**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**

****

**BÁO CÁO MÔN HỌC**

**TÌM KIẾM DỮ LIỆU ĐA PHƯƠNG TIỆN**

Đề tài:

**TRUY XUẤT HÌNH ẢNH DỰA TRÊN NỘI DUNG**

Học viên: Nguyễn Trọng Thái - CB190218

Trần Tất Huy - CB190213

Mã lớp: 19BKHMT

GVHD: TS. Nguyễn Tuấn Dũng

**HÀ NỘI - 2020**

1. **Giới thiệu chung**

Độ tương đồng giữa hai hình là phép đo giữa hai hình ảnh nhằm ước tính xem hai hình ảnh nhất định có tương đồng với nhau hay không, thông qua các điểm về màu sắc, hình dạng, kết cấu, cấu trúc không gian và thông tin ngữ nghĩa trong hình ảnh. Đây là một trong những bài toán trọng tâm trong lĩnh vực nhận dạng mẫu và thị giác máy tính, và được ứng dụng rộng rãi trong các bài toán tìm kiếm hình ảnh, so khớp hình ảnh, khử hình ảnh trùng lặp và các ứng dụng khác. Kết quả của phép đo độ tương đồng giữa hai hình ảnh phụ thuộc chủ yếu vào việc trích chọn đặc trưng đại diện, đã được các nhà nghiên cứu dữ liệu đa phương tiện nghiên cứu rộng rãi trong nhiều thập kỷ. Đặc trưng được trích chọn từ hình ảnh điển hình nhất là SIFT (Scale Invariant Feature Transform). Các đặc trưng này sau đó được mã hóa thành các vector đại điện thông qua các lược đồ khác nhau, chẳng hạn như BoW (Bag-of-Words). Mặc dù đã có nhiều kỹ thuật được đề xuất, nó vẫn là một trong những bài toánd thách thức nhất trong nghiên cứu hiện tại, chủ yếu là do vấn đề “khoảng cách ngữ nghĩa” nổi tiếng, chính là “khoảng cách” tồn tại giữa pixel hình ảnh (thông tin cấp thấp) được máy chụp lại và khái niệm ngữ nghĩa (thông tin cấp cao) do con người nhận thức được. Để giải quyết bài toán này, các nhà nghiên cứu hy vọng rằng máy tính có thể xử lý hình ảnh tương tự như con người. Do đó, cần phải tìm ra một cách tốt hơn để biểu diễn hình ảnh và có được những đặc trưng sâu hơn của chúng.

Trong những năm gần đây, học sâu đã được sử dụng rộng rãi để giải quyết các tác vụ học máy. Đặc biệt, mạng nơ-ron tích chập (Convolutional Neural Network - CNN) có hiệu suất tuyệt vời trong nhiều tác vụ thị giác máy tính, bao gồm phân loại hình ảnh, phát hiện đối tượng, truy xuất hình ảnh, phân vùng ngữ nghĩa, ... Với các mạng nơ-ron sâu, chúng có thể học được những khái niệm ngữ nghĩa trừu tượng gần gũi với nhận thức của con người. Một số nghiên cứu gần đây cho thấy rằng các đặc trưng của CNN được huấn luyện trên các bộ dữ liệu lớn và đa dạng như ImageNet có thể được sử dụng để giải quyết các tác vụ mà chúng từng chưa được huấn luyện. Ngoài ra, một số kiến trúc mạng nơ-ron phức tạp hơn đã được đề xuất, chẳng hạn như Alexnet, VGGNet, GoogLetNet, ResNet, ... Tất cả đều có hiệu suất đột phá trong cuộc thi ILSVRC (ImageNet Large Scale Visual Recognition Challenge).

Gần đây, mạng Triplet Net được giới thiệu là phương pháp để tìm hiểu sự tương đồng về đặc điểm bằng cách tối ưu hóa khoảng cách giữa các đặc trưng được trích chọn của các cặp hình ảnh. Kiến trúc này sử dụng 3 nhánh CNN giống hệt nhau và được chia sẻ trọng số của chúng để có được đặc trưng đại diện cấp cao của hình ảnh để các cặp hình ảnh tương đồng có thể được ánh xạ gần nhau trong không gian đặc trưng và các cặp hình ảnh khác nhau được ánh xạ xa nhau hơn. Do đó, phương pháp này có thể trích chọn các đặc trưng tốt hơn từ hình ảnh. Mạng Triplet Net được biết đến như một phương pháp tiên tiến nhất về đo độ tương đồng hình ảnh. Hiện nay, nó đã được sử dụng trong một số lĩnh vực ứng dụng như truy xuất và so khớp hình ảnh, xác thực khuôn mặt, nhận dạng đối tượng…

Phần tiếp theo của báo cáo được sắp xếp như sau: Phần 2 thảo luận về tổng quan lý thuyết liên quan. Phần 3 mô tả chi tiết về kiến trúc mạng Triplet Net sử dụng cho bài toán truy xuất hình ảnh. Phần 4 trình bày kết quả của thuật toán thông qua các thí nghiệm. Phần 5 đưa ra kết luận.

1. **Tổng quan lý thuyết**

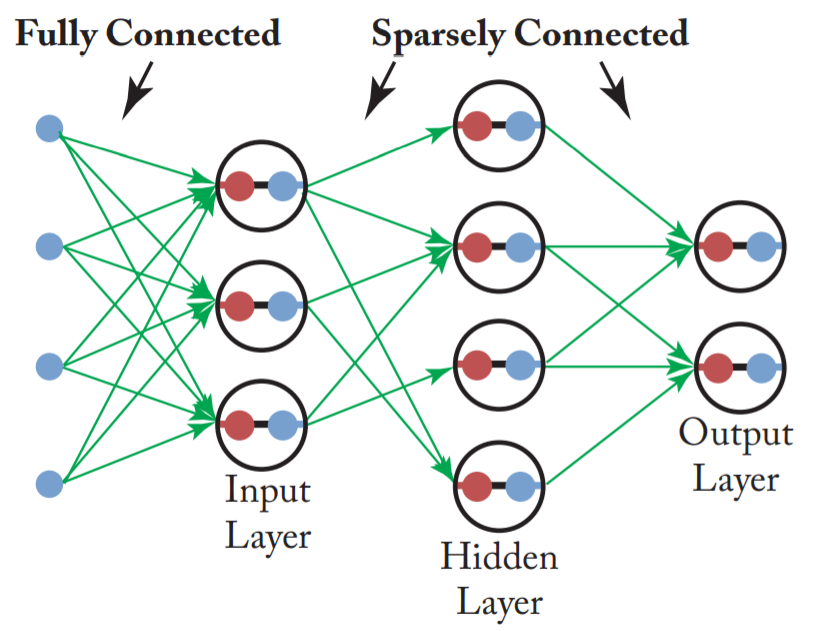
Mạng nơ-ron sâu (Deep Neural Networks - DNN) có nhiều hình dạng và kích thước khác nhau tùy thuộc vào ứng dụng. Các hình dạng và kích thước phổ biến cũng đang phát triển nhanh chóng để cải thiện độ chính xác và hiệu quả. Trong mọi trường hợp, đầu vào cho DNN là một tập hợp các giá trị đại diện cho thông tin cần được mạng nơ-ron phân tích. Ví dụ: những giá trị này có thể là pixel của một hình ảnh, biên độ được lấy mẫu của sóng âm thanh hoặc biểu diễn số về trạng thái của một số hệ thống hoặc trò chơi. Đối với các ứng dụng thị giác, mạng nơ-ron tích chập rất phổ biến và đạt đuọc hiệu suất hiện đại trong nhiều tác vụ thách thức.

