HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Phạm Tuấn Anh

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

HÀ NỘI – 2024

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Phạm Tuấn Anh

NGHIÊN CỨU GIẢI PHÁP AI TRÊN BIÊN MẠNG SỬ DỤNG CHO BÀI TOÁN CHUẨN ĐOÁN SÂU BỆNH

CHUYÊN NGÀNH : KỸ THUẬT VIỄN THÔNG

MÃ SỐ: 8.52.02.08 (Kỹ thuật Viễn thông)

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC: PSG. TS. HOÀNG TRỌNG MINH

LỜI CAM ĐOAN

Tôi cam đoan đây là công trình nghiên cứu của riêng tôi. Các số liệu, kết quả nêu trong đề án tốt nghiệp là trung thực và chưa từng được công bố trong bất kỳ công trình nào khác.

HỌC VIÊN CAO HỌC

Phạm Tuấn Anh

LỜI CẢM ƠN

Suốt hành trình nghiên cứu và hoàn thành đề án, tôi nhận được sự hỗ trợ và giúp đỡ vô cùng quý báu từ những người thầy cô bộ môn và giảng viên hướng dẫn khoa học. Trước tiên, tôi xin gửi lời tri ân sâu sắc đến Ban Giám hiệu và thầy cô Khoa Đào tạo sau đại học, Học viện Công nghệ Bưu chính Viễn thông đã tạo điều kiện thuận lợi cho tôi trong quá trình học tập và nghiên cứu.

Lời cảm ơn đặc biệt đến PGS.TS Hoàng Trọng Minh, người thầy kính mến, người đã tận tình hướng dẫn, động viên và tạo điều kiện cho tôi hoàn thành luận văn. Tôi cũng xin gửi lời cảm ơn chân thành đến gia đình, bạn bè và đồng nghiệp trong cơ quan đã luôn động viên, hỗ trợ, giúp tôi vượt qua những khó khăn trong học tập và nghiên cứu.

Mặc dù đã nỗ lực hết mình, nhưng do thời gian và kinh nghiệm nghiên cứu còn hạn chế, luận văn không thể tránh khỏi những thiếu sót. Tôi mong nhận được những góp ý quý báu từ thầy cô và bạn bè để hoàn thiện kiến thức và kỹ năng của bản thân.

Xin chân thành cảm ơn!

MỤC LỤC

DANH MỤC CÁC THUẬT NGỮ VIẾT TẮT

2G/3G	Mạng không dây thế hệ thứ 2 và 3 (2nd and 3rd Generation Wireless Network)
4G/5G	Mạng không dây thế hệ thứ 4 và 5 (4th and 5th Generation Wireless Network)
6G	Mạng không dây thế hệ thứ 6 (6th Generation Wireless Network)
AI	Trí tuệ Nhân tạo (Artificial Intelligence)
ANN	Mạng nơ-ron nhân tạo (Artificial Neural Network)
Big Data	Dữ liệu Lớn
boundary	Biên giới ảnh
centroids	Tâm cụm (Centroids)
CNN	Mạng nơ-ron tích chập (Convolutional Neural Network)
convolutional	Tích chập (Convolutional)
Deep learning/Machine Learning	Học sâu/Học máy (Deep learning/Machine Learning)
Edge / Fog	Biên / Sương Mù (Edge / Fog)
Edge AI	Trí tuệ Nhân tạo Biên (Edge AI)
FLOP	Số phép toán dấu phẩy động (Floating-Point Operations Per Second)
Gateways	Cổng kết nối
GPS	Định vị toàn cầu (Global Positioning System)
Internet	Mạng lưới Internet
IoT	Internet of Things (Vạn vật kết nối)

kernel	Nhân kernel
kNN	k-Nearest Neighbors (K Láng Giềng Gần Nhất)
LoRa	Công nghệ mạng diện rộng năng lượng thấp (Long Range)
LPWAN	Mạng diện rộng năng lượng thấp (Low-Power Wide-Area Network)
NB-IoT	Internet of Things băng thông hẹp (Narrowband Internet of Things)
neuron	Nơ-ron
ReLU	Rectified Linear Unit (Đơn vị Tuyến Tính Tuyến tính)
sensor-cloud-terminal	Cảm biến-đám mây-thiết bị đầu cuối (Sensor-Cloud-Terminal)
Service	Dịch vụ
SoC	Hệ thống trên một Chip (System on a Chip)
synaptic weight	Trọng số khớp thần kinh (Synaptic Weight)
vector	Véctơ
weight - bias	Trọng số - Biến thiên (Weight - Bias)
WiFi	Mạng không dây (Wireless Fidelity)

DANH SÁCH CÁC BẢNG

DANH SÁCH CÁC HÌNH

I. MỞ ĐẦU

1. Lý do chọn đề tài

Nông nghiệp là một ngành quan trọng đối với sự phát triển kinh tế và xã hội của mỗi quốc gia trong đó có Việt Nam. Tuy nhiên, việc giám sát và phát hiện sâu bệnh thường gặp nhiều khó khăn tại nước ta, đặc biệt là ở các vùng nông thôn, miền núi, nơi có điều kiện kinh tế - xã hội khó khăn. AI và IoT được hướng đến sử dụng để phân tích và giám sát dữ liệu từ cảm biến, từ đó phát hiện sớm các dấu hiệu sâu bệnh và đưa ra các cảnh báo kịp thời. Đề tài "Nghiên cứu giải pháp AI trên biên mạng sử dụng cho bài toán chuẩn đoán sâu bệnh" được triển khai vì đề tài có ý nghĩa thực tiễn, có thể ứng dụng để giải quyết vấn đề sâu bệnh trong nông nghiệp ở Việt Nam, góp phần nâng cao năng suất và chất lượng nông sản. Kết quả của đề tài sẽ là một giải pháp AI có khả năng giám sát và dự đoán sâu bệnh từ ảnh lá cây với độ chính xác cao, có thể được ứng dụng trong nông nghiệp, lâm nghiệp, môi trường, v.v.

2. Tổng quan về vấn đề nghiên cứu

Nhằm hướng đến chiến lược phát triển nông nghiệp thông minh giai đoạn 2021-2030, tầm nhìn đến năm 2045 và Phát triển nông nghiệp ứng dụng công nghệ cao [7-8] theo Bộ Nông nghiệp và Phát triển Nông thôn và Bộ Thông tin Truyền thông, nền nông nghiệp thông minh của Việt Nam hướng đến nền nông nghiệp cao đứng đầu thế giới. Từ đó, cần phải giải quyết những vấn đề chung, những khó khăn còn đang khúc mắc. Việc giám sát và phát hiện sâu bệnh thường gặp nhiều khó khăn, đặc biệt là ở các vùng nông thôn, miền núi, nơi có điều kiện kinh tế - xã hội khó khăn [2-5]. Các phương pháp giám sát và phát hiện sâu bệnh truyền thống đang gặp phải các hạn chế của riêng họ. Đã có rất nhiều những mô hình, hệ thống, hay các đề tài hay được triển khai [1][4] [6] nhưng đều gặp phải những vấn đề chung, để triển khai thành công đề tài, cần giải quyết các vấn đề về dữ liệu, hệ thống, mô hình, … Kết hợp với những nghiên cứu mới và tân tiến trên thế giới [9-16], việc giải quyết các vấn đề trên sẽ góp phần nâng cao hiệu quả của đề tài, giúp hệ thống IoT nhúng AI có khả năng giám sát và dự đoán sâu bệnh từ

ảnh lá cây với độ chính xác cao, đáp ứng được yêu cầu thực tiễn của ngành nông nghiệp.

3. Mục đích nghiên cứu

- Thu thập và xử lý thông tin từ hệ thống IoT cho giải pháp chuẩn đoán sâu bệnh tự động.
- Tìm kiếm giải pháp chuẩn đoán sâu bệnh dựa trên hình ảnh có độ chính xác cao và thuận tiện cho người sử dụng.

4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu

- Hệ thống thu thập thông tin cảm biến môi trường tự động qua tiếp cận IoT.
- Các mô hình dự báo dựa trên trí tuệ nhân tạo. Sử dụng mô hình AI tối ưu trên thiết bị biên mạng hiệu suất cao và đáp ứng yêu cầu thời gian thực để áp dụng vào bài toán chuẩn đoán sâu bệnh cây cà chua.

Phạm vi nghiên cứu

- Phạm vi nghiên cứu của đề án là các thiết bị của hệ thống IoT, các mô hình AI và các tập dữ liệu thử nghiệm để đưa ra giải pháp chuẩn đoán sâu bệnh qua hình ảnh lá cây cà chua.

5. Phương pháp nghiên cứu

Nghiên cứu lý thuyết

- Nghiên cứu các lý thuyết liên quan tới hệ thống tự động thu thập thông tin môi trường của các thiết bị IoT, lý thuyết về chuỗi thời gian và tiền xử lý dữ liệu. Các thuật toán và mô hình AI ứng dụng trong IoT nông nghiệp.

Nghiên cứu thực nghiệm

- Tiến hành các thí nghiệm để đánh giá hiệu quả của giải pháp bao gồm các bộ dữ liệu ảnh lá cây, nghiên cứu và ứng dụng mô hình AI thu thập dữ liệu từ hệ thống, đánh giá hiệu quả của hệ thống.

II. NỘI DUNG

CHƯƠNG 1: TỔNG QUAN VỀ HỆ THỐNG IOT NÔNG NGHIỆP

1.1 Tổng quan về nông nghiệp thông minh và thực trạng tại Việt Nam

1.1.1 Tổng quan về nông nghiệp thông minh

Nông nghiệp thông minh là một hệ thống sản xuất nông nghiệp ứng dụng các công nghệ tiên tiến như Internet vạn vật (IoT), trí tuệ nhân tạo (AI), dữ liệu lớn (Big Data), robot, cảm biến,... nhằm tối ưu hóa quy trình sản xuất, tăng hiệu quả và năng suất, đồng thời giảm thiểu tác động đến môi trường. Đầu tiên, lợi ích chính có thể thấy của nông nghiệp thông minh nhằm tăng năng suất và hiệu quả. Nông nghiệp thông minh giúp tối ưu hóa việc sử dụng tài nguyên, giảm chi phí sản xuất, tăng năng suất và chất lượng sản phẩm. Hơn nữa, nông nghiệp thông minh làm giảm thiểu tác động đến môi trường, giúp sử dụng hiệu quả các nguồn tài nguyên thiên nhiên, giảm thiểu sử dụng hóa chất và thuốc trừ sâu, bảo vệ môi trường, giải quyết vấn đề thiếu hụt lao động, giúp tự động hóa các quy trình sản xuất, giảm bớt sự phụ thuộc vào lao động, đặc biệt là trong bối cảnh già hóa dân số. Nền sản xuất nông nghiệp trên thế giới nói chung và tại Việt Nam nói riêng, đang phải chịu áp lực từ hai bài toán lớn là gia tăng dân số và giảm sút diện tích đất nông nghiệp [4]. Trong ngành nông nghiệp, để tận dụng hiệu quả và bền vững các nguồn tài nguyên, việc chuyển đổi số hóa là cần thiết [8, 9].

Công nghệ số có thể cải thiện khả năng ra quyết định, giúp quản lý rủi ro, kiểm soát sự biến động, từ đó tối ưu hóa sản lượng và gia tăng giá trị kinh tế của quá trình canh tác. Nhờ những tiến bộ trong điện tử, tự động hóa và công nghệ thông tin, hiệu quả canh tác và năng suất cây trồng đã được nâng cao. Tuy nhiên, cần có nhiều phương pháp tiếp cận thích hợp hơn để giúp người nông dân xử lý và hiểu tổng thể những thông số này. Ở các vùng nông thôn kém phát triển, thiếu hụt cơ sở hạ tầng làm cho canh tác và chăn nuôi trở nên lạc hậu, thu công, và ít hoặc không có kết nối với các thiết bị tiên tiến. Một ví dụ là tại Hoa Kỳ, dù là quốc gia tiên phong về công

nghệ kết nối, chỉ có khoảng 1/4 số nông trại sử dụng thiết bị kết nối để truy cập dữ liệu trên nền tảng mạng không dây thế hệ thứ 2 và 3 (sóng 2G/3G) hoặc các mạng IoT băng tần thấp.

Các công cụ theo dõi và kiểm soát truyền thống trên đồng ruộng và trong nhà lưới vẫn khá đơn giản, chưa khai thác hết giá trị tiềm năng mà công nghệ kết nối có thể mang lại. Điều này đòi hỏi ngành nông nghiệp phải tận dụng đầy đủ các ứng dụng và phân tích kỹ thuật số, yêu cầu độ trễ thấp, băng thông cao và độ tin cậy cao. Các phương thức kết nối cũng cần đòi hỏi giảm thiểu độ trễ, tốc độ phản hồi và cải thiện sự ổn định, giúp vận hành tự động và chính xác máy móc và các thiết bị bay không người lái. Tất cả những vận hành này, với sự kết nối của các phương thức truyền thông phù hợp, sẽ giúp quản lý, thực hiện và kiểm soát tốt các khâu trong hệ thống canh tác nông nghiệp chính xác.

1.1.2 Thực trạng nông nghiệp thông minh tại Việt Nam

Hiện nay, Việt Nam vẫn là một quốc gia chủ yếu dựa vào nông nghiệp với hơn 66,9% dân số sống ở vùng nông thôn và 42% lao động trong toàn xã hội làm việc trong ngành nông nghiệp [3]. Nông nghiệp thông minh đang là xu hướng phát triển tất yếu của ngành nông nghiệp Việt Nam. Chính phủ Việt Nam đã ban hành nhiều chính sách và chiến lược phát triển nông nghiệp thông minh, trong đó có Chiến lược phát triển nông nghiệp ứng dụng công nghệ cao giai đoạn 2021-2030, tầm nhìn đến năm 2045 [9].

Tuy nhiên, việc ứng dụng nông nghiệp thông minh đáp ứng các công nghệ và tiêu chuẩn trên tại Việt Nam còn gặp nhiều hạn chế về hạ tầng công nghệ thông tin, hệ thống internet và mạng lưới viễn thông chưa được phủ sóng rộng khắp, đặc biệt là ở khu vực nông thôn. Hơn nữa, việc thiếu hụt nguồn nhân lực có trình độ điển hình là nông dân Việt Nam chủ yếu là lao động già, trình độ học vấn và kỹ năng sử dụng công nghệ còn hạn chế. Cùng với đó là chi phí đầu tư cao, việc ứng dụng các công nghệ cao vào sản xuất nông nghiệp cần có vốn đầu tư lớn, trong khi nhiều hộ nông dân còn gặp khó khăn về tài chính [6]. Việc chuyển dịch cơ cấu kinh tế theo hướng công nghiệp và dịch vụ đã làm giảm nhân lực trong nông nghiệp, và dự báo rằng số

lượng này sẽ tiếp tục tăng trong những năm tới, đặt ra vấn đề về nhân lực trong ngành nông nghiệp [2-3]. Công nghệ được áp dụng trong việc chăm sóc và thu hoạch nông sản để khắc phục vấn đề thiên tai, môi trường, tiết kiệm nhân lực, tăng năng suất cây trồng và đơn giản hóa việc quản lý.

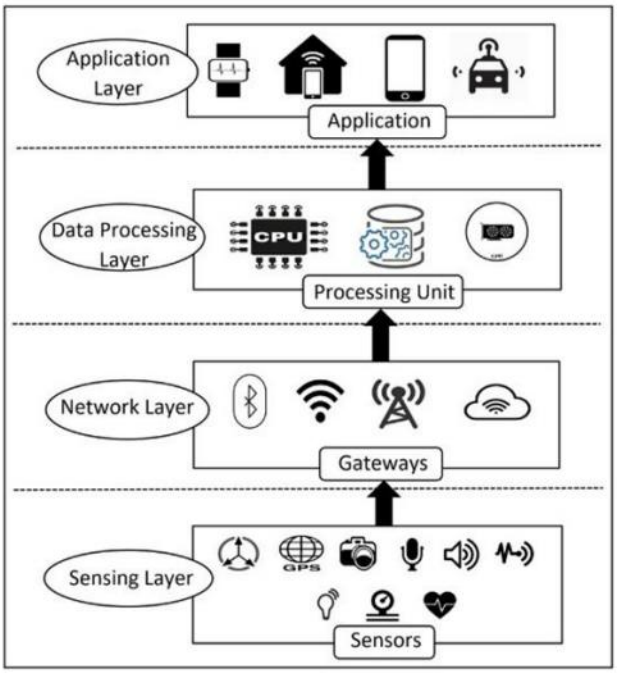
Một trong những ứng dụng công nghệ nổi bật trong nông nghiệp gần đây là Internet of Things (IoT) và Artificial Intelligence (AI), đã và đang mang lại nhiều kết quả thành công, dần dần được áp dụng và phổ biến trên nhiều diện tích canh tác nông nghiệp. Đề tài "Nghiên cứu giải pháp AI trên biên mạng sử dụng cho bài toán chuẩn đoán sâu bệnh" nhằm hiểu rõ hơn về tác động của công nghệ đối với khả năng phát triển của cây trồng và quản lý của người điều khiển, cũng như nghiên cứu thêm về các ứng dụng công nghệ điện tử được áp dụng.

