ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA KHOA HỌC MÁY TÍNH

BÁO CÁO ĐỒ ÁN

THỊ GIÁC MÁY TÍNH NÂNG CAO

CS331.N12.KHCL

ĐỀ TÀI: CROWD DETECTION

GIẢNG VIÊN BỘ MÔN: TS. MAI TIẾN DŨNG

SINH VIÊN THỰC HIỆN: HOÀNG NGỌC QUÂN - 17520934

TP. HỒ CHÍ MINH, 12/2022

MỤC LỤC

TÓM TẮT ĐỒ ÁN

Search Image (tìm kiếm hình ảnh) là một công nghệ cho phép người dùng tìm kiếm các hình ảnh trực tuyến thông qua các công cụ tìm kiếm. Công nghệ này sử dụng thuật toán phân tích hình ảnh để so sánh các đặc điểm của hình ảnh với những đặc điểm của các hình ảnh khác trong cơ sở dữ liệu của nó. Khi có một hình ảnh mới được tìm kiếm, thuật toán sẽ phân tích hình ảnh này và so sánh nó với các hình ảnh trong cơ sở dữ liệu để tìm kiếm các kết quả phù hợp nhất.

1. GIỚI THIỆU

1.1 Nội dung

Công nghệ tìm kiếm hình ảnh phát triển nhanh chóng trong những năm qua và trở thành một công cụ quan trọng trong nhiều lĩnh vực, bao gồm marketing, giáo dục, y tế, giải trí và nhiều lĩnh vực khác. Với sự phát triển của trí tuệ nhân tạo, các thuật toán phân tích hình ảnh ngày càng trở nên chính xác và đa dạng hơn, giúp người dùng tìm kiếm hình ảnh dễ dàng và nhanh chóng hơn.

Các công cụ tìm kiếm hình ảnh phổ biến hiện nay bao gồm Google Images, Bing Images, Yahoo Image Search và DuckDuckGo Images. Các công cụ này cho phép người dùng tìm kiếm hình ảnh trực tuyến bằng cách sử dụng từ khóa hoặc đường dẫn hình ảnh. Ngoài ra, người dùng cũng có thể tìm kiếm hình ảnh tương tự hoặc hình ảnh liên quan bằng cách sử dụng các tính năng tìm kiếm liên quan.

Tìm kiếm hình ảnh cũng được sử dụng rộng rãi trong lĩnh vực quảng cáo và tiếp thị trực tuyến. Các doanh nghiệp có thể sử dụng tìm kiếm hình ảnh để tìm kiếm hình ảnh liên quan đến sản phẩm hoặc dịch vụ của họ và sử dụng chúng trong các chiến dịch quảng cáo trực tuyến.

1.2 Nhiệm vụ đề tài

Từ nội dung nêu trên, đề tài sẽ bao gồm các nhiệm vụ sau:

- Tìm hiểu khái quát về trích xuất đặc trưng của ảnh
- Tìm hiểu về cách tìm hình ảnh có cùng đặc trưng với ảnh vừa xử lý
- Cuối cùng đánh giá và xuất ảnh ra có cùng đặc trưng

2. Tổng quan bài toán

2.1. Khái niệm Image Search

Search Image (tìm kiếm hình ảnh) là một công nghệ cho phép người dùng tìm kiếm các hình ảnh trực tuyến thông qua các công cụ tìm kiếm. Công nghệ này sử dụng thuật toán phân tích hình ảnh để so sánh các đặc điểm của hình ảnh với những đặc điểm của các hình ảnh khác trong cơ sở dữ liệu của nó. Khi có một hình ảnh mới được tìm kiếm, thuật toán sẽ phân tích hình ảnh này và so sánh nó với các hình ảnh trong cơ sở dữ liệu để tìm kiếm các kết quả phù hợp nhất.

2.2. Phương pháp giải quyết

Tìm kiếm dựa trên nội dung (Content-based image retrieval - CBIR): Phương pháp này sử dụng các đặc trưng của hình ảnh (như màu sắc, hình dạng, texture, v.v.) để tìm kiếm các hình ảnh tương tự trong cơ sở dữ liệu. Các thuật toán CBIR thông thường sử dụng các kỹ thuật như histogram, biểu đồ tính chất (feature vector), mô hình mạng neural, v.v.

