Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XIV về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), TP. HCM, ngày 23-24/12/2021
DOI: 10.15625/vap.2021.0050

PHÂN LOẠI ẢNH SÂU BỆNH VỚI EFFICIENTNET VÀ POWER MEAN SVM

Đoàn Thanh Nghị

Trường Đại học An Giang, Đại học Quốc gia Thành phố Hồ Chí Minh

dtnghi@agu.edu.vn

TÓM TẮT: *Phát hiện sớm và phân loại chính xác sâu bệnh hại cây trồng là rất hữu ích cho việc kiểm soát dịch hại, nâng cao chất lượng và sản lượng nông sản. Vì vậy nghiên cứu một hệ thống tự động phân loại hiệu quả sâu bệnh bằng ảnh là rất quan trọng và cần thiết. Hiện nay đa số các nghiên cứu tập trung vào sử dụng mô hình mạng nơron tích chập học sâu để phân loại ảnh côn trùng, trong đó hàm softmax được sử dụng để dự đoán phân lớp và tối thiểu hóa cross-entropy loss. Trong bài báo này, chúng tôi trình bày một phương pháp tiếp cận mới để phân loại ảnh côn trùng bằng cách kết hợp EfficientNet với Power Mean SVM. Trong đó mô hình mạng EfficientNet được tinh chỉnh với phương pháp phù hợp và huấn luyện lại trên tập dữ liệu mới. Mô hình mạng kết quả này sau đó được sử dụng để trích xuất các đặc trưng ảnh trong tập dữ liệu. Trong giai đoạn phân loại ảnh, chúng tôi thay thế hàm softmax bởi bộ phân lớp Power Mean SVM. Do đó quá trình học sẽ tập trung vào tối thiểu hóa margin-based loss thay vì cross-entropy loss. Phương pháp do chúng tôi đề xuất đã được đánh giá trên nhiều tập dữ liệu ảnh sâu bệnh kiểm chuẩn. Kết quả từ các thí nghiệm đã chứng tỏ phương pháp của chúng tôi là hiệu quả và có độ chính xác cao hơn các phương pháp phân loại ảnh sâu bệnh tốt nhất hiện nay. Cụ thể, độ chính xác của mô hình kết hợp EfficientNet với Power Mean SVM là 99% trên tập dữ liệu Xie24, 99% trên tập dữ liệu D0 và 71,84% trên tập dữ liệu IP102. Theo kiến thức của chúng tôi thì đây là kết quả tốt nhất hiện nay trên các tập dữ liệu này.*

Từ khóa: *Thị giác máy tính, máy học vector hỗ trợ, mạng nơron tích chập học sâu, phân loại ảnh sâu bệnh.*

I. GIỚI THIỆU

Côn trùng đóng vai trò quan trọng trong phát triển nông nghiệp bền vững. Hiện nay chỉ có trên một triệu loài được biết đến và mô tả trong tổng số khoảng hơn 30 triệu loài [1]. Tuy nhiên chỉ có khoảng 500 loài phá hoại mùa màng, số còn lại là côn trùng có ích, chúng tiêu diệt côn trùng có hại, bảo vệ cây trồng. Do thiếu kiến thức chuyên môn, nông dân khó xác định chính xác loài sâu bệnh gây hại, do đó họ sử dụng thuốc trừ sâu không phù hợp. Vì vậy, việc nghiên cứu một hệ thống tự động phân loại ảnh sâu bệnh hiệu quả là rất quan trọng. Trong những năm gần đây, nhiều hệ thống nhận dạng sâu bệnh với sự hỗ trợ của máy tính đã được nghiên cứu [2]. Trong đó việc rút trích đặc trưng ảnh để phân loại ảnh sâu bệnh là rất quan trọng. Chúng ta có thể phân loại thành hai nhóm sau: (1) Phương pháp rút trích đặc trưng ảnh được thiết kế thủ công và (2) Phương pháp rút trích đặc trưng ảnh tự động bằng cách sử dụng mạng nơron tích chập học sâu (DCNN). Các phương pháp tạo đặc trưng ảnh thủ công như SIFT [3], HOG [4] là hiệu quả khi biểu diễn các đặc trưng ảnh cấp thấp như màu sắc (color), cạnh (edge), kết cấu (texture). Samanta và cộng sự [5] đã sử dụng phương pháp lựa chọn đặc trưng ảnh dựa trên hệ số tương quan và mạng DCNN để chẩn đoán 8 loài sâu bệnh gây hại cây chè trên tập dữ liệu với 609 ảnh. Trong nghiên cứu [6], [7] bộ phân loại SVM được áp dụng để xác định ruồi trắng (whiteflies), rệp (aphids) và bọ trĩ (thrips) trong ảnh lá cây. Các phương pháp này trích xuất các đặc trưng ảnh nổi trội để tạo vector đại diện cho các ảnh sâu bệnh, sau đó đánh giá trên các tập dữ liệu nhỏ. Tuy nhiên, thực tế số lượng các loài sâu bệnh là rất lớn. Vì vậy, việc thiết kế các bộ rút trích đặc trưng ảnh thủ công là không hiệu quả và đặc trưng ảnh không có mức trừu tượng cao. Hiệu suất phân lớp của các phương pháp này vì vậy phụ thuộc nhiều vào đặc điểm của từng tập dữ liệu, thường là những tập dữ liệu nhỏ, dẫn đến khả năng tổng quát và độ chính xác kém.

