

DẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

KHOÁ LUẬN TỐT NGHIỆP
CỦ NHÂN CÔNG NGHỆ THÔNG TIN

**KHÓA LUẬN TỐT NGHIỆP
PHÂN LỚP VẬT LIỆU BỀ SỬ
DỤNG MẠNG HỌC SÂU**



GIẢNG VIÊN HƯỚNG DẪN:

TS. MAI TIẾN DŨNG
TS. LÊ ĐÌNH DUY

SINH VIÊN THỰC HIỆN:

TRƯƠNG PHÚC ANH - 14520040

6, 2018

Xin dành tặng quyển luận văn này cho ...

LỜI CẢM ƠN

Sau quá trình học tập và rèn luyện tại trường Đại học Công Nghệ Thông Tin, Khoa Khoa Học Máy Tính và 4 tháng thực hiện đề tài nghiên cứu này, em xin tỏ lòng cảm ơn chân thành đến các thầy, cô giảng viên, cán bộ các phòng, ban chức năng tại trường đã giúp đỡ em hoàn thành luận văn tốt nghiệp này.

Đặc biệt, em chân thành cảm ơn thầy Mai Tiến Dũng và Lê Đình Duy đã chỉ bảo, hướng dẫn và giúp đỡ em rất nhiều trong suốt quá trình thực hiện đề tài. Một lần nữa em chân thành cảm ơn và chúc các thầy dồi dào sức khoẻ.

Em cũng xin cảm ơn tất cả các bạn bè, anh chị đang học tập và làm việc tại trường, đặc biệt là MMLab đã nhiệt tình giúp đỡ em trong suốt thời gian qua.

Tuy nhiên vì kiến thức chuyên môn còn hạn chế và bản thân còn thiếu nhiều kinh nghiệm thực tiễn nên nội dung của báo cáo không tránh khỏi những thiếu sót, em rất mong nhận sự góp ý, chỉ bảo thêm của quý thầy cô để báo cáo này được hoàn thiện hơn.

TÓM TẮT

Luận văn trình bày quá trình nghiên cứu, thực nghiệm và áp dụng các kiến thức trong lĩnh vực máy học, thị giác máy tính để xây dựng một mô hình nhằm phát triển và cải thiện kết quả bài toán phân lớp bề mặt vật liệu trên những tập dữ liệu mới nhất. Sau quá trình nghiên cứu, nhóm đề xuất và hiện thực một số mô hình kết hợp deep feature và các hancrafted feature thích hợp theo nhiều cách khác nhau để huấn luyện các bộ phân lớp theo nhiều cách khác nhau nhằm tăng độ chính xác cho bài toán, phân tích, đánh giá và giải thích kết quả. Ngoài ra, nhóm còn tiếp tục thực hiện các thực nghiệm để cải thiện hiệu suất tính toán, giảm thiểu chi phí và tăng khả năng thích ứng của các bộ phân lớp đối với dữ liệu mới bằng cách thiết kế một mạng convolutional neural network duy nhất. Cuối cùng nhóm đề ra những hướng phát triển tiếp theo của mô hình.

Mục lục

Mục lục	iv
Danh sách hình vẽ	vi
Nomenclature	viii
1 Mở đầu	1
1.1 Giới thiệu bài toán	1
1.2 Tại sao cần phân lớp vật liệu?	2
1.3 Các thách thức của bài toán	4
2 Các công trình liên quan	5
2.1 Lab-based methods	6
2.2 Image-based methods	8
3 Cơ sở lý thuyết	12
3.1 Support Vector Machines (SVMs)	12
3.1.1 Các khái niệm liên quan	13
3.1.2 Ý tưởng chính của SVM	16
3.2 Neural Network	19
3.3 Convolutional Neural Networks (CNNs)	23
3.4 Transfer Learning	27
4 Phương pháp đề xuất	29
4.1 Mô hình 1: kết hợp probability predictions (posfusion)	30
4.2 Mô hình 2: kết hợp features (pre-fusion)	30

MỤC LỤC

4.3 Mô hình 3: kết hợp mô hình 1 và 2	31
4.4 Thiết kế một mạng CNN duy nhất tích hợp các mô hình trên vào một	31
5 Thực nghiệm và kết quả	37
5.1 Môi trường và công cụ thử nghiệm	37
5.2 Dataset	38
5.3 Evaluation measures	39
5.4 Quá trình thực nghiệm và kết quả	40
6 Kết luận và hướng phát triển	45
6.1 Kết quả đạt được	45
6.2 Hướng phát triển	45
Tài liệu tham khảo	47

Danh sách hình vẽ

1.1	Phân lớp vật liệu cho biết bề mặt trong ảnh thuộc vật liệu nào	1
1.2	Phân lớp vật liệu cho biết chính xác vật liệu của từng pixel trong ảnh	2
1.3	Ba chai đựng nước với hình dáng tương tự nhau được làm từ những vật liệu khác nhau - quyết định những tính chất vật lý khác nhau	3
1.4	Xe tự lái: "Xin lỗi, tôi không biết đó là nước :"	3
1.5	Những đối tượng với chất liệu khác nhau nhưng bề mặt lại có texture giống nhau (đều là sọc ca-rô)	4
1.6	Một ví dụ từ tập dữ liệu Open-Surface cho thấy bài toán này thách thức như thế nào	4
2.1	Các nghiên cứu được công bố tại hội nghị CVPR qua các năm từ 2010 đến nay	6
2.2	Thông tin về chiều sâu được dùng để phân lớp vật liệu trong một nghiên cứu 2017 [1]	7
2.3	Thông tin về sự phản chiếu ánh sáng được dùng để phân lớp vật liệu trong một nghiên cứu 2015 [2]	8
2.4	Các local features (cạnh, <i>micro structures</i> , SIFT - Scale-Invariant Feature Transform) được dùng để phân lớp vật liệu trong một nghiên cứu 2010 [3]	9
2.5	Thông tin về gradient, màu sắc, local binary patterns được dùng để phân lớp vật liệu trong một nghiên cứu 2011 [4]	9
2.6	Two-stream network được dùng trên tập GTOS năm 2017 [5]	10

DANH SÁCH HÌNH VẼ

2.7	Differential Angular Image được tạo thành bằng cách lấy hai ảnh của cùng một mẫu dữ liệu được chụp từ hai góc khác nhau (hai góc này chênh lệch không quá nhiều) trừ nhau [5]	10
2.8	Cấu trúc bên trong của DAIN (Differential Angular Image Network) [5]	11
3.1	SVMs được dùng để phân lớp tập dữ liệu thành hai lớp phân biệt	13
3.2	Magnitude của vector \vec{OA} là độ dài đoạn OA	14
3.3	Magnitude của vector \vec{OA} là độ dài đoạn OA	14
3.4	Hướng của vector u được thể hiện bởi góc mà u tạo với các trục tọa độ trong không gian của nó (Công thức 3.3)	15
3.5	Hai vector x và y	16
3.6	Hyperplane trong không gian hai chiều (đường thẳng) và ba chiều (mặt phẳng)	17
3.7	Dữ liệu được ánh xạ lên không gian cao hơn trước khi các bộ phân lớp được học	18
3.8	Margin được maximize để giảm thiểu phân lớp sai cho dữ liệu mới	20
3.9	Tín hiệu được truyền từ Axons của một neuron đến Dendrites của neuron tiếp theo	21
3.10	Cấu trúc của một Feedforward Perceptron Neural Network đơn giản.	21
3.11	Một trong những cách đơn giản thay đổi weights và thresholds của một node để có kết quả tốt hơn	23
3.12	Cấu trúc mạng LeNet-5, một mạng CNN dùng cho nhận diện chữ viết	24
3.13	Hai filter 3x3 được dùng để tính toán hai 4x4 2D feature map từ một ma trận đầu vào 6x6	25
3.14	Một max pooling layer với filter 2x2 và stride = 2 pixels	26
3.15	ReLU layer sử dụng hàm $f(x) = \max(0, x)$ để thay đổi tất cả các giá trị âm thành 0	27
4.1	Các đối tượng chính của hai ảnh đều có hình dạng giống nhau và có thể bị nhầm lẫn cả hai đều là đá	29
4.2	Các đối tượng có hình dạng khác nhau được làm từ vật liệu khác nhau nhưng lại giống nhau về texture (đều là sọc caro)	30

DANH SÁCH HÌNH VẼ

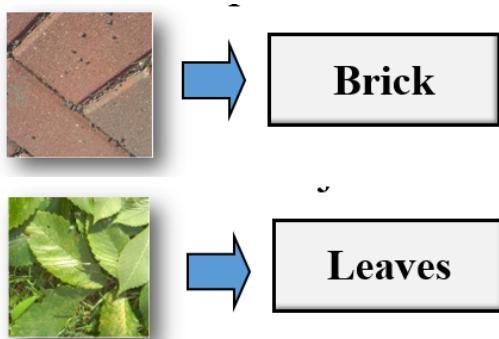
4.3	Deep features rút trích từ một mạng CNN đã được huấn luyện sẵn thể hiện cho cách con người học, kết hợp chúng với các handcrafted features thích hợp có thể giúp cải thiện kết quả phân lớp	33
4.4	Mô hình 1: kết hợp probability predictions	34
4.5	Mô hình 2: kết hợp features (pre-fusion)	35
4.6	Mô hình 3: kết hợp mô hình 1 và 2 (full-fusion)	36
5.1	Một cảnh ngoài trời được dùng để lấy các mẫu cho tập dữ liệu GTOS với nhiều góc nhìn nhìn, điều kiện chiếu sáng khác nhau [5]	39
5.2	Một tập ảnh từ FMD, đảm bảo sự đa dạng về điều kiện chiếu sáng, bố cục, màu sắc, kết cấu	39
5.3	Ảnh từ FMD (bên trái) trong lớp vật liệu "giấy" lại có background là một mặt đường khiến thông tin về texture không còn sự hiệu quả, trong khi ảnh từ GTOS là một bề mặt đồng nhất duy nhất (lớp "Gạch")	41
5.4	Normalized confusion matrix trên GTOS	42
5.5	Normalized confusion matrix trên FMD	43
5.6	Cấu trúc mạng VGG16 được nhóm dùng để rút trích đặc trưng ở layer 'fc2' và thực hiện các thử nghiệm ở phần thử nghiệm với VGG16	44

Chương 1

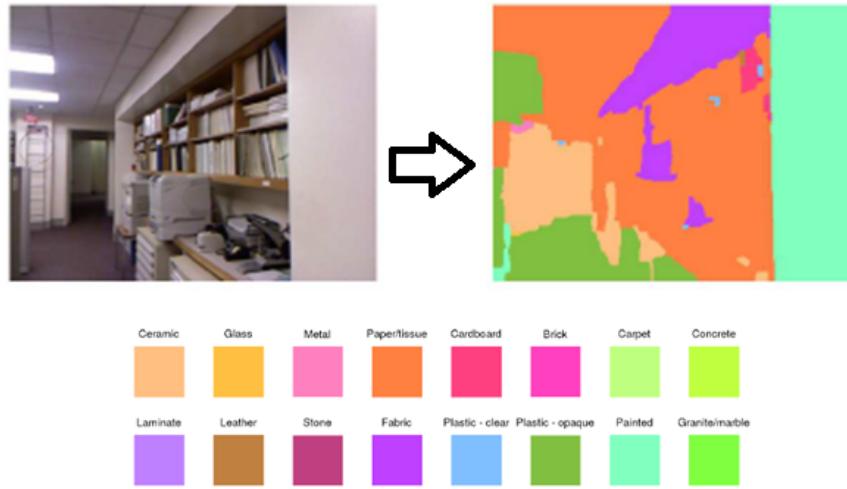
Mở đầu

1.1 Giới thiệu bài toán

Mục tiêu chính của bài toán phân lớp vật liệu là cung cấp thông tin vật liệu càng chi tiết càng tốt của một đối tượng hoặc một bề mặt trong ảnh. Hiểu một cách đơn giản, cho trước một ảnh I, máy tính cần trả lời câu hỏi "Đối tượng (hay bề mặt này) được làm từ vật liệu gì?" (chẳng hạn như gỗ, giấy, đá, kim loại, ...) (Hình 1.1). Ở một cấp độ cao hơn của bài toán, máy tính cần cho biết chính xác loại vật liệu của từng pixel trên ảnh (Hình 1.2). Trong luận văn này, nhóm giải quyết câu hỏi thứ nhất.



Hình 1.1: Phân lớp vật liệu cho biết bề mặt trong ảnh thuộc vật liệu nào



Hình 1.2: Phân lớp vật liệu cho biết chính xác vật liệu của từng pixel trong ảnh

1.2 Tại sao cần phân lớp vật liệu?

Vật liệu của một bề mặt hay đối tượng nào đó là một thông tin rất giá trị để máy tính của thợ hiểu được các thuộc tính của nó và từ đó có thể đưa ra các quyết định liên quan cũng như tương tác với chúng. Hình 1.3 và 1.4 là hai ví dụ đơn giản cho thấy máy tính có thể làm được rất nhiều việc khi có thể biết được thông tin về vật liệu. Trong hình 1.3, với thông tin về vật liệu của các chai nước này, đơn giản nhất máy tính có thể sắp xếp chúng theo cân nặng, ngoài ra còn có thể quyết định chai nào có thể được dùng để đựng nước nóng chai nào không thể hay thậm chí chai nào có thể sử dụng để gây sát thương cho người khác (chai thủy tinh)

Ngoài ra, vật liệu còn là một chìa khóa quan trọng để cải tiến bài toán "Scene understanding" trong Thị Giác Máy Tính [6] và còn được ứng dụng trong nhiều lĩnh vực khác nhau trong đời sống như các hệ thống xe tự lái (Advanced Driver-Assistance Systems) [7], Robotic Manipulation [8] hay Robotic Navigation [9].



Hình 1.3: Ba chai đựng nước với hình dáng tương tự nhau được làm từ những vật liệu khác nhau - quyết định những tính chất vật lý khác nhau



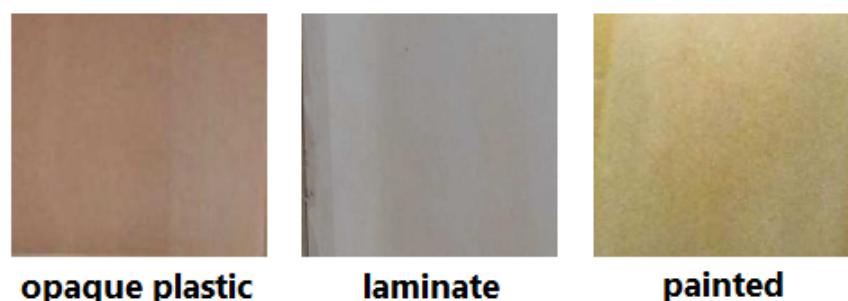
Hình 1.4: Xe tự lái: "Xin lỗi, tôi không biết đó là nước :("

1.3 Các thách thức của bài toán

Nhiều thách thức khác nhau kết hợp làm bài toán phân lớp vật liệu rất khó giải quyết triệt để. Sự đa dạng về hình dáng, kích thước và texture là một trong số đó, có rất nhiều đối tượng có hình dáng khác nhau nhưng lại cùng loại vật liệu, ngược lại có những đối tượng trông có vẽ rất giống nhau nhưng lại làm từ những vật liệu khác nhau. Ngoài ra, điều kiện chiếu sáng khác nhau cũng khiến việc phân biệt giữa các vật liệu trở nên rất khó khăn (đặc biệt đối với việc chỉ dùng một ảnh màu để phân biệt). Bên cạnh đó, sự chồng lấp giữa các đối tượng với nhau, giữa đối tượng với background cũng là một thách thức không nhỏ. Hình ?? và ?? là hai ví dụ cho thấy bài toán này thật sự rất thách thức.



