BÁO CÁO ĐỒ ÁN CUỐI KỲ

Môn học: CS519 - PHƯƠNG PHÁP LUẬN NCKH

Lớp: CS519.N11

GV: PGS.TS. Lê Đình Duy

Trường ĐH Công Nghệ Thông Tin, ĐHQG-HCM

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

Đào Tuấn Anh – 19520377

Cao Thanh Bình – 19520408

Đặng Phi Hùng – 19520573

Nguyễn Hữu Tân – 19520921

Tóm tắt

- Lớp: CS519.N11
- Link Github của nhóm: https://github.com/nhuongemconxe/CS519.N11
- Link YouTube video: https://www.youtube.com/watch?v=6uk4kavhjfM
- Ảnh + Họ và Tên của các thành viên

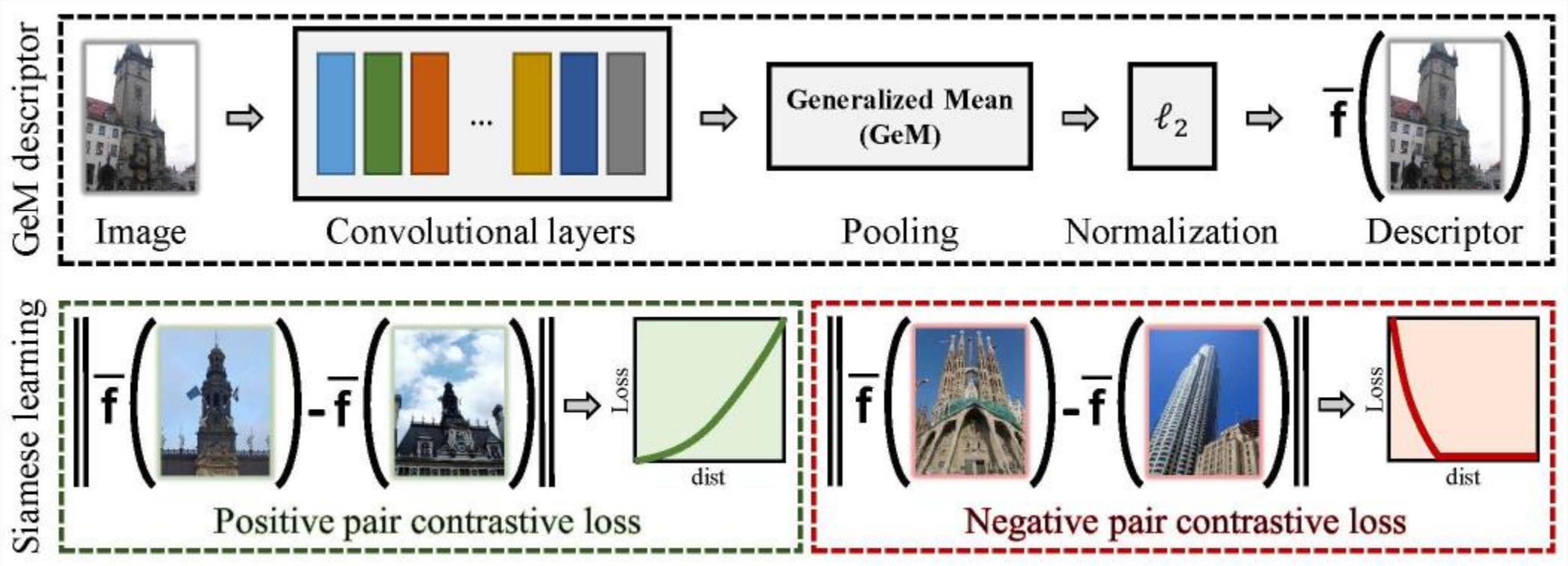
Đào Tuấn Anh

Cao Thanh Bình

Đặng Phi Hùng

Nguyễn Hữu Tân

Giới thiệu



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

Mục tiêu

- 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.

- Mở rộng nghiên cứu trước đó bằng cách tiến hành thêm thí nghiệm để tìm hiểu sâu hơn về vấn đề truy xuất hình ảnh.

- 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

❑ **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.

❑ **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 ∈ {1, ..., K}.

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

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

Nội dung và Phương pháp

- **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.

- **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).