Ngược lại, các mạng DCNN đã khắc phục được nhược điểm của các phương pháp trên. Trong đó việc rút trích đặc trưng ảnh được thực hiện tự động bởi DCNN, kết quả là đặc trưng ảnh có khả năng biểu diễn ảnh ở mức trừu tượng cao hơn. Các mạng GoogleNet [8] và ResNet [9] đã cho thấy độ chính xác rất cao trong các bài toán phân loại hình ảnh. Gần đây các phương pháp dựa trên DCNN đã được áp dụng nhiều hơn trong nông nghiệp như xác định cỏ dại, nhận dạng thực vật và phân loại bệnh cây trồng [10]. Phân loại ảnh sâu bệnh cũng là lĩnh vực mà DCNN cho kết quả tốt hơn so với các phương pháp truyền thống khác. Có thể sơ lược các nghiên cứu sau: [11] đề xuất mô hình dựa trên VGG-19 và Region Proposal Network (RPN) để phân loại ảnh sâu bệnh. Trong đó VGG-19 được dùng để trích xuất đặc trưng ảnh sâu bệnh, RPN được dùng để phát hiện vị trí sâu bệnh trong ảnh. Phương pháp của họ cho kết quả vượt trội hơn Single Shot Multibox và Fast Region-based Convolutional Neural Network về mAP. [12] đã đề xuất một DCNN được huấn luyện từ đầu và đánh giá trên ba tập dữ liệu sâu bệnh NBAIR, Xie24 [13], Xie40 [14]. Phương pháp của họ đã được so sánh với các DCNN khác được huấn luyện trên ImageNet [15]. Ngoài ra, họ đã phân tích các tham số như kích thước lô, số lần lặp và tỷ lệ học. [16] Đã thu thập ảnh sâu bệnh từ nhiều nguồn khác nhau để xây dựng một tập dữ liệu quy mô lớn IP102. Họ đã đánh giá tập dữ liệu này bằng các phương pháp máy học truyền thống và các kỹ thuật học sâu DCNN. Độ chính xác phân loại tốt nhất là 49,5% với mạng ResNet-50 [17]. Nghiên cứu [18] đã đề xuất một phương pháp tạo đặc trưng ảnh bằng cách tái sử dụng cấu trúc khối dư để cải thiện khả năng biểu diễn ảnh. Họ đã đánh giá hiệu suất trên tập dữ liệu IP102 với độ chính xác đạt được là 55,24%. [19] Đã thực hiện việc tinh chỉnh và học chuyển tiếp với 7 mô hình mạng trên tập dữ liệu D0 [14]. Sau đó ba mô hình có độ chính xác tốt nhất được tích hợp thông qua chiến lược tổng xác suất tối đa để tăng hiệu suất phân loại. Phương pháp của họ đã được đánh giá trên tập dữ

liệu IP102 với độ chính xác lên đến 67,13%. Mặc dù DCNN đã được áp dụng thành công trong phân loại ảnh sâu bệnh, nhưng vẫn còn nhiều thách thức trong nghiên cứu các mô hình mạng hiệu quả với độ chính xác cao hơn.