1.2 Khái quát các hệ thống IoT nông nghiệp sử dụng hiện nay và xu hướng

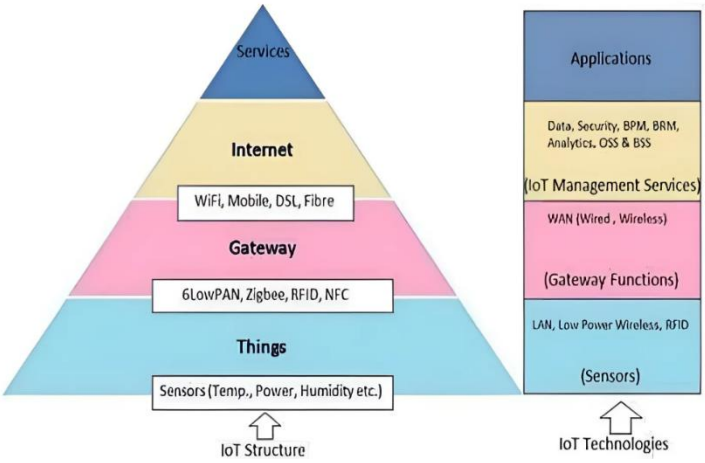
1.2.1 Tổng quan về hệ thống IoT nông nghiệp

Từ những khó khăn hiện đang gặp phải đã phân tích ở trên, những mô hình IoT trong nông nghiệp hiện tại Việt Nam nhìn chung mới phổ biến bao gồm các mô IoT truyền thống, ví dụ như các hệ thống giám sát thời tiết, đo lường nhiệt độ, độ ẩm, ánh sáng, ... kết hợp với ứng dụng di động và quản lý dữ liệu tập trung trên máy chủ đám mây. Một số ứng dụng của IoT trong nông nghiệp hiện nay đang được sử dụng phổ biến trong các hệ thống lớn và nhỏ.

Cấu trúc của một hệ thống IOT gồm bốn thành phần cơ bản chính gồm: Các cảm biến (Things), Trạm kết nối (Gateways), Hạ tầng mạng (Internet) và cuối cùng là lớp dịch vụ (Service).



Hình 1.1. Mô hình cụ thể của một hệ thống IoT



Hình 1.2. Bốn thành phần cơ bản của hệ thống IoT

Giống như trong các ngành công nghiệp khác, ứng dụng IoT trong nông nghiệp hứa hẹn hiệu quả hơn nhiều so với phương pháp thủ công trước đây, giúp giảm tài nguyên và chi phí, tự động hóa dựa trên phân tích dữ liệu, và tối ưu hóa quy trình. Tuy nhiên, riêng đối với ngành nông nghiệp, vai trò của IoT là vô cùng quan trọng. Nó sẽ mang tới các giải pháp bước ngoặt, giải quyết những vấn đề cấp bách liên quan tới sự sinh tồn và phát triển của loài người. Giúp nâng cao chất lượng nông phẩm, cải

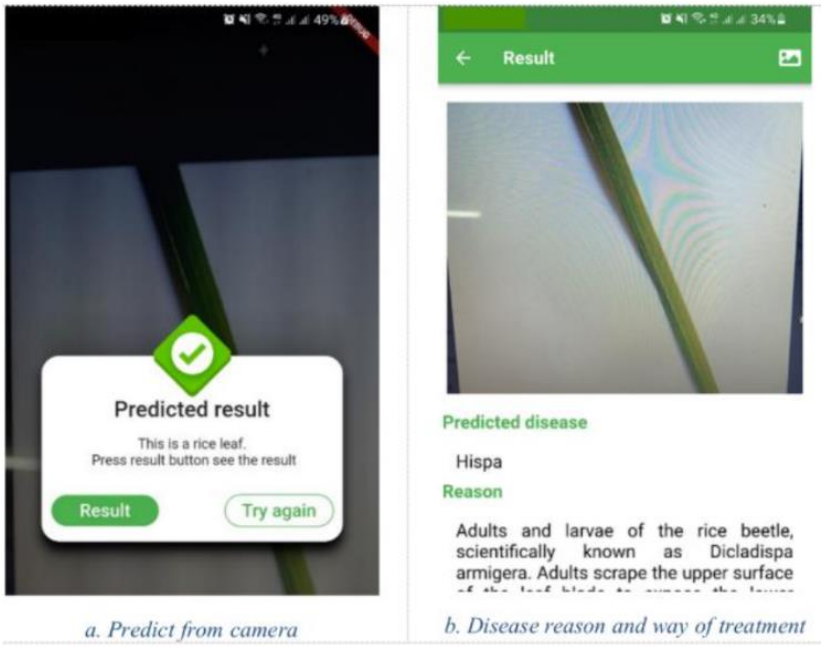
thiện năng suất canh tác, bảo đảm vệ sinh an toàn thực phẩm, xây dựng một hệ thống trồng trọt bền vững và chính xác trước các biến đổi khó lường của khí hậu hiện nay.

1.2.2 Khảo sát mô hình nông nghiệp thông minh tại Việt Nam

Một vài nghiên cứu như "Ứng dụng công nghệ IoT và AI giám sát và điều khiển nhà nuôi chim yến thông minh" [1], ứng dụng những công nghệ mới như công trình nghiên cứu [3] với đề tài "Ứng dụng công nghệ xử lý ảnh kết hợp IoT để theo dõi và phân tích tình trạng quả trên cây cà chua" (2022) và một vài nghiên cứu khác là một trong những nghiên cứu mới ứng dụng AI và IoT dẫn tới một xu hướng mới ứng dụng khoa học dữ liệu và internet vạn vật vào canh tác nông nghiệp. Mục tiêu của đề tài [3] là xây dựng được một hệ thống camera giám sát kết hợp IoT trên cây cà chua có khả năng giám sát nhiệt độ, độ ẩm (thông qua các cảm biến), ổn định điều kiện môi trưởng(thông qua bơm nước..vv). Đề tài phát triển một Kit Esp8266, ESP32 CAM để giám sát tình hình cây trồng và xử lý bằng thuật toán xử lý ảnh truyền thống. Hệ thống này cho phép thực hiện các thao tác giám sát – điều khiển trên firebase thông qua WiFi và một ứng dụng Android [3]. Tuy nhiên, nghiên cứu gặp phải một vài hạn chế như thời gian và tốc độ xử lý còn chậm, hệ thống camera chưa xử lí được điều kiện thiếu ánh sáng, và chưa ứng dụng được khoa học dữ liệu và học sâu, máy học, từ đó mới chỉ dừng lại ở quy mô nghiên cứu mà chưa thể ứng dụng rộng rãi. Một vài công trình nghiên cứu [5,7] như "Hệ thống so màu lá lúa trên thiết bị di động" (2016) đã ứng dụng máy học bằng kỹ thuật so khớp ảnh và kỹ thuật kNN (k-Nearest Neighbors) nhằm mục đích so màu lá lúa tự động từ ảnh chụp trên thiết bị di động nhằm xác định lượng phân đạm cần thiết (tương đối) để bón cho cây lúa dựa trên độ đậm của lá lúa. Từ tập ảnh chụp từ điện thoại, hệ thống tiến hành tiền xử lý, khử nhiễu và sau đó thực hiện so màu lá lúa bằng phương pháp so khớp ảnh và kỹ thuật máy học.

Cùng với khảo sát tại đề tài [7], Nghiên cứu mới đã mang đến một phương pháp đột phá giúp phát hiện bệnh lá lúa chỉ bằng ảnh chụp từ thiết bị di động. Phương pháp này sử dụng kỹ thuật học sâu kết hợp học chuyển giao, cho phép xác định ba loại bệnh phổ biến trên lá lúa: đốm nâu, cháy bìa lá và đạo ôn lá với độ chính xác lên

đến 95%. Mô hình được phát triển bởi các nhà khoa học đã được huấn luyện trên 1.790 hình ảnh và tích hợp vào ứng dụng Android (Hình 2). Ứng dụng này có thể phát hiện bệnh và đưa ra giải pháp điều trị chỉ trong 1,7 giây, hứa hẹn trở thành công cụ hữu ích cho người nông dân trong việc bảo vệ mùa màng. Các nghiên cứu trên là một vài trong những nghiên cứu đi đầu và tiên phong trong các nghiên cứu ứng dụng học máy và xử lý ảnh trong nông nghiệp.



Hình 1.3. Ứng dụng xử lý ảnh trên điện thoại hỗ trợ nông dân trồng trọt [7]

Lấy cảm hứng từ những hệ thống hiện có đang phát triển nhanh chóng, đề án kết hợp với giải pháp AI tại biên mạng sẽ xử lý và chia sẻ tài nguyên tính toán trên nhiều điểm biên và ra quyết định khi kết nối đa thiết bị, có tính kết nối vạn vật và kết nối tập trung trên máy chủ đám mây, thu thập dữ liệu và cải tiến thuật toán/mô hình theo thời gian sử dụng.

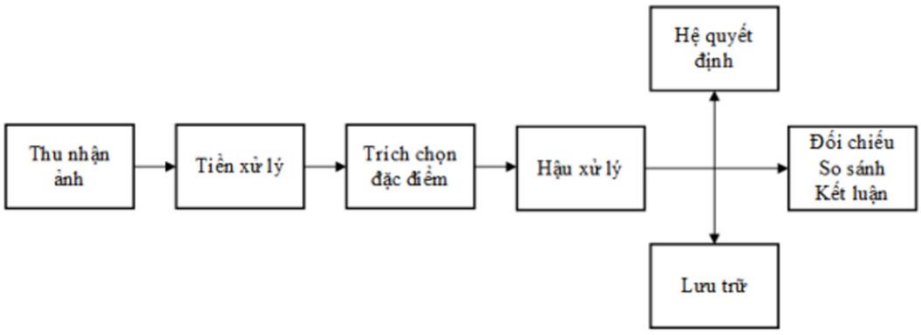
1.3 Lý thuyết xử lý ảnh và phương pháp xử ý nhằm chuẩn đoán sâu bệnh

1.3.1 Thông tin biểu diễn dưới dạng ảnh và lý thuyết xử lý ảnh truyền thống

Thực chất của thông tin là sự thông báo, trao đổi, giải thích về một đối tượng cụ thể nào đó và được thể hiện thông qua các dạng tín hiệu như âm thanh, chữ số, chữ viết, hình ảnh… nhằm mang lại một sự hiểu biết nào đó cho đối tượng cụ thể. Máy

tính thông thường được thiết kế các thuật toán nhằm xử lý hình ảnh được biểu diễn dưới dạng kỹ thuật số, phân tích và nhận dạng các hình ảnh và trích xuất dữ liệu đa chiều từ thế giới thực để cho ra các thông tin số hóa. Học máy có khả năng tương tự như thị giác con người bởi sự nhận diện và hiểu biết thông tin từ một hình ảnh số đối với thiết bị điện tử hay hiểu được thế giới quan và các dạng thông tin. Thị giác máy tính cũng được mô tả là sự tổng thể của một lượng lớn các quá trình tự động và tích hợp, các thể hiện cho các nhận thức thị giác về thế giới thực tế.

Xử lý ảnh là một lĩnh vực khoa học máy tính tập trung vào việc chuyển đổi và cải thiện hình ảnh thông qua các thuật toán và phương pháp tính toán. Nó có thể được hiểu đơn giản là quá trình biến đổi ảnh đầu vào thành ảnh đầu ra với các đặc điểm mong muốn hoặc trích xuất thông tin từ ảnh. Ở phạm vi đề tài này sẽ tìm hiểu về vấn đề nhận dạng ảnh trên cở sở màu sắc ảnh thu được từ cảm biến camera. Quá trình xử lý ảnh theo nhận dạng ảnh thực hiện các bước như trong (Hình 1.4)



Hình 1.4. Quá trình trích xuất và xử lý, nhận diện ảnh

Mức xám của một điểm ảnh thể hiện cường độ sáng tại điểm đó, được gán bằng một giá trị số. Các mức ảnh xám thông thường bao gồm 16, 32, 64, 128 và 256. Mức được sử dụng phổ biến nhất là 256, sử dụng 1 byte để biểu diễn mức xám.

- Ảnh nhị phân: chỉ có 2 mức trắng và đen, tương ứng với giá trị 0 và 1, sử dụng 1 bit dữ liệu cho mỗi điểm ảnh.

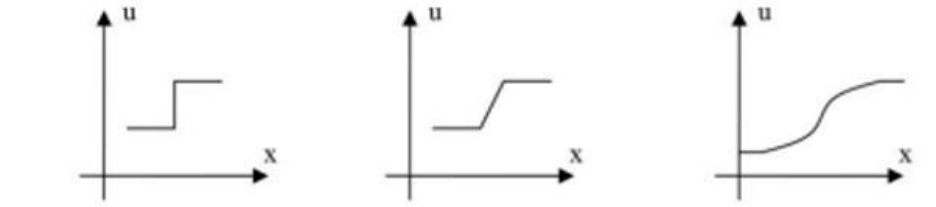
- Ảnh đen trắng: có hai màu đen và trắng (không chứa màu khác) với mức xám tại các điểm ảnh có thể khác nhau.
- Ảnh màu: kết hợp 3 màu cơ bản (đỏ, lục, lam) để tạo ra nhiều màu sắc khác nhau. Người ta thường dùng 3 byte để mô tả mức màu, tạo ra khoảng 16,7 triệu mức màu.
- Ảnh thu nhận được thường bị nhiễu, do đó cần phải loại bỏ nhiễu để cải thiện chất lượng ảnh.

Lọc ảnh là kỹ thuật sử dụng các bộ lọc để loại bỏ nhiễu hoặc làm nổi bật các chi tiết trong ảnh.

Có hai loại lọc chính:

- Lọc tuyến tính: bao gồm lọc trung bình, lọc thông thấp, v.v. Lọc trung bình thay thế mỗi điểm ảnh bằng trung bình trọng số của các điểm ảnh lân cận. Lọc thông thấp cho phép tín hiệu có tần số thấp (tương ứng với các chi tiết ảnh) đi qua và loại bỏ tín hiệu có tần số cao (tương ứng với nhiễu).
- Lọc phi tuyến: bao gồm lọc trung vị, lọc ngoài, v.v. Lọc trung vị thay thế mỗi điểm ảnh bằng trung vị của các điểm ảnh lân cận. Lọc ngoài loại bỏ các điểm ảnh có giá trị khác biệt so với các điểm ảnh lân cận.
- Lọc nhiễu có thể được thực hiện bằng lọc tuyến tính hoặc phi tuyến. Lọc tuyến tính thường được sử dụng để làm trơn nhiễu, trong khi lọc phi tuyến thường được sử dụng để làm nổi bật các cạnh trong ảnh.

Lọc ảnh là một kỹ thuật quan trọng trong xử lý ảnh, giúp cải thiện chất lượng ảnh và trích xuất thông tin từ ảnh một cách hiệu quả.



a, Đường biên lý tưởng b, Đường biên bậc thang c, Đường biên thực

Hình 1.5. Đường biên ảnh

Điểm biên là điểm ảnh có sự thay đổi nhanh chóng hoặc đột ngột về mức xám (hoặc màu). Trong ảnh nhị phân, điểm đen được coi là điểm biên nếu có ít nhất một điểm trắng lân cận. Đường biên (đường bao) là tập hợp các điểm biên liên tiếp tạo thành một đường cong khép kín. Đường biên là đặc trưng quan trọng trong phân tích và nhận dạng ảnh. Biên được sử dụng để phân cách các vùng xám (màu) cách biệt. Ngược lại, các vùng ảnh cũng có thể được sử dụng để tìm đường phân cách.

Điểm ảnh có sự biến đổi mức xám u(x) đột ngột.

Phát hiện biên là một phần trong phân tích ảnh, được thực hiện sau khi lọc ảnh (hay tiền xử lý ảnh).

Các bước bao gồm:

- Lọc ảnh
- Phát hiện điểm biên
- Liên kết các điểm biên để tạo thành đường biên
- Dò và tìm biên ảnh là một trong các đặc trưng thuộc khối trích chọn đặc trưng.