* 1. **Mô tả các lớp DNN phổ biến**

Trong phần lớn các lớp DNN, phép tính chủ yếu là phép tính *tổng trọng số* (weighted sum). Có một số loại lớp khác nhau, chủ yếu khác nhau về cách đầu vào và đầu ra được kết nối với nhau như thế nào. Có hai thuộc tính chính của các kết nối trong một lớp:

* Nếu một lớp mà mọi kích hoạt đầu vào đều được kết nối với mọi kích hoạt đầu ra, thì lớp đó được gọi là lớp kết nối đầu đủ (fully connected). Mặt khác, nếu một lớp chỉ có một tập hợp con của các kích hoạt đầu vào được kết nối với kích hoạt đầu ra, thì lớp đó được gọi là kết nối thưa (sparsely connected). Đối với kết nối thưa, kích hoạt đầu vào có thể kết nối với bất kỳ kích hoạt đầu ra nào (toàn cục) hoặc chỉ trong một vùng lân cận của chúng (cục bộ). Hệ quả là xuất hiện một hàm sliding window xác định vùng lận kích hoạt đầu vào cho mạng, được gọi là trường tiếp thụ (receptive field).
* Giá trị trọng số liên quan đến mỗi kết nối: trường hợp phổ biến nhất trọng số có thể nhật bất kỳ giá trị nào (mỗi trọng số có thể có một giá trị duy nhất). Một trường hợp khác là cùng một giá trị được chia sẻ bởi nhiều trọng số, được gọi là cơ chế chia sẻ trọng số (weight sharing).

Các lớp có thuộc tính kết nối đầy đủ được gọi là lớp FC. Một lớp khác được sử dụng rộng rãi là lớp CONV, lớp kết nối thưa cục bộ và chia sẻ trọng số. Tính toán trong lớp FC và CONV là phép tính tổng trọng số, ngoài ra còn có thể thực hiện các phép tính khác.



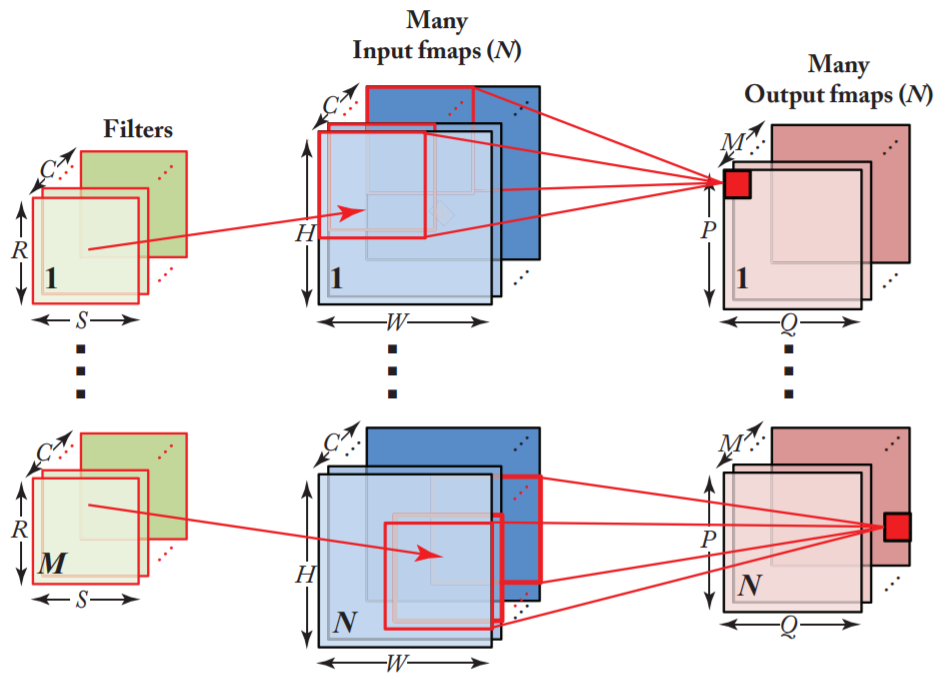
Hình 1. Kết nối đầy đủ vs Kết nối thưa

* 1. **Các loại lớp phổ biển trong DNN**

Phần này mô tả về các loại lớp phổ biến được sử dụng trong DNN, các lớp này có thể được xem như lớp nguyên thuỷ, thường được kết hợp với nhau để tạo thành khối phức hợp.

* + 1. **Lớp tích chập (CONV)**

Các lớp CONV chủ yếu được cấu tạo bởi các phép tích chập nhiều chiều. Trong tính toán này, các kích hoạt đầu vào của một lớp được cấu trúc dưới dạng bản đồ đặc trưng đầu vào dạng 3-D (), trong đó kích thước là chiều cao (), chiều rộng () và số kênh của đầu vào (). Trọng số của một lớp được cấu trúc như một bộ lọc 3-D, trong đó kích thước là chiều cao (), chiều rộng () và số kênh đầu vào () (số lượng kênh của bản đồ đặc trưng đầu vào và bộ lọc là như nhau). Đối với mỗi kênh đầu vào, bản đồ đặc trưng đầu vào trải qua phép tích chập 2-D với kênh tương ứng trong bộ lọc. Kết quả của phép tích chập tại mỗi điểm được tính tổng trên tất cả các kênh đầu vào để tạo ra đầu ra. Kết quả của quá trình này là bản đồ đặc trưng đầu ra (). Các bộ lọc 3-D bổ sung có thể được sử dụng trên cùng một bản đồ đặc trưng đầu vào để tạo thêm các kênh đầu ra (tức là áp dụng bộ lọc cho bản đồ đặc trưng đầu vào để tạo ra kênh đầu ra trong bản đồ đặc trưng đầu ra). Cuối cùng, nhiều bản đồ đặc trưng đầu vào () có thể được xử lý cùng nhau như một *batch* để có khả năng cải thiện việc tái sử dụng các trọng số của bộ lọc.



