Báo cáo cuối kỳ môn Thị giác máy

Giảng viên: TS. Nguyễn Thị Hồng Thịnh

Thành viên thực hiện: Nguyễn Trường Sơn - MHV: 20025058
Trần Chính Đoàn - MHV: 20025054
Cao Huy Nhật - MHV: 21025118

1. Giới thiệu

Ngày nay, chúng ta đang sống trong thời đại của dữ liệu. Với sự phát triển của Internet of Things (IoTs) và trí tuệ nhận tạo (AI – Artificial Inteligence), chúng ta đang có một khối lượng dữ liệu khổng lồ được tạo ra. Dữ liệu được sinh ra với nhiều dạng khác nhau, từ dữ liệu về giọng nói, văn bản, hình ảnh hoặc là sự kết hợp của các kiểu dữ liệu trong số này. Trong đó dữ liệu dạng hình ảnh chiếm một phần đáng kể trong kho dữ liệu này. Với một lượng lớn dữ liệu được thu thập từ rất nhiều nguồn khác nhau như vậy, dữ liệu cần phải được xử lý và phân tích một cách hiệu quả. Và một trong những khâu không thể thiếu được của quá trình này đó là phân loại ảnh (Image Classification). Với việc sử dụng các mô hình học sâu dựa trên AI để phân tích hình ảnh, chúng ta có thể đưa ra những kết quả với độ chính xác vượt qua con người (ví dụ như nhận dạng khuôn mặt).

Về cơ bản, Image Classification là việc phân loại một tập hợp hình ảnh theo các danh mục đã được quy định trước. Với con người, việc này khá dễ dàng vì chúng ta nhận thức được đặc điểm của từng bức ảnh, tuy nhiên, với máy tính, ảnh chỉ là một ma trận số. Để giúp máy tính "hiểu" được nội dung của các bước ảnh, chúng ta sẽ phải áp dụng các Image Descriptors và Deep Learning methods. Cuối cùng, dựa trên các tính chất của bức ảnh, chúng ta có thể áp dụng Machine Learning để "dạy" cho máy tính cách phân loại hình ảnh.

Trong báo cáo này, chúng tôi sẽ trình bày một vài kỹ thuật được sử dụng trong quá trình phân loại hình ảnh và thử nghiệm các kỹ thuật này trên dữ liệu chúng tôi thu thập được, từ đó đánh giá và so sánh các kết quả sử dụng các kỹ thuật này. Các kỹ thuật sẽ được sử dụng bao gồm:

- Sử dụng Bag of Visual Words (BoVW) (Hand-crafted feature)
- Sử dụng Deep Learning model – ResNet18

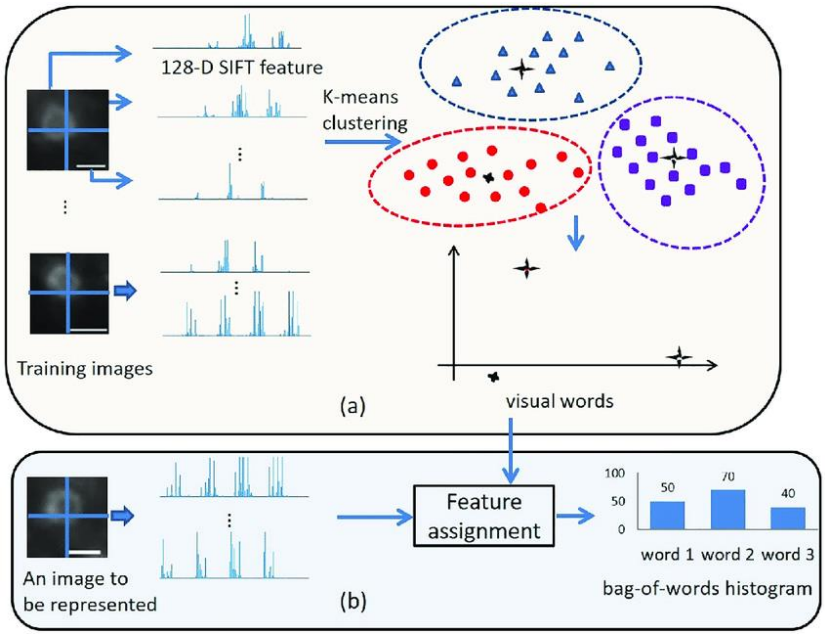
2. Phương pháp

Trong phần này, chúng tôi sẽ trình bày cơ bản về kỹ thuật sử dụng và quy trình thực hiện của các kỹ thuật này.