2.3. Input và Output

2.3.1. Input

Tập ảnh (định dạng .jpg/.png)

2.3.2. Output

Chọn một ảnh bất kỳ để tìm kiếm và xuất ra các ảnh tương tự với ảnh được tìm kiếm

2.4. Giới thiệu phương pháp

2.4.1. VGG16

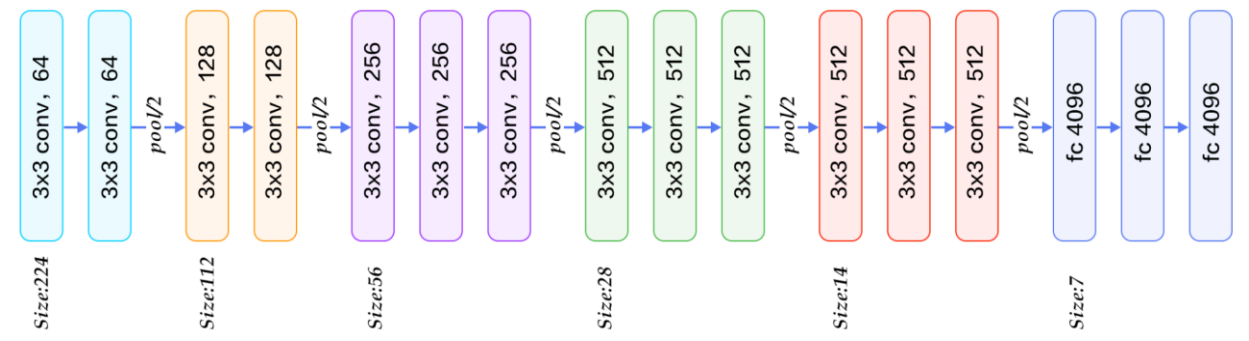
VGG16 là một mô hình mạng nơ-ron tích chập (Convolutional Neural Network - CNN) được đề xuất bởi nhóm nghiên cứu Visual Geometry Group tại Đại học Oxford vào năm 2014. VGG16 được đặt tên theo số lượng lớp của mô hình, bao gồm 16 lớp, trong đó có 13 lớp tích chập và 3 lớp fully connected.

Mô hình VGG16 có kiến trúc đơn giản và dễ hiểu, với tất cả các lớp đều sử dụng các bộ lọc nhỏ (3x3 hoặc 1x1) và các lớp max pooling để giảm kích thước đầu vào của ảnh. VGG16 có khả năng học được các đặc trưng phức tạp của hình ảnh, giúp nó trở thành một trong những mô hình CNN hiệu quả nhất cho việc phân loại hình ảnh trên các tập dữ liệu lớn.

Cụ thể, kiến trúc của VGG16 gồm:

- Lớp đầu vào: nhận ảnh đầu vào với kích thước 224x224x3.
- 13 lớp tích chập (convolutional layer): với các bộ lọc 3x3 và stride 1, đầu ra của mỗi lớp là các feature map với số lượng channel tăng dần từ 64 đến 512.
- 5 lớp max pooling: với kích thước pooling là 2x2 và stride là 2, giúp giảm kích thước của feature map đầu vào.

- 3 lớp fully connected: lớp cuối cùng là lớp phân loại với 1000 node (tương ứng với 1000 lớp phân loại trong ImageNet).