Lợi ích của việc phát hiện biên:

Giúp trích xuất các đặc trưng quan trọng của ảnh

Hỗ trợ phân loại và nhận dạng ảnh

Có nhiều ứng dụng trong các lĩnh vực như y tế, công nghiệp, an ninh, v.v.

Hiện nay, có nhiều phương pháp phát hiện biên khác nhau:

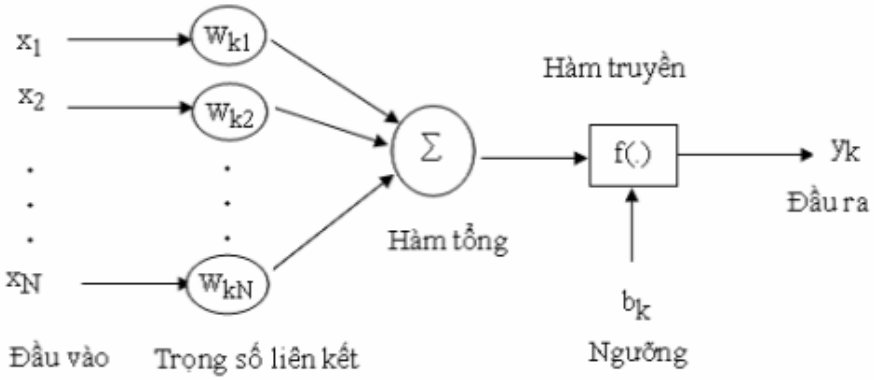
- Phương pháp dựa trên gradient
- Phương pháp dựa trên Laplace
- Phương pháp dựa trên Canny
- Phương pháp dựa trên Watershed

Phát hiện biên là một kỹ thuật quan trọng trong xử lý ảnh, đóng vai trò quan trọng trong việc phân tích và nhận dạng ảnh.

1.3.2 Phương pháp xử lý ảnh và trích xuất đặc trưng nhằm chuẩn đoán sâu bệnh

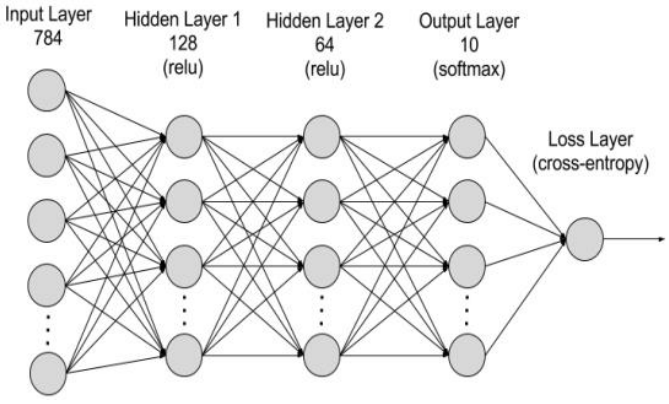
Dựa trên các lý thuyết của thông tin biểu diễn dưới dạng ảnh và sự phát triển của khoa học dữ liệu và học sâu, cơ sở lý thuyết và thuật toán tự động trích xuất đặc trưng của dữ liệu ảnh trở nên phổ biến và ngày càng hiệu quả. Học sâu (Deep learning) là một phương pháp tiên tiến trong lĩnh vực trí tuệ nhân tạo, nhằm giúp máy tính học và xử lý dữ liệu theo mô phỏng quá trình tư duy trong não bộ con người. Mô hình học sâu có khả năng nhận diện và hiểu được nhiều mẫu phức tạp trong hình ảnh, văn bản, âm thanh và các dữ liệu khác, từ đó đưa ra thông tin và dự đoán chính xác.

Cấu trúc của một mạng neuron cơ bản:



Hình 1.6: Cấu tạo một neuron

Năm 2012, Alex Krizhevsky, Ilya Sutskever và Geoffrey Hinton với công trình nghiên cứu "Phân loại ImageNet bằng mạng nơ ron tích chập sâu" được coi là một công trình đầy tầm quan trọng trong lĩnh vực Deep learning.



Hình 1.7 Mô hình neuron network

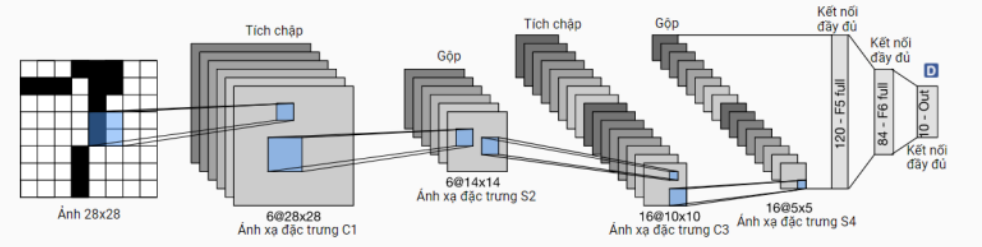
Các tác giả đề xuất một kiến trúc mạng nơ ron sâu được gọi là AlexNet, đạt được thành công đột phá đáng kể trong cuộc thi "ImageNet Large Scale Visual Recognition Challenge" năm 2012. Kiến trúc AlexNet đạt được tỷ lệ lỗi là 15,3%, vượt qua phương pháp tốt nhất thứ hai một khoảng cách hơn 10%. Các đóng góp của bài báo bao gồm việc phát triển kiến trúc AlexNet, sử dụng một số kỹ thuật đổi mới như hàm kích hoạt ReLU, chuẩn hóa đáp ứng cục bộ, chế độ giảm thiểu và tăng lượng dữ liệu để cải thiện hiệu suất mạng, mở đường cho việc áp dụng rộng rãi deep learning trong các ứng dụng thị giác máy tính. Thị giác máy tính là khả năng của máy tính trích xuất thông tin và dữ liệu chuyên sâu từ hình ảnh và video. Các kỹ thuật học sâu được sử dụng để giúp máy tính hiểu và phân tích hình ảnh như con người. Các thuật toán phân tích và xử lý dựa trên các đặc trưng trong ảnh và video, từ đó đưa ra các dự đoán. Thị giác máy tính đã được ứng dụng trong nhiều lĩnh vực, bao gồm:

- Nhận dạng khuôn mặt để xác định khuôn mặt và nhận biết các đặc điểm như mở mắt, đeo kính và có râu.
- Phân loại hình ảnh để xác định quần áo và các chi tiết khác trong hình ảnh.
- Kiểm duyệt nội dung để tự động loại bỏ các nội dung không an toàn hoặc không phù hợp khỏi kho lưu trữ hình ảnh và video.

Một kiến trúc mạng nơ ron được sử dụng phổ biến trong thị giác máy tính đó là mạng nơ ron tích chập (Convolutional Neural Network - CNN). Đây là một kiến trúc đa tầng được sử dụng để xử lý dữ liệu đầu vào và tạo ra một tập hợp số để so sánh với các dữ liệu đã biết, nhằm định nghĩa và phân loại dữ liệu đó.

Có nhiều loại kiến trúc CNN nổi tiếng như AlexNet, VGGNet, GoogleNet, và còn nhiều kiến trúc khác. Mạng nơ ron tích chập là một trong những mô hình học sâu tiên tiến và được sử dụng rộng rãi. Nó rất hiệu quả trong việc xử lý dữ liệu hình ảnh và đã đóng góp đáng kể vào việc xây dựng các hệ thống thông minh hiện đại với độ chính xác cao. Đó cũng là lý do chính mà CNN được sử dụng phổ biến trong xử lý hình ảnh. Một hình ảnh là một ma trận các pixel, nhưng thường không biến đổi ma trận thành một vector và xử lý nó bằng cách sử dụng kiến trúc mạng nơ ron truyền thống.

Lý do là vì ngay cả với hình ảnh đơn giản nhất, các pixel liền kề có sự phụ thuộc lẫn nhau, việc biến đổi thành vector sẽ làm mất đi thông tin phụ thuộc này và làm thay đổi ý nghĩa của bức hình. Ví dụ, biểu tượng của mắt, miệng con người hoặc thậm chí là cạnh của một đối tượng khác được xây dựng từ một số pixel được bố trí theo một cách nhất định. Nếu xử lý hình ảnh thành một vector, những phụ thuộc này bị mất và làm giảm độ chính xác của mô hình.



Hình 1.8. Mô phỏng mạng neuron tích chập CNN

Các đặc trưng chính của mạng CNN

- Sử dụng phương pháp tích chập: Các mạng nơ ron tích chập đều sử dụng phương pháp tích chập để trích xuất các đặc trưng từ dữ liệu. Do đó, chúng được gọi chung là mạng nơ ron tích chập.

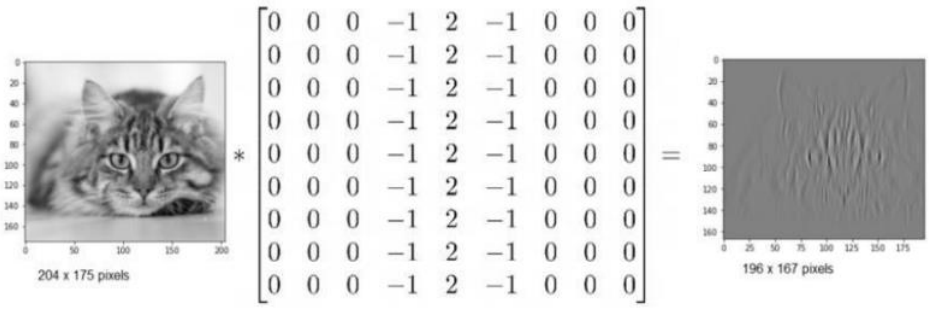
- Kiến trúc phân tầng: Kiến trúc phân tầng trong mạng CNN giúp cho quá trình học của mạng xảy ra ở nhiều cấp độ khác nhau, từ cấp độ thấp đến cao. Khi đó, mạng CNN có khả năng tiếp cận và học các đặc trưng từ mức độ thấp như đường chéo, ngang, dọc, rìa và cạnh, tới các đặc trưng chi tiết hơn. Các đặc trưng này được tổng hợp và trích xuất thông qua các tầng tích chập của mạng.

- Được huấn luyện trên những bộ dữ liệu lớn: Sự khác biệt quan trọng giữa mô hình học sâu nhiều tầng và các phương pháp học máy truyền thống chỉ trở nên rõ rệt khi huấn luyện trên các bộ dữ liệu lớn. Trên bộ dữ liệu nhỏ, phương pháp truyền thống có thể đủ để phân biệt các nhãn với nhau, và không có sự ưu thế đáng kể của mạng nơ ron đa tầng. Tuy nhiên, trên các bộ dữ liệu lớn, kiến trúc học sâu nhiều tầng đã chứng tỏ sự vượt trội về độ chính xác và khả năng biểu diễn. Điều này dễ hiểu bởi

mạng nơ ron có thể có hàng chục triệu tham số, vượt xa số lượng tham số của các phương pháp học máy truyền thống, cho phép nó có khả năng biểu diễn tốt hơn.

- Kích thước layers giảm dần: Nghiên cứu đã chứng minh rằng việc giảm kích thước layers có thể giảm số lượng tham số của mô hình một cách đáng kể, từ đó tạo ra các mạng nhẹ hơn và tăng tốc độ dự đoán. Đồng thời, sự giảm này không ảnh hưởng đáng kể đến độ chính xác của mô hình.

- Độ sâu tầng layers tăng dần: Bằng việc tăng số lượng bộ lọc theo cấp số nhân, độ sâu của các layers trong mạng CNN được gia tăng. Điều này giúp mô hình học được nhiều đặc trưng đa dạng hơn. Các layers đầu tiên thường tạo ra những đặc trưng chung có hình dạng và phương hướng tương tự, do đó không cần nhiều bộ lọc. Tuy nhiên, khi đến các layers sau, yêu cầu độ chi tiết tăng cao hơn, điều này đồng nghĩa với việc cần nhiều bộ lọc hơn để phân biệt được các chi tiết đặc trưng. Từ mạng CNN, quá trình tích chập tạo ra các đặc trưng hai chiều. Để sử dụng những đặc trưng này trong quá trình phân loại của mạng CNN, cần được chuyển thành đặc trưng một chiều thông qua phương pháp flatten và được truyền qua các layers kết nối đầy đủ. Mỗi layer được kích hoạt bằng hàm phi tuyến để tăng khả năng biểu diễn và cải thiện kết quả phân loại.



Hình 1.9. Ví dụ cơ bản về kernel trong CNN

Lớp tích chập: Khối xây dựng chính trong một mạng lưới thần kinh tích chập là lớp Convolutional layer. Một lớp tích chập như nhiều mẫu vuông nhỏ, được gọi là hạt nhân kernel, trượt qua hình ảnh và tìm kiếm các mẫu. Trong trường hợp phần đó của hình ảnh khớp với mẫu của kernel, sẽ trả về một giá trị dương lớn và khi không

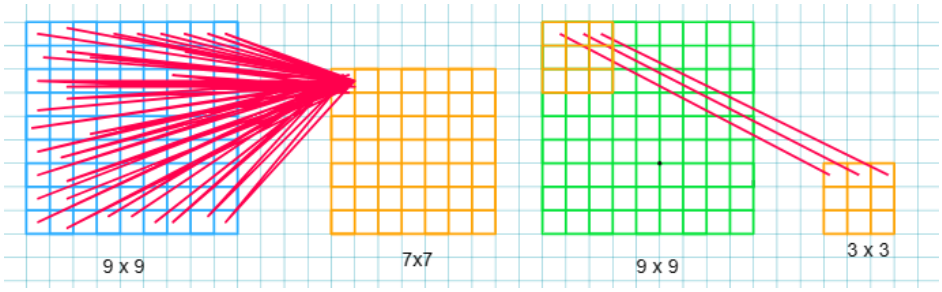
có khớp, hạt nhân trả về bằng không hoặc một giá trị nhỏ hơn. Mặc dù đơn giản, khả năng phát hiện các đường thẳng đứng hoặc ngang, góc, đường cong và các tính năng đơn giản khác là một thuộc tính cực kỳ mạnh mẽ của hạt nhân phức tạp. Các lớp sau này trong mạng thần kinh có thể xây dựng trên các đặc trưng được phát hiện bởi các lớp trước đó và xác định các hình dạng phức tạp hơn bao giờ hết.

Hàm kích hoạt: Sau khi truyền hình ảnh qua một lớp phức tạp, đầu ra thường được truyền qua một chức năng kích hoạt. Các chức năng kích hoạt phổ biến bao gồm hàm sigmoid và hàm Relu. Hàm kích hoạt có tác dụng thêm tính phi tuyến tính vào mạng lưới thần kinh tích chập. Nếu chức năng kích hoạt không có, tất cả các lớp của mạng thần kinh có thể được cô đọng xuống một phép nhân ma trận duy nhất. Trong trường hợp hình 18, ảnh con mèo ở trên, việc áp dụng hàm ReLU cho đầu ra lớp đầu tiên dẫn đến độ tương phản mạnh hơn làm nổi bật các đường thẳng đứng và loại bỏ nhiễu bắt nguồn từ các đặc trưng không phải đường thẳng đứng khác.

Cấu trúc của mạng lưới thần kinh CNN

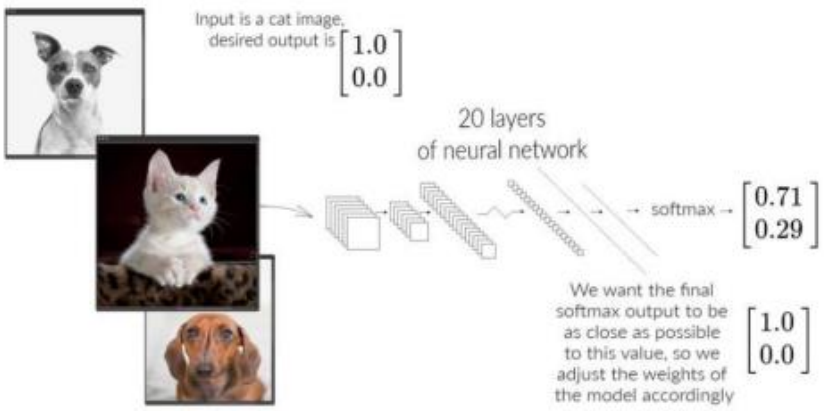
Một mạng lưới thần kinh phức tạp cơ bản có thể được xem như hàng loạt các lớp phức tạp, tiếp theo sẽ là một chức năng kích hoạt, tiếp theo nữa là một lớp gộp (downscaling) lặp đi lặp lại nhiều lần. Với việc kết hợp lặp đi lặp lại của các hoạt động này, lớp đầu tiên phát hiện các tính năng đơn giản như các cạnh trong hình ảnh và lớp thứ hai bắt đầu phát hiện các đặc trưng cấp cao hơn. Đến lớp thứ mười, một mạng lưới thần kinh phức tạp có thể phát hiện các hình dạng phức tạp hơn như mắt. Đến lớp thứ hai mươi, nó thường có thể phân biệt khuôn mặt của con người với nhau. Khả năng này đến từ việc phân lớp các hoạt động lặp đi lặp lại, mỗi đặc trưng có thể phát hiện các đặc trưng có thứ tự cao hơn một chút so với người tiền nhiệm trước đó.