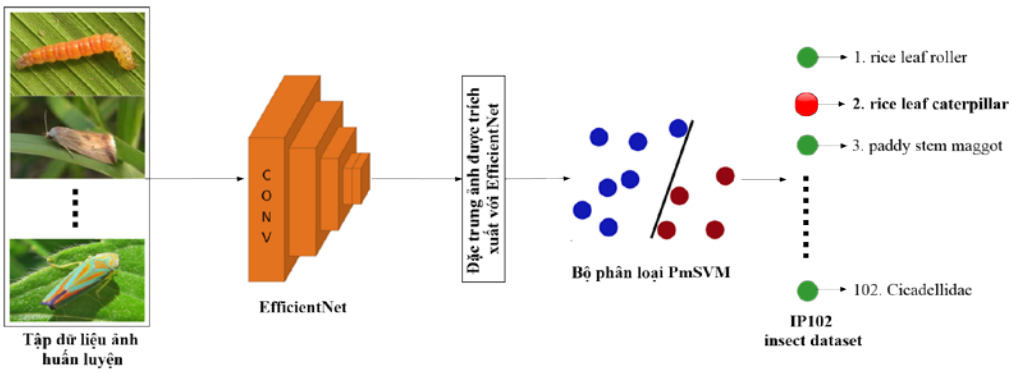
Trong bài báo này chúng tôi đề xuất một phương pháp mới nhằm cải thiện hơn nữa độ chính xác phân loại ảnh sâu bệnh, đặc biệt với các tập dữ liệu lớn. Phương pháp của chúng tôi sử dụng mạng EffcientNet [20] được tinh chỉnh và học chuyển tiếp trên các tập dữ liệu mới. Mô hình mạng kết quả này sau đó được sử dụng để trích xuất đặc trưng ảnh của các tập dữ liệu. Trong đó kỹ thuật xử lý dữ liệu theo lô được sử dụng để tránh nạp toàn bộ dữ liệu vào bộ nhớ chính của máy tính. Cuối cùng chúng tôi sử dụng bộ phân lớp Power Mean SVM (PmSVM) [21] để thay thế cho hàm *softmax* trong giai đoạn phân loại ảnh sâu bệnh. Cách tiếp cận của chúng tôi vì vậy sẽ tập trung vào tối thiểu hóa *margin-based loss* thay vì *cross-entropy loss*, hơn nữa chúng tôi sử dụng mô hình SVM với *kernel* phi tuyến tính thay vì tuyến tính như trong [22]. Các đóng góp của chúng tôi trong bài báo này bao gồm:

- Một phương pháp tinh chỉnh và huấn luyện EfficientNet trên tập dữ liệu ảnh sâu bệnh. Quá trình huấn luyện gồm hai giai đoạn, giai đoạn đầu chỉ huấn luyện các đầu ra lớp FC (full connected layer), giai đoạn hai một số lớp tích chập được rã băng và huấn luyện với tỷ lệ học nhỏ hơn. Nhờ đó mạng hội tụ nhanh hơn, cải thiện được hiệu suất.

- Một phương pháp kết hợp mạng EfficientNet với PmSVM, trong đó bộ phân lớp DCNN với hàm *softmax* được thay thế bằng bộ phân lớp SVM với Power mean kernel [21]. Điều này giúp quá trình học tập trung vào việc tối thiểu hóa *margin-based loss* thay vì *cross-entropy loss*, kết quả đã cải thiện được hiệu suất của mô hình phân loại.

- Một phương pháp trích xuất đặc trưng ảnh trên các tập dữ liệu quy mô lớn. Phương pháp này chia tập dữ liệu thành nhiều lô nhỏ và xử lý trực tuyến. Do đó, cách tiếp cận này có thể dễ dàng áp dụng trên các tập dữ liệu lớn với nhiều lớp và dữ liệu huấn luyện lớn hơn dung lượng bộ nhớ chính của máy tính.

Phương pháp của chúng tôi đã được đánh giá trên các tập dữ liệu Xie24 [13], D0 [14] và IP102 [16]. Thực nghiệm cho thấy phương pháp của chúng tôi tốt hơn các phương pháp tốt nhất hiện nay với độ chính xác 99% trên Xie24, 99% trên D0 và 71,84% trên IP102. Theo kiến thức của chúng tôi thì đây là kết quả tốt nhất hiện nay trên các tập dữ liệu này. Ngoài ra phương pháp này có thể dễ dàng áp dụng trên các tập dữ liệu lớn, có kích thước lớn hơn dung lượng bộ nhớ chính của máy tính. Hình 1 mô tả sơ đồ tổng quát hệ thống phân loại ảnh sâu bệnh do chúng tôi đề xuất.



Hình 1. Sơ đồ tổng quát hệ thống phân loại ảnh sâu bệnh với sự kết hợp của EfficientNet và Power mean SVM

Phần còn lại của bài báo được tổ chức như sau. Phần II trình bày cơ sở lý thuyết và phương pháp của chúng tôi, trong đó trình bày các tập dữ liệu ảnh sâu bệnh, các mô hình DCNN và phương pháp học chuyển tiếp được sử dụng để đánh giá mô hình, trực quan hóa các đặc trưng ảnh được trích xuất bởi các DCNN, máy học vector hỗ trợ PmSVM và sự kết hợp của nó với EfficientNet. Phần III trình bày kết quả các thí nghiệm đạt được trong nghiên cứu và thảo luận. Kết luận và hướng phát triển của bài báo được trình bày trong Phần IV.