2.4.2. RESNet-50

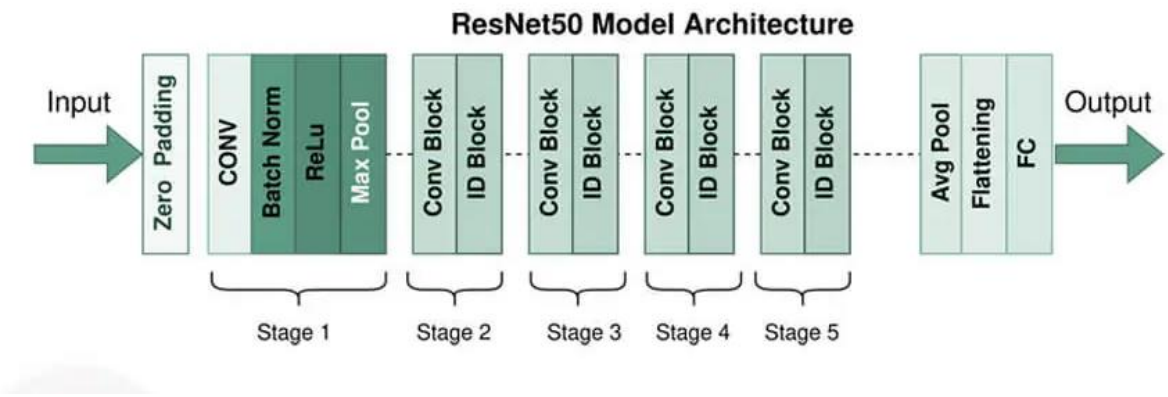
ResNet-50 là một trong những mô hình mạng nơ-ron tích chập (Convolutional Neural Network - CNN) phổ biến nhất hiện nay, được đề xuất bởi nhóm nghiên cứu Microsoft Research vào năm 2015. ResNet-50 có kiến trúc sâu và phức tạp, với hơn 50 lớp tích chập (convolutional layer) và hàm kích hoạt phi tuyến (non-linear activation function).

ResNet-50 được đặt tên theo khả năng của nó trong việc "skip connection" hay còn gọi là "residual connection", điều này giúp cho mô hình có khả năng xử lý các vấn đề về độ sâu của mạng. Thông thường, khi ta xây dựng các mạng nơ-ron sâu, việc huấn luyện mô hình gặp khó khăn khi số lượng lớp tăng cao. Điều này dẫn đến hiện tượng "vanishing gradient", làm cho các thông tin của hình ảnh không thể truyền xuống các lớp tiếp theo của mạng. Để khắc phục vấn đề này, ResNet-50 sử dụng các skip connection để bỏ qua một số lớp và truyền các thông tin trực tiếp từ lớp đầu vào đến các lớp

đầu ra, qua đó giữ lại thông tin quan trọng và giúp mô hình có khả năng học các đặc trưng phức tạp của hình ảnh.

Cụ thể, kiến trúc của ResNet-50 gồm:

- Lớp đầu vào: nhận ảnh đầu vào với kích thước 224x224x3.
- 50 lớp tích chập (convolutional layer): với các bộ lọc 3x3 hoặc 1x1, và có các skip connection để giữ lại thông tin quan trọng.
- 1 lớp fully connected: lớp cuối cùng để phân loại hình ảnh.



3. Thực thi các phương pháp và đánh giá

3.1. Tập dữ liệu

Caltech 101 là một tập dữ liệu (dataset) gồm 101 lớp ảnh khác nhau, mỗi lớp bao gồm khoảng 40-800 ảnh. Tập dữ liệu này được tạo ra bởi Đại học California, Los Angeles (Caltech) nhằm mục đích phát triển và đánh giá các thuật toán nhận dạng đối tượng và phân loại ảnh. Các lớp ảnh trong tập dữ liệu Caltech 101 bao gồm các đối tượng khác nhau như động vật, phương tiện giao thông, đồ vật và con người.

3.2. Phương pháp VGG16

Khi đưa một ảnh vào mạng VGG16, quá trình xử lý sẽ diễn ra theo các bước sau:

1. Đầu tiên, ảnh sẽ được tiền xử lý bằng cách chuyển đổi kích thước ảnh về độ phân giải 224x224x3 pixel. Sau đó, ảnh được chuẩn hóa để các giá trị pixel nằm trong khoảng [0,1] hoặc [-1,1], tùy thuộc vào cách tiền xử lý được sử dụng.