Hình 1.5: Những đối tượng với chất liệu khác nhau nhưng bề mặt lại có texture giống nhau (đều là sọc ca-rô)



Hình 1.6: Một ví dụ từ tập dữ liệu Open-Surface cho thấy bài toán này thách thức như thế nào

Chương 2

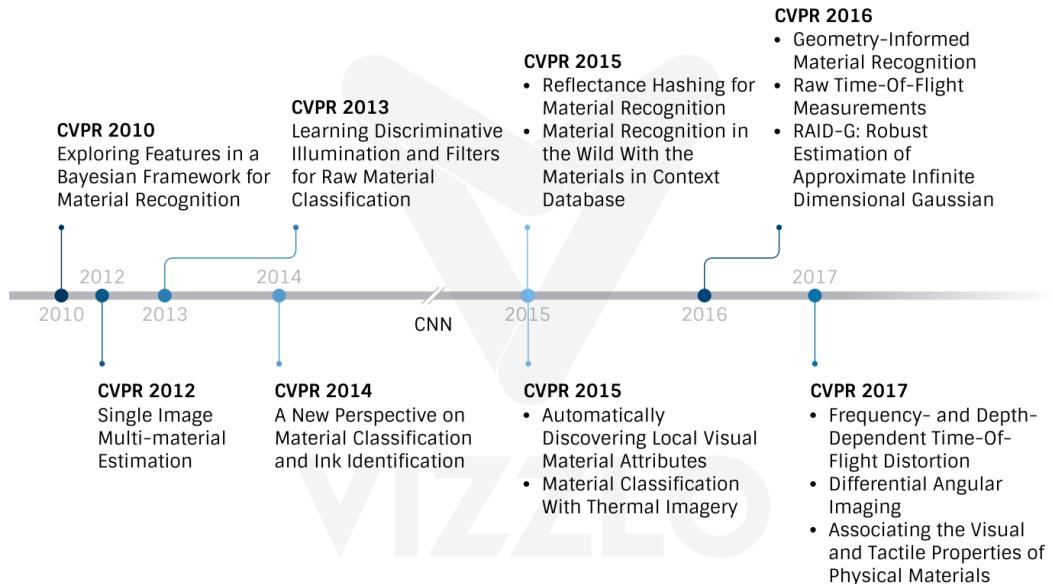
Các công trình liên quan

Từ khi ra đời, bài toán phân lớp vật liệu nhận được sự chú ý của rất nhiều nhóm nghiên cứu. Cùng với sự phát triển đó, rất nhiều tập dữ liệu mới cũng được xây dựng để giải quyết nhiều vấn đề khác nhau của bài toán. Hình 2.1 là tóm tắt các công trình nghiên cứu liên quan được công bố tại hội nghị CVPR (Conference on Computer Vision and Pattern Recognition) từ năm 2010 tới nay. Bên cạnh đó bảng 2.1 là tổng hợp một số tập dữ liệu được sử dụng.

Một điểm chung có thể thấy, sau khi CNN bắt đầu được dùng nhiều trong Thị giác máy tính và chứng minh sự hiệu quả của chúng, số lượng và chất lượng của các nghiên cứu cũng tăng lên đáng kể (kết quả phân lớp rất chính xác, gần như bằng với khả năng đoán của con người), các tập dữ liệu được dùng cũng lớn hơn, đa dạng hơn.

Name	Samples	Classes	Views	Illumination	In scene	Scene image	Camera parameters	Year
CURET [10]	61	61	205	205	No	No	No	1999
KTH-TIPS [11]	11	11	27	3	No	No	No	2004
UBO2014 [12]	84	7	151	151	No	No	No	2014
Reflectance disk [2]	190	19	3	3	No	No	Yes	2015
4D Light-field [13]	1200	12	1	1	Yes	No	No	2016
NISAR [14]	100	100	9	12	No	No	No	2016
GTOS [5]	606	40	19	4	Yes	Yes	Yes	2016

Bảng 2.1: Một số tập dữ liệu được dùng cho phân lớp dữ liệu [5]



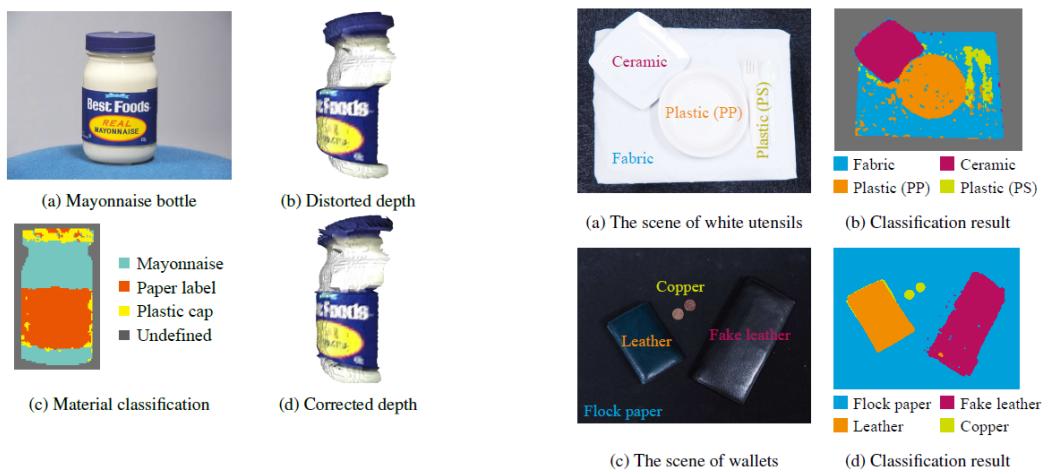
Hình 2.1: Các nghiên cứu được công bố tại hội nghị CVPR qua các năm từ 2010 đến nay

Có rất nhiều phương pháp khác nhau để giải quyết bài toán phân lớp vật liệu. Các phương pháp này có thể được chia thành hai nhóm chính: **Lab-based** và **Image-based**

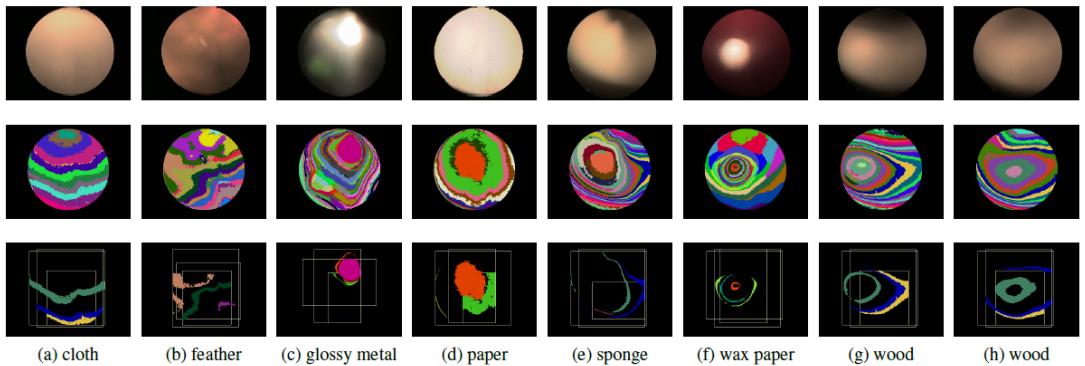
2.1 Lab-based methods

Đây là các phương pháp sử dụng chính các thông tin vật lý của đối tượng để phân lớp. Các thông tin vật lý có thể là độ đàn hồi (elasticity) [15], sự thấm nước (water permeation) [16], phản ứng với ánh sáng (optical response) hay độ phản chiếu (reflectance) [1]. Đây là các thông tin rất giá trị, vì vậy độ chính xác của các phương pháp này thường rất cao. Tuy nhiên quá trình xây dựng các bộ dữ liệu để thực hiện các phương pháp này rất kỳ công, đòi hỏi các thiết bị đặc biệt và một môi trường lý tưởng (vì vậy nên chúng có tên lab-based). Ví dụ, GelFabric [17] - tập dữ liệu ra đời năm 2016 chứa thông tin về chiều sâu và chỉ dùng để phân biệt các loại vải khác nhau; Ground Terrain in Outdoor Scenes (GTOS) [5] - năm

2016 dùng để lấy thông tin về các góc nhìn khác nhau cho cùng một mẫu trong dữ liệu; Reflectance Disk Database [2] - năm 2015 với thông tin về sự phản chiếu của bề mặt. Thêm vào đó, các phương pháp khác nhau trong nhóm này thường đòi hỏi các thông tin vật lý khác nhau, chính vì thế tập dữ liệu của phương pháp khác nhau thường không thể dùng lại được. Ví dụ, một phương pháp dùng độ đàm hồi làm feature chính để phân lớp thì không thể dùng một tập dữ liệu chỉ có thông tin về độ thấm nước được. Vì vậy, các phương pháp đạt kết quả rất cao trên tập dữ liệu này có thể dễ dàng thất bại trên tập dữ liệu khác (hoặc không thể sử dụng trên dữ liệu khác vì thiếu thông tin). Hình 2.2 và 2.3 là hai ví dụ được lấy từ hai nghiên cứu năm 2017 và 2015 sử dụng các thông tin vật lý khác nhau (thông tin về chiều sâu, độ phản chiếu ánh sáng) để phân biệt các loại vật liệu. Cả hai đều đạt kết quả khá tốt trên những tập dữ liệu có đầy đủ thông tin trên [1] [2]



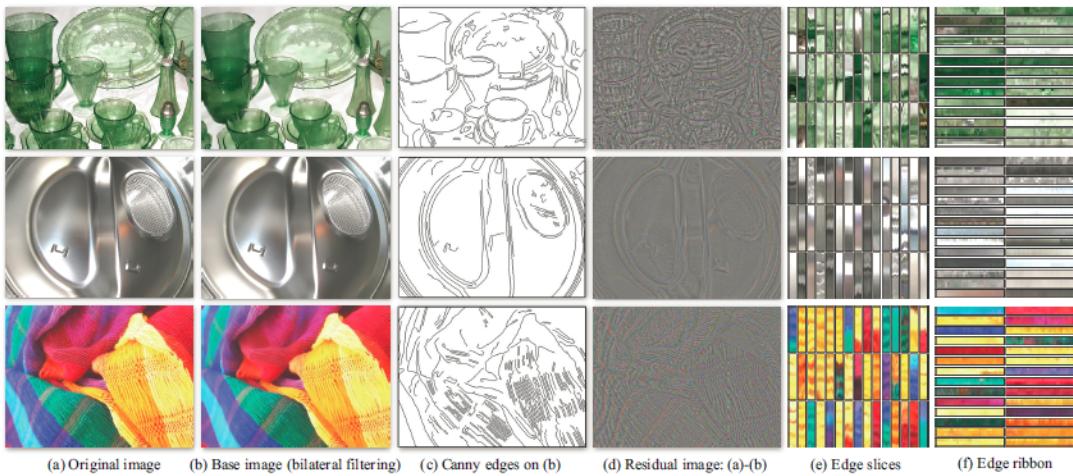
Hình 2.2: Thông tin về chiều sâu được dùng để phân lớp vật liệu trong một nghiên cứu 2017 [1]



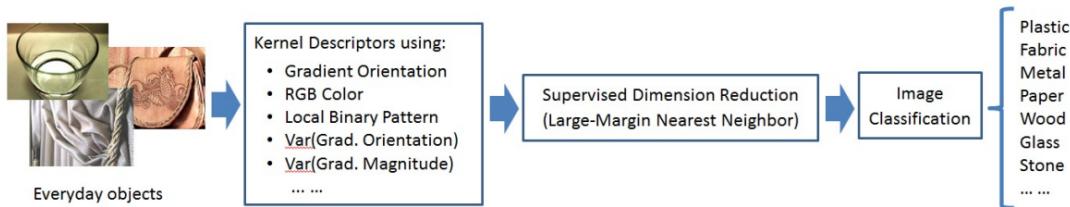
Hình 2.3: Thông tin về sự phản chiếu ánh sáng được dùng để phân lớp vật liệu trong một nghiên cứu 2015 [2]

2.2 Image-based methods

Ngược lại với các phương pháp Lab-base, Image-base là các phương pháp chỉ sử dụng ảnh màu RGB để phân lớp. Vì thế các tập dữ liệu cũng dễ dàng thu thập hơn, và các phương pháp này nếu có độ chính xác cao thì rất dễ áp dụng vào thực tế. Dựa trên các thông tin trực quan (chẳng hạn như màu sắc) thông qua ảnh và thường dựa trên bài toán phân loại đối tượng để tìm ra vật liệu của ảnh (vì các phương pháp hiện tại dùng nhiều các mạng CNN được huấn luyện sẵn cho bài toán phân loại đối tượng để làm gốc), vấn đề chính của các hệ thống dùng phương pháp này chính là sự thiếu thông tin và vì thế chúng dễ dàng bị "lừa" bởi các ảnh có các đối tượng tương tự nhau nhưng lại khác nhau về chất liệu. Hình 2.4 và 2.5 là hai ví dụ cho các phương pháp thuộc nhóm này được lấy từ các nghiên cứu năm 2010 [3] và 2011 [4]. Trong những năm gần đây với sự phát triển của Deep Learning, các mạng CNN (Convolutional Neuron Network) được dùng thay thế cho các dạng local features này và nhiều nghiên cứu khác sử dụng các mạng CNN đã huấn luyện cho bài toán phân lớp đối tượng và đạt kết quả khá cao, tuy nhiên vẫn còn một số hạn chế như đã nêu bên trên. Nhóm tập trung giải quyết các hạn chế này và kế thừa các thành công từ CNN để cải thiện kết quả của bài toán.



