{L}(i,j) = \begin{cases} \frac{1}{2}\left\|\bar{f}(i) - \bar{f}(j)\right\|^2 & if\ Y(i,j) = 1 \\ \frac{1}{2}\left(max\left\{0, \tau - \left\|\bar{f}(i) - \bar{f}(j)\right\|\right\}\right)^2 & if\ Y(i,j) = 0 \end{cases} \quad (3)$$

6. Làm trắng và giảm kích thước (Whitening and dimensionality reduction)

Tiếp theo, về các phương pháp xử lý hậu kỳ cho các vector GeM được tinh chỉnh. Để có thể sử dụng dữ liệu đã được gán nhãn từ các mô hình 3D, chúng tôi sẽ triển khai các phép chiếu phân biệt tuyến tính [14]. Hình chiếu bao gồm hai phần riêng biệt: làm trắng và xoay. Thành phần làm trắng thu được bằng cách lấy căn bậc hai nghịch đảo của ma trận phương sai C_S cho các cặp khớp với nhau, được thể hiện trong Phương trình 4. Phần xoay được áp dụng bằng phương pháp Phân tích thành phần chính [15] (Principal Component Analysis - PCA) vào ma trận phương sai của các cặp chưa từng có trong vùng làm trắng, được thể hiện trong Phương trình 5.

$$C_S = \sum_{Y(i,j)=1} \left(\bar{f}(i) - \bar{f}(j)\right)\left(\bar{f}(i) - \bar{f}(j)\right)^T \quad (4)$$

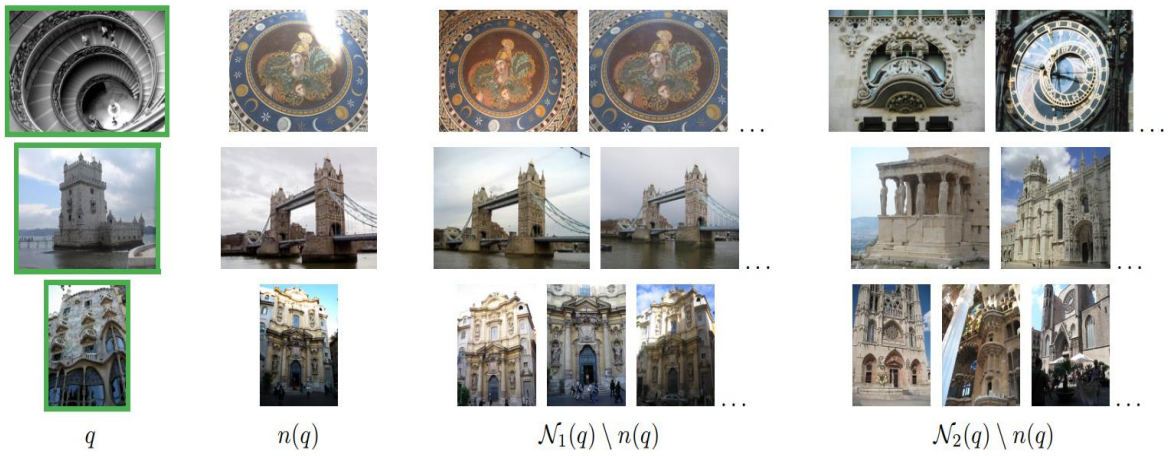
$$C_D = \sum_{Y(i,j)=0} \left(\bar{f}(i) - \bar{f}(j)\right)\left(\bar{f}(i) - \bar{f}(j)\right)^T \quad (5)$$

Cách tiếp cận này mặc dù không được tối ưu hoá hoàn toàn và thực hiện mà không

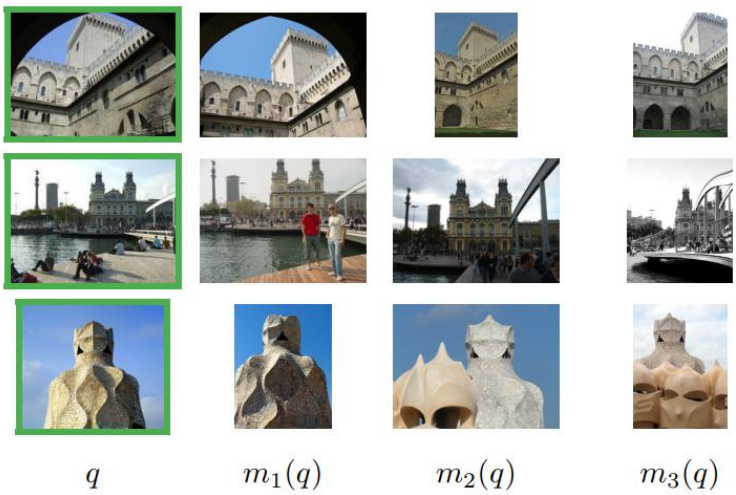
cần xử lí hàng loạt, nhưng đã được sử dụng hiệu quả tất cả các cặp đã có sẵn trong việc tối ưu hoá việc làm trắng. Chúng tôi tập trung tối ưu hoá bộ GeM và sau đó tiến hành quá trình làm trắng.