2. Ảnh được đưa vào mạng và truyền qua các lớp tích chập (convolutional layer) của VGG16. VGG16 có kiến trúc gồm 13 lớp tích chập và 3 lớp fully connected. Các lớp tích chập này sử dụng các bộ lọc có kích thước 3x3 và ảnh đầu vào sẽ được truyền qua các bộ lọc này để tìm ra các đặc trưng của ảnh. Các lớp tích chập này có thể được nhóm lại thành các khối, mỗi khối bao gồm nhiều lớp tích chập và một lớp pooling.

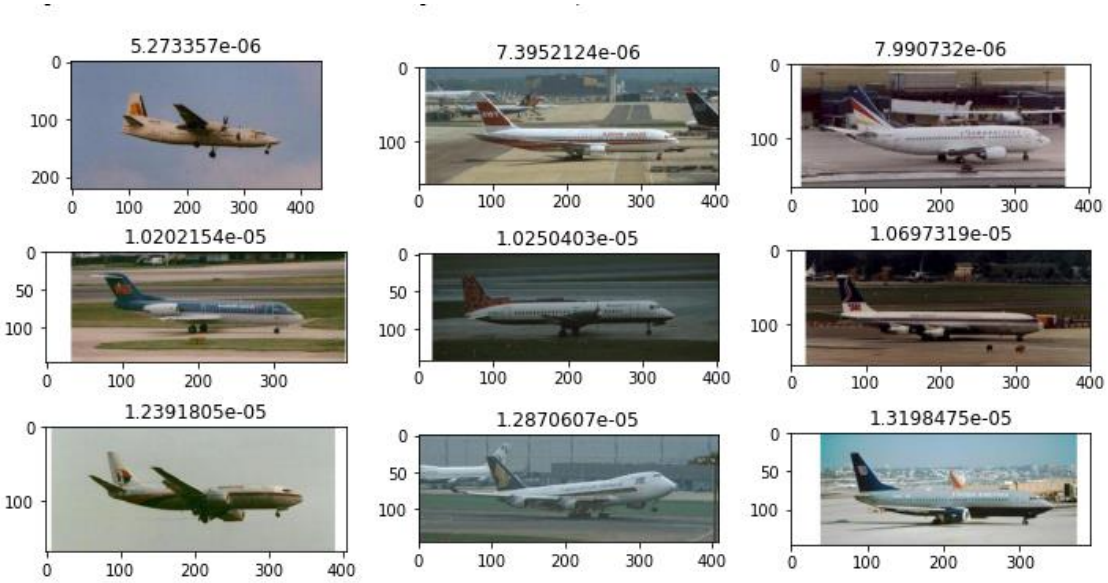
3. Sau khi ảnh đã đi qua các lớp tích chập, các feature map của ảnh được truyền vào các lớp fully connected. Những lớp này sẽ giúp mô hình học được các quan hệ phức tạp giữa các đặc trưng của ảnh để phân loại ảnh vào các lớp khác nhau.

4. Cuối cùng, mô hình sử dụng softmax để tính toán xác suất của ảnh thuộc về mỗi lớp phân loại.

3.3. Thực nghiệm

Ảnh muốn tìm kiếm:

Kết quả các ảnh muốn tìm kiếm:



3.4. Phương pháp Resnet-50

RESNet-50 là một kiến trúc mạng nơ-ron sâu (deep neural network) được sử dụng cho các tác vụ phân loại ảnh. Khi đưa một ảnh vào mạng RESNet-50, quá trình xử lý sẽ diễn ra theo các bước sau:

1. Tiền xử lý ảnh: Ảnh được chuẩn hóa để các giá trị pixel nằm trong khoảng [-1, 1] hoặc [0, 1]. Sau đó, ảnh được chuyển về độ phân giải 224x224x3 pixel.

2. Đưa ảnh vào mạng: Ảnh được đưa vào mạng và truyền qua các lớp tích chập (convolutional layer) của RESNet-50. RESNet-50 có kiến trúc sâu với 50 lớp tích chập, được nhóm lại thành các khối (blocks). Mỗi khối bao gồm các lớp tích chập và lớp skip connection. Các lớp skip connection giúp tránh tình trạng mất mát độ sâu của thông tin khi đưa ảnh qua các lớp tích chập sâu.