Một mạng lưới thần kinh tích chập là một loại mạng lưới thần kinh đặc biệt với trọng lượng ít hơn so với một mạng được kết nối đầy đủ.



Hình 1.10 So sánh phép chập trong ảnh và mạng thần kinh kết nối đầy đủ

Trong một mạng lưới thần kinh được kết nối đầy đủ, mọi nút trong đầu vào được gắn vào mọi nút trong lớp đầu tiên, cứ thế tiếp tục các lớp sau cũng gắn đầy đủ và không có hạt nhân kernel tích chập. Vì vậy, ví dụ ảnh 9x9 trong đầu vào và hình ảnh 7x7 là đầu ra lớp đầu tiên, nếu điều này được thực hiện với một mạng lưới thần kinh được kết nối đầy đủ, sẽ có: 9x9x7x7 = 3969 kết nối. Tuy nhiên, khi điều này được thực hiện với một lớp convolutional có một hạt nhân kernel 3x3 duy nhất, thì sẽ chỉ có 3x3 = 9 kết nối. Rõ ràng là một mạng lưới thần kinh tích chập sử dụng ít thông số hơn rất nhiều so với mạng lưới thần kinh được kết nối đầy đủ tương đương với cùng kích thước lớp. Điều này là do các tham số mạng được sử dụng lại khi hạt nhân trượt trên hình ảnh do đó một mạng lưới thần kinh tích chập sẽ có thể phát hiện các đặc trưng trong hình ảnh bất kể chúng nằm ở đâu.



Hình 1.11 Mô tả chi tiết cấu tạo thuật toán CNN

1.2 Kết luận chương

Nông nghiệp thông minh đang dần trở thành xu hướng tất yếu cho sự phát triển của ngành nông nghiệp Việt Nam, có tiềm năng giải quyết các vấn đề như năng suất thấp, chất lượng sản phẩm chưa cao, sử dụng nhiều hóa chất, thiếu hụt lao động, và biến đổi khí hậu. Tuy nhiên, để ứng dụng hiệu quả nông nghiệp thông minh tại Việt Nam cần phải có những giải pháp phù hợp với điều kiện thực tế. Một trong những giải pháp tiềm năng là sử dụng trí tuệ nhân tạo (AI) trên biên mạng, xử lý ảnh và chuẩn đoán sâu bệnh. Giải pháp này đóng vai trò quan trọng trong nông nghiệp thông minh, được sử dụng để phát hiện sâu bệnh, xác định tình trạng dinh dưỡng của cây trồng, và đánh giá năng suất cây trồng. Với nhiều phương pháp xử lý ảnh được sử dụng để chuẩn đoán sâu bệnh, bao gồm phân loại, phân đoạn, trích xuất đặc trưng và nổi bật là kiến trúc mạng nơ-ron tích chập (CNN) là một kiến trúc mạng nơ-ron nhân tạo có hiệu quả cao trong xử lý ảnh, có khả năng học các đặc trưng trực tiếp từ ảnh và xác định các đối tượng trong ảnh với độ chính xác cao. Nông nghiệp thông minh là giải pháp đầy tiềm năng và hứa hẹn cho các vấn đề trong ngành nông nghiệp Việt Nam trong tương lai.

CHƯƠNG 2: GIẢI PHÁP AI TRÊN BIÊN MẠNG

2.1 Mô hình triển khai hệ thống AI trên biên mạng

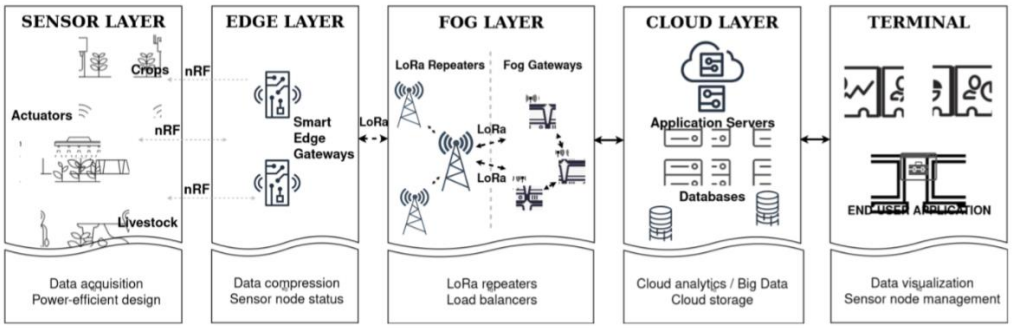
2.1.1 Tổng quan hạ tầng mạng trong nông nghiệp thông minh

Trong thời đại Internet vạn vật (IoT), một giải pháp là triển khai hệ thống giám sát và quản lý từ xa cho các trang trại và hệ thống nông nghiệp. Các trang trại áp dụng công nghệ IoT thường được gọi là trang trại thông minh. Tuy nhiên, việc giám sát và kiểm soát trang trại từ xa có tác động hạn chế ở những khu vực có kết nối Internet không ổn định hoặc kém. Điều này xảy ra không chỉ ở các nước đang phát triển mà còn ở các nước phát triển. Internet of Things (IoT) có thể được định nghĩa là một nền tảng nơi các đối tượng ảo và vật lý được kết nối với nhau và giao tiếp với nhau. Hệ thống IoT bao gồm các công nghệ khác nhau như mạng cảm biến không dây, điện toán đám mây và trí thông minh nhúng. Các hệ thống này cung cấp các dịch vụ tiên tiến như giám sát từ xa theo thời gian thực, phân tích trực tuyến và quản lý từ xa.

Tuy nhiên, chỉ dựa vào các kiến trúc IoT tập trung vào đám mây truyền thống để giám sát và quản lý trang trại từ xa không thể đảm bảo rằng các hệ thống hoạt động bình thường vì IoT vẫn còn một số thách thức. Ví dụ: các ứng dụng IoT tập trung vào đám mây không thể được triển khai ở các vùng sâu vùng xa, nơi Internet không ổn định hoặc phạm vi phủ sóng bị hạn chế. Trong những trường hợp như vậy, dữ liệu không thể được theo dõi theo thời gian thực và các hành động đối với sự bất thường có thể không được thực hiện đúng hạn. Ví dụ, nếu một đám cháy bất ngờ xảy ra hoặc một nhóm động vật hoang dã phá hoại cây trồng, hệ thống không thể phản ứng kịp thời.

Điện toán biên và sương mù có thể được minh họa như một đám mây nhỏ gần biên của mạng hơn. Nói cách khác, điện toán Edge và Fog đại diện cho sự hội tụ của các lớp mạng khác nhau thành các cổng thông minh được kết nối với nhau. Điện toán biên và sương mù có thể giúp khắc phục một số hạn chế của các hệ thống IoT có thể tập trung vào truyền thống. Ví dụ, điện toán biên và sương mù cung cấp nhiều lợi thế như hiệu quả năng lượng, lưu trữ cục bộ phân tán, khả năng tương tác và bảo mật nâng cao. Chi tiết hơn, điện toán biên và sương mù có thể giúp giảm tải mạng và gánh

nặng tính toán và lưu trữ của các máy chủ đám mây. Điều này được thực hiện bằng cách di chuyển nhiều quy trình tính toán chuyên sâu từ đám mây sang các lớp và cổng Edge và Fog, đồng thời cho phép các nút cảm biến tiết kiệm năng lượng hơn khi chúng phụ thuộc nhiều hơn vào các cổng thông minh mạng cục bộ. So với các ứng dụng IoT truyền thống thường dựa vào kiến trúc 3 lớp (sensor-cloud-terminal), ứng dụng IoT hỗ trợ sương mù có thêm các lớp giữa các nút cảm biến và đám mây. Tùy thuộc vào ứng dụng và loại dữ liệu thu được, một số lớp Edge / Fog khác nhau có thể được triển khai.



Hình 2.1. Mô hình cơ bản của hệ thống IoT trong nông nghiệp tiên tiến

Mặc dù điện toán biên và sương mù có thể cung cấp nhiều dịch vụ tiên tiến, các hệ thống dựa trên sương mù vẫn không thể hoạt động bình thường ở các vùng sâu vùng xa, nơi Internet không ổn định hoặc không được phủ sóng, vì chúng thường dựa vào mạng cục bộ tốc độ cao để xử lý thời gian thực và các ứng dụng quan trọng về độ trễ. Một giải pháp trong các tình huống này là triển khai các công nghệ mạng diện rộng công suất thấp, chẳng hạn như LoRa, cho phép truyền tầm xa với nhược điểm là tốc độ dữ liệu giảm. LoRa là một trong những giao thức LPWAN phổ biến nhất cho lớp vật lý, cung cấp giao tiếp công suất thấp và tầm xa lên đến 10 hoặc 20 km trong truyền dẫn mở và tầm nhìn. Tuy nhiên, LoRa không thể được sử dụng để gửi dữ liệu với tốc độ dữ liệu cao do các quy định của địa phương và giới hạn đối với chu kỳ nhiệm vụ truyền tải ở hầu hết các khu vực trên thế giới là 0.1%, 1% hoặc 10%. Do đó, LoRa một mình không thể giúp giải quyết các vấn đề hiện có của các ứng dụng IoT ở vùng sâu vùng xa.

Sự bùng nổ của mạng diện rộng công suất thấp (LPWAN) trong thập kỷ qua, với các công nghệ như LoRa hoặc NB-IoT, đã cung cấp một cơ sở hạ tầng tương đối rẻ cho phép truyền tải công suất thấp và tầm xa. Tuy nhiên, những lợi ích mà công nghệ LPWAN mang lại có nhược điểm là truyền băng thông thấp. Do đó, việc tích hợp điện toán biên và sương mù, di chuyển phân tích dữ liệu và nén các thiết bị gần cuối, là chìa khóa để mở rộng chức năng. Bằng cách tích hợp trí tuệ nhân tạo ở lớp mạng cục bộ, hoặc Edge AI, nội dung chương này trình bày kiến trúc và triển khai hệ thống giúp mở rộng khả năng của các ứng dụng nông nghiệp và nông nghiệp thông minh với điện toán Edge và Fog nhằm hỗ trợ xử lí dữ liệu lớn cho các công nghệ truyền thông cho vùng phủ sóng diện rộng. Edge Intelligence hoặc Edge AI là sự kết hợp giữa AI và Edge Computing, nó cho phép triển khai các thuật toán học máy đến thiết bị biên nơi dữ liệu được tạo ra. Edge Intelligence có tiềm năng cung cấp trí tuệ nhân tạo cho mọi người và mọi tổ chức ở bất kỳ đâu có hạ tầng. Ngày nay, một số lượng lớn các cảm biến và thiết bị thông minh tạo ra một lượng lớn dữ liệu và đòi hỏi sức mạnh tính toán ngày càng tăng, đang thúc đẩy cốt lõi của các nhiệm vụ tính toán và dịch vụ từ đám mây đến biên của mạng. Bằng cách kết hợp IoT với AI, dữ liệu được thu thập bởi các nút có thể được sử dụng bằng cách áp dụng các kỹ thuật AI như học máy và học sâu. Kết quả là, khả năng học máy được di chuyển gần hơn với nguồn dữ liệu. Khái niệm này được gọi là Edge AI, hoặc Edge Intelligence, và nó cho phép khả năng mở rộng lớn hơn, mạnh mẽ và hiệu quả. Do đó, các mô hình học máy trong các hệ thống AI được kết hợp với khả năng kết nối và truyền dữ liệu của Internet vạn vật IoT. Nói cách khác, với sự kết hợp của AI trong các hệ thống IoT, chức năng của chúng không chỉ giới hạn trong việc thu thập và truyền thông tin mà thực sự hiểu và phân tích được dữ liệu.

2.1.2 Tổng quan về trí tuệ nhân tạo và hệ ra quyết định trong hạ tầng mạng

Ban đầu, Machine Learning và Deep Learning bị giới hạn ở Cloud, chủ yếu là do tính sẵn có và khả năng mở rộng của các tài nguyên yêu cầu tính toán cao cần thiết để xử lý các tác vụ ML. Việc kết hợp điện toán đám mây và IoT mang đến những lợi ích thiết thực, đặc biệt khi sử dụng các cảm biến thông minh. Trước đây, dữ liệu thu

thập từ các thiết bị IoT cơ bản (như camera hay micrô) được gửi đến đám mây để phân tích. Quá trình này tốn thời gian và gây tắc nghẽn mạng do lượng dữ liệu lớn.

Với sự ra đời của các thiết bị biên sử dụng cảm biến thông minh (như camera tích hợp thị giác máy tính hay micrô có chức năng xử lý ngôn ngữ tự nhiên), việc phân tích dữ liệu có thể diễn ra nhanh hơn ngay tại thiết bị. Với lợi ích bảo mật rõ ràng, người dùng có thể yên tâm sử dụng các thiết bị IoT thu thập dữ liệu cá nhân mà không lo bị xâm phạm quyền riêng tư. Bằng cách kết hợp điện toán đám mây và IoT, chúng ta có thể tận dụng tối đa tiềm năng của cả hai công nghệ này để mang lại những lợi ích thiết thực cho cuộc sống. Một động lực chính ở đây là lượng điện năng tiêu thụ ngày càng tăng của máy tính có thể thay thế bằng các thiết bị ngày càng nhỏ và tiết kiệm điện hơn, nhờ vào thiết kế giao diện cho người dùng và pin hiệu quả hơn. Vào năm 2022, khi nhiều tổ chức tiếp tục hướng tới hệ sinh thái đám mây kết hợp để cung cấp các dịch vụ IoT cho khách hàng của họ, điện toán biên sẽ ngày càng trở thành một phần quan trọng của giải pháp khi có yêu cầu cung cấp thông tin chi tiết nhanh và an toàn. Nhờ những tiến bộ về phần cứng mà học máy đã đẩy nhanh việc triển khai trên hàng tỷ thiết bị kết nối thông minh và thích ứng trong các cơ sở hạ tầng quan trọng như y tế, kiểm soát môi trường, hậu cần, giao thông vận tải và nông nghiệp. Chuyển xử lý AI từ Đám mây sang các thiết bị biên được phân tán, kết nối cung cấp một giải pháp để khắc phục các tắc nghẽn, độ trễ và các vấn đề về quyền riêng tư của các ứng dụng AI dựa trên đám mây. So với các thiết bị IoT công suất thấp truyền thống, AIoT yêu cầu các thiết bị biên có đủ tài nguyên để thực hiện các tác vụ học máy trên thiết bị. Tuy nhiên, khả năng tài nguyên và năng lượng của các thiết bị biên tự nhiên bị hạn chế. Do đó, các ứng dụng AIoT dựa trên các thách thức cần tối ưu hóa để cân bằng:

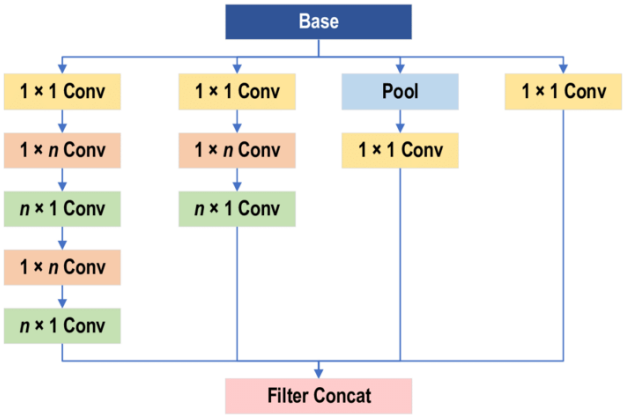
- Chi phí phần cứng và hiệu suất
- Mô hình được tối ưu hóa cho nền tảng thiết bị.