Hình 2. Phép tích chập trong lớp CONV.

Tính toán của lớp CONV được định nghĩa như sau:

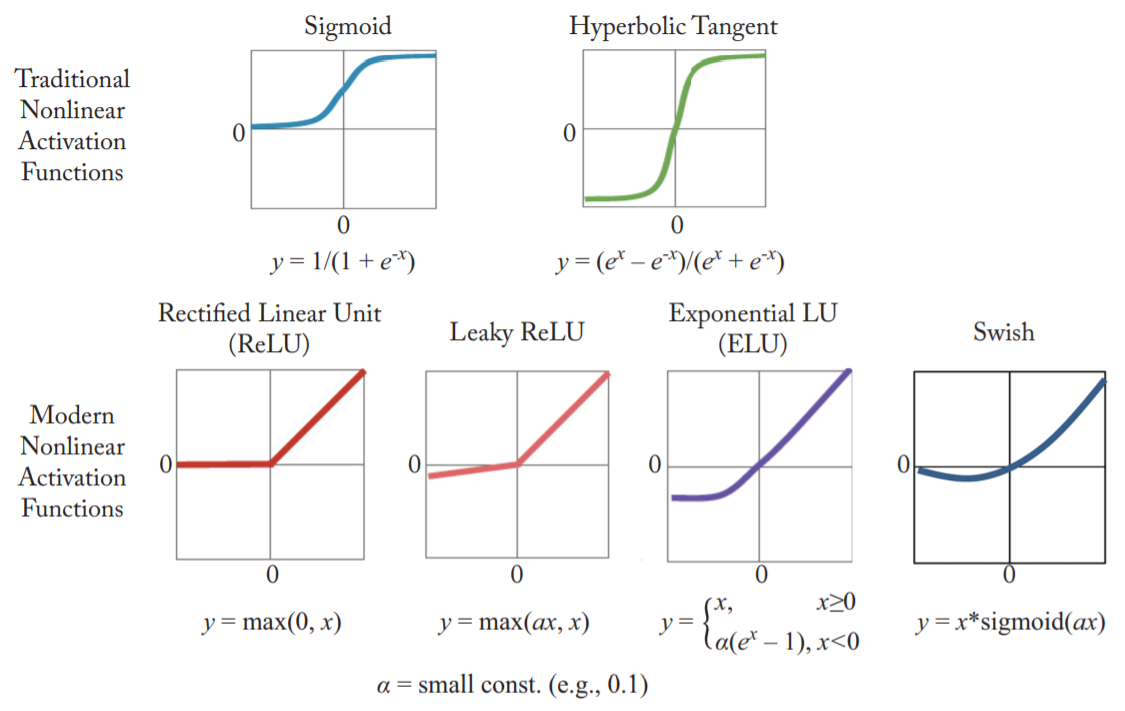
Trong đó lần lượt là tensor của các , và bộ lọc. là kích thước một stride đã biết trước.

* + 1. **Lớp kết nối đầy đủ (FC)**

Trong lớp FC, mọi giá trị trong bản đồ đặc trưng đầu ra là tổng trọng số của mọi giá trị đầu vào trong bản đồ đặc trưng đầu vào (tức là nó được kết nối đầy đủ). Hơn nữa, các lớp FC thường không thể hiện sự chia sẻ trọng số và kết quả là việc tính toán có xu hướng bị giới hạn bởi bộ nhớ. Các lớp FC thường được xử lý dưới dạng một phép nhân ma trận. Lớp FC cũng có thể được xem như một trường hợp đặc biệt của lớp CONV. Cụ thể nó là lớp CONV với các bộ lọc có cùng kích thước với bản đồ đặc trưng đầu vào. Do đó, nó không có tính chất cục bộ, kết nối thưa thớt và chia sẻ trọng số của các lớp CONV. Công thức vẫn giữ nguyên cho việc tính toán các lớp FC với một vài ràng buộc bổ sung đối với các tham số: , , và .

* + 1. **Hàm kích hoạt phi tuyến**

Một hàm kích hoạt phi tuyến thường được áp dụng sau mỗi lớp CONV hoặc FC. Các hàm phi tuyến khác nhau được sử dụng để đưa tính phi tuyến vào DNN, như thể hiện trong Hình 3. Chúng bao gồm các hàm phi tuyến thông thường trước đây như sigmoid hoặc tiếp tuyến hyperbol (tanh). Những hàm này rất phổ biến vì chúng tạo điều kiện cho phân tích / chứng minh toán học. ReLU đã trở nên phổ biến trong những năm gần đây do tính đơn giản và khả năng cho phép huấn luyện nhanh, đồng thời đạt được độ chính xác tương đương. Các biến thể của ReLU, chẳng hạn như leaky ReLU, parametric ReLU, exponential LU, và Swish cũng đã được giới thiệu để cải thiện độ chính xác.