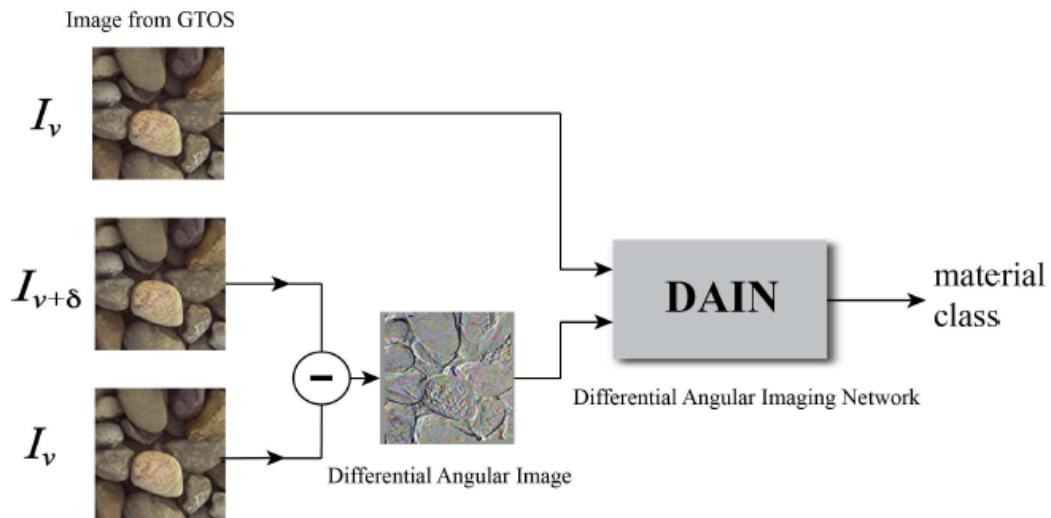
Hình 2.4: Các local features (cạnh, *micro structures*, SIFT - Scale-Invariant Feature Transform) được dùng để phân lớp vật liệu trong một nghiên cứu 2010 [3]



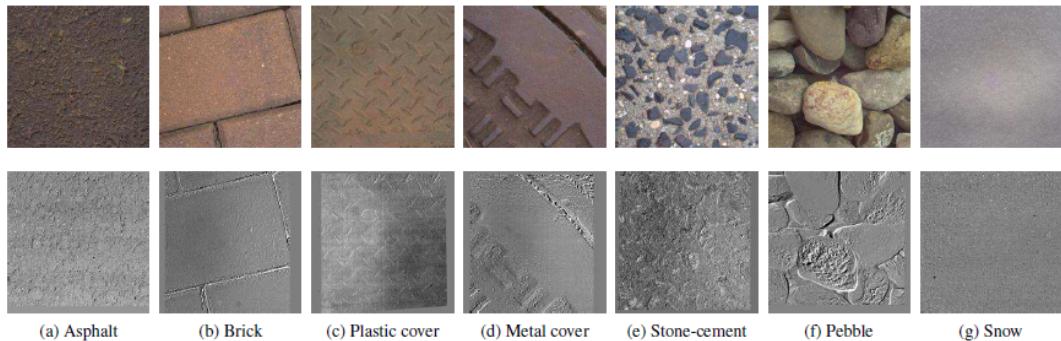
Hình 2.5: Thông tin về gradient, màu sắc, local binary patterns được dùng để phân lớp vật liệu trong một nghiên cứu 2011 [4]

Trong đề tài này, phương pháp mà nhóm sử dụng thuộc loại Image-based vì đối với lab-based, các tập dữ liệu sẽ được xây dựng dựa trên phương pháp (đã có phương pháp X và xây dựng dữ liệu để phục vụ cho phương pháp X), cho nên các thông tin vật lý này thường như đã được sử dụng một cách triệt để nhất và nhóm sẽ có ít cơ hội hơn để cải tiến chúng. Ngoài ra, nhóm còn được thúc đẩy bởi một two-stream network (mạng với 2 nhánh chính) được dùng trên tập GTOS năm 2017 [5]. Hình 2.6 thể hiện ý tưởng chính của nghiên cứu này, sử dụng một hệ thống mạng CNN với hai nhánh đầu vào (một là ảnh gốc, hai là ảnh thể hiện thông tin khác nhau về góc nhìn (view-point) - hay được tác giả gọi là Differential Angular Image). Differential Angular Image được tạo thành bằng cách lấy hai ảnh của cùng một mẫu dữ liệu được chụp từ hai góc khác nhau (hai

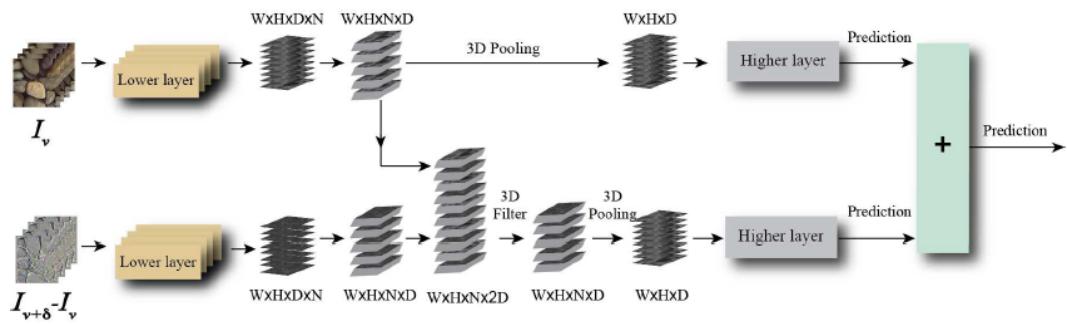
góc này chênh lệch không quá nhiều) trừ nhau (Hình 2.7).



Hình 2.6: Two-stream network được dùng trên tập GTOS năm 2017 [5]



Hình 2.7: Differential Angular Image được tạo thành bằng cách lấy hai ảnh của cùng một mẫu dữ liệu được chụp từ hai góc khác nhau (hai góc này chênh lệch không quá nhiều) trừ nhau [5]



Hình 2.8: Cấu trúc bên trong của DAIN (Differential Angular Image Network) [5]

Chương 3

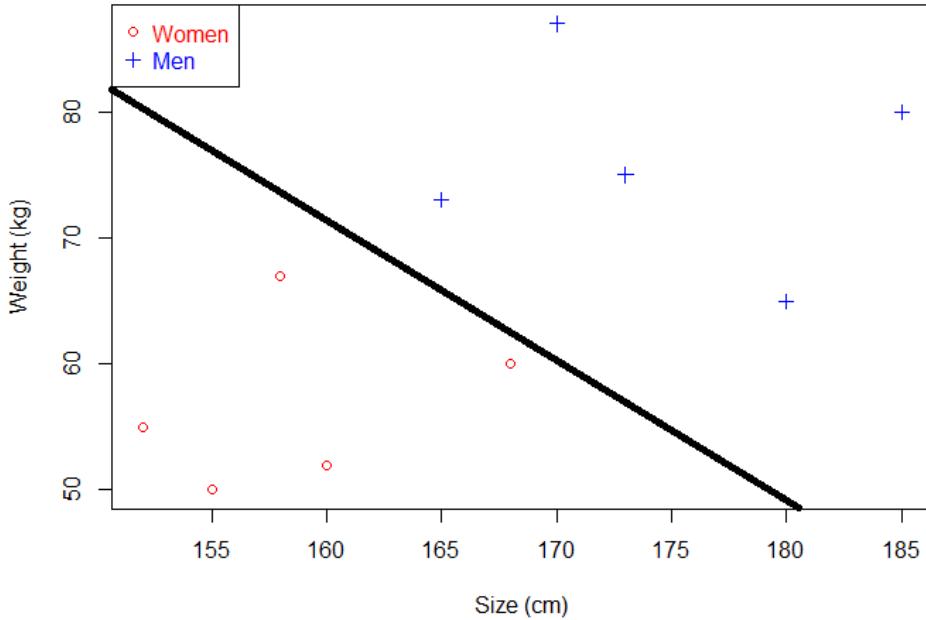
Cơ sở lý thuyết

3.1 Support Vector Machines (SVMs)

Support Vector Machine là một trong những thuật toán học máy có giám sát (supervised learning) hiệu quả nhất. Được xây dựng dựa trên một nền tảng toán học vững chắc, SVM thường cho kết quả tốt trong nhiều bài toán mà không cần điều chỉnh quá nhiều. Tuy nhiên, cũng chính vì điều này mà nó thường được xem như một hộp đen (black box). Có nhiều thuật toán SVM khác nhau nên chúng thường được gọi chung là SVMs. Ở dạng chuẩn, SVM là thuật toán phân lớp nhị phân (binary classification). Từ dữ liệu huấn luyện, SVM xây dựng (learn) một siêu phẳng (hyperplane) để phân lớp (classify) tập dữ liệu thành hai lớp riêng biệt (Hình 3.1).

SVMs là kết quả của một quá trình nghiên cứu xuyên suốt nhiều năm của rất nhiều người. Thuật toán SVM đầu tiên thuộc về Vladimir Vapnik năm 1963. Sau này, ông làm việc với Alexey Chervonenkis về cái được gọi là lý thuyết VC (VC theory - Vapnik–Chervonenkis theory), giải thích quá trình học từ góc nhìn của thống kê, và cả hai đều đóng góp rất nhiều cho SVM. Lịch sử chi tiết về SVM có thể được tìm thấy ở đây (<http://www.svms.org/history.html>).

Trên thực tế, SVMs đã chứng minh được sự hiệu quả của mình trong rất nhiều lĩnh vực khác nhau: phân loại văn bản (text categorization), nhận diện ảnh (image recognition) và bioinformatics (Cristianini & ShaweTaylor, 2000). Trong



Hình 3.1: SVMs được dùng để phân lớp tập dữ liệu thành hai lớp phân biệt

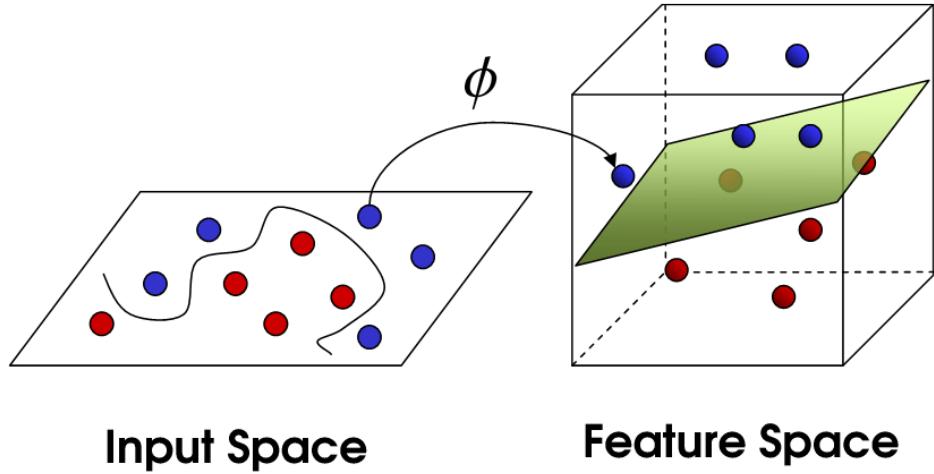
phần này, nhóm sẽ trình bày ý tưởng chính đằng sau và các khái niệm cơ bản được sử dụng trong SVM: vector, linear separability và hyperplanes.

3.1.1 Các khái niệm liên quan

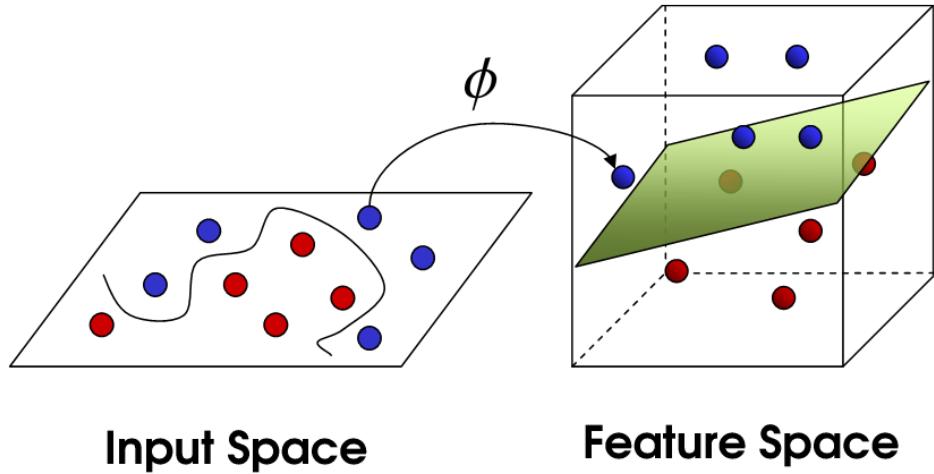
Vector

Vector là một đối tượng toán học có thể được biểu diễn bằng một mũi tên (3.2). Vector được đặc trưng bởi hai thuộc tính chính là: độ dài (**vector magnitude**) và hướng (**vector direction**). Trong SVM, vector được dùng để biểu diễn cho các điểm dữ liệu (feature vector).

Vector magnitude Magnitude của một vector $x = (x_1, \dots, x_n)$ (kí hiệu là $\|x\|$) là độ dài của vector x được tính dựa trên định lý Pythagorean (Hình 3.3). Công thức 3.1 là công thức chung để tính magnitude của vector x được.



Hình 3.2: Magnitude của vector \vec{OA} là độ dài đoạn OA



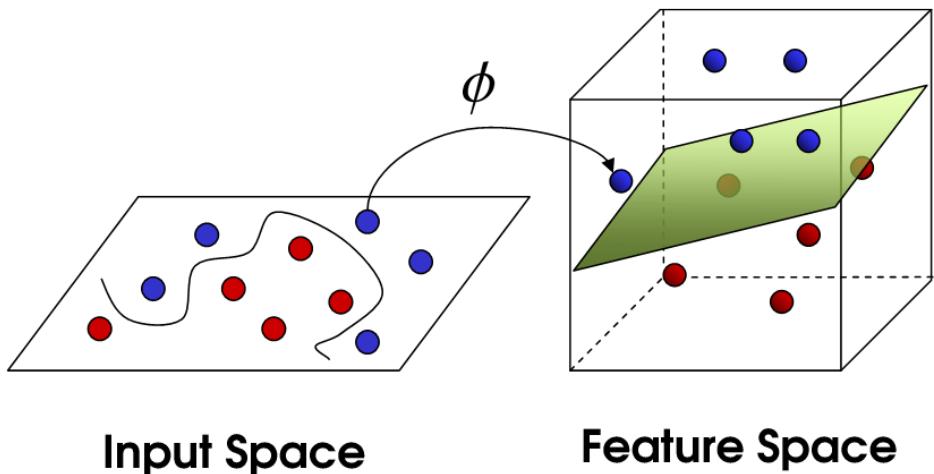
Hình 3.3: Magnitude của vector \vec{OA} là độ dài đoạn OA

$$\|x\| = \sqrt{x_1^2 + \dots + x_n^2} \quad (3.1)$$

Vector direction Hướng của vector $x = (x_1, \dots, x_n)$ là một vector:

$$w = \left(\frac{x_1}{\|x\|}, \dots, \frac{x_n}{\|x\|} \right) \quad (3.2)$$

Về mặt hình học, các giá trị của vector w (w_1, \dots, w_n) chính là giá trị cosin của các góc mà vector x lần lượt tạo với các trục tọa độ trong không gian của nó. Hình 3.5 thể hiện góc của vector x với hai trục tọa độ trong không gian hai chiều.