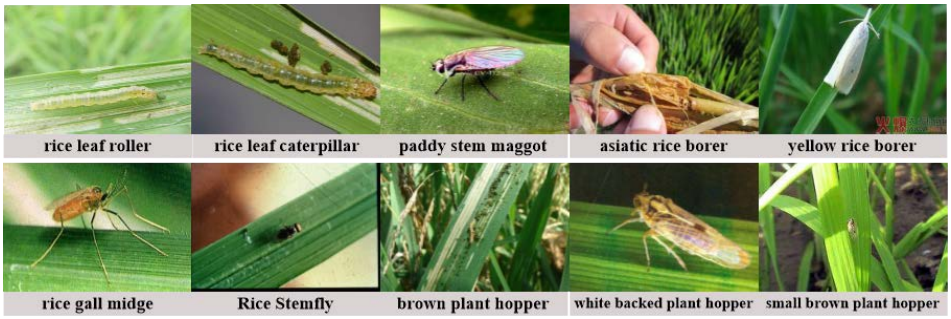
II. CƠ SỞ LÝ THUYẾT VÀ PHƯƠNG PHÁP

Trong nghiên cứu này, một phương pháp phân loại ba giai đoạn được đề xuất cho hệ thống phân loại ảnh sâu bệnh, như được trình bày trong hình 1. Đầu tiên, các mô hình DCNN với các trọng số đã đào tạo trước trên tập dữ liệu ImageNet [15] được tinh chỉnh và đào tạo lại bằng cách sử dụng phương pháp học chuyển tiếp trên các tập dữ liệu sâu bệnh Xie24, D0, IP102. Trong giai đoạn thứ hai, mô hình có độ chính xác tốt nhất được chọn để trích xuất đặc trưng ảnh từ các tập dữ liệu. Cuối cùng là giai đoạn xây dựng mô hình phân loại ảnh sâu bệnh, chúng tôi sử dụng PmSVM để huấn luyện mô hình phân loại phi tuyến tính dựa trên tập đặc trưng ảnh đã trích xuất ở giai đoạn hai, trong đó có so sánh với các bộ phân lớp SVM khác để lựa chọn mô hình có độ chính xác tốt nhất.

A. Các tập dữ liệu ảnh sâu bệnh

Chúng tôi sử dụng ba tập dữ liệu ảnh côn trùng kiểm chuẩn sau đây để đánh giá mô hình đề xuất: 1) Tập dữ liệu Xie24 được đề xuất bởi Xie và cộng sự [13], nó bao gồm 1600 hình ảnh RGB của 24 loài côn trùng khác nhau, các ảnh đã được tiền xử lý để có độ phân giải 227 x 227. Tương tự như Thenmozhi và cộng sự [12], trước khi huấn luyện một mạng DCNN, chúng tôi sử dụng kỹ thuật *tăng cường dữ liệu* để tăng số lượng và sự đa dạng của các tập dữ liệu ảnh. Điều này đã giúp cải thiện hiệu suất mô hình và tránh *overfitting* trong khi huấn luyện [23], [24]. Tập dữ liệu Xie24 sau

khi biến đổi có số lượng 6892 ảnh. Tập dữ liệu này được chia thành ba tập con: 70% mẫu dùng để huấn luyện, 10% để kiểm thử và còn lại để đánh giá mô hình. Kết quả chúng tôi có 4653 ảnh huấn luyện, 516 ảnh kiểm thử và 1723 ảnh kiểm tra; 2) Tập dữ liệu D0 được đề xuất bởi [14] bao gồm 40 lớp côn trùng khác nhau, chứa 4508 ảnh RGB có độ phân giải 200 × 200. Chúng tôi cũng chia tập dữ liệu ảnh này thành ba tập con với tỷ lệ tương tư như Xie24. Kết quả có 2682 ảnh huấn luyện, 473 ảnh kiểm thử và 1353 ảnh kiểm tra; 3) IP102 [16] là tập dữ liệu côn trùng quy mô lớn với 102 loài sâu bệnh hại cây trồng. Tập dữ liệu này chứa 45.095 ảnh huấn luyện, 7.508 ảnh kiểm thử và 22.619 ảnh để đánh giá mô hình. Hình 2 trình bày một số ảnh mẫu từ IP102. Như mô tả trong [16], tập dữ liệu này chứa nhiều yếu tố ảnh hưởng đến hiệu suất của mô hình phân loại ảnh khi so sánh với hai tập dữ liệu Xie24 và D0. Thứ nhất, các loài sâu bệnh rất khó để phân biệt vì màu sắc của chúng và hậu cảnh tương tự nhau. Thứ hai, mỗi lớp chứa hình ảnh của cả vòng đời của một loài sâu bệnh và vì vậy rất khó phân loại, đặc biệt là giai đoạn ấu trùng. Thứ ba, nhiều loài sâu bệnh khác nhau nhưng có ảnh tương đồng nhau. Đây là những yếu tố gây nhiều khó khăn và thách thức khi thiết kế các thuật toán phân loại ảnh trên tập dữ liệu này. Độ chính xác phân lớp tốt nhất hiện nay trên tập dữ liệu IP102 là 67,13% [19].