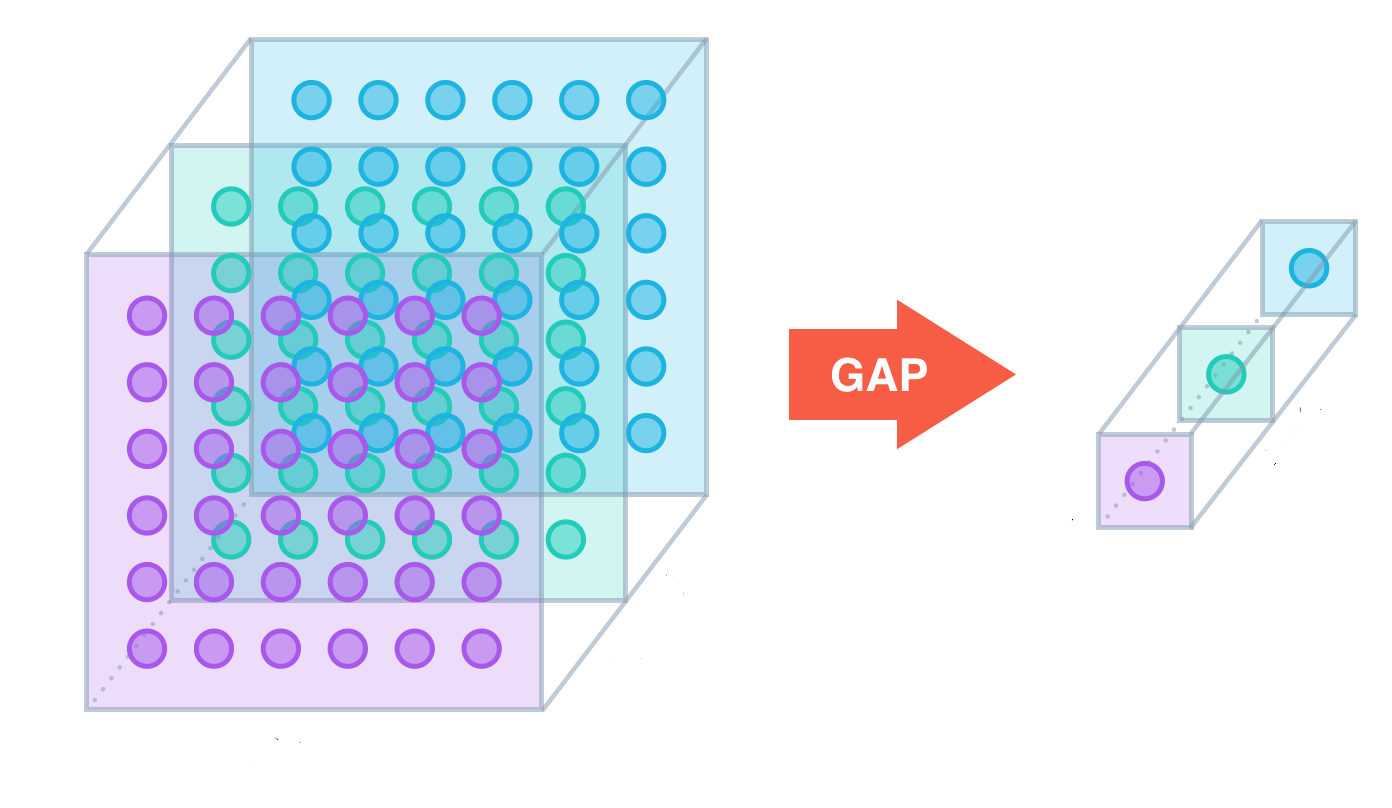
Hình 3. Các dạng hàm kích hoạt phi tuyến khác nhau.

* + 1. **Pooling**

Có nhiều phép tính có thể được sử dụng để thay đổi độ phân giải không gian (tức là và hoặc và ) của bản đồ đặc trưng tùy thuộc vào ứng dụng. Đối với các ứng dụng như truy xuất hình ảnh, mục tiêu là tóm tắt toàn bộ hình ảnh vào vector đại diện; do đó, giảm độ phân giải không gian có thể hữu ích. Giảm độ phân giải không gian của bản đồ đặc trưng được gọi là pooling hoặc nói chung là giảm số lượng mẫu (downsampling). Pooling được áp dụng cho từng kênh riêng biệt, cho phép mạng nơ-ron mạnh mẽ hơn và bất biến đối với những thay đổi và biến dạng nhỏ. Một tập hợp các giá trị trong trường tiếp thụ của pooling được ánh xạ thành một số lượng giá trị nhỏ hơn. Thông thường, pooling xảy ra trên các khối không chồng chéo (tức là stride bằng kích thước của pooling).

Gần đây, một hình thức pooling khác đã trở nên phổ biển, hình thức pooling này được thực hiện như một đơn lớp tại đầu ra của lớp CONV, được goi là GAP (Global Average Pooling). Mục đích của GAP khá khác biệt với mục đích của pooling thông thường. Nó không được sử dụng để tăng tính bất biến đối với các biến dạng nhỏ hoặc giảm độ phân giải của bản đồ đặc trưng, thay vào đó nó thay thế một phần hoặc toàn bộ lớp FC thường được sử dụng trong các mạng nơ-ron phức hợp. Các lớp FC thường được sử dụng như một lớp cao nhất trong mạng nơ-ron phức hợp. Trong khi lớp CONV trích chọn đặc trưng từ kích hoạt đầu vào, lớp FC thực hiện các tác vụ cấp cao hơn, như phân loại, nhận dạng hoặc truy xuất đối trong các tác vụ thị giác. Các lớp FC không sử dụng cơ chế chia sẻ trọng số, bởi vậy số lượng tham số của nó tăng lên rất nhanh chóng khi mạng tăng kích thước đầu vào, yêu cầu một chi phí lớn về sức mạnh tính toán và ngoài ra dẫn đến overfitting thường rất khó kiểm soát. Do vậy, việc sử dụng GAP cho phép các mạng nơ-ron phức hợp hoạt đôngj với các tác vụ cấp cao mà cần ít sức mạng tính toán hơn do số lượng tham số tăng không đáng kể, mạng nơ-ron đạt được mức tống quát hoá cao hơn và hiệu suất tăng lên.

Có hai dạng GAP. Trong một dạng, một lớp GAP duy nhất thay thế hoàn toán lớp FC. Ở một dạng khác, một lớp GAP cung cấp đầu ra của nó cho một hoặc nhiều lớp FC. Tuỳ thuộc vào mục đích, GAP và FC có thể hoán đổi ở bất cứ đâu trong mạng nơ-ron, thông thường với mục đích fine tuning sẽ sử dụng lớp FC, còn các mục đích cho các tác vụ cấp cao đầu ra thường sử dụng GAP thay vì lớp FC.



Hình 4. Lớp Global Average Pooling.

* + 1. **Chuẩn hoá**

Kiểm soát phân phối đầu vào giữa các lớp có thể giúp tăng tốc đáng kể việc huấn luyện mạng nơ-ron và cải thiện độ chính xác. Theo đó, phân phối của các lớp kích hoạt đầu vào được chuẩn hóa sao cho nó có giá trị trung bình bằng 0 và độ lệch chuẩn đơn vị. Trong BN (Batch Normalization), giá trị chuẩn hóa được chia tỷ lệ và dịch chuyển hơn nữa, trong đó các tham số được học từ quá trình huấn luyện:

Trong đó là một hằng số nhỏ để tránh các vấn đề về số (chia cho 0). BN hiện được coi là tiêu chuẩn trong thiết kế CNN. BN thường được thực hiện giữa lớp CONV hoặc FC và hàm kích hoạt phi tuyến.