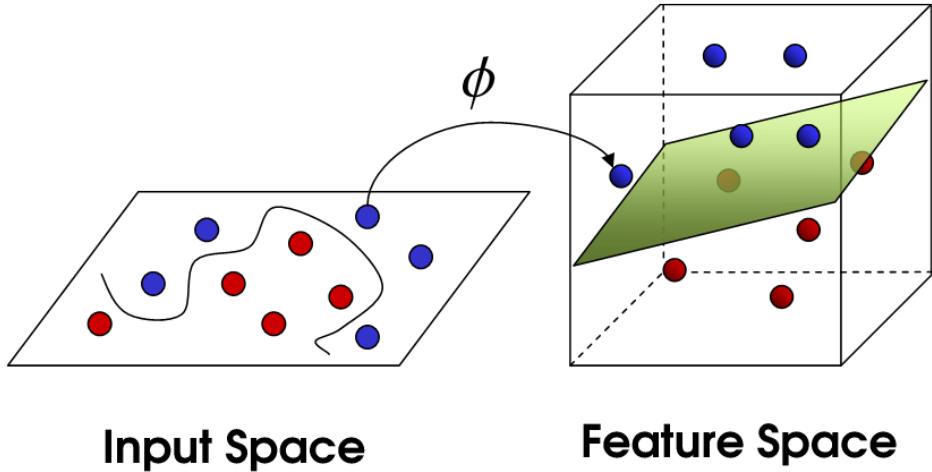


Hình 3.4: Hướng của vector u được thể hiện bởi góc mà u tạo với các trục tọa độ trong không gian của nó (Công thức 3.3)

$$w = \left(\frac{u_1}{\|u\|}, \frac{u_2}{\|u\|} \right) = (\cos(\beta), \cos(\alpha)) \quad (3.3)$$

Tích vô hướng (dot product) Dot product là phép toán phổ biến được thực hiện trên hai vector và cho kết quả là một số thực (còn được gọi là scalar - vì vậy dot product còn có tên khác là scalar product). Dot product thể hiện mối quan hệ của hai vector. Về mặt hình học, dot product là tích của magnitude hai vector và cosin của góc tạo bởi hai vector này (Hình ??, công thức 3.4). Về mặt đại số, dot product có thể định nghĩa bằng công thức 3.5. Dot product là phép toán cơ bản được dùng nhiều trong việc xây dựng SVM, đặc biệt là tìm hyperplane.

$$x \cdot y = \|x\| \|y\| \cos(\theta) \quad (3.4)$$



Hình 3.5: Hai vector x và y

$$x.y = \sum_{i=1}^n (x_i.y_i) \quad (3.5)$$

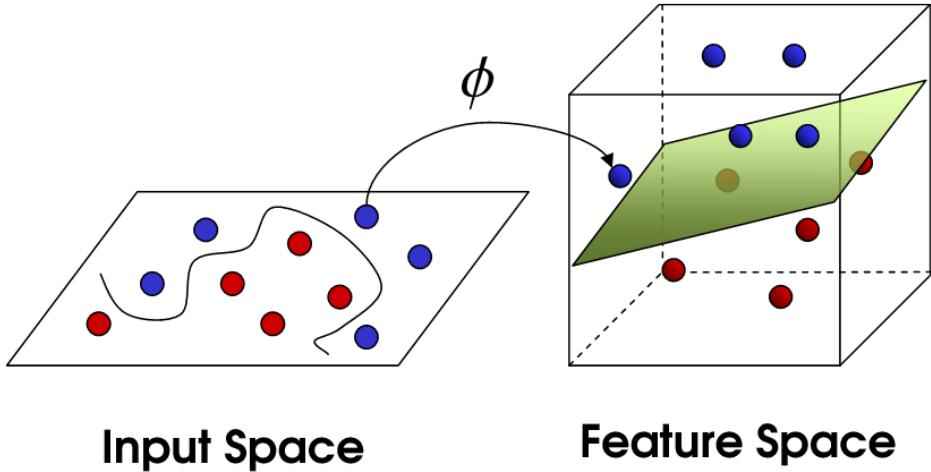
Hyperplane (Siêu phẳng)

Về mặt hình học, có thể hiểu hyperplane là một không gian con có số chiều nhỏ hơn không gian chứa nó. Trong không gian hai chiều, hyperplane là đường thẳng. Trong không gian ba chiều, hyperplane là một mặt phẳng (Hình 3.6). Với phương trình đường thẳng $y = ax + b$, ta có thể viết lại thành $x_2 = ax_1 + b$ hay $ax_1 - x_2 + b = 0$. Nếu đặt hai vector $x = (x_1, x_2)$ và $w = (a, -1)$ thì phương trình đường thẳng trở thành phương trình 3.7. Đây cũng chính là phương trình chung của hyperplane trên không gian bất kỳ.

$$wx + b = 0 \quad (3.6)$$

3.1.2 Ý tưởng chính của SVM

Với bài toán phân lớp, mục tiêu của SVM là tìm một hyperplane "tối ưu" nhất để chia các điểm dữ liệu thành hai phần (binary classification). Để so sánh



Hình 3.6: Hyperplane trong không gian hai chiều (đường thẳng) và ba chiều (mặt phẳng)

giữa hai hyperplane để chọn cái tốt hơn, chúng ta cần một thước đo. Ở đây chúng ta có hai tiêu chí để chọn một hyperplane, một là hyperplane này phải nằm giữa phân tách hai lớp dữ liệu và hai là hyperplane phải cách xa điểm dữ liệu nằm gần nó nhất (hay chính là bài toán maximize margin) (Hình ??). Cho tập dữ liệu $D = \{(x_i, y_i) | x_i \in \mathbb{R}, y_i \in [-1; 1]\}_{i=1}^m$. Chúng ta cần tìm một hyperplane $wx + b = 0$ (hay chính là tìm w và b) sao cho M đạt giá trị lớn nhất với:

$$M = \min_{i=1}^m y_i \left(\frac{wx + b}{\|w\|} \right) \quad (3.7)$$

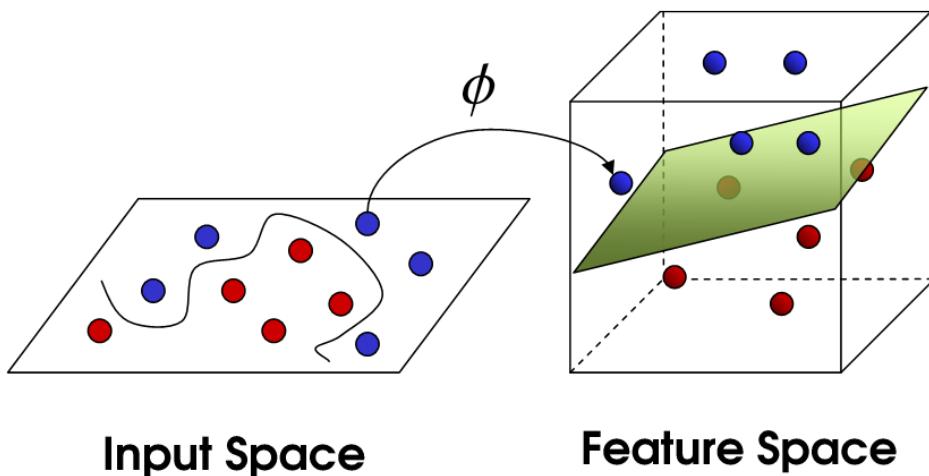
Đây chính là bài toán maximize khoảng cách từ điểm dữ liệu gần hyperplane nhất đến hyperplane (maximize margin). Lưu ý $y_i(\frac{wx+b}{\|w\|})$ sẽ cho kết quả dương nếu điểm dữ liệu được phân lớp đúng và ngược lại âm nếu nó bị phân lớp sai. Với bài toán tối ưu (Optimization Problem) này, để có thể sử dụng các phương pháp có sẵn từ toán học để giải, trước tiên chúng ta cần biến đổi nó một chút để về dạng đơn giản hơn. Bài toán chúng ta cần giải là:

$$\begin{aligned} &\text{Maximize } M \text{ theo } w \text{ và } b \\ &\text{Sao cho } y_i \left(\frac{wx+b}{\|w\|} \right) \geq M, i = 1, \dots, m \end{aligned}$$

Dầu tiên vì ta có thể scale vector w và b một cách thích hợp mà không làm thay đổi phương trình hyperplane (ví dụ: hyperplane $5x_1 + x_2 + 1 = 0$ cũng có thể scale thành $10x_1 + 2x_2 + 2 = 0$), nên ta có thể đặt $F = \min_{i=1}^m y_i(wx + b) = 1$ mà không ảnh hưởng đến kết quả bài toán. Khi đó bài toán được viết lại:

$$\begin{aligned} & \text{Maximize } \frac{1}{\|w\|} \text{ theo } w \text{ và } b \\ & \text{Sao cho } y_i(wx + b) \geq 1, i = 1, \dots, m \end{aligned}$$

Sau khi dùng các phương pháp toán học đã có sẵn để giải quyết bài toán tối ưu này, chúng ta sẽ tìm được hyperplane phân tách các điểm dữ liệu thành hai lớp, với dữ liệu mới ta chỉ cần thay vào phương trình của hyperplane và quyết định xem nó thuộc lớp nào.



Hình 3.7: Dữ liệu được ánh xạ lên không gian cao hơn trước khi các bộ phân lớp được học

Tuy nhiên dữ liệu thực tế rất phức tạp và đan xen lẫn nhau khiến chúng ta không thể tìm được hyperplane nào có thể phân tách chúng trên không gian hiện tại, khi đó một phép ánh xạ sẽ được thực hiện để đưa các điểm dữ liệu này vào không gian với số chiều lớn hơn và hy vọng tìm được một hyperplane ở không gian này. Hình 3.7 là ví dụ cho phép ánh xạ này. Nếu chúng ta xem các chấm màu xanh và đỏ là những quả bóng bóng, trong hình bên trái, các quả bóng này

được đặt trên bàn, nếu chúng phân bố không quá đan xen vào nhau, ta có thể dùng một cây que dài để chia các quả bóng thành hai tập xanh và đỏ mà không động đến các quả bóng. Lúc này, khi đưa một quả bóng mới đặt lên mặt bàn, bằng cách xác định nó nằm bên phía nào của cây que, ta có thể dự đoán màu sắc của quả bóng đó. Các quả bóng tượng trưng cho các điểm dữ liệu, màu xanh và đỏ tượng trưng cho 2 lớp. Cái bàn tượng trưng cho một mặt phẳng. Cây que tượng trưng cho một siêu phẳng đơn giản đó là một đường thẳng. Tuy nhiên khi các quả bóng nằm đan xen nhau trên bàn (dữ liệu phức tạp hơn), lúc này không thể dùng cái que để phân tách chúng, để làm được điều này chúng ta cần hất các quả bóng bay lên (chính là phép ánh xạ), từ đó có thể sử dụng một tờ giấy để phân tách chúng (tờ giấy chính là siêu phẳng). Trên thực tế, SVMs thực hiện ánh xạ bằng việc sử dụng các kernel thích hợp. Việc lựa chọn kernel với các tham số thích hợp sẽ giúp SVM học được các bộ phân lớp tốt hơn trên những tập dữ liệu khác nhau

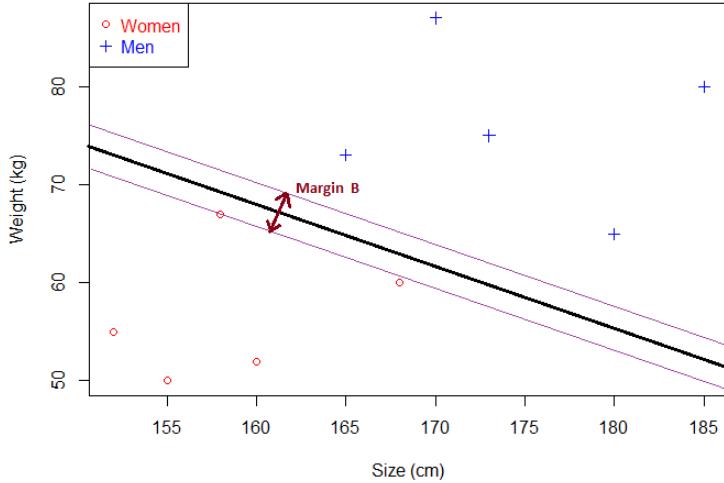
Margin

Margin là một thuật ngữ quan trọng trong SVM chỉ khoảng cách giữa siêu phẳng đến 2 điểm dữ liệu gần nhất tương ứng với các phân lớp (3.8). Để có được bộ phân lớp tốt nhất, SVM cố gắng maximize margin này, từ đó thu được một siêu phẳng tạo khoảng cách xa nhất với hai lớp, nhờ vậy có thể giảm thiểu việc phân lớp sai (misclassification) đối với điểm dữ liệu mới đưa vào.

Trong luận văn này, nhóm sử dụng SVM để huấn luyện các bộ phân lớp dựa trên các đặc trưng rút trích từ ảnh để phân lớp các loại vật liệu.

3.2 Neural Network

Neural Network là một loại hệ thống máy tính đặc biệt [18], được lấy cảm hứng từ cấu trúc não của con người. Một tế bào trong não người được gọi là nơron có hai thành phần chính: Axons và Dendrites. Những tế bào thần kinh này hoạt động bằng cách xử lý các tín hiệu điện. Dựa trên tín hiệu nhận được thông qua Dendrites, Axons thực hiện mã hóa thần kinh (Neural coding) [19], sau đó được

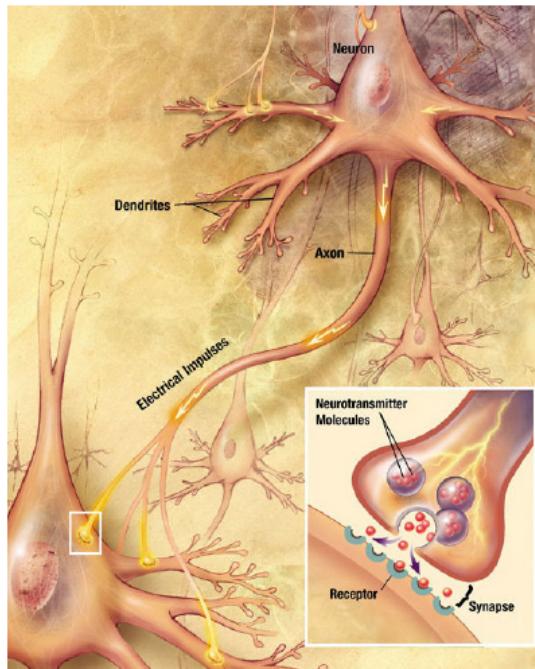


Hình 3.8: Margin được maximize để giảm thiểu phân lớp sai cho dữ liệu mới

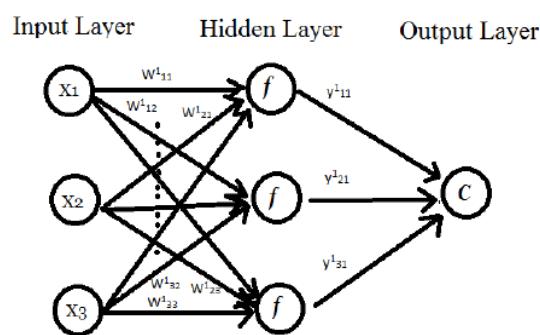
phân phối tín hiệu này qua các Neurons và có thể được coi là đầu ra của hệ thần kinh. Axon của một Neuron được kết nối với Dendrites của Neurons khác, cứ như vậy các Neurons liên kết với nhau và hình thành Neurons Network trong não người (Hình 3.9).

Với các tiến bộ trong khoa học máy tính, hiện nay chúng ta đã có thể phát triển hệ thống mạng hoạt động dựa trên cách lan truyền tín hiệu của các neurons trong hệ thần kinh. Thành phần cơ bản của hệ thống này là các nút (node) tương tác với nhau thông qua các trọng số và hàm lan truyền. Các node tương tự như các neurons trong não. Một mạng neuron là sự kết hợp của nhiều layer của các node. Các layers được chia thành input layer, output layer và intermediate layer (còn được gọi là hidden layer). Đầu ra của mỗi node được tính toán thông qua hàm kích hoạt (activation function). Mỗi node gắn với mỗi giá trị trọng số (weight) và độ lệch (bias) riêng, hai giá trị. Trong nhiều trường hợp, mạng neuron là một hệ thống thích ứng (adaptive system) tự thay đổi cấu trúc của mình (các trọng số) dựa trên các thông tin bên ngoài hay bên trong đi qua mạng trong quá trình học.