Đây là lý do tại sao các xu hướng gần đây tập trung vào tối ưu hóa mô hình AI để giảm thiểu kích thước mô hình và tìm cách tăng hiệu quả mô hình. Nén mô hình AI

được sử dụng để thực hiện suy luận mô hình có độ trễ thấp và tiết kiệm năng lượng ở biên. Các mô hình ML "nhẹ" nhỏ hơn và hiệu quả hơn nhiều có thể chạy trên các thiết bị năng lượng thấp như điện thoại di động, SoC hoặc máy tính nhúng. Ví dụ phổ biến là phiên bản mô hình ML trên thiết bị TensorFlow Lite, OpenVino của Intel hoặc Lightweight OpenPose, …

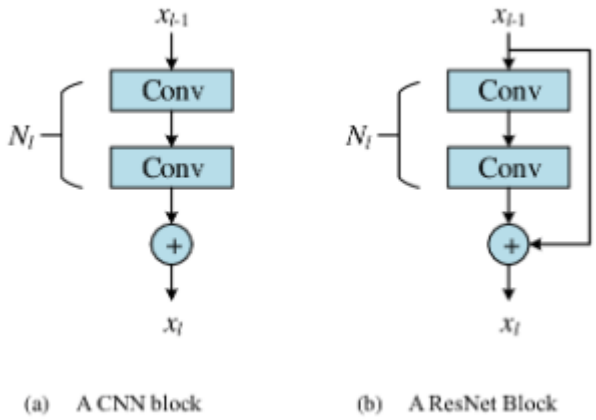
2.1.3 Các mô hình AI đã được tối ưu trên biên

InceptionNets: Biến thể GoogLeNet với các mô-đun Inception đã được giới thiệu vào năm 2016 [7]. Inception-v3 đã đạt được hiệu suất phân loại tốt trong một số ứng dụng y sinh sử dụng học tập chuyển giao. Nó đề xuất một mô hình khởi đầu kết hợp nhiều bộ lọc tích chập có kích thước khác nhau vào một bộ lọc mới. Mục tiêu của mô-đun khởi động là hoạt động như một "trình trích xuất tính năng đa cấp" bằng cách tính toán các kết cấu 1×1, 3×3 và 5×5 trong cùng một mô-đun của mạng. Thiết kế như vậy làm giảm số lượng tham số được đào tạo và do đó làm giảm độ phức tạp tính toán.



Hình 2.3 Inception module của biến thể InceptionNet V3

ResNets: Các mô hình ResNet, dựa trên các kiến trúc sâu đã cho thấy các hành vi hội tụ tốt và độ chính xác hấp dẫn, được phát triển bởi He et al. [7]. ResNet được xây dựng bởi một số đơn vị còn lại xếp chồng lên nhau và được phát triển với nhiều số lớp khác nhau: 18, 34, 50, 101, 152 và 1202. Các đơn vị còn lại bao gồm tích chập, gộp và lớp. ResNet 50 chứa 49 lớp tích chập và một lớp được kết nối hoàn toàn ở cuối mạng. Để tiết kiệm tài nguyên máy tính và thời gian đào tạo, ResNet 50 đã được chọn để so sánh trong phần sau.

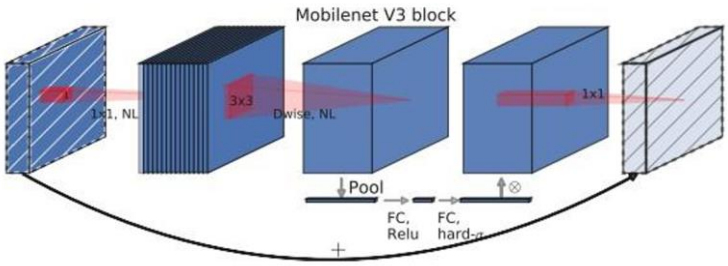


Hình 2.4 Ý tưởng chính trong tải tiến model Resnet

MobileNets: Có ba phiên bản MobileNets, mới nhất, MobileNet V3. Kiến trúc cốt lõi của MobileNetV1 dựa trên một kiến trúc được sắp xếp hợp lý sử dụng các lớp phức tạp có thể tách rời chiều sâu để xây dựng các mạng thần kinh sâu nhẹ. MobileNetV2 đã giới thiệu hai tính năng mới cho kiến trúc: nút cổ chai tuyến tính giữa các lớp và kết nối phím tắt giữa các bottlenecks. MobileNetV3 là phiên bản thứ ba của kiến trúc, cung cấp khả năng phân tích hình ảnh của nhiều ứng dụng di động phổ biến. Đóng góp chính của MobileNetV3 là sử dụng AutoML để tìm kiếm kiến trúc mạng lưới thần kinh tốt nhất có thể cho một vấn đề nhất định. Điều này trái ngược với thiết kế thủ công của các phiên bản kiến trúc trước đó.

MobileNetV3 là một mô hình DNN nhẹ được điều chỉnh cho phù hợp với các CPU của điện thoại di động hay các thiết bị biên thông qua sự kết hợp của tìm kiếm kiến trúc mạng nhận biết phần cứng (NAS) được bổ sung bởi thuật toán NetAdapt và sau đó được cải thiện thông qua các tiến bộ kiến trúc mới. MobileNetV3 được định nghĩa có hai mô hình: MobileNetV3- Large và MobileNetV3-Small. Các mô hình này được nhắm mục tiêu vào các trường hợp sử dụng tài nguyên cao và thấp tương ứng. MobileNets là một loạt các mạng nơ-ron sâu có trọng lượng nhẹ dựa trên các Depthwise Separable Convolutions. Tiếp sau đó phiên bản cải tiến từ Version 1 là MobileNetV2. MobileNetV2 tiếp tục sử dụng Depthwise Separable Convolutions, ngoài ra còn đề xuất thêm: Linear bottlenecks và Inverted Residual Block (shortcut connections giữa các bottlenecks). MobileNetV3 đạt được hiệu suất tốt hơn với ít

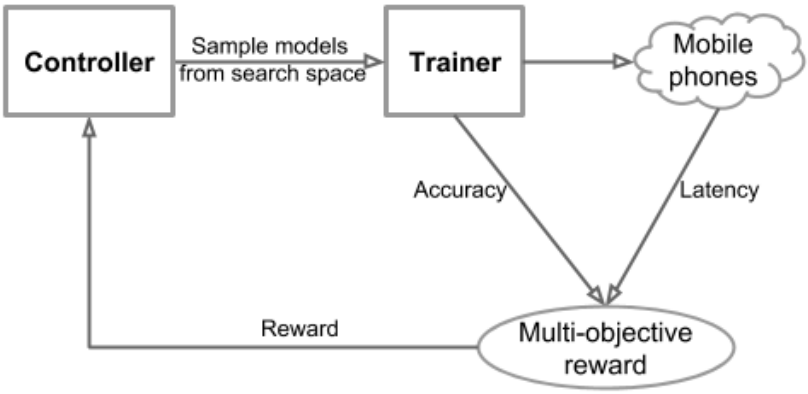
FLOP hơn với các cải tiến mới so với các mô hình tiền nhiệm với khối kiến trúc mới như Hình 27 dưới đây. Trái ngược với phiên bản MobileNet trước đó được thiết kế thủ công, MobileNetV3 có thể tự tìm kiếm kiến trúc tốt nhất có thể trong không gian tìm kiếm phù hợp với các tác vụ thị giác máy tính di động.



Hình 2.5. Sơ đồ khối của MobileNet V3

Để khai thác hiệu quả nhất không gian tìm kiếm, hai kỹ thuật được triển khai theo trình tự là MnasNet và NetAdapt. Đầu tiên, tìm kiếm một kiến trúc thô bằng MnasNet, sử dụng tính năng học tăng cường để chọn cấu hình tối ưu từ một tập hợp các lựa chọn rời rạc. Sau đó, tinh chỉnh kiến trúc bằng cách sử dụng NetAdapt, một kỹ thuật bổ sung giúp cắt bỏ các kênh kích hoạt chưa được sử dụng theo mức độ nhỏ. Để cung cấp hiệu suất tốt nhất có thể trong các điều kiện khác nhau, ta có thể tạo các mô hình lớn hoặc nhỏ. Ngoài ra cải tiến mạng bằng cách thiết kế lại các lớp tốn nhiều chi phí tính toán và sửa đổi hàm phi tuyến tính thành hard-swish (h-swish) dựa trên hàm phi tuyến tính của Swish để có thể khắc phục hạn chế lớn nhất của hàm Swish là nó rất kém hiệu quả khi tính toán trên phần cứng di động.

- MnasNet: Khối xây dựng chính của MnasNet là một khối còn lại đảo ngược (từ MobileNet V2 đề cập ở trên). Lấy cảm hứng từ sự tiến bộ trong tìm kiếm kiến trúc thần kinh AutoML, cách tiếp cận tìm kiếm kiến trúc MnasNet để thiết kế các mô hình di động bằng cách sử dụng học tăng cường. Tổng thể của phương pháp này bao gồm chủ yếu là ba thành phần: bộ điều khiển dựa trên RNN để học hỏi và lấy mẫu kiến trúc mô hình, một huấn luyện viên xây dựng và đào tạo các mô hình để có được độ chính xác và một động cơ suy luận để đo tốc độ mô hình trên điện thoại di động thực. MnasNet thực hiện một vấn đề tối ưu hóa đa ngôn từ nhằm mục đích đạt được cả accuracy cao và tốc độ cao.



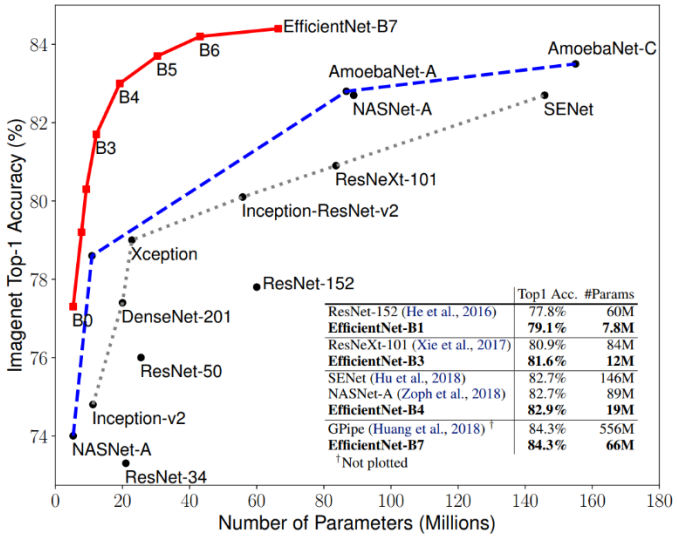
Hình 2.6. Công cuộc đổi mới trong việc thiết kế mạng Mnasnet

- EfficientNets Lite: EfficientNet-Lite mang lại sức mạnh của EfficientNet cho các thiết bị biên và có năm biến thể, cho phép người dùng chọn từ tùy chọn độ trễ / kích thước mô hình thấp (EfficientNet-Lite0) đến độ chính xác cao(EfficientNet-Lite4). Một số hoạt động trong EffcientNet không được hỗ trợ tốt bởi một số máy gia tốc nhất định. Để giải quyết vấn đề không đồng nhất, EfficientNets ban đầu được điều chỉnh với các sửa đổi đơn giản sau:

• Loại bỏ một vài các layer mạng vì chúng không được hỗ trợ tốt.

• Thay thế tất cả các kích hoạt swish bằng RELU6, điều này cải thiện đáng kể chất lượng định lượng sau đào tạo.

• Cố định tham số và scale-down mô hình để giảm kích thước và tính toán của các mô hình thu nhỏ.



Hình 2.7: Hiệu năng vượt trội của EfficientNet so với các mạng tối ưu khác

2.2 Các phương pháp xử lý dữ liệu và cải tiến hệ thống

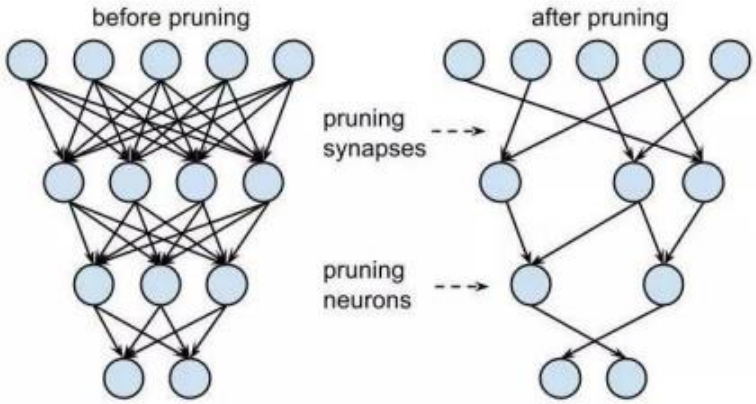
2.2.1 Ý tưởng chính trong việc cải tiến và tối ưu mô hình

Đầu tiên, có thể nói các thuật toán Neural Network Compression (nén mạng) là một nhánh nhỏ trong tập thuật toán tối ưu hóa mô hình (model optimization), nó được sinh ra với mục đích giúp giải quyết bài toán khi triển khai (deploy) các model Deep Learning trên các thiết bị phần cứng không được mạnh mẽ như (mobile devices, edge devices ...).

Với một mô hình deep learning, sẽ luôn có 1 câu hỏi thường trực là liệu model này có khả năng ứng dụng thực tế hay không, có khả năng chạy thời gian thực trên một thiết bị đời thường để khi triển khai, ai cũng có thể sử dụng được không. Thật sự là không dễ dàng gì khi mà chúng ta luôn phải đánh đổi giữa độ chính xác và tốc độ xử lý của 1 mô hình, thông thường độ chính xác cao như ở một số công trình nghiênn cứu hiện đại nhất thường chứa 1 lượng tham số rất lớn (chục triệu đến hàng trăm triệu tham số) dẫn đến việc lưu trữ và tính toán trở lên khó khăn hơn rất nhiều nếu không có các thiết bị hỗ trợ (như GPU), còn một số mô hình quá ít tham số thì dẫn đến việc đôi khi lại không đủ sâu để học được hết các đặc trưng và trả về độ chính xác đủ tốt. Các phương pháp tối ưu mô hình ra đời giúp các mô hình lớn này trở lên gọn nhẹ hơn, nhỏ hơn nhưng vẫn đủ mạnh mẽ như mô hình ban đầu khi đưa chúng chạy trên các ứng dụng thực tế, điều này là rất cần thiết vì không ai trong chúng ta muốn một mô hình tiêu tốn nhiều công đào tạo lại không thể áp dụng vào bất kì ứng dụng nào.

2.2.2 Pruning (Cắt tỉa)

Cắt tỉa mạng được lấy cảm hứng bắt nguồn từ sự cắt tỉa liên kết nơ ron trong não người, nơi các liên kết thần kinh giữa các nơron(axon) bị phân giã hoàn toàn và chết đi xảy ra giữa thời thơ ấu và sự khởi đầu của dậy thì.



Hình 2.8: Cắt tỉa neuron

Bộ não con người lưu trữ thông tin bằng cách tạo ra các liên kết thần kinh. Khi một liên kết không được sử dụng trong một thời gian dài, nó sẽ bị cắt tỉa đi. Việc cắt tỉa này giúp não bộ tiết kiệm năng lượng và tăng hiệu quả hoạt động. Kỹ thuật Pruning trong học máy cũng dựa trên nguyên tắc tương tự. Kỹ thuật này loại bỏ các thành phần dư thừa trong mô hình, giúp mô hình nhỏ gọn và hiệu quả hơn.

Cách thức hoạt động của Pruning:

Xác định ngưỡng: Pruning cần xác định một ngưỡng để phân biệt các kết nối quan trọng và không quan trọng. Ngưỡng này có thể được định nghĩa thủ công hoặc tự động dựa trên độ lệch chuẩn của tập trọng số.

Loại bỏ các kết nối dư thừa: Các kết nối có trọng số nhỏ hơn ngưỡng sẽ bị loại bỏ. Việc này không ảnh hưởng nhiều đến khả năng suy luận của mô hình.

Cập nhật mô hình: Sau khi loại bỏ các kết nối dư thừa, mô hình được cập nhật để phản ánh những thay đổi này.

Lợi ích của Pruning:

- Giảm kích thước mô hình: Pruning giúp giảm kích thước mô hình, giúp tiết kiệm bộ nhớ và tăng tốc độ suy luận.
- Tăng hiệu quả mô hình: Pruning giúp mô hình tập trung vào các kết nối quan trọng, giúp tăng độ chính xác và hiệu quả của mô hình.
- Giảm chi phí đào tạo: Pruning giúp giảm chi phí đào tạo mô hình bằng cách giảm số lượng tham số cần được đào tạo.