* + 1. **Dropout**

Các mạng nơ-ron có số lượng tham số lớn thường có thể gặp vấn đề overfitting. Điều này ảnh hưởng rõ ràng đến khả năng khái quát hoá của mạng nơ-ron, dẫn đến hiệu suất kém trên tập kiểm thử. Một cách tiếp cận hiệu quả nhưng đơn giản là sử dụng dropout. Trong quá trình huấn luyện một số đặc trưng đầu ra có thể bị “bỏ qua” một cách ngẫu nhiên, với tỉ lệ ngẫu nhiên là . Điều này khiến mạng nơ-ron không bị phụ thuộc vào các đặc trưng có giá trị lớn hơn nhiều so với các đặc trưng còn lại (nếu có), vì mạng sẽ cho rằng đây là đặc trưng quan trọng và đánh trọng số lớn cho nó, trong khi các đặc trưng mang thông tin khác bị cho là nhiễu và bị bỏ qua (nguyên nhân chính của overftting). Do đặc trưng bị dropout (giá trị gán bằng 0) nên nó không tham gia vào quá trình cập nhật trọng số trong thuật toán gradient descent, nên mạng có thể giữ ổn định khi tối ưu

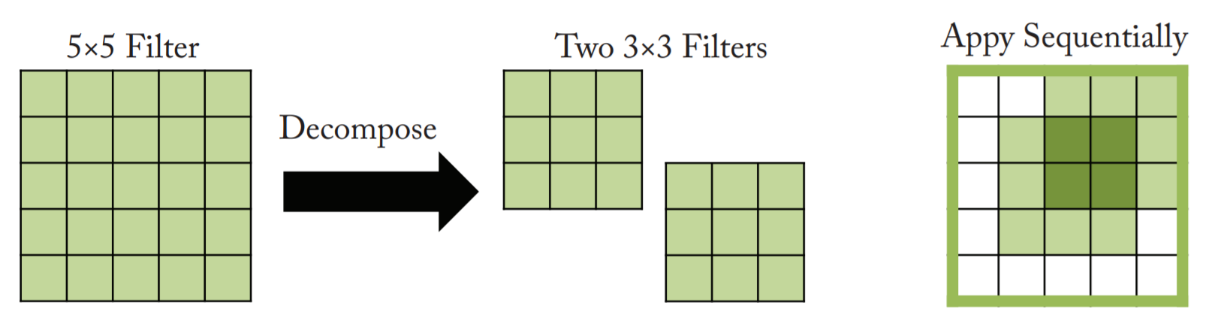
Điều cần nhớ là việc sử dụng dropout chắc chắn đã làm mất đi một phần thông tin được trích chọn, do đó để áp dụng dropout mà không làm giảm đáng kể hiệu suất của mạng, vị trí của dropout trong kiến trúc mạng sẽ đóng vai trò quan trọng.

* 1. **Mạng nơ-ron tích chập (CNN)**

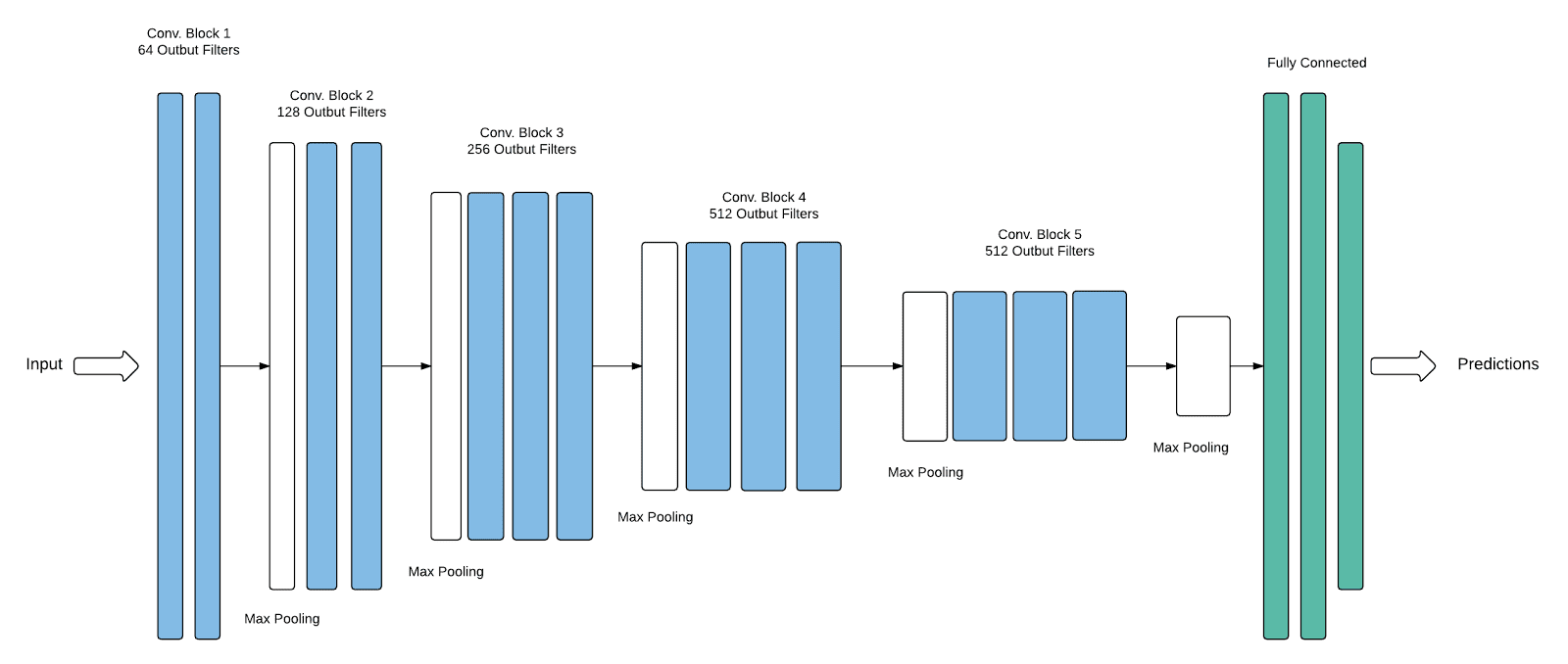
CNN là một dạng DNN phổ biến được cấu tạo bởi nhiều lớp CONV. Trong các mạng nơ-ron như vậy, mỗi lớp tạo ra một phần trừu tượng cấp cao hơn liên tiếp trên dữ liệu đầu vào, được gọi là bản đồ đặc trưng (), lưu giữ thông tin thiết yếu nhưng duy nhất. CNN hiện đại có thể đạt được hiệu suất vượt trội bằng cách sử dụng hệ thống phân cấp các lớp rất sâu.