Hình 3.10 là cấu trúc của một mạng neural đơn giản. Thuật ngữ "Feedforward"



Hình 3.9: Tín hiệu được truyền từ Axons của một neuron đến Dendrites của neuron tiếp theo



Hình 3.10: Cấu trúc của một Feedforward Perceptron Neural Network đơn giản.

được dùng để chỉ cách dữ liệu đi qua mạng, trong trường hợp này dữ liệu chỉ đi qua mạng một chiều. Layer đầu tiên lấy các thông tin từ đầu vào, được gọi là input layer. Khi các giá trị đầu vào được truyền qua input layer, mỗi node sẽ tính toán giá trị output thông qua weights và biases, sau đó truyền output này tới hidden layer. Tại đây, các nodes dùng activation function để tính toán đầu ra

rồi tiếp tục truyền tới layer cuối. Layer này được gọi là output layer, tính toán tất cả các "score" từ giá trị lan truyền từ layer trước. Với bài toán phân lớp, kết quả cuối cùng chính là lớp với giá trị "score" cao nhất sau khi qua output layer.

Giá trị đầu ra của mỗi node trong mạng được tính toán theo phương trình 3.8. Trong đó, y_j là giá trị đầu ra của node thứ j của layer hiện tại, n là số lượng node của layer trước truyền thông tin vào node hiện tại, f là hàm kích hoạt của node, w_{mn} là trọng số được truyền từ node thứ m của layer trước đến node thứ n của layer hiện tại, b_j là bias của node hiện tại (đang cần tính đầu ra)

$$y_j = f\left(\sum_{i=1}^n w_{ij}x_i + b_j\right) \quad (3.8)$$

Node đầu ra C sẽ nhận giá trị của các y_j sau đó chọn lớp có score lớn nhất là output cuối cùng. Phương trình 3.9, 3.10 và 3.11 là cách tính giá trị đầu ra của từng node trong hidden layer

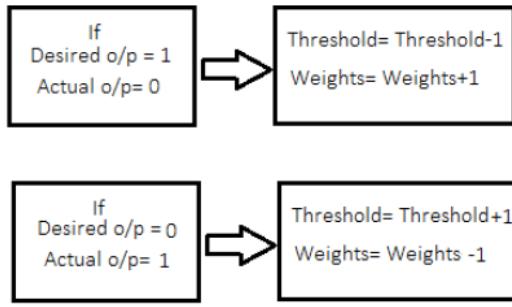
$$y_1 = f(w_{11}x_1 + w_{21}x_2 + w_{31}x_3 + b_1) \quad (3.9)$$

$$y_2 = f(w_{12}x_1 + w_{22}x_2 + w_{32}x_3 + b_2) \quad (3.10)$$

$$y_3 = f(w_{13}x_1 + w_{23}x_2 + w_{33}x_3 + b_3) \quad (3.11)$$

Một hệ thống mạng neural với khả năng học tập hạn chế đã được phát triển trong công trình của Lugar và Stubble [20] và được gọi là Perceptron. Mỗi node sẽ có đầu ra là 0 hoặc 1 được tính toán dựa trên một giá trị ngưỡng đặc biệt. Mạng neural này sẽ được học trên một tập dữ liệu huấn luyện. Mỗi mẫu trong tập huấn luyện gồm các giá trị đầu vào và giá trị đầu ra mong muốn tương ứng với nó. Các giá trị đầu vào được truyền vào input layer, đi qua mạng neural và cho một giá trị đầu ra (ở đây là 0 hoặc 1). Nếu giá trị đầu ra của mạng không giống với giá trị đầu ra mong muốn ban đầu thì giá trị các weights và thresholds sẽ được điều chỉnh để có kết quả tốt hơn. Hình 3.11 là một cách thay đổi weights và thresholds để có kết quả tốt hơn.

Perceptron sau đó được phát triển thành mạng với nhiều hidden layer (multi-layer). Cấu trúc này được học dựa trên thuật toán lan truyền ngược (backpropagation) [18]. Thuật toán này tuân theo quy tắc delta [?? để tính toán độ lỗi tại



Hình 3.11: Một trong những cách đơn giản thay đổi weights và thresholds của một node để có kết quả tốt hơn

nút đầu ra. Độ lỗi của nút cuối cùng trong mạng (output node) được lan truyền ngược qua mạng sau mỗi lần huấn luyện, giá trị các weights sẽ được cập nhật dựa trên hàm lỗi. Cách học này còn được gọi là gradient descent. Giá trị đầu ra của một node phụ thuộc vào hàm kích hoạt và trong mạng lan truyền ngược là một hàm sigmoid (Phương trình ??) với $f_j = \sum_{i=1}^n w_{ij}x_i$

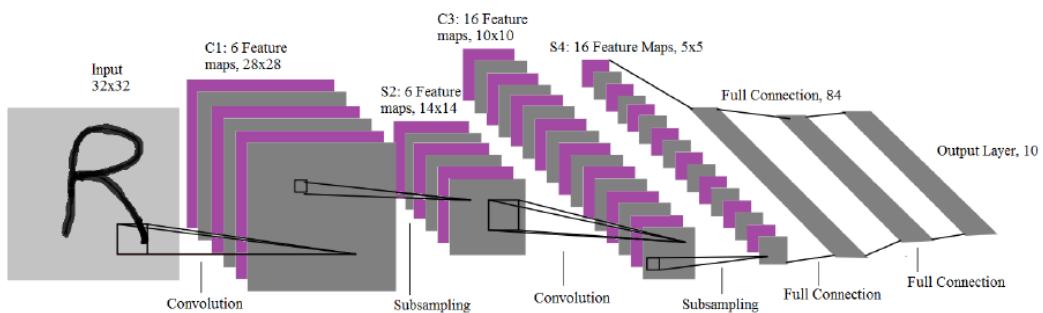
$$\sigma(f_x) = \frac{1}{1 + e^{-f_x}} \quad (3.12)$$

Hàm sigmoid áp dụng cho tất cả các node trừ các input node và giá trị đầu ra được giới hạn trong khoảng [0,1].

3.3 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) là một loại Neural Network đặc biệt. Nguyên tắc hoạt động của CNNs tương tự như một Neural Network bình thường: lấy input tại input layer, truyền giá trị này qua các hidden layer, dùng activation function tính toán giá trị đầu ra tại mỗi node. Score tại các node ở layer cuối (output layer) được lan truyền ngược và các trọng số của mạng được cập nhật sau mỗi lần huấn luyện. Trong thực tế, CNNs khó huấn luyện hơn các Neural Network bình thường do sự phức tạp của nó. Nhưng điều gì khiến CNNs đặc biệt hiệu quả trong các vấn đề của Thị Giác Máy Tính? Hãy cùng tìm hiểu bên dưới.

Các Neural Network bình thường không thích ứng tốt với các ảnh đầu vào khác nhau. Với ảnh đầu vào 32×32 RGB, input layer sẽ có $32 * 32 * 3 = 3072$ weights. Nhưng trên thực tế, kích thước của ảnh đầu vào lớn hơn nhiều, cho một ảnh với kích thước 600×300 RGB, sẽ có tới $600 * 300 * 3 = 540000$ weights chỉ tính input layer - một con số không hề nhỏ! Để khắc phục điều này nhận cả bức ảnh làm đầu vào. Năm 2012, CNNs bắt đầu nhận được sự chú ý của nhiều nhà nghiên cứu trong lĩnh vực phân lớp ảnh khi Alex Krizhevsky thắng giải trong cuộc thi của Imagenet bằng việc thiết kế một mạng CNN và giảm độ lỗi trong việc phân lớp đối tượng từ 26.2% xuống còn 15.3% [21]. CNN tận dụng lợi thế của bất kỳ cấu trúc đầu vào nào thể hiện mối tương quan không gian (chẳng hạn như hình ảnh) và sắp xếp các neural theo không gian này. Sự sắp xếp này cho phép thông tin đi qua mạng hiệu quả và giảm đáng kể số lượng tham số của mạng. Cấu trúc của một mạng CNN gồm nhiều phần khác nhau: Convolutional Layer, Pooling Layer, Rectified Linear Unit hoặc ReLU Layer và Fully Connected Layer. Một trong những kiến trúc CNN cơ bản (LeNet-5) [22] dùng cho nhận diện chữ viết được thể hiện trong hình 3.12



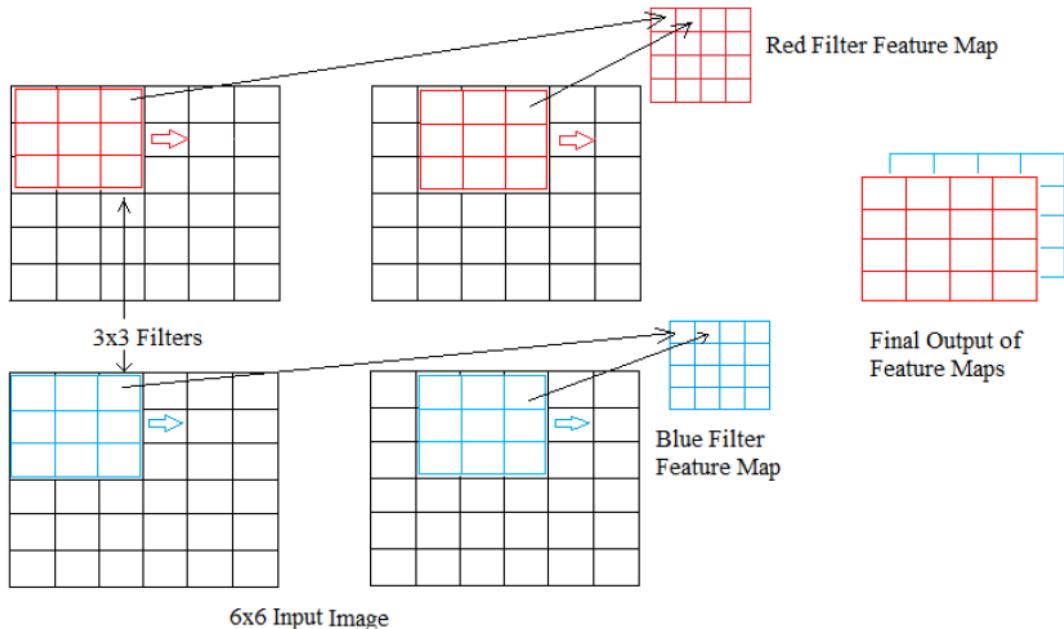
Hình 3.12: Cấu trúc mạng LeNet-5, một mạng CNN dùng cho nhận diện chữ viết

Convolutional layer

Convolutional layer là thành phần chính thực hiện hầu hết các tính toán trong một mạng CNN. Mỗi convolutional layer chứa một tập các filters (hay kernels) với kích thước ứng với các weights và biases. Trong quá trình lan truyền của một

mạng CNN, mỗi kernel trượt qua tất cả các vùng của ảnh đầu vào và thực hiện phép toán dot product trên mỗi vùng đó. Mỗi kernel rút trích các đặc trưng khác nhau từ ảnh. Các phép toán này làm giảm dần số chiều (kích thước) của đầu vào nhưng cũng làm tăng độ sâu của mạng.

Ba tham số quyết định đầu ra của một convolutional layer bao gồm: depth, stride và zero-padding. **Depth** thay đổi đầu ra của convolutional layer dựa vào số lượng kernel của layer đó. Số lượng kernel càng lớn thì đầu ra của layer sẽ càng "sâu". **Stride** là tham số thể hiện bước nhảy của kernel khi trượt trên ảnh để thực hiện dot product. Nếu stride có giá trị là 1 thì kernel sẽ trượt một pixel sau mỗi phép dot product. Bằng cách này, stride sẽ quyết định kích thước của đầu ra. **Zero-padding** cũng ảnh hưởng đến kích thước của đầu ra thông qua việc thêm 0 vào các cạnh của ảnh hay ma trận đầu vào.



Hình 3.13: Hai filter 3x3 được dùng để tính toán hai 4x4 2D feature map từ một ma trận đầu vào 6x6

Hình 3.13 thể hiện quá trình tính toán của một convolutional layer sử dụng 2 filters 3x3 cho ma trận đầu vào 6x6 (không dùng zero-padding). Sau khi các phép toán dot product hoàn thành, kết quả đầu ra của layer này là hai 2D feature map

4x4 (hay có thể nói đầu ra có kích thước 4x4x2). Đây chính là ví dụ cho việc giảm kích thước đầu vào nhưng tăng chiều sâu của layer trong mạng CNN. Một cách tổng quát, với ma trận đầu vào kích thước NxN đi qua một convolutional layer với Y filters MxM (stride = 1 và không dùng zero-padding) thì đầu ra có kích thước $(N - M + 1) \times (N - M + 1) \times Y$

Pooling Layer

Là một loại layer quan trọng khác trong CNN, pooling layer thường được dùng ở giữa hai convolutional layers. Pooling layer được dùng để giảm kích thước của ma trận đầu vào (vì vậy nó còn có tên khác là downsampling layer). Kiểu pooling layer phổ biến nhất là max pooling layer, thường sử dụng với filter có kích thước 2x2 và stride = 2 pixels. Hình 3.14 cho thấy tác dụng của layer này, giảm kích thước ma trận đầu vào từ 4x4 xuống còn 2x2.

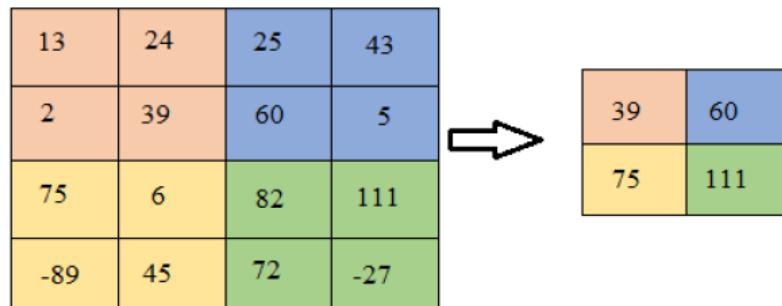


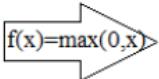
Fig. 1.6. Example of a Max Pooling Layer with 2x2 filter with stride of 2 pixels.

Hình 3.14: Một max pooling layer với filter 2x2 và stride = 2 pixels

Những phương pháp gần đây thường không dùng pooling layer ở giữa hai convolutional layer. Thay vào đó, để đạt được tác dụng tương tự pooling layer (giảm kích thước), stride của convolutional layer thường được dùng với giá trị cao hơn ??.

ReLU (Rectified Linear Unit) Layer

Rectified Linear Unit hay ReLU layer thường được dùng sau convolutional layer và trước pooling layer. Thực nghiệm cho thấy layer này được dùng để tăng hiệu quả tính toán của mạng CNN [16]. ReLU layer không thay đổi kích thước của ma trận đầu vào. Trong ví dụ Hình 3.15, layer này sử dụng hàm $f(x) = \max(0, x)$ để thay đổi tất cả các giá trị âm thành 0.