2.2.3 Quantization (Lượng tử hóa)

Quantization là kỹ thuật tối ưu hóa việc lưu trữ trọng số trong mạng nơ-ron. Thay vì tập trung vào việc tối ưu hóa giá trị của trọng số, Quantization hướng đến việc giảm số lượng bit cần thiết để biểu diễn chúng mà vẫn đảm bảo độ chính xác của mô hình. Ý nghĩa của việc sử dụng Quantization như sau:

- Giảm kích thước mô hình: Việc sử dụng ít bit hơn để biểu diễn trọng số sẽ giúp giảm kích thước mô hình, tiết kiệm bộ nhớ và tăng tốc độ suy luận.
- Tăng hiệu quả mô hình: Quantization có thể giúp tăng hiệu quả mô hình bằng cách giảm chi phí tính toán và năng lượng tiêu thụ.
- Giảm chi phí triển khai: Mô hình có kích thước nhỏ hơn sẽ dễ dàng triển khai trên các thiết bị có tài nguyên hạn chế.

Cách thức hoạt động của Quantization:

Chuyển đổi định dạng: Quantization chuyển đổi các số dấu phẩy động sang định dạng số dấu phẩy tĩnh. Ví dụ, có thể chuyển đổi từ FP32 (32 bit) sang FP16 (16 bit) hoặc INT8 (8 bit).

Phân cụm trọng số: Quantization sử dụng kỹ thuật phân cụm để nhóm các trọng số có giá trị gần nhau lại với nhau. Mỗi nhóm sẽ được biểu diễn bằng giá trị trung bình, gọi là centroid.

Fine-tuning: Sau khi chuyển đổi định dạng và phân cụm, mô hình cần được fine-tuning để đảm bảo độ chính xác. Quá trình fine-tuning sẽ điều chỉnh các centroid để tối ưu hóa hiệu suất của mô hình.

Lợi ích chính của việc Quantization là giảm kích thước mô hình, Quantization có thể giúp giảm kích thước mô hình xuống 4-8 lần. Tăng tốc độ suy luận, Quantization có thể giúp tăng tốc độ suy luận lên 2-4 lần. Giảm chi phí, Quantization có thể giúp giảm chi phí đào tạo và triển khai mô hình. Lý do mà phương pháp lượng tử hóa được sử dụng đầu tiên, từ suy luận và đào tạo Mạng lưới thần kinh đều chuyên sâu về tính toán. Vì vậy, việc biểu diễn hiệu quả các giá trị số là đặc biệt quan trọng. Thứ hai, hầu hết các mô hình Mạng thần kinh hiện tại đều được tham số hóa quá mức, do đó, có nhiều cơ hội để giảm độ chính xác của bit mà không ảnh hưởng đến độ

chính xác. Tuy nhiên, một sự khác biệt rất quan trọng là NN rất mạnh mẽ đối với quá trình lượng tử hóa tích cực và sự rời rạc hóa cực độ. Mức độ tự do mới ở đây liên quan đến số lượng tham số liên quan, tức là chúng ta đang làm việc với các mô hình được tham số hóa quá mức. Điều này có ý nghĩa trực tiếp đối với việc liệu chúng ta có đang giải quyết tốt các vấn đề được đặt ra hay không, liệu chúng ta có quan tâm đến lỗi tiến hay lùi, v.v. đang được giải quyết. Thay vào đó, người ta quan tâm đến một số loại chỉ số lỗi chuyển tiếp (dựa trên chất lượng phân loại, độ phức tạp, v.v.), nhưng do việc tham số hóa quá mức nên có nhiều mô hình rất khác nhau tối ưu hóa chính xác hoặc gần đúng chỉ số này. Do đó, có thể có lỗi/khoảng cách cao giữa mô hình lượng tử hóa và mô hình không lượng tử hóa ban đầu, trong khi vẫn đạt được hiệu suất tổng quát hóa rất tốt.

2.3 Kết luận chương

Nhìn chung, mục tiêu của các kĩ thuật tối ưu mô hình là có thể nén được các mô hình học sâu phức tạp, chuyển chúng sang các thiết bị phần cứng cơ bản, chấm dứt sự phụ thuộc của chúng vào các tài nguyên tính toán khổng lồ. Đạt được điều này có thể giúp chúng ta nhúng được các mô hình AI và mọi hệ thống chip, nhúng, các thiết bị IoT nhỏ quanh ta. Điều này làm tăng đáng kể tốc độ tính toán của model và giúp nó có thể dễ dàng triển khai trên nhiều thiết bị khác nhau.

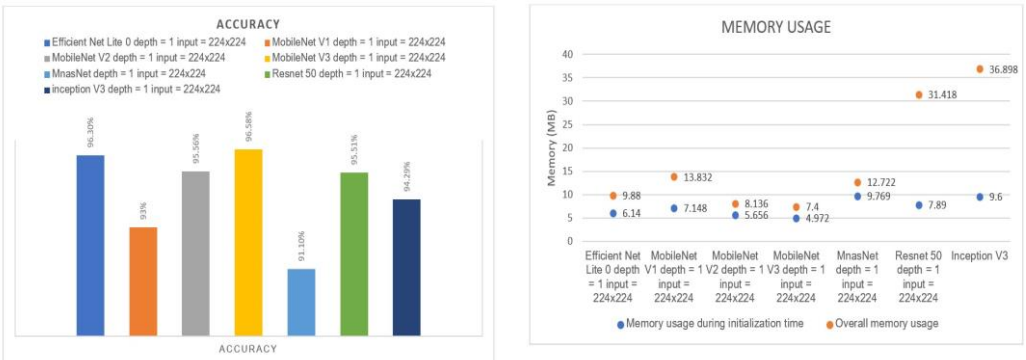
CHƯƠNG 3: ĐỀ XUẤT GIẢI PHÁP CHUẨN ĐOÁN SÂU BỆNH SỬ DỤNG MÔ HÌNH AI TẠI BIÊN

3.1 Đề xuất giải pháp cải tiến mô hình và dữ liệu chuẩn đoán sâu bệnh thông qua mô hình AI tại biên

3.1.1 Đánh giá hiệu năng các mô hình AI được cải tiến để thực thi trên biên mạng

Với mục tiêu tập trung vào các thiết bị biên công suất thấp, đề án khảo sát và đánh giá một số mẫu máy hiện đại và nhẹ nhất hiện nay. Trong đề án, hiệu suất của các mô hình DL đã được đánh giá trên Raspberry Pi 3 Model B. Mặc dù nguồn lực tính toán hạn chế, nền tảng nhúng chi phí thấp này có đủ sức mạnh tính toán để suy luận DNN theo thời gian thực. Với CPU ARM Cortex-A53 1.2GHz 64-bit lõi tứ có thể hoạt động ở tần số từ 700 MHz đến 1.2 GHz. Hệ thống tích hợp RAM 1GB LPDDR2 ở tốc độ 900MHz.

Tương tự như vậy, để giảm tác động của hệ điều hành đến hiệu suất, quá trình khởi động của RPi đã được cấu hình để ngăn chặn các quy trình và dịch vụ không cần thiết được khởi động. Tất cả các thiết bị ngoại vi trong quá trình mô phỏng cũng được ngắt kết nối.

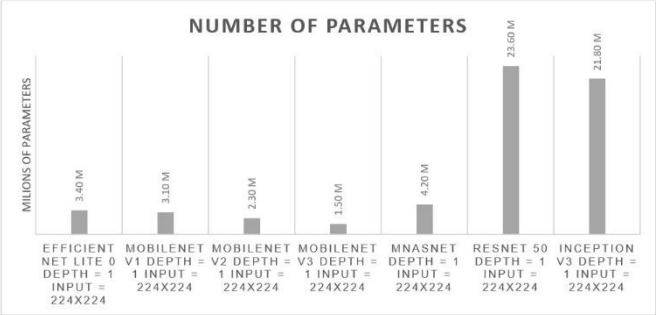


Hình 3.1. So sánh độ chính xác và tài nguyên sử dụng của các mô hình [13]

Độ chính xác: Kết quả được thể hiện trong hình 3.1. Kết quả cho thấy MobileNetV3 đạt độ chính xác cao nhất là 96,58%, cao hơn 0,28% so với vị trí thứ

hai, EfficientNet Lite 0 (96,3%). Điểm chung của hai mạng này là có các thuật toán Neural Architecture Search (NAS), NetAdapt được thiết kế tự động và kế thừa các khối với cấu trúc ngày càng được tối ưu hóa cho Edge / Mobile Device. Do đó, hai kiến trúc mạng phổ biến được thiết kế thủ công khác như InceptionNets 94,29%, ResNets 95,51% hay các phiên bản tiền nhiệm của MobileNetV3 là MobileNetV2 và MobileNetV1 lần lượt đạt 95,56% và 93%. Trong khi đó, EfficientNets Lite sử dụng mô hình EfficientNet B0 nhẹ nhất và kỹ thuật mở rộng mô hình để thu nhỏ và tìm các biến thể của EfficientNets Lite trong khi vẫn đạt được độ chính xác cao và thông số FLOPS được tối ưu hóa. Với MobileNetV1, nó chỉ đơn giản là cải thiện tích chập cổ điển mà không cần nhiều tối ưu hóa về kiến trúc mạng. MobileNetV2 đã được cải thiện với Linear Bottleneck Block và Residual Block nhưng được thiết kế thủ công và không đảm bảo là tối ưu. Với MnasNet, kết quả thu được về độ chính xác khá hạn chế mặc dù kiến trúc mạng được NAS tìm kiếm, nhưng vì phần thưởng cho bộ điều khiển không được tối ưu trong Multi-Objective. Cũng không có phương pháp nào để đánh giá mô hình phù hợp nhất thu được từ kết quả tìm kiếm như trong NetAdapt.

Sử dụng bộ nhớ: Hình 3.1 cho thấy bộ nhớ sử dụng các mô hình được đánh giá. Tất cả các mô hình trong bài báo được đánh giá, cụ thể với MobileNetV3 vượt trội hơn các kiến trúc khác khi chỉ sử dụng 4.972MB trong quá trình khởi tạo model và tổng cộng 7,4MB trong RAM 1GB của RPi 3B. MobileNetV2 và EfficientNet Lite 0 cũng cho kết quả đáng tin cậy khi chúng chỉ chiếm lần lượt 8.136MB và 9,88MB tổng lượng bộ nhớ sử dụng.

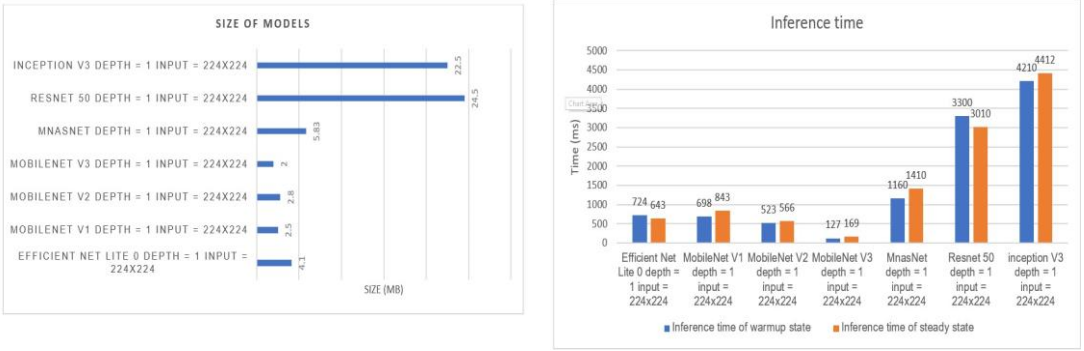


Hình 3.2 Đánh giá số lượng tham số của các mô hình [13]

Số lượng tham số: Hình 3.2 cho thấy kết quả số lượng tham số của mỗi mạng. Số lượng tham số của MobileNetV3 là nhỏ nhất so với các kiến trúc mạng khác (1,5 triệu), vượt trội so với các kiến trúc mạng phổ biến không được tối ưu hóa cho Edge/ Mobile như ResNet 50 (23,6 triệu) và InceptionV3 (21,8 triệu).

Kích thước của mô hình: Lượng tử hóa tham số có thể được sử dụng để giảm kích thước của mô hình. Các mô hình nhỏ hơn có những lợi ích sau:

- Kích thước lưu trữ / tải xuống nhỏ hơn: Các mô hình nhỏ hơn chiếm ít dung lượng lưu trữ hơn trên thiết bị của người dùng do đó yêu cầu ít thời gian và băng thông hơn để tải xuống.
- Sử dụng ít bộ nhớ hơn: Các mô hình nhỏ hơn sử dụng ít RAM hơn khi chúng chạy, giúp giải phóng bộ nhớ để các phần khác trong ứng dụng của bạn sử dụng và có thể chuyển thành hiệu suất và độ ổn định xác định tốt hơn.

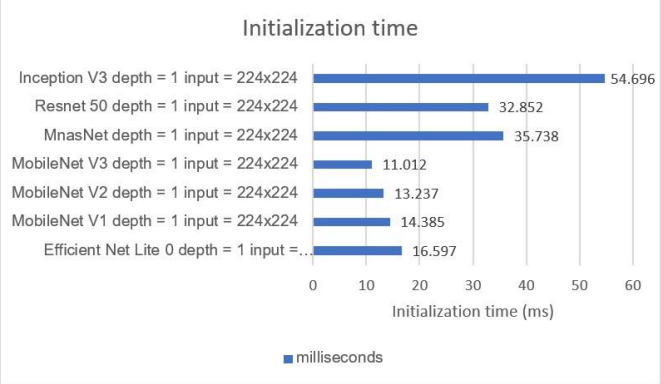


Hình 3.3: Đánh giá kích thước mô hình và tốc độ suy luận của mô hình [13]

Kết quả của lượng tử hóa là kích thước của các mô hình DNN như thể hiện trong hình 5. Các mô hình được thiết kế đặc biệt cho Edge / Mobile là MobileNets, MnasNets, EfficientNets Lite được tham số hóa và tối ưu hóa cho bộ nhớ và kích thước mô hình tốt hơn so với các DNN thường được sử dụng.

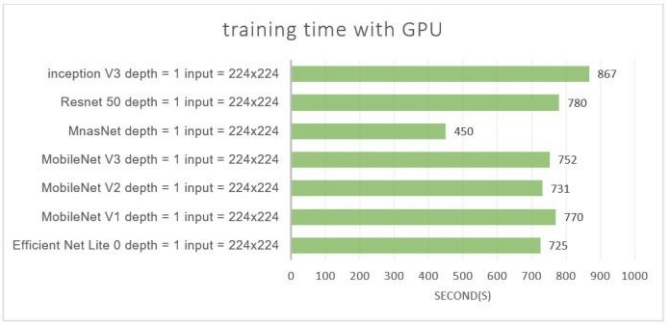
Thời gian suy luận: Độ trễ là khoảng thời gian cần thiết để chạy một suy luận duy nhất với một mô hình nhất định. Một số hình thức tối ưu hóa có thể làm giảm số lượng tính toán cần thiết để chạy suy luận bằng cách sử dụng mô hình, dẫn đến độ trễ thấp hơn. Độ trễ cũng có thể có tác động đến mức tiêu thụ điện năng. Lượng tử hóa không chỉ là một kỹ thuật chuyển đổi có thể làm giảm kích thước mô hình mà còn cải thiện độ trễ của CPU và bộ tăng tốc phần cứng, trong khi bù lại chúng ta chỉ bị suy giảm nhỏ về độ chính xác của mô hình. Hiện tại, Lượng tử hóa mô hình có thể được sử dụng để giảm độ trễ bằng cách đơn giản hóa các tính toán xảy ra trong quá trình suy luận, có khả năng phải trả giá bằng một số độ chính xác. Kết quả dưới đây trong hình 6, MobileNetV3 vẫn cho kết quả đáng kể nhất.

Thời gian khởi tạo: Thời gian khởi tạo mô hình của mobbileNet V3 là nhanh nhất với 11.012 ms, kết quả so sánh được thể hiện bằng hình 3.4



Hình 3.4 So sánh thời gian khởi tạo mô hình trên thiết bị [13]

Thời gian đào tạo: Hình 3.5 so sánh thời gian đào tạo của các mô hình



Hình 3.5: Thời gian đào tạo của các mô hình [13]

Thu nhỏ/tối ưu mô hình: Như vậy sau quá trình đánh giá, đề án đã chứng minh mô hình MobileNetV3 đạt độ hiệu quả cao trong công việc. Để triển khai hiệu quả MobileNetV3 trên các thiết bị cạnh / di động, đề án tiếp tụ giảm độ sâu và độ phân giải ảnh đầu vào. So sánh giữa mô hình thu nhỏ và mô hình bình thường được thể hiện trong Bảng 3.1. Mô hình thu nhỏ trong khi vẫn có độ chính xác thỏa đáng là 94,4%, chỉ bằng một nửa kích thước và mất ít thời gian hơn đáng kể để đào tạo.