* + 1. **VGG Net**

VGG-16 đạt độ sâu đến 16 lớp bao gồm 13 lớp CONV tiếp theo là 3 lớp FC. Để cân bằng chi phí, các bộ lọc lớn hơn (ví dụ: ) được xây dựng từ nhiều bộ lọc nhỏ hơn (ví dụ: ), có số lượng tham số ít hơn, để đạt được cùng các trường tiếp thụ hiệu quả, như thể hiện trong Hình 5. Kết quả là tất cả các lớp CONV có cùng kích thước bộ lọc là . Tổng cộng, VGG-16 yêu cầu 138M tham số và 15.5G MACs để xử lý một hình ảnh đầu vào . VGG có hai mẫu khác nhau: VGG-16 và VGG-19. VGG-19 cho tỷ lệ lỗi Top-5 thấp hơn 0.1% so với VGG-16 với chi phí thêm gấp 1.27 MAC.



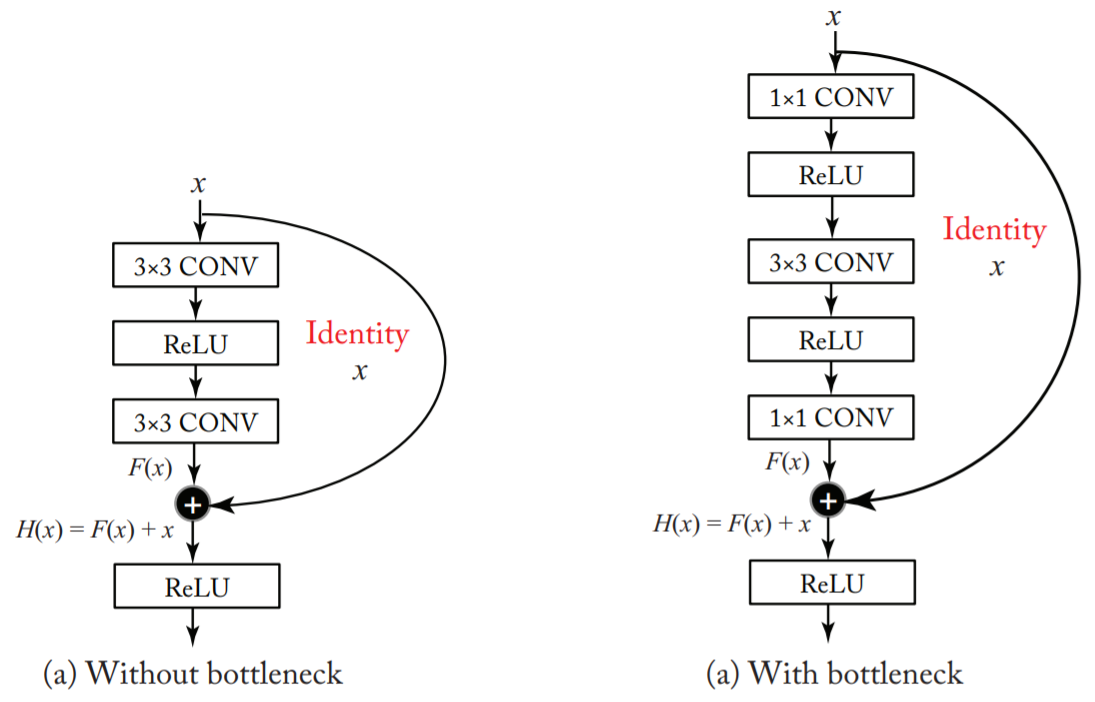
Hình 5. Xây dựng bộ lọc từ hai bộ lọc .



Hình 6. Kiến trúc VGG-16

* + 1. **ResNet**

ResNet còn được gọi là Residual Net, sử dụng kết nối feed forward để kết nối với các lớp bên ngoài lớp ngay kế tiếp (thường được gọi là kết nối residual), những kết nối này cho phép một mạng CNN sâu (ví dụ từ 34 lớp trở lên) có thể được huấn luyện hiệu quả. Đây là kiến trúc đầu tiên trong ImageNet Challenge vượt quá độ chính xác cấp độ con người với tỉ lệ lỗi Top-5 dưới 5%. Một trong những thách thức của mạng nơ-ron sâu vấn đề gradient biến mất trong quá trình huấn luyện, khi tối ưu thông qua lan truyền ngược, gradient sẽ bị thu hẹp lại và tiệm cận về 0, do đó quá trình cập nhật tham số bị dừng lại, gây ảnh hưởng đến khả năng fine tuning trong các mạng rất sâu. ResNet giới thiệu kiến trúc shortcut module chứa các kết nối residual, như trong Hình 7, cho phép huấn luyện mạng rất sâu mà giảm thiểu sự tác động từ gradient biến mất. ResNet cũng sử dụng cách tiếp cận bottleneck với bộ lọc để giảm số lượng tham số. Có nhiều phiên bản ResNet khác nhau với các độ sâu (18, 34, 50, 101, 152).



Hình 7. Shortcut module trong ResNet.

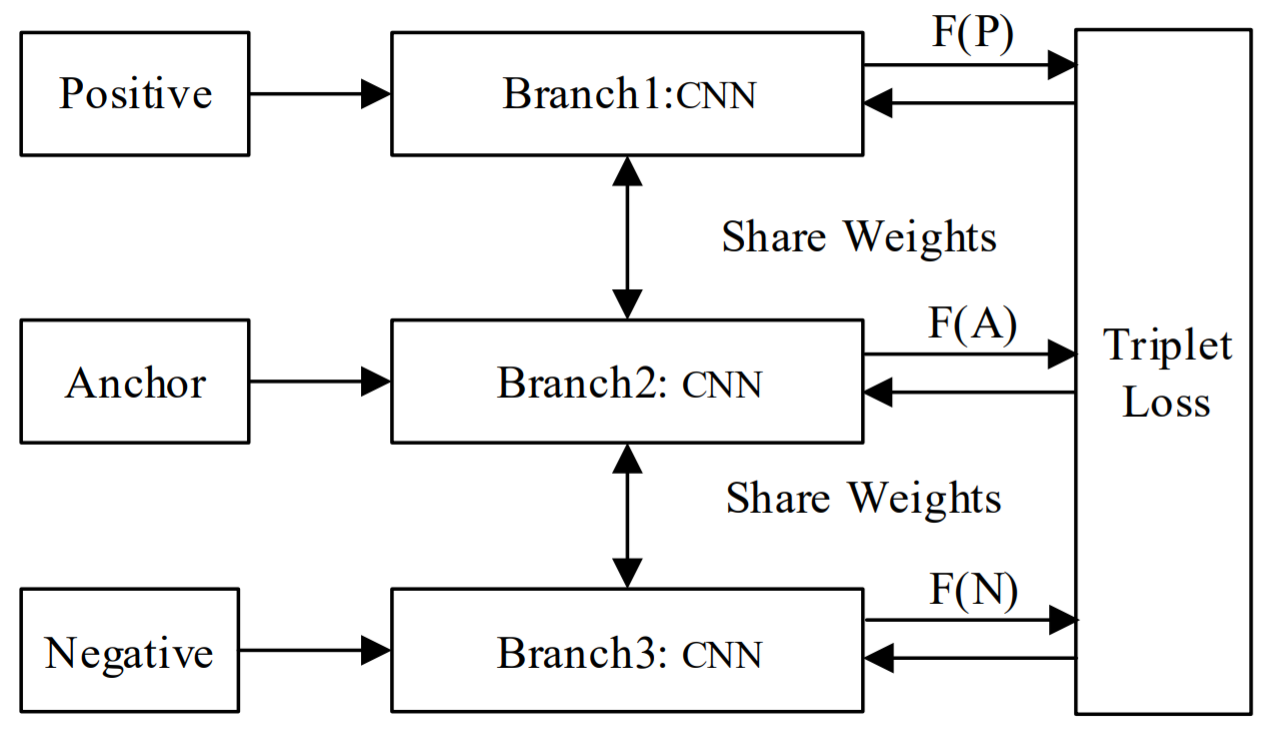
1. **Triplet Net cho bài toán truy xuất hình ảnh**