15	-25	7	49	65
33	62	-102	55	18
75	11	-17	-25	78
34	41	93	24	5
-62	9	37	-71	19

15	0	7	49	65
33	62	0	55	18
75	11	0	0	78
34	41	93	24	5
0	9	37	0	19

Hình 3.15: ReLU layer sử dụng hàm $f(x) = \max(0, x)$ để thay đổi tất cả các giá trị âm thành 0

Fully Connected Layer

Fully connected layer được dùng như layer cuối cùng của một mạng CNN (output layer). Tại đây, mỗi node trong layer trước liên kết với tất cả các node của layer tiếp theo, Layer này được dùng để chuyển đổi feature map từ layer trước thành classification score.

3.4 Transfer Learning

Với bài toán phân lớp ảnh, để đạt kết quả tốt đòi hỏi một cấu trúc mạng thích hợp, điều này phụ thuộc nhiều vào tập dữ liệu huấn luyện. Trong thực tế, dữ liệu huấn luyện rất lớn và đa dạng. Một cấu trúc mạng tốt có thể học được khối lượng dữ liệu lớn thường rất phức tạp và đòi hỏi khá nhiều về phần cứng để thực hiện huấn luyện. Alexnet [21] sử dụng hơn 1.2 triệu ảnh thuộc 1000 lớp khác nhau để huấn luyện trên hai GTX 580 3GB GPUs trong gần một tuần.

Với các bài toán trên tập dữ liệu khác, kỹ thuật transfer learning được dùng để sử dụng lại các mạng phức tạp đã được huấn luyện sẵn này, và chỉ huấn luyện lại một phần nhỏ để thích hợp với bài toán mới. Cách tiếp cận này đã được được rất nhiều kết quả tốt trong nhiều bài toán phân lớp trên những tập dữ liệu khác nhau.

Một cách đơn giản nhưng hiệu quả để áp dụng kỹ thuật này là sử dụng lại toàn bộ một mạng CNN được train trên tập dữ liệu lớn, thay layer cuối cùng bằng một Softmax Regression nhưng với số lượng units bằng với số lượng class ở bộ cơ sở dữ liệu mới và chỉ huấn luyện lại tham số cho layer này để giảm thiểu chi phí tính toán và thời gian

Ngoài ra, chúng ta cũng có thể sử dụng feature map ở những layer gần cuối (fully connected layers) rút ra từ các mạng được huấn luyện sẵn như một feature (deep feature) và sử dụng các thuật toán máy học như SVMs (Support Vector Machines) để huấn luyện các bộ phân lớp cho bài toán mới.

Chương 4

Phương pháp đề xuất

Như đã đề cập bên trên, các phương pháp phân lớp vật liệu dựa trên ảnh RGB thường dựa trên việc phân lớp đối tượng để phân loại vật liệu của chúng, điều này khiến những hệ thống này dễ phân loại sai khi những đối tượng có hình dạng hoặc texture tương tự nhau (Hình 4.1 và 4.2). Nhóm đề xuất phương pháp kết hợp deep feature rút trích từ một CNN được huấn luyện sẵn (transfer learning) với các handcrafted features thích hợp trên những tập dữ liệu khác nhau (thể hiện thông tin về hình dạng và texture) để giải quyết sự nhầm lẫn này và cải thiện kết quả phân lớp (Hình 4.3).



Hình 4.1: Các đối tượng chính của hai ảnh đều có hình dạng giống nhau và có thể bị nhầm lẫn cả hai đều là đá



Hình 4.2: Các đối tượng có hình dạng khác nhau được làm từ vật liệu khác nhau nhưng lại giống nhau về texture (đều là sọc caro)

4.1 Mô hình 1: kết hợp probability predictions (posfusion)

Cấu trúc của mô hình này gồm 3 nhánh song song nhau. Nhánh thứ nhất nhận ảnh đầu vào sau đó dùng một mạng CNN được huấn luyện sẵn để rút deep feature và huấn luyện bộ phân lớp thứ 1. Nhánh thứ 2 và thứ 3 lần lượt nhận cùng ảnh đầu vào đó và dùng cái bộ lọc tương ứng để rút các thông tin về hình dạng và texture trong ảnh (handcrafted features) sau đó được dùng để huấn luyện bộ phân lớp thứ 2 và 3.

Sau khi 3 bộ phân lớp này đã được huấn luyện, chúng được dùng để phân lớp các ảnh trong tập test và cho ra 3 kết quả khác nhau (3 vector of scores), 3 kết quả này sau đó được kết hợp với nhau để cho ra kết quả cuối cùng (Hình 4.4).

4.2 Mô hình 2: kết hợp features (pre-fusion)

Với mô hình này, đầu vào cũng có ba nhánh tương tự mô hình thứ nhất, tuy nhiên chúng sẽ không đi song song, thay vào đó các feature rút trích từ ba nhánh sẽ được kết hợp thành một và chỉ huấn luyện một bộ phân lớp duy nhất trên feature đã được kết hợp này.

Trong quá trình test, các feature từ 3 nhánh cũng được kết hợp tương tự và sau đó đi qua bộ phân lớp đã huấn luyện để có kết quả cuối cùng (Hình 4.5).

4.3 Mô hình 3: kết hợp mô hình 1 và 2

Mô hình này là sự kết hợp của pre-fusion và post fusion bên trên để có thể cho kết quả tốt nhất. Đầu tiên, ảnh huấn luyện sẽ được rút deep feature và các handcrafted features tương tự như trên, sau đó deep feature sẽ dùng để huấn luyện một bộ phân lớp riêng, cùng với đó, deep feature cũng sẽ được kết hợp với các handcrafted features còn lại để huấn luyện một bộ phân lớp riêng (Hình 4.6).

Với mô hình này, nhánh trên (chỉ dùng deep feature) có thể coi là transfer learning từ một mạng CNN đã được huấn luyện sẵn trên ImageNet và tương ứng với cách dùng bài toán phân lớp đối tượng để giải bài toán phân lớp vật liệu, cùng với đó nhánh 2 kết hợp với các handcrafted features thích hợp sẽ giải quyết được các trường hợp nhập nhằng khi hai đối tượng có cùng hình dạng hay texture (Hình 4.1, 4.2), chính vì thế kết quả phân lớp sẽ tốt hơn.

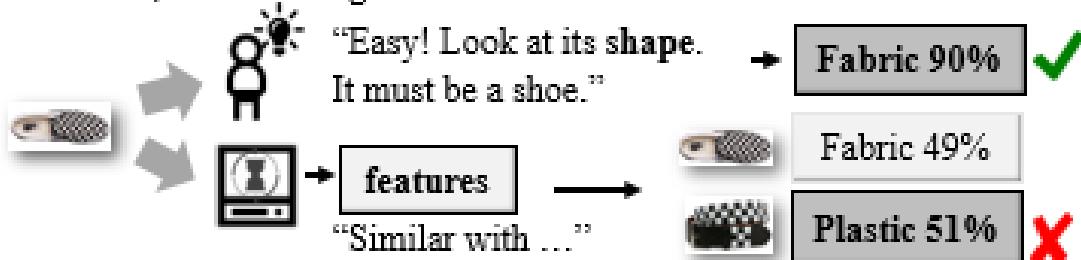
4.4 Thiết kế một mạng CNN duy nhất tích hợp các mô hình trên vào một

Cách rút trích features và huấn luyện bằng SVMs như các mô hình trên cần lưu trữ một lượng lớn dữ liệu (các feature đã được rút trích). Bên cạnh đó quá trình huấn luyện (bằng SVMs) mất rất nhiều thời gian, tuy nhiên để có một kết quả tốt nhất nhóm cần phải thử nghiệm với rất nhiều bộ tham số khác nhau (của SVMs) nên thời gian và khối lượng tính toán không hề nhỏ, thêm vào đó các bộ phân lớp sau khi đã huấn luyện không thể dùng lại trong lần huấn luyện sau. Thế nên, nhóm đã đề xuất phương pháp kết hợp các mô hình trên vào một mạng CNN duy nhất. Mục tiêu được đề ra là với mạng CNN nhóm sẽ có thể:

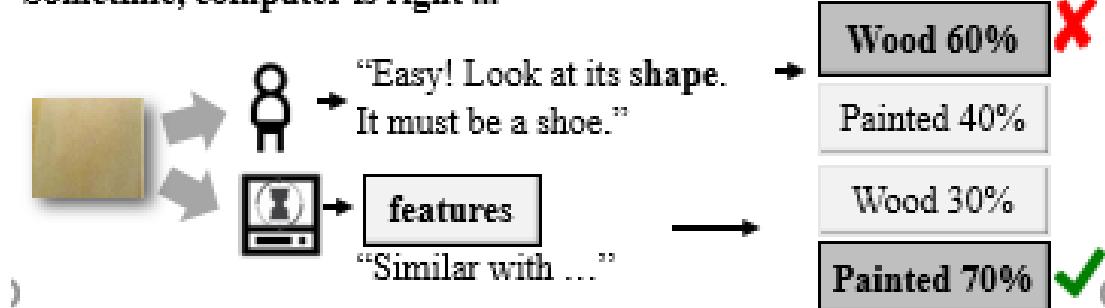
1. Gộp cả 3 quá trình rút trích feature, huấn luyện và test vào một quá trình duy nhất là huấn luyện mạng CNN này.
2. Giảm chi phí (không gian lưu trữ, khối lượng và thời gian tính toán)
3. Khả năng mở rộng và sử dụng lại: mạng CNN có khả năng tự thay đổi tham số để thích ứng tốt hơn với dữ liệu mới (các bộ phân lớp của SVMs thì không thể).

Ý tưởng chung của mạng được trình bày trong Hình ??

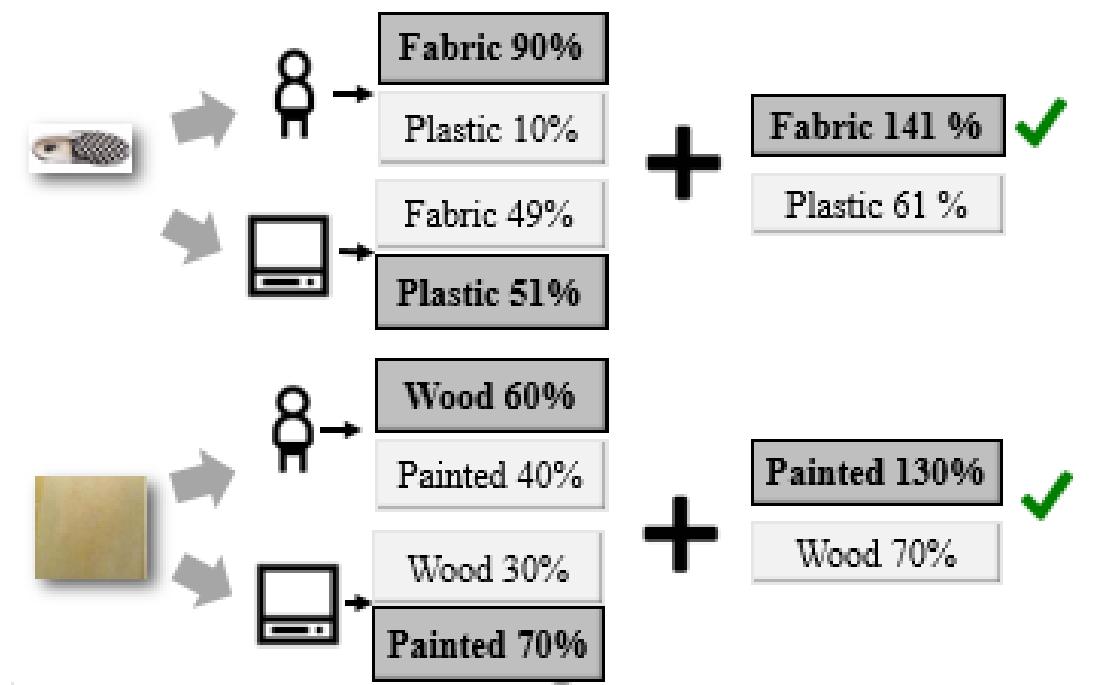
Sometime, human is right ...



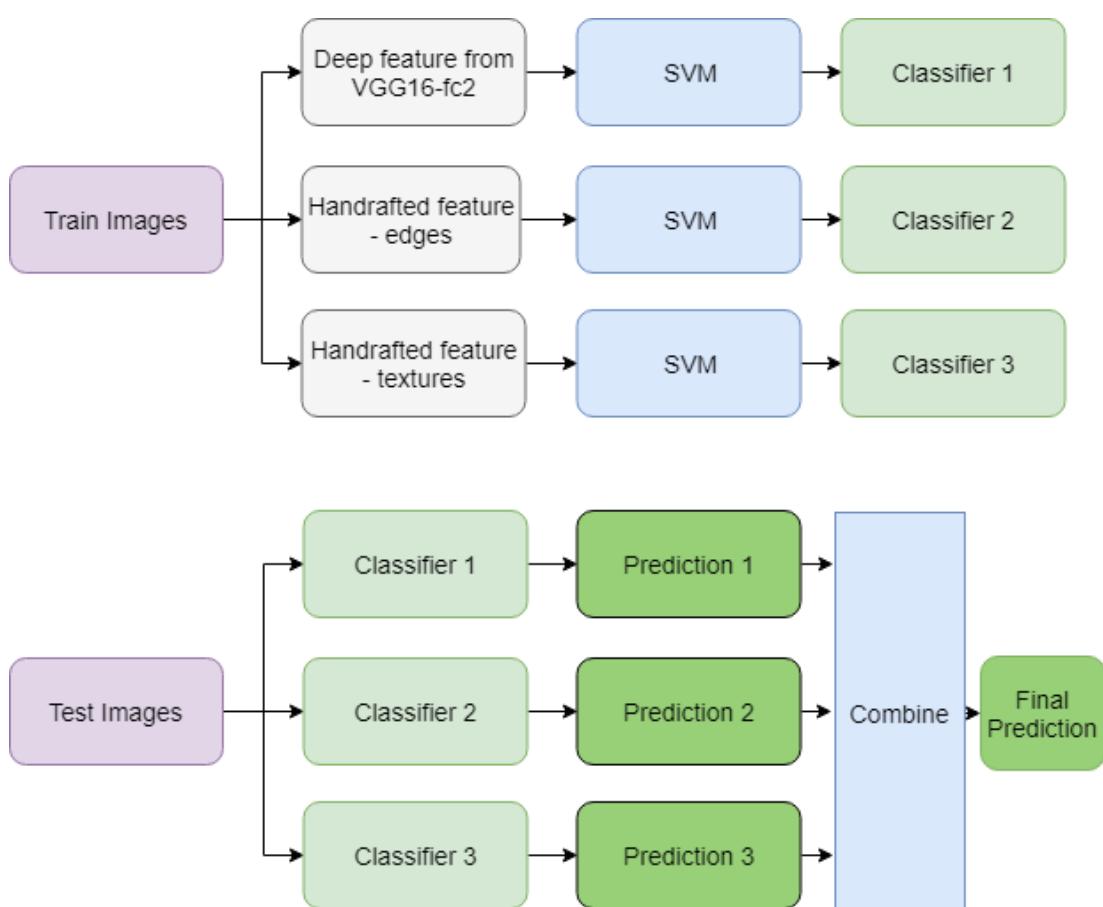
Sometime, computer is right ...