2.1. Bag of Visual Words

Bag of visual words (BoVW) là một kỹ thuật thường được sử dụng trong phân loại hình ảnh. Ban đầu kỹ thuật này được sử dụng trong xử lý ngôn ngữ tự nhiên (NLP). Ý tưởng chung của BoVW là coi các ảnh như là một tập hợp của các feature. Những feature này bao gồm keypoint và descriptor. Keypoint ở đây được hiểu là các điểm nột bật trong ảnh, là những điểm mà dù cho ảnh có bị xoay, thu nhỏ hay mở rộng thì vẫn sẽ giống nhau. Descriptor là mô tả của những điểm này. Dựa trên các đặc điểm trên của feature, các ảnh đầu vào sẽ được trích xuất điểm đặc trưng, xây dựng bộ từ điển feature và mô tả ảnh lại dưới dạng histogram của các feature trong

ảnh. Từ biểu đồ tần xuất này, chúng ta có thể tìm kiếm một hình ảnh tương tự hoặc phân loại ảnh dựa theo dữ liệu được train.



Hình 1: Quy trình thực hiện sử dụng BoVW

Các bước thực hiện:

Step 1: Chia tệp data ban đầu thành hai tệp train và test (sử dụng file split.py)

Step 2: Trích xuất feature từ tệp train và xây dựng từ điển feature sử dụng kmean (SIFT feature)

Step 3: Trích xuất feature từ tệp train/test và xây dựng histogram của các ảnh trong tệp train/test

Step 4: Sử dụng SVM để train dựa trên label data

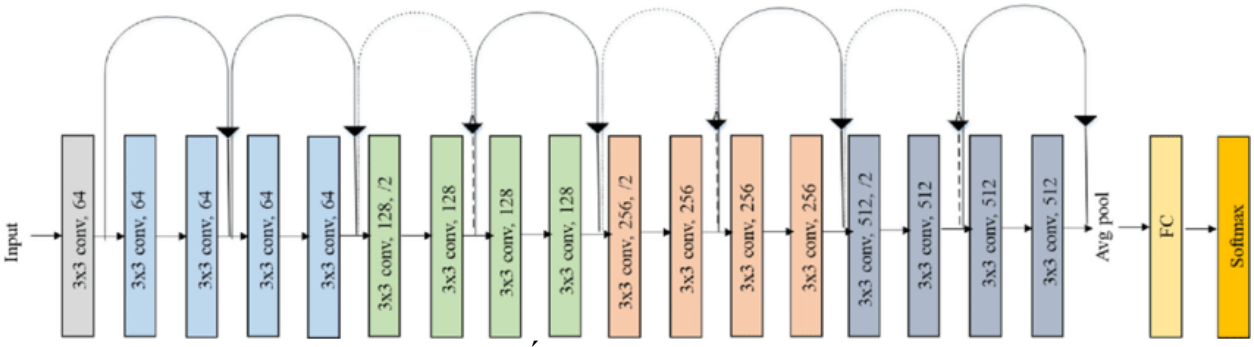
Step 5: Kiểm tra trên tệp test

Step 6: Đưa vào ảnh test và đưa ra k ảnh gần giống nhất (sử dụng SVM để predict label của ảnh, sau đó sử dụng chi-square distance để so sánh histogram của các ảnh để đưa ra các ảnh giống nhất)

2.2. ResNet18

ResNet (Residual Network) được giới thiệu đến công chúng vào năm 2015 và thậm chí đã đạt được top 1 trong cuộc thi ILSVRC 2015 với tỷ lệ lỗi top 5 chỉ 3.57%. Không những thế nó còn đứng vị trí đầu tiên trong cuộc thi ILSVRC and COCO 2015 với ImageNet Detection, ImageNet localization, Coco detection và Coco segmentation. Hiện tại thì có rất nhiều biến thể của kiến trúc ResNet với số lớp khác

nhau như ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-152, ...Với tên là ResNet theo sau là một số chỉ kiến trúc ResNet với số lớp nhất định.