Để giải quyết bài toán truy xuất hình ảnh, mục tiêu là thiết kế một kiến trúc mạng nơ-ron có khả năng trích chọn đặc trưng cấp cao phù hợp cho việc tính độ tương đồng một cách hiệu quả. Hai hình ảnh và được tính toán độ tương đồng theo khoảng cách Euclid bình phương trong không gian vector đại diện của chúng:

trong đó là hàm tính toán vector đại diện, ánh xạ một ảnh đầu vào tới một điểm trong không gian Euclid và là khoảng cách Euclid bình phương trong không gian này. Khoảng cách càng nhỏ thì hai ảnh và càng tương đồng. Định nghĩa này hình thành bài toán truy xuất hình ảnh tương tự như bài toán tìm kiếm láng giềng gần nhất trong không gian Euclide, có thể được giải một cách hiệu quả thông qua thuật toán tìm kiếm láng giềng gần nhất.

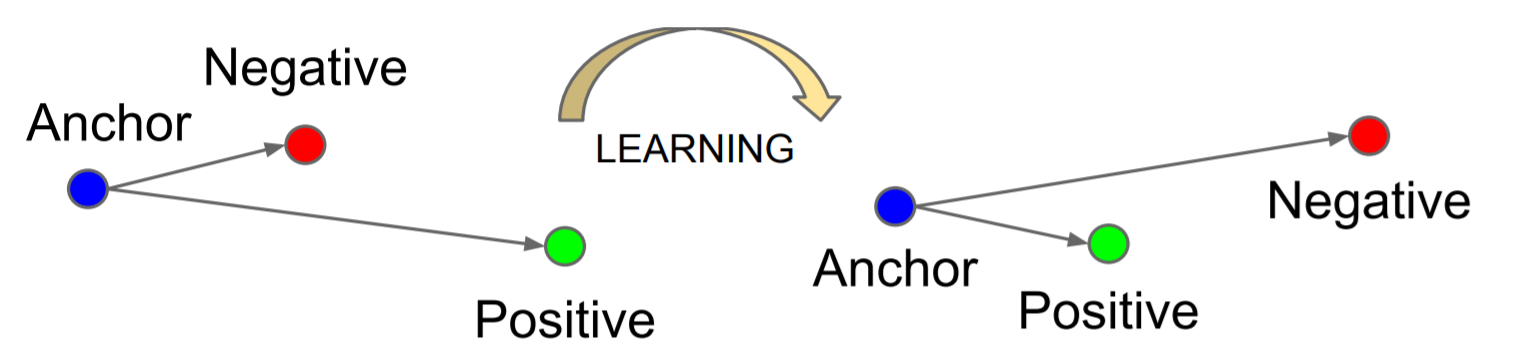
Giả sử có một tập các hình ảnh mà có thể chọn ra bộ ba hình ảnh tương ứng là ảnh truy vấn (anchor), ảnh dương (positive) và ảnh âm (negative). Trong đó với mỗi ảnh truy vấn đã xác định được tập hợp các ảnh dương (là ảnh có nội dung tương đồng với ảnh truy vấn), tức và là cặp ảnh tương đồng. Ngoài ra và là các ảnh khác danh mục hoặc có nội dung khác nhau. Thông qua mối quan hệ về khoảng các giữa 3 hình ảnh, mạng Triplet Net thiết kế một hàm mục tiêu Triplet Contrastive Loss:

trong đó tham số đại diện cho ngưỡng (margin) giữa và , biểu thị số lượng mẫu bộ 3.



Hình 8. Kiến trúc mạng Triplet Net

Mục tiêu tối ưu của mô hình là các cặp ảnh tương đồng sẽ ở gần nhau và các cặp hình ảnh không giống nhau ở xa nhau. Nghĩa là, mạng Triplet Nét muốn cực tiểu hoá khoảng cách giữa và , cả hai đều là những hình ảnh tương đồng với nhau và cực đại hóa khoảng cách giữa và là những hình ảnh khác nhau. Từ mối quan hệ khoảng cách giữa các hình ảnh, hàm mục tiêu trở thành khoảng cách giữa và nhỏ hơn khoảng cách giữa và . Do đó, có một mối quan hệ khoảng cách như được chỉ ra cho tất cả các bộ 3:



Hình 9. Hàm mục tiêu Triplet Contrastive Loss

Giá trị càng lớn thì khả năng trích chọn đặc trưng của mạng để tính toán tương đồng càng hiệu quả, tuy nhiên khả năng học của mạng sẽ trở nên khó khăn hơn. Trong các thử nghiệm trong báo cáo, giá trị .

Kiến trúc Triplet Net gồm ba nhánh CNN giống nhau và được chia sẻ trọng số, như trong Hình 8. Mỗi nhánh sẽ xử lý một ảnh đầu vào trong mẫu bộ ba. Các vector đại diện được trích chọn từ ba nhánh này được đưa vào tính toán tối ưu với hàm mục tiêu Triplet Loss. Kiến trúc cụ thể của mỗi nhánh như sau. Hình ảnh được sử dụng làm đầu vào cho bộ trích chọn đặc trưng (CNN) để trích chọn các đặc trưng đại diện tương ứng của hình ảnh. Các kiến trúc mạng nơ-ron cơ bản được sử dụng trong báo cáo là Resnet và VGGNet, và hoàn toàn có thể mở rộng với bất kỳ bộ trích chọn đặc trưng nào tương tự.