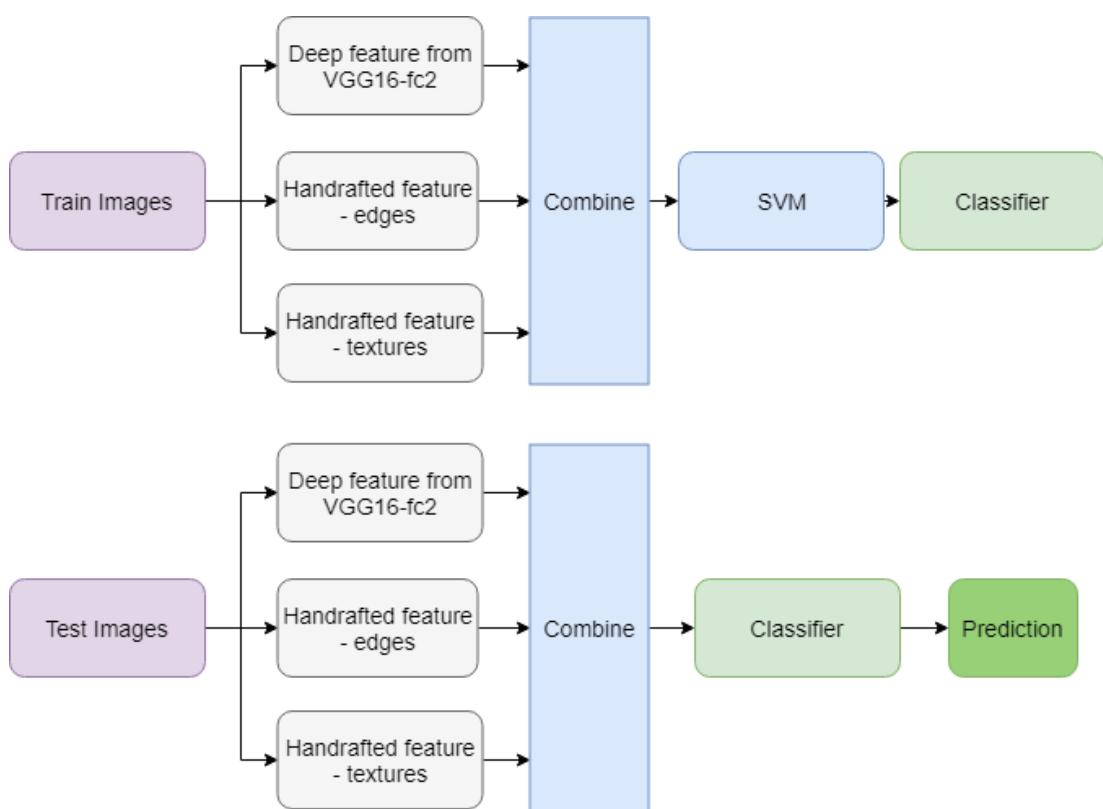
So, why don't we combine it?



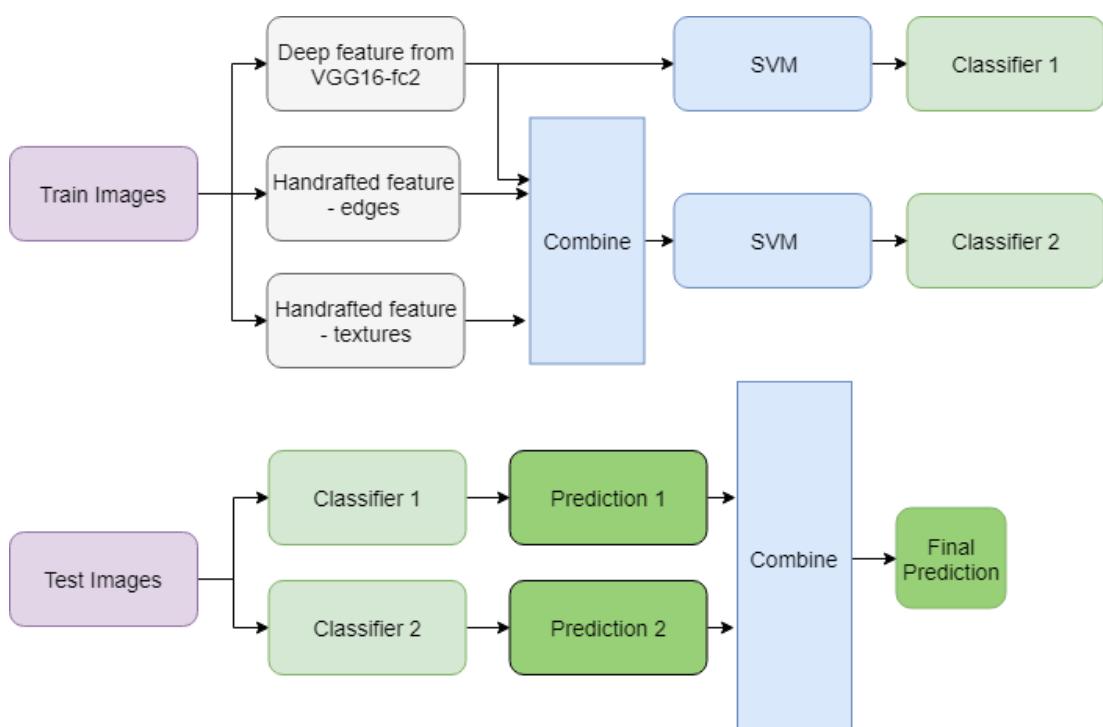
Hình 4.3: Deep features rút trích từ một mạng CNN đã được huấn luyện sẵn thể hiện cho cách con người học, kết hợp chúng với các handcrafted features thích hợp có thể giúp cải thiện kết quả phân lớp



Hình 4.4: Mô hình 1: kết hợp probability predictions



Hình 4.5: Mô hình 2: kết hợp features (pre-fusion)



Hình 4.6: Mô hình 3: kết hợp mô hình 1 và 2 (full-fusion)

Chương 5

Thực nghiệm và kết quả

Nhóm thực hiện thực nghiệm trên hai tập dữ liệu Ground Terrain in Outdoor Scenes (GTOS) [5] và Flicker Material Dataset (FMD) [23], sau đó so sánh kết quả với các phương pháp state-of-the-art để xem xét sự ảnh hưởng của các handcrafted features trên những dataset khác nhau. Trên mỗi tập dữ liệu, nhóm đầu tiên thực nghiệm với từng cấu trúc mạng nhóm đề xuất trong phần trên và đánh giá sự hiểu quả của chúng, sau đó so sánh kết quả khi sử dụng các handcrafted feature khác nhau để chỉ ra rằng việc kết hợp handcrafted features càng thích hợp với tập dữ liệu, kết quả sẽ càng tốt. Sau cùng, để có thể thiết kế một CNN có thể tích hợp các phần trên lại với nhau, nhóm thực hiện một số thực nghiệm nhằm hiểu rõ hơn cấu trúc, cách hoạt động của từng layer của một mạng CNN phổ biến và hiểu quả - VGG16 ??, từ đó có thể dựa trên cấu trúc của mạng này cũng như chọn được những layer thích hợp nhất để xây dựng một CNN riêng.

5.1 Môi trường và công cụ thử nghiệm

Các thực nghiệm trong luận văn này được thực hiện trên PC có cấu hình:

1. Processor: Intel(R) Core(TM) i7 - 6200U CPU @ 2.40GHz 3.2GHz
2. RAM: 16GB
3. : Opera system: Windows 10 (x64)

Các công cụ được sử dụng bao gồm:

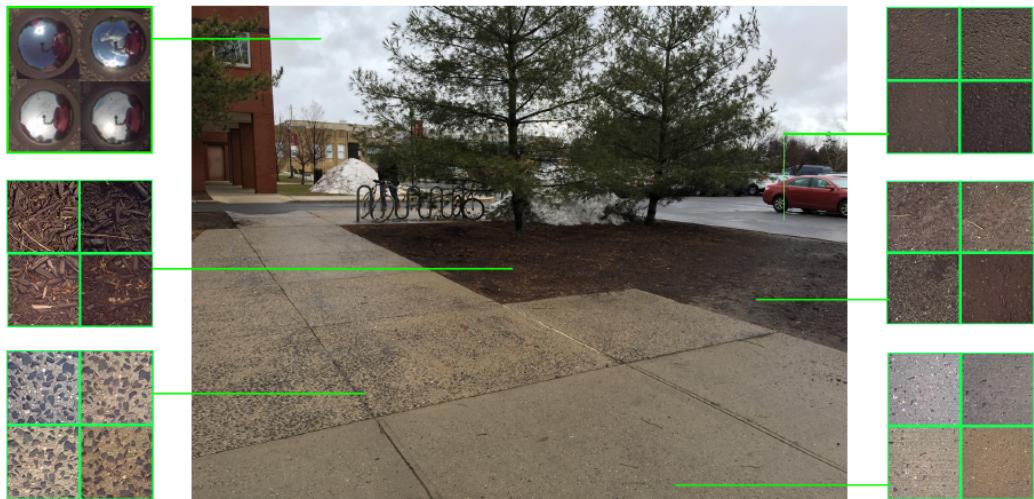
-
1. Python 3.6+: Ngôn ngữ lập trình chính để hiện thực các bước thực nghiệm
 2. Matlab: Hỗ trợ rút trích thông tin về texture và các thực nghiệm nhỏ liên quan
 3. scikit-learn: Hỗ trợ quá trình huấn luyện các bộ phân lớp với SVM, visualize kết quả
 4. Keras: Hỗ trợ việc rút trích deep feature từ CNN và thực hiện các thực nghiệm trên VGG16
 5. Tensorflow: back-end cho Keras.
 6. OpenCV: Dùng cho việc rút trích các handcrafted features và visualize kết quả
 7. Một số công cụ hỗ trợ khác

5.2 Dataset

Như đã đề cập bên trên, nhóm sử dụng hai tập dữ liệu để thực nghiệm: GTOS và FMD.

GTOS là tập dữ liệu mới nhất cho bài toán phân lớp vật liệu, được hoàn thành vào năm 2017 bởi Rutgers ECE Vision Lab. Bao gồm hơn 34000 ảnh thuộc 40 lớp khác nhau được thu thập dưới nhiều điều kiện thời tiết, chiếu sáng khác nhau, đây là một trong những tập dữ liệu đa dạng nhất cho bài toán phân lớp vật liệu.

Ở chiều hướng ngược lại, FMD là một trong những tập dữ liệu phổ biến nhất cho phân loại vật liệu ngay từ những ngày đầu ra đời của bài toán này khoảng đầu năm 2009. Tất cả các ảnh dùng trong hai tập dữ liệu này đều có kích thước 244 x 244.



Hình 5.1: Một cảnh ngoài trời được dùng để lấy các mẫu cho tập dữ liệu GTOS với nhiều góc nhìn nhìn, điều kiện chiếu sáng khác nhau [5]



Hình 5.2: Một tập ảnh từ FMD, đảm bảo sự đa dạng về điều kiện chiếu sáng, bố cục, màu sắc, kết cấu

5.3 Evaluation measures

Trong cả hai tập dữ liệu, độ chính xác trung bình (mean accuracy) được dùng để đánh giá kết quả cho tất cả các thực nghiệm.

5.4 Quá trình thực nghiệm và kết quả

Nhóm cài đặt các thực nghiệm bằng ngôn ngữ Python với sự hỗ trợ của framework scikit-learn và keras (trên nền tensorflow). Mạng VGG16 được huấn luyện sẵn trên ImageNet cũng được dùng cho việc transfer learning và các thực nghiệm khác nhằm hiểu rõ hơn về cách hoạt động của các layer trong CNN.

Tập dữ liệu được chia thành hai tập training và test theo tỷ lệ như các nghiên cứu trước sử dụng: 70% cho training và 30% cho test trên tập GTOS, 80% cho training và 20% cho test trên tập FMD.

Thực nghiệm với ba mô hình đề xuất

Đầu tiên nhóm thực hiện rút các handcrafted feature từ ảnh gốc sử dụng các filter thích hợp, bên cạnh đó deep feature cũng được rút từ layer "fc2" của mạng VGG16, sau đó thực hiện huấn luyện các bộ phân lớp ứng với các features khác nhau sử dụng SVMs với nhiều bộ tham số khác nhau, sau cùng dùng các bộ phân lớp này để phân lớp ảnh trên tập test. Probability prediction (trong mô hình 1 và 3) được kết hợp với nhau bằng phép toán trung bình cộng. Và các features được kết hợp với nhau (trong mô hình 2 và 3) bằng phép nối vector (vector concatenation). Các bộ tham số ban đầu (của SVMs) dùng để huấn luyện các bộ phân lớp được ghi nhận lại, từ đó nhóm đánh giá kết quả, điều chỉnh các tham số này thích hợp theo thời gian để có kết quả tốt nhất.

Method	Deep feature	Deep+edges	Deep+edges+textures
Combine prediction	75 ± 1.8	75.5 ± 3.0	76.3 ± 1.9
Combine features	75 ± 1.8	77.8 ± 2.5	79.5 ± 2.5
Kết hợp cả hai	75 ± 1.8	78.9 ± 2.2	82.2 ± 2.3

Bảng 5.1: Kết quả thực nghiệm với các mô hình trên tập GTOS

Bảng ?? thể hiện kết quả trên tập GTOS và FMD với ba mô hình đề xuất. Kết quả cho thấy mô hình thứ 3 hoạt động tốt nhất. Hai handcrafted features được dùng thể hiện thông tin về hình dạng và texture hoạt động hiệu quả trên hai tập dữ liệu này. Tuy nhiên, có thể thấy ảnh hưởng của hai features này trên hai tập

Method	Deep feature	Deep+edges	Deep+edges+textures
Combine prediction	74.2 ± 1.4	75.1 ± 2.2	75.5 ± 2.3
Combine features	74.2 ± 1.4	75.3 ± 1.5	76.3 ± 1.7
Mixed two above	74.2 ± 1.4	75.5 ± 1.2	77.2 ± 0.9

Bảng 5.2: Kết quả thực nghiệm với các mô hình trên tập FMD

dữ liệu là khác nhau, trên GTOS hiệu quả của chúng lớn hơn so với trên FMD (đối với cả ba mô hình). Lý do cho điều này được cho là do nhiều ảnh trong FMD là một đối tượng chính nằm trên một background còn GTOS ngược lại, toàn bộ hình đều đồng nhất là một bề mặt duy nhất (Hình 5.3). Chính background này khiến thông tin về hình dạng và texture bị ảnh hưởng khác nhiều.