Bảng 3.1. Mô hình thu nhỏ [13]

MobileNet V3	Độ sâu = 1,0 Đầu vào = 224x224	Độ sâu = 0,35 Đầu vào = 96x96
Độ chính xác	96.58%	94.4%
Sử dụng bộ nhớ (init/overall)	4.972 MB / 7.4 MB	4.464 MB / 5.644 MB
Thời gian suy luận (khởi tạo/tổng thể)	0,127 giây / 0,169 giây	0, 045 giây / 0, 043 giây
Thời gian khởi tạo	11.012 mili giây	0.815 mili giây
Thời gian đào tạo mô hình	752 giây	166 giây
Kích thước mô hình	2MB	797 KB
Số lượng tham số	1,5 triệu	0,4 triệu

Có thể thấy như kết quả thu nhỏ mô hình ở bảng 3.1, với mô hình MobileNet V3 cơ bản với độ sâu của mạng là 100% và đầu vào của kích thước ảnh là 224 x 224

điểm ảnh đã được cắt giảm bớt các lớp và rút gọn còn 35% độ sâu của mạng, giảm kích thước điểm ảnh đầu vào tới 96 x 96 điểm ảnh. Mô hình sau các quá trình rút gọn, tinh chỉnh và tối ưu, đã đánh đổi chỉ khoảng hơn 2% độ chính xác, từ 96.58% còn 94.4% nhưng giảm thiểu đáng kể bộ nhớ sử dụng (giảm đi khoảng 23.7% từ 7.4MB còn 5.644MB). Song song với đó, thời gian suy luận của mô hình trên thiết bị cũng được cải thiện đáng kể khi chỉ còn 0.045 giây (khi khởi tạo mô hình cho suy luận đầu tiên), sau đó ổn định ở mức 0.043 giây cho toàn bộ quá trình dự đoán. Như vậy, thời gian khởi tạo của mô hình khi chưa rút gọn và sau khi rút gọn lần lượt là 11.012 ms và 0.815 ms. Các thông số tương ứng của hai mô hình cũng có sự chênh lệch rõ rệt về kích thước mô hình giảm còn 797KB so với 2MB ban đầu, số lượng tham số của mô hình giảm chỉ còn 0,4 triệu / 1,5 triệu tham số.

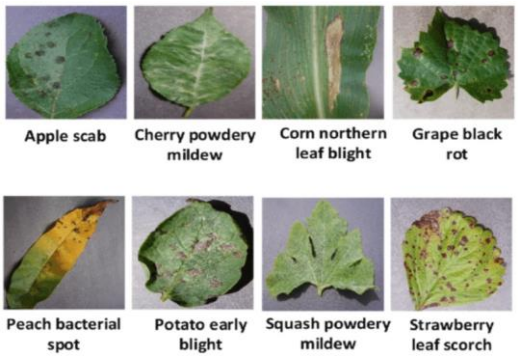
Công việc đánh giá mô hình đã trình bày ở trên được dự đoán sẽ đóng vai trò là tài liệu tham khảo sơ bộ khi lựa chọn các mô hình DNN trong hệ sinh thái rộng lớn của các thành phần học sâu có sẵn. Đề án đã chứng minh rằng một máy tính nhúng giá rẻ như Raspberry Pi 3 Model B có khả năng thực hiện suy luận thời gian thực trên cơ sở các mô hình DNN phức tạp. Qua đó, nội dung phần này thúc đẩy nghiên cứu mới và sâu rộng hơn về việc tích hợp vi điều khiển trong hệ thống Edge Intelligence. Hiệu suất của các mô hình khác nhau cho từng loại phần cứng có thể khác nhau, do đó, cấu trúc mô hình tương thích với phần cứng sẽ dẫn đến hiệu suất tốt hơn. Theo kết quả đánh giá, MobileNetV3 cho kết quả tốt hơn so với các mô hình được đánh giá với bộ dữ liệu dịch hại trên Raspberry Pi 3 Model B. Đề án cũng thử nghiệm thu nhỏ MobileNetV3, cung cấp một mô hình thỏa đáng cho các thiết bị cạnh / biên mạng.

3.1.2 Giải pháp cải tiến mô hình và dữ liệu chuẩn đoán sâu bệnh

Kế thừa từ kết quả đánh giá các mô hình của phần trước, đề án đã chọn ra mô hình tốt nhất để phát triển giải pháp là MobileNetV3. Nội dung phần tiếp theo tập trung vào việc phát triển một phương pháp để tối ưu hóa mô hình phân loại bệnh thực vật bằng DCNN (Mạng nơ-ron tích chập sâu) có tên là MobileNetV3 đã được chứng minh ở phần trước. Các bộ dữ liệu khác nhau được sử dụng bởi các mô hình DCNN

khác nhau có thể tạo ra các kết quả khác nhau. Để chứng minh tính hiệu quả của các mô hình DCNN, các mô hình thường sử dụng bộ dữ liệu phòng thí nghiệm (PlantVillage Dataset) và bộ dữ liệu thực tế (CropedPlantDoc). Mỗi tập dữ liệu đều có những đặc điểm cũng như ưu nhược điểm của nó. Đề án đã làm phong phú dữ liệu bằng cách kết hợp dữ liệu từ hai bộ dữ liệu công khai bao gồm Bộ dữ liệu PlantVillage (PVD) và Bộ dữ liệu CroppedPlant (CPD). Sau đó, mô hình đề xuất sẽ được đào tạo bằng phương pháp học chuyển hai bước.

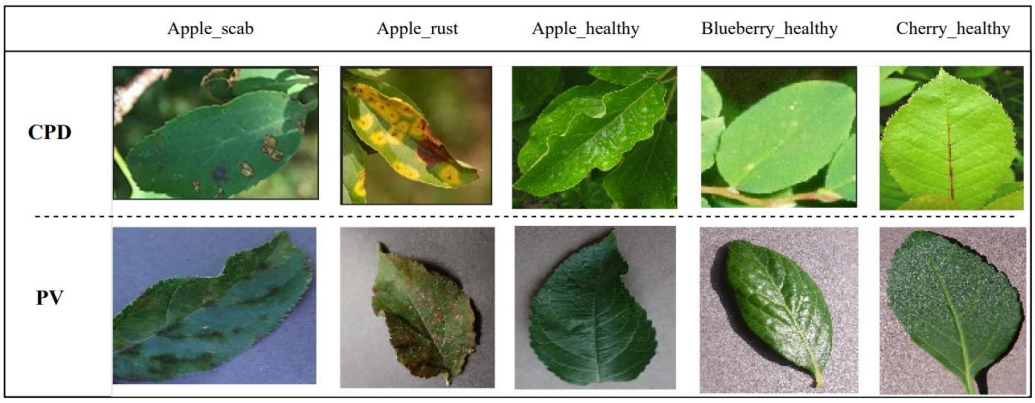
Bộ dữ liệu PlantVillage: Các bộ dữ liệu được sử dụng trong nông nghiệp thường yêu cầu một cơ sở dữ liệu rộng rãi, được xác nhận về hình ảnh của cây khỏe mạnh và bị hư hại để phát triển các bộ phân loại hình ảnh chính xác cho các ứng dụng chuẩn đoán bệnh thực vật. Một tập dữ liệu như vậy đã không tồn tại cho đến gần đây và thậm chí các bộ dữ liệu nhỏ hơn cũng không thể truy cập công khai. Để giải quyết vấn đề này, dự án PlantVillage Dataset đã khởi xướng việc thu thập hàng ngàn hình ảnh về những cây khỏe mạnh và bị bệnh có thể truy cập miễn phí cho công chúng. Tất cả các hình ảnh trong cơ sở dữ liệu Tập dữ liệu PlantVillage được chụp tại các trạm nghiên cứu và phòng thí nghiệm thực nghiệm, với độ sáng, môi trường khác nhau và các cài đặt do người dùng chỉ định khác. Cuối cùng, các thiết bị cuối (người dùng điện thoại thông minh) sẽ chụp ảnh trong các điều kiện "ngẫu nhiên" khác nhau. Hình 3.6 cho thấy một ví dụ về hình ảnh trong tập dữ liệu Planvillage.



Hình 3.6. Một số bệnh thực vật từ Tập dữ liệu PlantVillage

Hơn 50.000 hình ảnh trong bộ dữ liệu Planvillage hiện đang được lưu trữ trên trang web www. PlantVillage Dataset.org và bộ dữ liệu này có thể được truy cập thông qua các trường đại học Hoa Kỳ (Penn State, Florida State, Cornell và các trường khác). Bộ dữ liệu chứa 54.303 hình ảnh của những chiếc lá khỏe mạnh và không khỏe mạnh, được phân loại thành 38 loại dựa trên loài và bệnh. Các loại cây như Táo, Quả việt quất, Anh đào, Ngô, Nho, Cam, Đào, Ớt chuông, Khoai tây, Quả mâm xôi, Đậu nành, Bí ngô, Dâu tây và Cà chua đều được bao gồm trong tập dữ liệu đó. Ngoài ra, hình minh họa của 17 bệnh nấm, bốn bệnh do vi khuẩn, hai bệnh nấm mốc (oomycetes), hai bệnh do virus và một bệnh do ve gây ra cũng được hiển thị trong tập dữ liệu đó. Có hình ảnh của những chiếc lá khỏe mạnh trên 12 loài thực vật sạch bệnh và tổng số lớp trong bộ dữ liệu là 38.

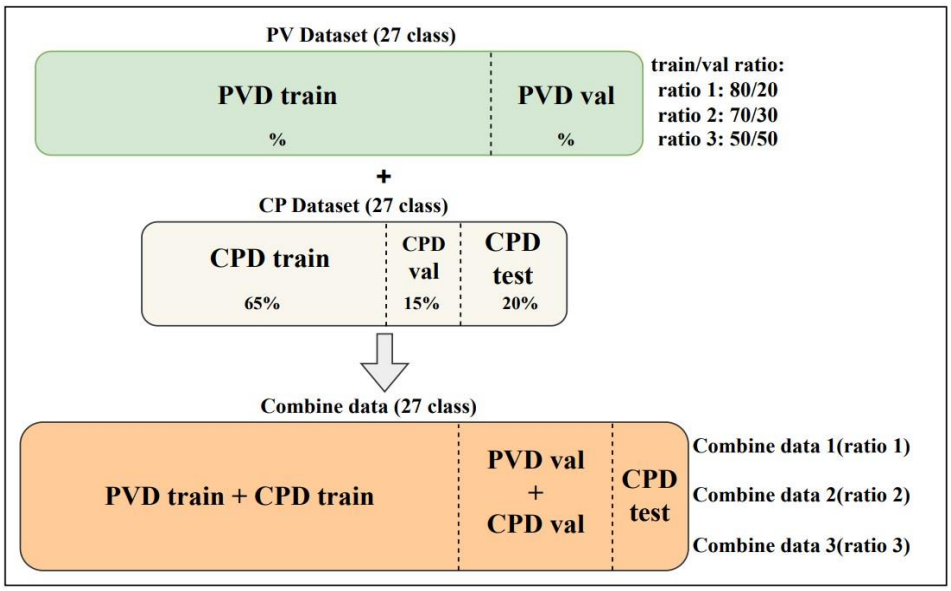
Bộ dữ liệu Cropped-PlantDoc: Singh và các cộng sự đã tạo ra Cropped-PlantDocdataset, chứa 13 loài thực vật và 27 lớp. Tương tự như tập dữ liệu Plant Village, bộ dữ liệu PlantDoc ban đầu bao gồm hình ảnh của mỗi lá. Tuy nhiên, những hình ảnh đó cũng cho thấy nền phức tạp và khu vực được bao phủ bởi các lá mục tiêu khác nhau, điều này khiến việc phân loại khó khăn hơn nhiều so với hình ảnh Tập dữ liệu PlantVillage. Các tác giả cắt thủ công các vùng hình ảnh có chứa các lá đích để giải quyết thiếu sót này. Điều này tạo ra những chiếc lá có khung thuận tiện trong khi tăng đáng kể số lượng mẫu (khoảng 9K) vì một số mẫu từ mỗi hình ảnh PlantDoc gốc có thể được trích xuất (khoảng 2,6K). Hình 3.7 trình bày một số ví dụ về hình ảnh lá trong Tập dữ liệu PlantDoc.



Hình 3.7. Hình ảnh ví dụ về CPD và PVD

Tiền xử lý dữ liệu:

Đề án kết hợp hai bộ dữ liệu trong bài báo này: bộ dữ liệu phòng thí nghiệm (PVD) và bộ dữ liệu được thu thập tự nhiên (CPD). Tuy nhiên, có một vấn đề với số lượng lớp trong hai bộ dữ liệu; tức là PVD có 38 lớp, nhưng CPD chỉ có 27 lớp. Vì 27 lớp trong CPD đều được bao gồm trong PVD và nghiên cứu này nhằm mục đích làm cho mô hình có thể triển khai trong thực tế, đề án sẽ kiểm tra hiệu suất mô hình trên CPD kết hợp với 27 trong số 38 lớp của PVD. Sau khi kết hợp, tập dữ liệu mới sẽ có 27 lớp tương tự như CPD.

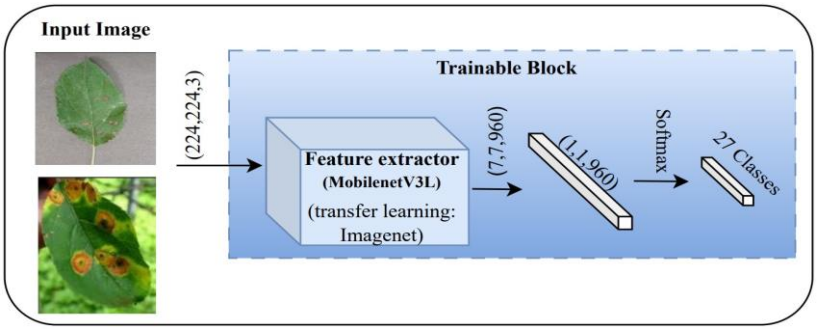


Hình 3.8. Bộ dữ liệu kết hợp của CPD và PVD

Trong nghiên cứu này, đề án kết hợp hai bộ dữ liệu bao gồm PVD và CPD theo cách minh họa trong Hình 3.8 Đầu tiên, PVD được chia ngẫu nhiên thành hai tập con: PVD train và PVD val. Việc phân chia này được thực hiện trong ba trường hợp khác nhau với tỷ lệ phân chia lần lượt là 80:20, 70:30 và 50:50. Tương tự, tập dữ liệu CPD được chia ngẫu nhiên một lần để tạo thành ba tập con: CPD train, CPD val và CPD test với tỷ lệ 65:15:20.

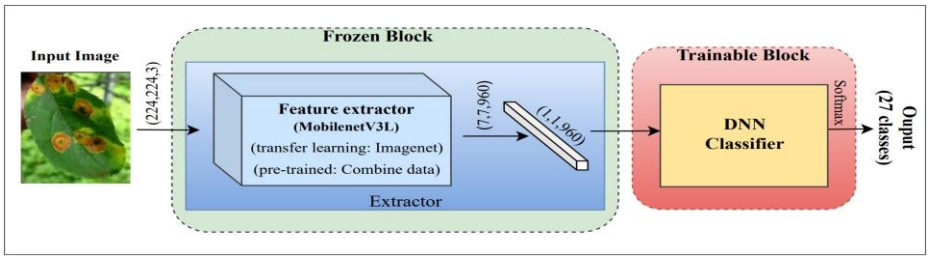
Quá trình hợp nhất dữ liệu chỉ được áp dụng trong các giai đoạn đào tạo và xác nhận. Khi tiến hành thử nghiệm mô hình cuối cùng, đề án tập trung vào tập dữ liệu phức tạp hơn, tức là CPD. húng tôi thực hiện kết hợp theo cặp: tập train CPD được kết hợp với tập train PVD, tập CPD val được kết hợp với PVD val và tập test CPD được giữ nguyên. Kết quả của sự kết hợp này tạo ra một tập dữ liệu có tên là "Combine Dataset", bao gồm ba trường hợp tương ứng với ba tỷ lệ phân chia PVD: Kết hợp dữ liệu 1, Kết hợp dữ liệu 2 và Kết hợp dữ liệu 3. Mỗi dữ liệu Kết hợp này bao gồm ba tập hợp con dữ liệu, được sử dụng để đào tạo, xác thực và thử nghiệm tương ứng. Bằng cách so sánh và phân tích kết quả của mô hình trên mỗi trong ba tỷ lệ phân chia tập dữ liệu PVD, đề án muốn xác định trường hợp phân vùng tối ưu nhất của mô hình của đề án.