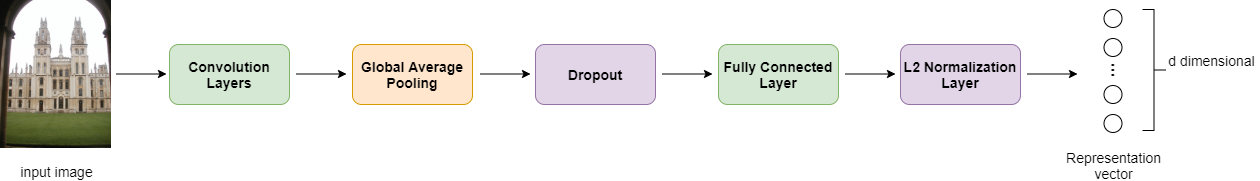
Trong đó ký hiệu của bộ trích chọn đặc trưng CNN. trong báo cáo này sử dụng kiến trúc VGG-16 và ResNet-34 (đã loại bỏ các lớp FC trên cùng) và ngoài ra hoàn toàn có thể tương thích với bất kỳ CNN nào khác. là đầu vào của mạng dưới dạng tensor 3 chiều, có kích thước đối với ảnh RGB; là tensor bản đồ đặc trưng đầu ra từ . Giá trị thường khi kèm với cấu hình riêng biệt cho từng mạng rơ-ron, phổ biến nhất là .

Ma trận không được hữu ích khi tính toán, do đó nó được chuyển sang dạng vector (tensor 1 chiều) một cách hiệu quả bằng cách sử dụng .

Trong đó là vector đại diện thu được. Trong quá trình huấn luyện, mô hình thường gặp phải vấn đề overfitting, một vấn đề quan trọng cần giải quyết trong việc thiết kế mạng nơ-ron hiệu quả. Để tránh overftting, toán tử dropout được áp dụng cho vector đại diện , về cơ bản giúp mạng nơ-ron tối ưu hoá một cách ổn định.

Với là tỉ lệ dropout (tức tỉ lệ bỏ ngẫu nhiên các giá trị trong vector). Trong các thực nghiệm .

Cuối cùng, để giảm chiều vector xuống một không gian nhỏ gọn hơn và đồng thời tiến hành fine tuning cho phù hợp với yêu cầu bài toán, sử dụng lớp FC để tính toán.

Trong đó ma trận là tham số của mạng nơ-ron dày đặc, là vector đại diện đầu ra sau khi giảm chiều, với được sử dụng trong các thực nghiệm. Toán tử chuẩn hoá được áp dụng để đưa các giá trị trong vector về cùng phân phối.

Hình 10. Quá trình trích chọn đặc trưng.

Huấn luyện mô hình Triplet Net và tối ưu hóa các tham số bằng cách lan truyền ngược và thuật toán gradient descent theo từng minibatch. Trong quá trình huấn luyện, hàm mục tiêu trở thành công thức và được định nghĩa là :

Trong đó đại diện cho kích thước của minibatch với các mẫu bộ ba được chọn ngẫu nhiên.

Khi , tính riêng các gradient cho hàm ánh xạ đặc trưng đại diện (ký hiệu là ) theo , và như sau:

Theo quy tắc chuỗi tính toán gradient cho hàm hợp, tính toán gradient của hàm mục tiêu cho các tham số :

Trong đó:

Bỏ qua giá trị bias không được đề cập trong tính toán cũng như tối ưu.

Do đó, mạng cập nhật trọng số như sau:

Trong đó là giá trị tốc độ học tập (learning rate), đại diện cho lớp.

Sau khi đã huấn luyện mô hình Triplet Net, chọn một trong ba nhánh mạng và tải hình ảnh đầu vào và thu được một vector đại diện được trích chọn từ nó. Cho hai ảnh và , sau đó thu được các vector đại diện tương ứng của chúng là và , với là số chiều của vector. Do đó, độ tương đồng giữa hai hình ảnh và có thể được xác định thông qua hàm tương đồng cosine :

Sử dụng thuật toán – láng giềng gần (KNN) đơn giản có thể thu được danh sách các hình ảnh có độ tương đồng cao nhất so với hình ảnh truy vấn.

1. **Triển khai và đánh giá**

Trong việc triển khai, để tránh sự ảnh hưởng của các kiến trúc CNN có hiệu suất hiện đại, việc triển khai Triplet Net sẽ trên các kiến trúc baseline như VGG-16 và Resnet-34 (đã loại bỏ 3 lớp FC trên cùng), và so sánh nó với các phương pháp học máy cổ điển như Bag of Words với các đặc trưng SIFT và SURF. Đánh giá dự trên hai bộ dữ liệu hình ảnh Oxford5k và Paris.

* Bộ dữ liệu Oxford5k bao gồm 5062 hình ảnh thu thập từ Flickr, được chú thích thủ công với 11 nhãn và 55 truy vấn mà hệ thống truy xuất hình ảnh có thể dùng để đánh giá.
* Bộ dữ liệu Paris bao gồm 6412 hình ảnh, có định dạng giống như bộ dữ liệu Oxford5k.

Kết quả đánh giá dựa trên hai bộ dữ liệu như trong Bảng 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Oxford5k** | | **Paris** | |
| ***mAP(%)*** | ***Time(s)*** | ***mAP(%)*** | ***Time(s)*** |
| BoW (SIFT) | 15.87 | 0.209 | 16.82 | 0.241 |
| BoW (SURF) | 10.08 | 0.236 | 9.76 | 0.296 |
| VGG - TripletNet | 90.56 | 0.247 | 92.32 | 0.260 |
| Resnet - TripletNet | 91.00 | 0.256 | 94.60 | 0.294 |

Bảng 1. Kết quả đánh giá các phương pháp trên hai bộ dữ liệu

Oxford5k và Paris

1. **Kết luận**

Báo cáo trình bày kiến trúc Triplet Net, một mạng nơ-ron hiệu quả cho bài toán truy xuất hình ảnh. Theo kết quả phân tích lý thuyết và thực nghiệm, Triplet Net mặc dù đơn giản nhưng có hiệu suất tốt trong lĩnh vực đo độ tương đồng ảnh. Mạng nơ-ron sâu có thể thể hiện thông tin hình ảnh một cách hiệu quả bằng cách mô phỏng nhận thức của con người, nó thu hẹp khoảng cách giữa máy móc và con người trong hiểu biết về ngữ nghĩa và xử lý tốt hơn bài toán “khoảng cách ngữ nghĩa”.