Hình 5.3: Ảnh từ FMD (bên trái) trong lớp vật liệu "giấy" lại có background là một mặt đường khiến thông tin về texture không còn sự hiệu quả, trong khi ảnh từ GTOS là một bề mặt đồng nhất duy nhất (lớp "Gạch")

Thực nghiệm với VGG16 (nền tảng để xây dựng một CNN mới)

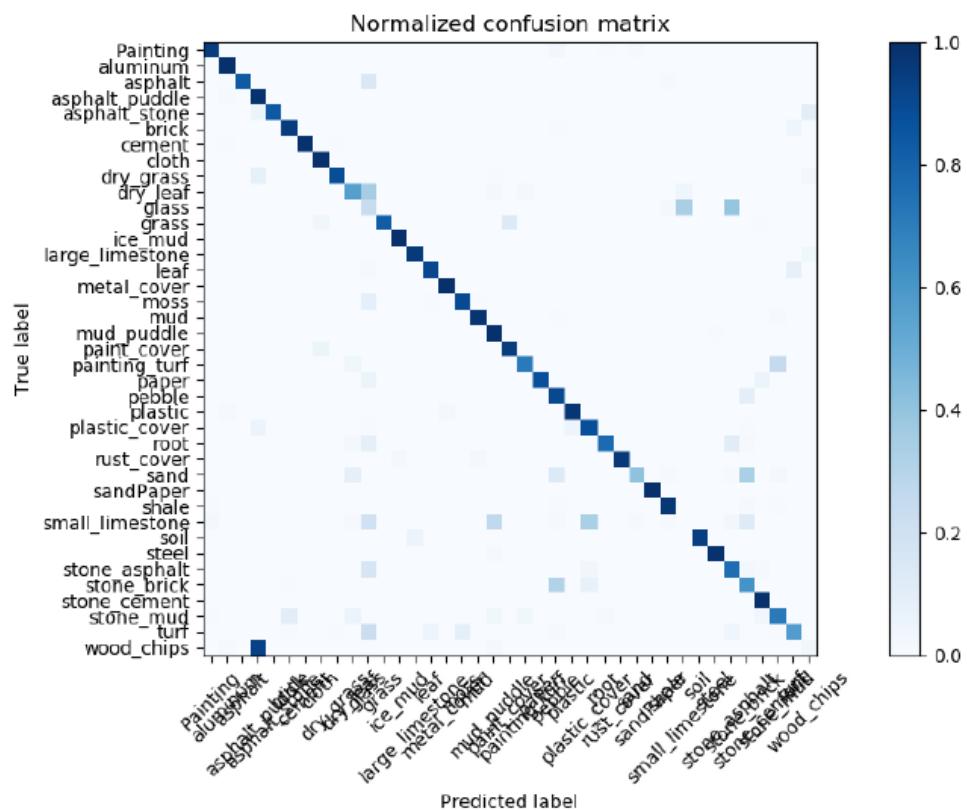
Quá trình thực nghiệm với VGG16 qua các bước sau đây:

1. Fine-tune VGG16: Thay fully-connected layer cuối cùng để đầu ra là 39 lớp (trên GTOS) thay vì 1000 lớp (trên ImageNet) và huấn luyện lại giá trị

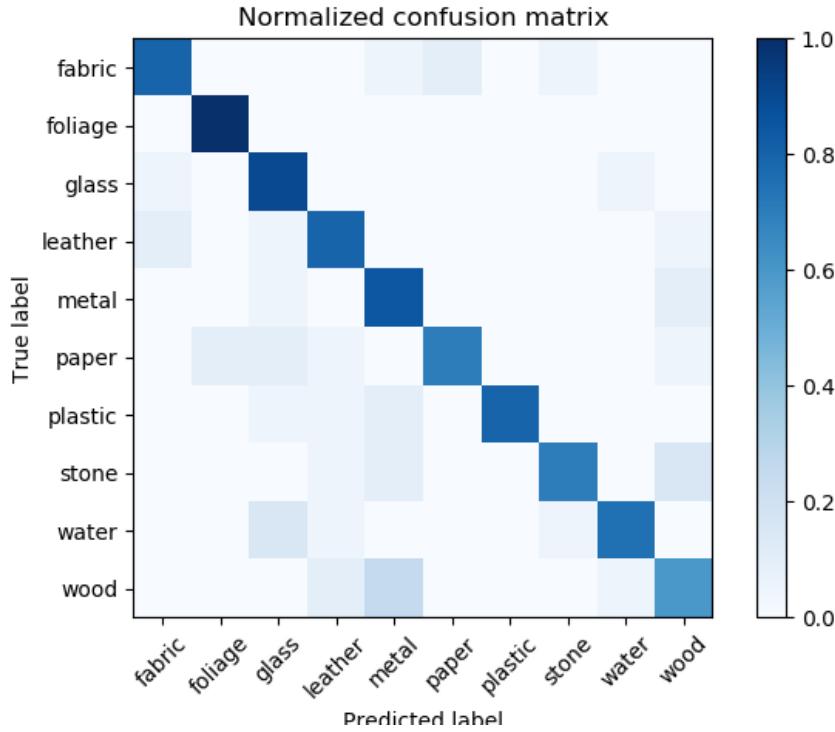
Method	GTOS	FMD
DAIN	81.2 ± 1.7	
Reflectance		65.5
SIFT IFV+fc7		69.6 ± 0.3
Ours	82.2	77.2

Bảng 5.3: So sánh kết quả với các phương pháp khác trên GTOS và FMD [5] [2]

[24]



Hình 5.4: Normalized confusion matrix trên GTOS



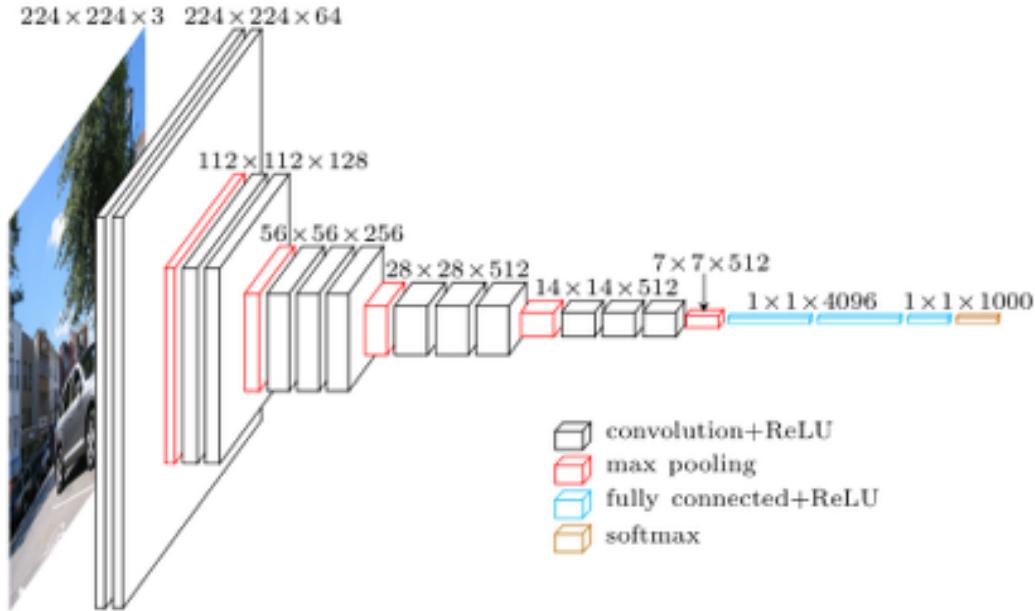
Hình 5.5: Normalized confusion matrix trên FMD

Classes	Acc.	Classes	Acc.	Classes	Acc.	Classes	Acc.
painting	0.95	glass	0.83	painting turf	0.96	small limestone	0.87
aluminum	0.82	grass	0.98	paper	0.90	soil	0.77
asphalt	0.90	ice mud	0.84	pebble	0.91	steel	0.96
asphalt puddle	0.76	large limestone	0.95	plastic	0.89	stone asphalt	0.40
asphalt stone	0.60	leaf	0.99	plastic cover	0.98	stone brick	0.92
brick	0.99	metal cover	0.92	root	0.99	stone cement	0.97
cement	0.71	moss	0.88	rust cover	0.94	stone mud	0.97
cloth	0.57	mud	0.56	sand	0.71	turf	0.94
dry grass	0.93	mud puddle	0.23	sand paper	0.87	wood chips	0.98
dry leaf	0.95	paint cover	0.85	shale	0.97		

Bảng 5.4: Kết quả cho từng lớp vật liệu trên tập GTOS

tham số của layer này, sau đó huấn luyện lại tất cả các layer để so sánh kết quả.

2. Tìm hiểu sự ảnh hưởng của các layer trong mạng bằng cách bỏ bớt layer



Hình 5.6: Cấu trúc mạng VGG16 được nhóm dùng để rút trích đặc trưng ở layer 'fc2' và thực hiện các thực nghiệm ở phần thực nghiệm với VGG16

trong mạng, quan sát ảnh hưởng của các layer này trên kết quả thu được.

3. Qua quan sát từ bước 2 và dựa trên nền của VGG16 chọn các layer thích hợp để xây dựng một mạng CNN ít layer hơn nhưng đạt kết quả xấp xỉ.

Hiện tại, khi đang viết luận văn này, nhóm vẫn đang thực nghiệm ở bước một, kết quả sau khi fine-tune trên GTOS cho thấy thời gian giảm 1/3 so với sử dụng SVM (bao gồm thời gian rút trích đặc trưng, huấn luyện và test), không cần tốn không gian lưu trữ đặc trưng và độ chính xác giảm từ 82.2% xuống còn 79.8%.

Chương 6

Kết luận và hướng phát triển

6.1 Kết quả đạt được

Sau giai đoạn luận văn, nhóm đã áp dụng được các kiến thức chuyên ngành đã được học vào bài toán phân lớp vật liệu. Sau quá trình tìm hiểu, tổng hợp các nghiên cứu trước, nhóm đã nắm được các phương pháp cơ bản hiện có cho bài toán, dựa vào đó đưa ra một số mô hình nhằm cải thiện kết quả. Báo cáo đã trình bày những mô hình nhóm đề xuất cũng như quá trình cài đặt, thử nghiệm đánh giá các mô hình này trên những tập dữ liệu khác nhau.

6.2 Hướng phát triển

Với kết quả hiện tại của đề tài, nhóm nhận thấy các mô hình đề xuất còn có nhiều hướng phát triển, chỉnh sửa nhằm cải thiện cả về độ chính xác, hiệu suất, tài nguyên sử dụng. Bên dưới là một số hướng phát triển tiếp theo cho đề tài mà nhóm dự định thực hiện.

1. Tiếp tục các thực nghiệm với mạng VGG16 để có một cái nhìn sâu hơn về cách các layer của nó hoạt động, từ đó có thể thiết kế một mạng mới thích hợp cho bài toán hiện tại. Điều này giúp tích kiệm chi phí huấn luyện các mô hình trên đồng thời cũng tăng độ thích ứng của mô hình với dữ liệu mới (so với việc dùng SVMs để huấn luyện các bộ phân lớp)

-
2. Thay đổi cách kết hợp features và probability prediction như đã trình bày trong ba mô hình trên. Có thể kết hợp theo trọng số thay vì dùng phép nối vector và trung bình cộng như hiện tại.
 3. Kết hợp thêm các thông tin khác có thể lấy từ ảnh (ví dụ như thông tin về góc nhìn - một mẫu với nhiều góc nhìn khác nhau, nghiên cứu [5] đã cho thấy thông tin này rất giá trị trong bài toán phân lớp vật liệu)

Tài liệu tham khảo

- [1] K. Tanaka, Y. Mukaigawa, T. Funatomi, H. Kubo, Y. Matsushita, and Y. Yagi, “Material classification using frequency-and depth-dependent time-of-flight distortion,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 79–88. [vi](#), [6](#), [7](#)
- [2] H. Zhang, K. Dana, and K. Nishino, “Reflectance hashing for material recognition,” in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015, pp. 3071–3080. [vi](#), [5](#), [7](#), [8](#), [42](#)
- [3] C. Liu, L. Sharan, E. H. Adelson, and R. Rosenholtz, “Exploring features in a bayesian framework for material recognition,” in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 239–246. [vi](#), [8](#), [9](#)
- [4] D. Hu, L. Bo, and X. Ren, “Toward robust material recognition for everyday objects.” in *BMVC*, vol. 2. Citeseer, 2011, p. 6. [vi](#), [8](#), [9](#)
- [5] J. Xue, H. Zhang, K. Dana, and K. Nishino, “Differential angular imaging for material recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 5, 2017. [vi](#), [vii](#), [viii](#), [5](#), [6](#), [9](#), [10](#), [11](#), [37](#), [39](#), [42](#), [46](#)
- [6] S. Corbett-Davies, “Real-world material recognition for scene understanding.” [2](#)
- [7] H. Lay, “Toyota to add wrong way driving alert to navigation systems, autoguide.com news,” May 2011.

TÀI LIỆU THAM KHẢO

- [Online]. Available: <http://www.autoguide.com/auto-news/2011/05/toyota-to-add-wrong-way-driving-alert-to-navigation-systems.html> 2
- [8] M. W. Spong, S. Hutchinson, and M. Vidyasagar, *Robot modeling and control*. Wiley New York, 2006, vol. 3. 2
- [9] J.-H. Kim, E. T. Matson, H. Myung, and P. Xu, *Robot Intelligence Technology and Applications 2012: An Edition of the Presented Papers from the 1st International Conference on Robot Intelligence Technology and Applications*. Springer Science & Business Media, 2013, vol. 208. 2
- [10] K. J. Dana, B. Van Ginneken, S. K. Nayar, and J. J. Koenderink, “Reflectance and texture of real-world surfaces,” *ACM Transactions On Graphics (TOG)*, vol. 18, no. 1, pp. 1–34, 1999. 5
- [11] E. Hayman, B. Caputo, M. Fritz, and J.-O. Eklundh, “On the significance of real-world conditions for material classification,” in *European conference on computer vision*. Springer, 2004, pp. 253–266. 5
- [12] M. Weinmann, J. Gall, and R. Klein, “Material classification based on training data synthesized using a btf database,” in *European Conference on Computer Vision*. Springer, 2014, pp. 156–171. 5
- [13] T.-C. Wang, J.-Y. Zhu, E. Hiroaki, M. Chandraker, A. A. Efros, and R. Ramamoorthi, “A 4d light-field dataset and cnn architectures for material recognition,” in *European Conference on Computer Vision*. Springer, 2016, pp. 121–138. 5
- [14] G. Choe, S. G. Narasimhan, and I. So Kweon, “Simultaneous estimation of near ir brdf and fine-scale surface geometry,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2452–2460. 5
- [15] A. Davis, K. L. Bouman, J. G. Chen, M. Rubinstein, F. Durand, and W. T. Freeman, “Visual vibrometry: Estimating material properties from small motion in video,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5335–5343. 6

TÀI LIỆU THAM KHẢO

- [16] P. Saponaro, S. Sorensen, A. Kolagunda, and C. Kambhamettu, “Material classification with thermal imagery.” in *CVPR*, 2015, pp. 4649–4656. [6](#)
- [17] W. Yuan, S. Wang, S. Dong, and E. Adelson, “Connecting look and feel: Associating the visual and tactile properties of physical materials,” in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR17), Honolulu, HI, USA*, 2017, pp. 21–26. [6](#)
- [18] G. Sampson, “Parallel distributed processing: Explorations in the microstructures of cognition,” 1987. [19](#), [22](#)
- [19] S. J. Thorpe, “Spike arrival times: A highly efficient coding scheme for neural networks,” *Parallel processing in neural systems*, pp. 91–94, 1990. [19](#)
- [20] G. F. Luger, *Artificial intelligence: structures and strategies for complex problem solving*. Pearson education, 2005. [22](#)
- [21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105. [24](#), [27](#)
- [22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. [24](#)
- [23] J. DeGol, M. Golparvar-Fard, and D. Hoiem, “Geometry-informed material recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1554–1562. [37](#)
- [24] S. Bell, P. Upchurch, N. Snavely, and K. Bala, “Material recognition in the wild with the materials in context database (supplemental material),” in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015. [42](#)