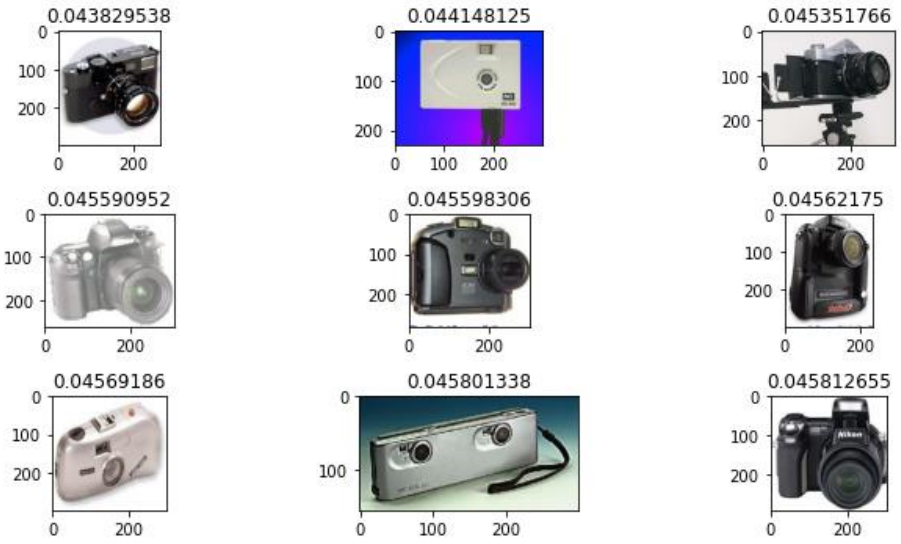
3. Tính toán feature maps: Các feature maps của ảnh được tính toán sau khi đã đi qua các lớp tích chập và lớp skip connection. Các feature maps này chứa các đặc trưng của ảnh.

4. Tính toán kết quả phân loại: Các feature maps của ảnh được đưa vào các lớp fully connected để học được các quan hệ phức tạp giữa các đặc trưng của ảnh và phân loại ảnh vào các lớp khác nhau. Cuối cùng, mô hình sử dụng softmax để tính toán xác suất của ảnh thuộc về mỗi lớp phân loại.

3.5. Thực nghiệm

Ảnh muốn tìm kiếm:

Kết quả các ảnh muốn tìm kiếm:



4. Kết luận

Từ những kết quả thu được ở trên em nhận thấy các phương pháp nhận diện đám đông có những ưu và khuyết điểm sau:

Ưu điểm:

- Về cả 2 phương pháp VGG16 và Resnet-50 thì nhận xét chung đều có tính đơn giản.
- Cả 2 phương pháp đều hoạt động tốt trên tập dữ liệu Caltech101.
- Cả 2 đều cho phép điều chỉnh các trọng số mô hình trên tập dữ liệu mới, giúp nó phù hợp với nhiều ứng dụng khác nhau.
- Riêng về Resnet-50 thì mặt hiệu suất lại tốt hơn so với VGG16

Nhược điểm:

- Về VGG16 thì rất dễ bị overfit vì tập dữ liệu thực hiện ở đây khá nhỏ so với thực tế.

- Về mặt tốc độ train để đánh giá mô hình với 2 thuật toán thì Resnet-50 nhanh hơn hẳn VGG16 vì Resnet-50 có sử dụng kỹ thuật skip connection.

5. Tài liệu tham khảo

[1] Very Deep Convolutional Networks for Large-Scale Image Recognition: https://arxiv.org/abs/1409.1556

[2] VGG16 trên trang chủ của thư viện deep learning Keras: https://keras.io/api/applications/vgg/

[3] Deep Residual Learning for Image Recognition: https://arxiv.org/abs/1512.03385

[4] ResNet-50 trên trang chủ của thư viện deep learning Keras: https://keras.io/api/applications/resnet/

[5] https://github.com/tensorflow/models/tree/master/research/slim#pre-trained-models