Hình 2: Kiến trúc mạng ResNet-18

Với việc sử dụng mạng ResNet-18 này, chúng tôi sử dụng 3 kỹ thuật phổ biến trong Deep Learning là (2 kỹ thuật sau thuộc Transfer Learning):

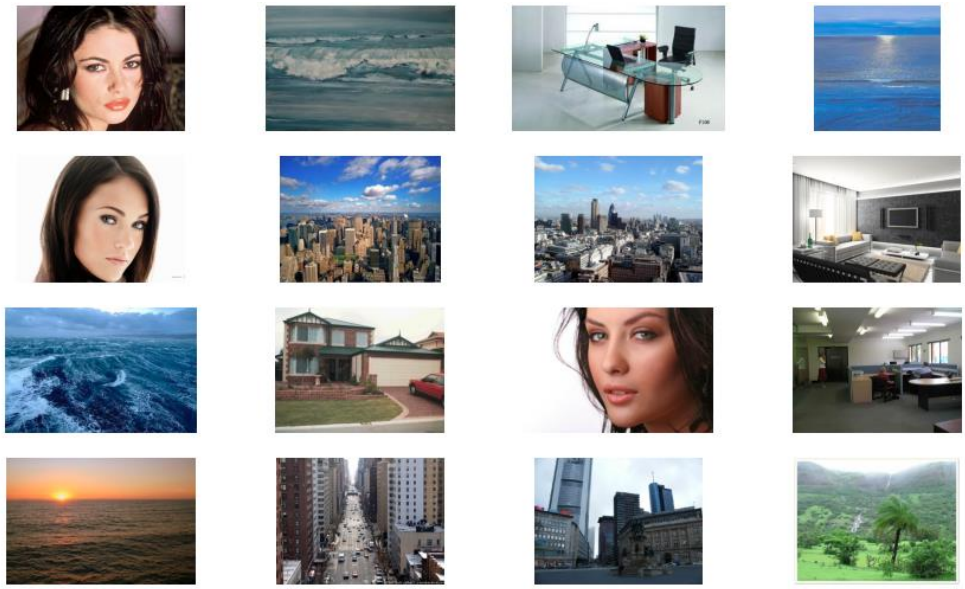
- Sử dụng kiến trúc ResNet-18 để train data (Train from Scratch)
- Giữ nguyên phần feature extraction (đã được pretrain với tệp dữ liệu ImageNet) và cập nhật lại phần phân loại dựa trên data của mình (Feature Extraction method)
- Sử dụng lại trọng số của model pretrain với ImageNet và tiếp tục train trên data của mình (Finetune method)

Về phần xác định k ảnh gần giống nhất, chúng tôi sử dụng phương pháp feature matching với SIFT feature trên các ảnh với label được predict qua mạng.

2.3. Tập dữ liệu

Tập dữ liệu được sử dụng trong báo cáo gồm 1030 ảnh với 7 class (city, face, green, building, house indoor, office, sea). Tập dữ liệu được chia làm 2 tệp train và test (với tỷ lệ 8:2).