Ngoài ra, việc tăng cường dữ liệu cũng được áp dụng trước khi cho huấn luyện mô hình đào tạo. Các kỹ thuật tăng cường được sử dụng bao gồm lật ngẫu nhiên hình ảnh theo chiều ngang, xoay ngẫu nhiên hình ảnh với góc tối đa 30 độ, phóng to ngẫu nhiên hình ảnh với hệ số tối đa 30% và Thay đổi độ tương phản của hình ảnh với hệ số tối đa là 30%. Đào tạo trước của đề án sử dụng trình trích xuất MobileNetV3 được rút gọn để trích xuất các tính năng từ dữ liệu hình ảnh đầu vào (Hình 3.9). Trọng số ban đầu của bộ trích xuất được học từ tập dữ liệu ImageNet. Cuối cùng, bộ trích xuất của MobileNetV3 được đào tạo lại trên tập dữ liệu kết hợp để cập nhật trọng số.



Hình 3.9. Khối trích xuất tính năng của lá

Khi kết thúc quá trình đào tạo trước, mô hình sẽ được tinh chỉnh. Đầu ra của model sẽ được trích xuất và tách hàm softmax ở khối “Trainable Block” trên hình. Điều này giúp kết quả có được tính năng của hình ảnh được trích xuất sau khi đi qua trình trích xuất để xử lý phân loại thêm. Sau khi được đào tạo trước, bộ phân loại DNN được đặt ở đầu ra của mô hình. Cụ thể, 4 lớp Dense với số lượng nút 512, 512, 128, 128 và 27 tương ứng với 27 lớp trong tập dữ liệu, sử dụng chức năng kích hoạt "RELU" được tăng cường nhằm nâng cao khả năng phân loại.

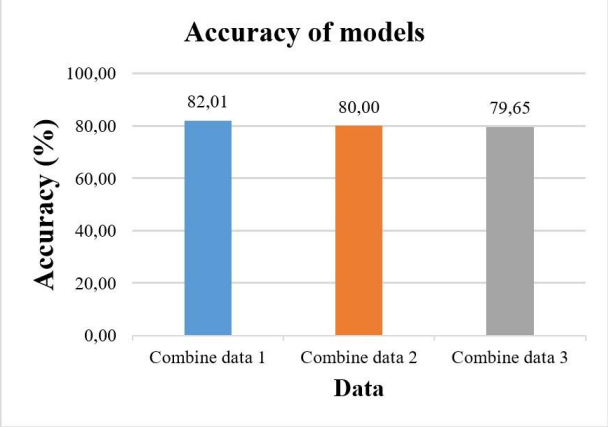


Hình 3.10. Đề xuất mô hình đào tạo với máy trích xuất được đào tạo trước và phân loại DNN

Sau khi thiết lập xong mô hình, đề án đóng băng khối trích xuất đặc trưng đã được tiền huấn luyện phần trước với tập dữ liệu được tăng cường, làm giàu, nhằm giữ nguyên những đặc trưng tốt nhất, sau đó bắt đầu quá trình tinh chỉnh khối phân loại DNN. Trong quá trình tinh chỉnh, đề án tinh chỉnh mô hình trên tập dữ liệu ảnh lá cây trên đồng, ruộng thực tế, chỉ với dữ liệu CPD. Các layer trước đó trong bộ trích xuất đặc trưng sẽ bị đóng băng. Quá trình xác nhận và thử nghiệm mô hình cuối cùng cũng sẽ được thực hiện trên tập dữ liệu ảnh lá cây ngoài thực tế cánh đồng, ruộng, tập test CPD.

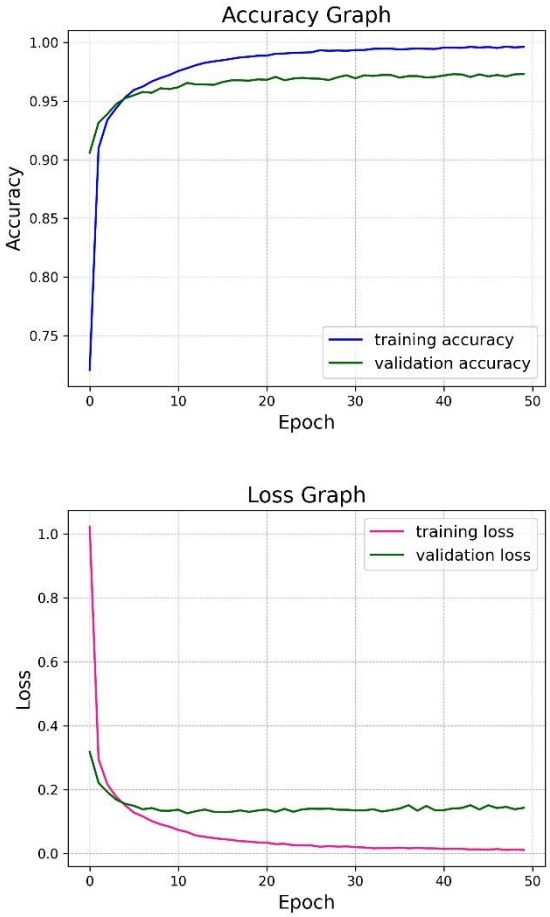
3.2 Các tham số và chỉ tiêu đánh giá & kết quả thu được từ thực nghiệm

Thử nghiệm của đề án được triển khai bằng Python 3.7 với nền tảng TensorFlow và thư viện Keras cho các tác vụ Deep Learning. Các thí nghiệm được thực hiện trên một máy tính có Intel ® Core i9 10900K, GPU Nvidia® RTX A4000 và RAM 48 GB. Sau khi đào tạo, kết quả mô hình trên 3 trường hợp của dữ liệu Kết hợp được trình bày trong Hình 3.11.



Hình 3.11. Độ chính xác của mô hình trên 3 bộ Dữ liệu kết hợp

Như thể hiện trong Hình 3.11, mô hình có thể đạt được độ chính xác cao nhất với tập dữ liệu Kết hợp dữ liệu 1, tương ứng với việc chia dữ liệu trong tập PVD với tỷ lệ 80:20. Vì mô hình đạt được độ chính xác tốt nhất trên Kết hợp dữ liệu 1, trong phần tiếp theo, đề án sẽ chọn kết quả của mô hình trên tập dữ liệu này để so sánh với kết quả của các tác giả khác. Hình 3.12 cho thấy kết quả đào tạo và xác nhận của bộ trích xuất đặc trưng.



Hình 3.12. Độ chính xác và hàm loss của quá trình đào tạo và xác nhận

Biểu đồ chính xác cho thấy quá trình đào tạo diễn ra suôn sẻ trong 10 epoch đầu tiên, do đó mô hình đã học từ khoảng 70% lên đến 96%. Sau 10 epoch đó, độ chính xác đào tạo tăng nhẹ và ổn định ở mức 98%. Tương tự như vậy, độ chính xác xác thực sau 10 epoch cũng bắt đầu ổn định khoảng 96%. Tương tự như đồ thị Độ chính xác, biểu đồ hàm loss trong 8 trong quá trình đào tạo mô hình sau 1- epoch cũng nhanh chóng đạt đến mức tối ưu. Sau epoch thứ 10, giá trị hàm mất mát của cả quá trình đào tạo và xác nhận đã giảm nhẹ và ổn định. Hơn nữa, sau 50 epoch, cả Độ chính xác và hàm Mất mát của mô hình đã đạt đến ngưỡng tối ưu.

Model	Year	Parameters	Accuracy
EfficientNet	2021	664K	64.39%
kEffNet-B0 32ch	2022	1.08M	65.74%
InceptionResNetV2	2020	“	70.53%
Color-Aware Two-Branch	2022	5M	76.91%
MobilenetV3-Large	2022	5M	77.71%

Bảng 3.2 Kết quả đánh giá mô hình base của mobileNet V3 so với các mô hình khác dựa trên tập dữ liệu kết hợp

Kết quả ở Bảng 3.2 đã chỉ ra rằng mô hình mô hình cơ bản mobileNetV3 khi không được tinh chỉnh và làm giàu dữ liệu đạt được kết quả chính xác là 77,71%, cao hơn các mô hình trước đó. Tuy nhiên, mức độ chính xác và hiệu suất của mô hình cơ bản MobileNetV3 vẫn chưa thực sự đáp ứng được dữ liệu thực tế. Do đó cần có một giải pháp huấn luyện hai giai đoạn, với kết quả đạt được trong bảng so sánh sau:

Model	Parameters	Accuracy	F1-Score
MobilenetV3large	5M	77.71%	0.7723
Proposed Model	3.8M	82.01%	0.8194

Bảng 3.3 Kết quả đánh giá mô hình base của mobileNet V3 so với các mô hình được cải tiến dựa trên tập dữ liệu kết hợp

Kết quả so sánh giữa mô hình được đề xuất và mô hình MobileNetV3 được thiết kế cơ bản chưa tinh chỉnh và làm giàu dữ liệu được trình bày ở bảng 3.3. Kết quả thí nghiệm cho thấy độ chính xác của mô hình là 82%, cao hơn so với các nghiên cứu trước đây. Mô hình được đề xuất trong bài viết này đã đạt được các thông số, độ chính xác và điểm F1 tốt hơn so với mô hình cơ bản. Cụ thể, độ chính xác của mô hình được đề xuất tốt hơn 5%, điểm F1-Score của mô hình này cũng cao hơn so với mô hình cơ bản.

3.3 Kết luận chương

Nội dung chương này trình bày một phương pháp để tăng cường dữ liệu đầu vào và phương pháp học tập chuyển giao cho mô hình DCNN dành riêng cho phân loại bệnh lá cây trồng. Công việc đã thực hiện phương pháp học chuyển hai bước bằng cách kết hợp hai bộ dữ liệu có sẵn công khai, PVD và CPD, ở các tỷ lệ khác nhau. Trong quá trình này, các tham số khởi tạo của trình trích xuất tính năng đã được chuyển từ Imagenet.

Sau đó, đề án đã đào tạo trình trích xuất tính năng trên tập dữ liệu kết hợp và áp dụng học tập chuyển giao cho mô hình cuối cùng với bộ phân loại DNN để tinh chỉnh. Kết quả thí nghiệm đã chứng minh rằng phương pháp của đề án cải thiện đáng kể độ chính xác của phân loại bệnh lá cây trồng trong khi vẫn duy trì hiệu quả trong việc sử dụng thông số mô hình, hứa hẹn mở ra những hướng nghiên cứu mới trong tương lai.

III. KẾT LUẬN

Tổng kết lại, đề án tốt nghiệp hướng đến và đã đạt được các kết quả về việc chi tiết hóa và thực hiện xử lý dữ liệu ảnh lá cây bao gồm lá cây cà chua, nghiên cứu và ứng dụng mô hình AI thích hợp trên thiết bị IoT tại biên mạng, từ đó triển khai hệ thống thực nghiệm mẫu đề đưa ra các đối sánh. Nghiên cứu này sử dụng kỹ thuật học chuyển hai bước để giảm chi phí tính toán kết hợp với phương pháp làm giàu dữ liệu bằng cách trộn tập dữ liệu, một cách tiếp cận để tăng tính đa dạng của các bộ dữ liệu và tăng cường khái quát hóa mô hình, trước khi đưa dữ liệu vào mô hình DCNN tối ưu. Đáng chú ý, kết quả đạt được với ít thông số hơn trong khi vẫn duy trì hiệu suất ổn định so với nghiên cứu trước đó. Điều này chứng tỏ rằng mô hình này sử dụng hiệu quả các nguồn lực tính toán hạn chế. Do đó, mô hình được đề xuất có thể được triển khai trên các thiết bị biên để tối ưu hóa tính khả dụng và hiệu quả trong môi trường thực tế, đồng thời góp phần triển khai các dịch vụ nông nghiệp và điện toán biên mới.

Để tiếp tục phát triển, đề án có thể được tiếp tục thực hiện các đề xuất và hướng nghiên cứu tương lai như nghiên cứu và phát triển các giải pháp phần cứng và phần mềm có khả năng tự động hóa các quá trình thu thập dữ liệu, phân tích dữ liệu, v.v., nghiên cứu và phát triển các giải pháp IoT nhúng AI cho các loại cây trồng khác nhau, các điều kiện môi trường khác nhau. Phương pháp luận của đề án có tiềm năng cho các ứng dụng thực tế, chẳng hạn như hỗ trợ nông dân phát hiện và kiểm soát dịch bệnh, hứa hẹn mở ra những hướng nghiên cứu mới trong tương lai về thiết bị biên và cảm biến thông minh.

IV. DANH MỤC CÁC TÀI LIỆU THAM KHẢO

[1] Lương Công Bình. "Ứng dụng công nghệ IoT và AI giám sát và điều khiển nhà nuôi chim yến thông minh." (2023).

[2] Nguyễn Trung Dũng, Nguyễn Tuấn Anh. (2023). "Nông nghiệp thông minh với biến đổi khí hậu và phát triển bền vững: Phân tích chi phí - lợi ích trong trồng hồng không hạt ở tỉnh Hà Giang". Tạp chí Khoa học và Công nghệ Nông nghiệp, 10(2), 227-234.

[3] Nguyễn, Hiếu Nghĩa. "Ứng dụng công nghệ xử lý ảnh kết hợp IoT để theo dõi và phân tích tình trạng quả trên cây cà chua". Diss. 2022.

[4] Lưu Thị Quỳnh Trang, Vương Quang Huy, Vũ Minh Trung, Nguyễn Trường Sơn, Chu Đức Hà, La Việt Hồng, Phạm Minh Triển. (2022). "Công nghệ kết nối trong sản xuất nông nghiệp thông minh và định hướng cho Việt Nam". Tạp chí Khoa học và Công nghệ, 5(2), 85-92.

[5] Bùi Xuân Thiện, Đặng Khuê Văn Nguyễn, and Thanh Hoàng Trần. "Nghiên cứu mô hình nông nghiệp công nghệ cao hỗ trợ nông trại trồng măng cụt." (2022).

[6] Nguyễn Nhuần Hữu. "Chuyển đổi số trong hợp tác xã quốc tế và bài học kinh nghiệm cho Việt Nam." Tạp chí Kinh tế và Phát triển 305 (2) (2022): 58-68.

[7] Nghe, N. T., Ngôn, N. C., & Hòa, N. H. (2022). "Một số mô hình ứng dụng công nghệ 4.0 hỗ trợ nông nghiệp, thủy sản thông minh". Tạp chí Khoa học Đại học cần Thơ, 58(SDMD), 42-47.

[8] Bộ Nông nghiệp và Phát triển Nông thôn. (2022). "Chiến lược phát triển nông nghiệp thông minh giai đoạn 2021-2030, tầm nhìn đến năm 2045". Hà Nội: Bộ Nông nghiệp và Phát triển Nông thôn.

[9] Bộ Thông tin và Truyền thông. (2023). "Phát triển nông nghiệp ứng dụng công nghệ cao". Báo cáo, Bộ Thông tin và Truyền thông.

[10] Martin Otieno, "An extensive survey of smart agriculture technologies: Current security posture", World Journal of Advanced Research and Reviews, 2023, 18(03), 1207–1231.

[11] Hoang Trong Minh, Tuan Pham Anh, et al. "A novel light-weight dcnn model

for classifying plant diseases on internet of things edge devices.” In MENDEL, volume 28, pages 41–48, 2022

[12] .Van-Nhan Nguyen, Tuan-Anh Pham, Thanh-Tra Nguyen, Thu-Anh Pham, Trong-Minh Hoang, "An Efficiency Edge-based Plant Disease Detection Model Using Enriched Dataset and Deep Convolutional Neural Network", 2023 RIVF International Conference on Computing and
Communication Technologies (RIVF) - Image, Computer Vision, Pattern Recognition (R2).

[13] . Anh T. Pham (Presenter), Duc T.M. Hoang, “A Benchmark of Deep Learning Models for Multi-leaf Diseases for Edge Devices”, 2021 International Conference on Advanced Technologies for Communications (ATC), Ho Chi Minh City, Vietnam October 14-16, 2021.

[14] Tuan Nguyen gia, Qingqing L., Jorge Peña Queralta, Zhuo Zou, “Edge AI in Smart Farming IoT: CNNs at the Edge and Fog Computing with LoRa.”, IEEE AFRICON 2019.

[15] Jules Degila, Ida Sèmévo Tognisse, Anne-Carole Honfoga, A Survey on Digital Agriculture in Five West African Countries, Agriculture 2023, 13, 1067.

[16] Indira, P., Arafat, I.S., Karthikeyan, R. et al. Fabrication and investigation of agricultural monitoring system with IoT & AI. SN Appl. Sci. 5, 322 (2023).

[17] Silke Hemming * , Feije de Zwart, Anne Elings , Anna Petropoulou and Isabella Righin, Cherry Tomato Production in Intelligent Greenhouses—Sensors and AI for Control of Climate, Irrigation, Crop Yield, and Quality, Sensors 2020, 20(22), 6430.