- **Hàm thất thoát (Loss function)**

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

Nội dung và Phương pháp

❑ **Làm trắng và giảm kích thước (Whitening and dimensionality reduction)**

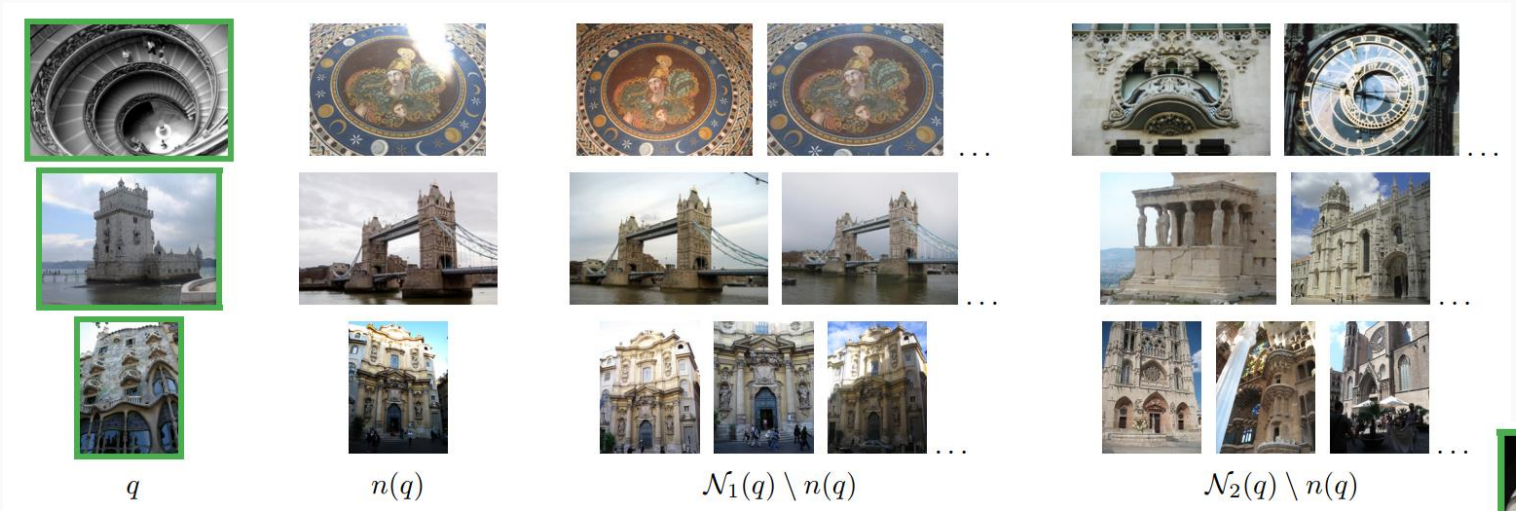
Về các phương pháp xử lý hậu kỳ cho các vector GeM được tinh chỉnh, để có thể sử dụng dữ liệu đã được gán nhãn từ các mô hình 3D, chúng tôi sẽ triển khai các phép chiếu phân biệt tuyến tính [14]. Hình chiếu bao gồm hai phần riêng biệt: làm trắng và xoay. Thành phần làm trắng thu được bằng cách lấy căn 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.

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

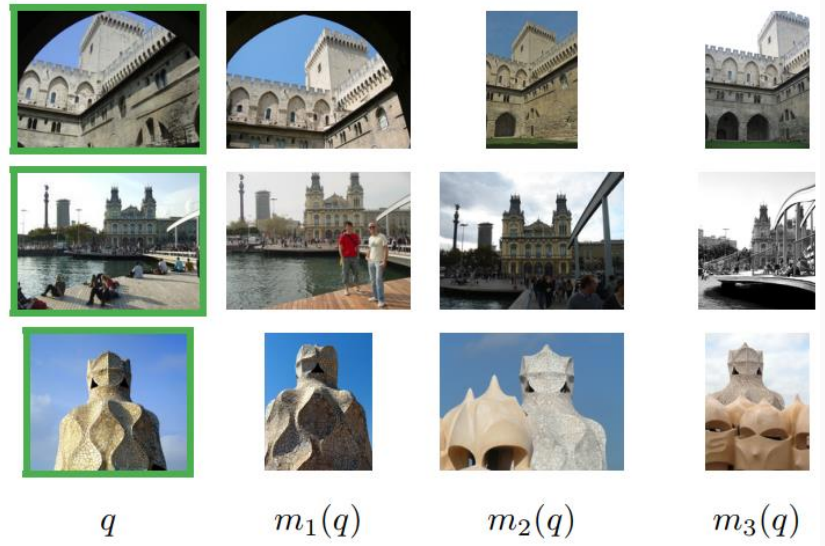
$$C_D = \sum_{Y(i,j)=0} \left(\bar{f}(i) - \bar{f}(j)\right)\left(\bar{f}(i) - \bar{f}(j)\right)^T \qquad (5)$$

Nội dung và Phương pháp

- **Tái tạo mô hình 3D (3D reconstruction)**



Hình 2: Ví dụ về query training hình ảnh q (màu 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 dự kiến

- 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.

- 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.

- 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

[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.

Tài liệu tham khảo

[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.