Train: 814 ảnh
Test: 210 ảnh

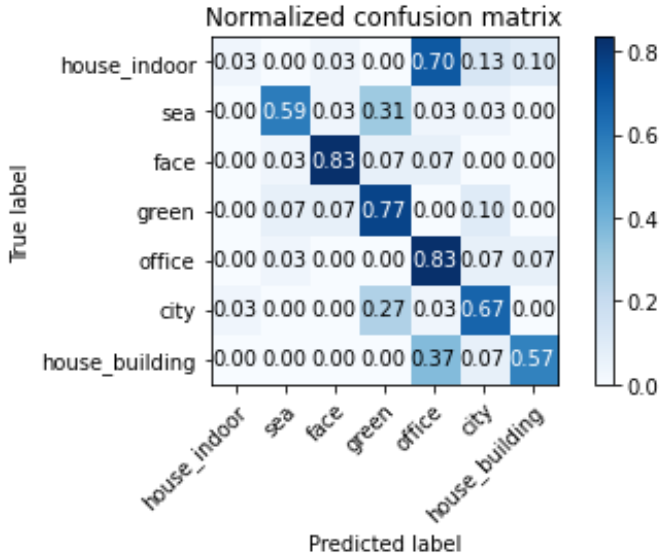


Hình 3: Một vài hình ảnh trong tệp dữ liệu

3. Kết quả

3.1. Bag of Visual Words

Số lượng từ trong từ điển: 500

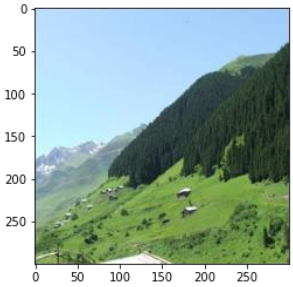


Hình 4: Bag of Visual Words Confusion Matrix

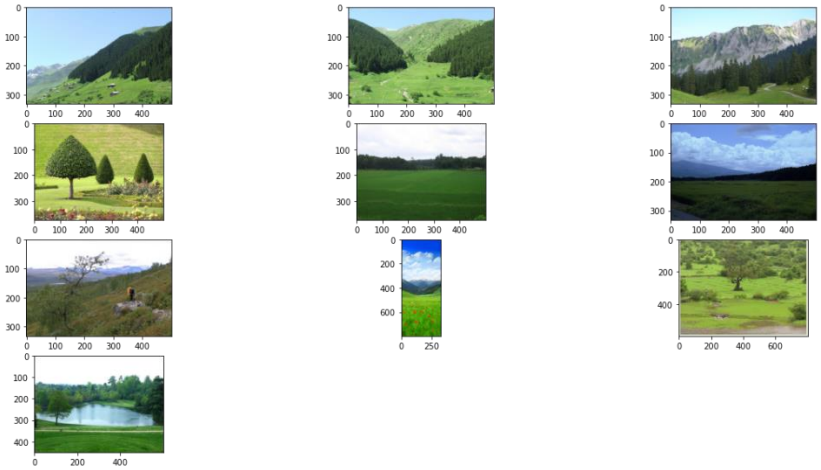
Accuracy: **0.622**

Thử nghiệm tìm kiếm ảnh gần giống nhất (k = 10):

Input:

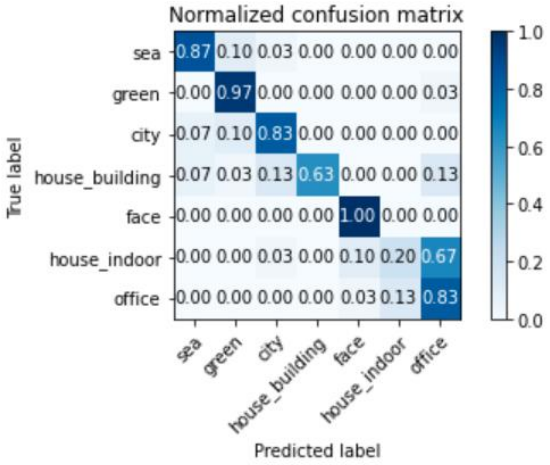


Output:



3.2. ResNet-18

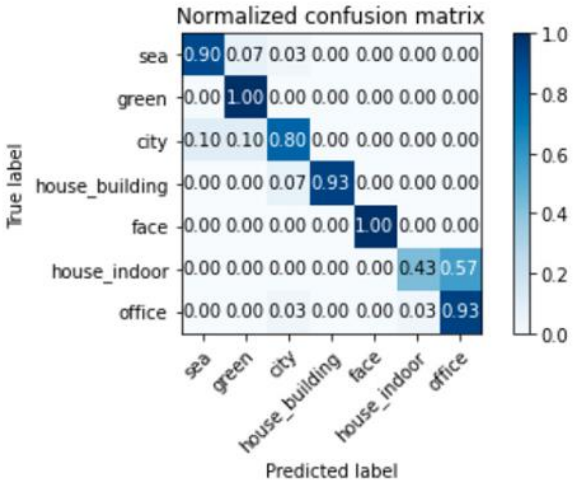
- **Train from Scratch**



Hình 5: Train from Scratch Confusion matrix

Accuracy: **0.7619**

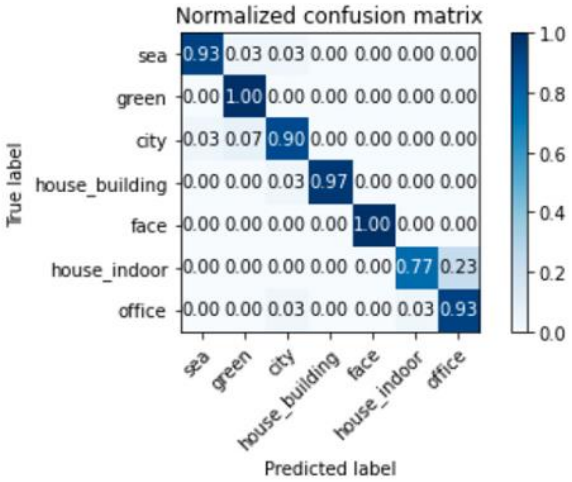
- **Feature Extraction method**



Hình 6: Feature Extraction method Confusion matrix

Accuracy: **0.8571**

- **Finetune method**

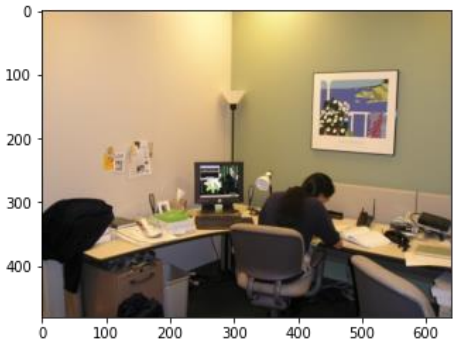


Hình 7: Finetune method Confusion matrix

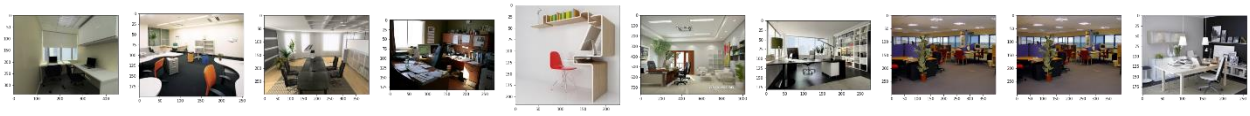
Accuracy: **0.9286**

Thử nghiệm tìm kiếm ảnh gần giống nhất (k = 10):

Input:



Output:



4. Kết luận

Như vậy, trong báo cáo này, chúng tôi đã thực hiện 4 phương pháp sử dụng học máy trong bài toán phân loại hình ảnh là: Bag of Visual Words (Hand-carfted feature), sử dụng model Deep Learning (Train from Scratch), Feature Extraction (Transfer Learning), Finetune (Transfer Learning). Độ chính xác đạt được tương ứng là **62.2%**, **76.19%**, **85.71%** và **92.86%**. Độ chính xác này cũng khá phù hợp với đặc tính của các phương pháp kể trên.

Dựa vào các confusion matrix, chúng ta có thể nhận thấy rằng với những class có đặc trưng khá riêng biệt và rõ ràng (như **green**, **face**, **sea**) cả 4 phương pháp đều đạt được kết quả rất tốt (đặc biệt có thể lên đến 100% với Deep Learning) trong khi đó 2 class **house_indoor** và **office** có tỷ lệ sai khá cao. Việc này có thể lý giải do đặc trưng của 2 class này khá giống nhau.

Với kết quả nói trên, chúng ta có thể đưa ra kết luận rằng các phương pháp Deep Learning đã cải thiện một cách rõ ràng độ chính xác của thuật toán phân loại ảnh. Tuy nhiên, hiện tại chúng tôi chưa sử dụng các phương pháp augmentation với dữ liệu, trong tương lai có thể thử nghiệm với các model khác và sử dụng thêm các phương pháp làm giàu dữ liệu để cải thiện độ chính xác của hệ thống.