7. Tái tạo mô hình 3D (3D reconstruction)

Mục tiêu của công việc này là phát triển và tiếp cận để chọn dữ liệu đã được train để tìm kiếm hình ảnh mà không dựa vào dữ liệu có chú thích của con người hoặc đưa ra bất kỳ giả định nào về trainning dataset. Để có thể đạt được điều này, chúng tôi kết hợp truy xuất hình ảnh Bag-of-Words (BoW) với Structure-from-Motion (SfM) bằng cách sử dụng quy trình truy xuất SfM hiện đại. Quy trình này lấy một bộ sưu tập hình ảnh không có thứ tự làm đầu vào và xây dựng càng nhiều mô hình 3D càng tốt trong khi lọc ra các hình ảnh không khớp. Để tạo ra hiệu quả cho quá trình này hơn, chúng tôi cũng áp dụng phân cụm hình ảnh nhanh dựa trên tính năng cục bộ. Các kết quả phù hợp và vị trí máy ảnh sau đó được sử dụng để tự động chọn dữ liệu đào tạo cho hệ thống tìm kiếm hình ảnh. Cách tiếp cận này hoàn toàn tự động và không yêu cầu đến bất kỳ sự can thiệp thủ công nào.



Hình 2: Ví dụ về query training hình ảnh q (viền xanh lục) và các phủ định tương ứng của chúng được chọn bởi các chiến lược khác nhau [5]



Hình 3: Ví dụ về query training hình ảnh q (viền xanh lục) và hình ảnh phù hợp được chọn làm ví dụ tích cực theo các phương pháp m1, m2, m3 [5]

KẾT QUẢ MONG ĐỢI

1. Dự kiến sẽ đạt được hiệu suất truy xuất hình ảnh được cải thiện bằng cách tinh chỉnh các CNN bằng cách sử dụng phương pháp học không giám sát với thông tin SfM và lớp tổng hợp Generalized-Mean có thể đào tạo được.
2. Nghiên cứu sử dụng kiến trúc VGG và ResNet, đồng thời tiến hành thử nghiệm trên nhiều điểm chuẩn khác nhau như Oxford5k và Paris6k. Các kết quả có khả năng chứng minh tính hiệu quả của các kỹ thuật được đề xuất và góp phần vào sự tiến bộ của việc truy xuất hình ảnh dựa trên CNN.
3. Hoàn thiện bài báo khoa học với mô tả chi tiết cấu trúc mô hình và kèm theo một chương trình demo để minh họa nghiên cứu của trên một cách trực quan.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1] S. I. H. Krizhevsky, Alex and G. E, “Imagenet classification with deep convolutional neural networks,” Advances in neural information processing systems, vol. 25, pp. 1097–1105, 2012.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.

[3] A. F. Agarap, “Deep learning using rectified linear units (relu),” arXiv preprint

arXiv:1803.08375, 2018.

[4] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, “Neural codes for image retrieval,” in European conference on computer vision. Springer, 2014, pp. 584–599.

[5] Filip Radenović, Giorgos Tolias, Ondřej Chum, “Fine-tuning CNN Image Retrieval with No Human Annotation” TPAMI 2018.

[6] Filip Radenović, Giorgos Tolias, Ondřej Chum, “CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples” ECCV 2016.

[7] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2015.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” 2021.

[9] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proceedings of the 36th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 6105–6114. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2015.

[11] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” 2018.

[12] O. Morere, J. Lin, A. Veillard, L.-Y. Duan, V. Chandrasekhar, and T. Poggio, “Nested invariance pooling and rbm hashing for image instance retrieval,” in Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, 2017, pp. 260–268.

[13] S. Chopra, R. Hadsell, and Y. LeCun, “Learning a similarity metric discriminatively, with application to face verification,” in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), vol. 1. IEEE, 2005, pp. 539–546.

[14] K. Mikolajczyk and J. Matas, “Improving descriptors for fast tree matching by optimal linear projection,” in 2007 IEEE 11th International Conference on Computer Vision. IEEE, 2007, pp. 1–8.
[15] K. P. F.R.S., “Liii. on lines and planes of closest fit to systems of points in space,” The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 2, no. 11, pp. 559–572, 1901.

----- *Trang này cố tình để trống* -----