Hình 2. Các hình ảnh mẫu từ tập dữ liệu sâu bệnh quy mô lớn IP102

B. Mạng nơron tích chập học sâu

1. ResNet

ResNet [17] là một kiến trúc mạng được đề xuất để giải quyết vấn đề một số lớp phi tuyến không học được các *activation map* của ảnh. ResNet được thiết kế với mô hình mạng dựa trên nhiều đơn vị dư được xếp chồng lên nhau. Các đơn vị dư này được sử dụng như các khối nguyên liệu để xây dựng mạng. Các đơn vị này bao gồm các lớp tích chập và *pooling*. Kiến trúc này sử dụng bộ lọc 3 × 3 với ảnh đầu vào có kích thước là 224 × 224 điểm ảnh. Trong ResNet, quá trình lan truyền ngược đã khắc phục được sự suy thoái *gradient*. ResNet có các *shortcut* song song với các lớp tích chập thông thường sẽ giúp mạng hiểu được các đặc trưng ảnh toàn cục. *Shortcut* được sử dụng để thêm vector đầu vào x của lớp trước đó vào đầu ra của lớp kế tiếp sau một vài lớp trọng số. Các *shortcut* này cho phép mạng bỏ qua các lớp không hữu ích, dẫn đến số lượng lớp được điều chỉnh tối ưu hơn, qua đó giúp quá trình huấn luyện nhanh hơn. ResNet đã được đánh giá để phân loại sâu bệnh trong nhiều nghiên cứu [19], [25].

2. DenseNet

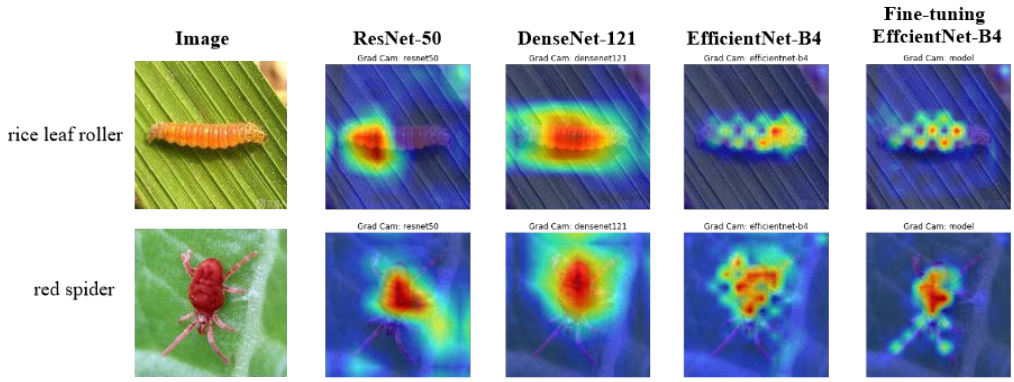
DenseNet [26] là một mạng tích chập lớn có hiệu suất cao, mỗi lớp đều được kết nối với tất cả các lớp trước đó. Nó đề xuất các kết nối trực tiếp giữa hai lớp bất kỳ với cùng kích thước *feature map*. DenseNet có khả năng mở rộng quy mô đến hàng trăm lớp mà không gặp khó khăn trong quá trình tối ưu hóa. Trong các thí nghiệm, DenseNet thường có độ chính xác ổn định khi số lượng tham số ngày càng tăng và không có dấu hiệu suy giảm hiệu suất hoặc *overfitting* trong quá trình huấn luyện. Trong nhiều trường hợp, DenseNet vẫn đạt được kết quả tốt nhất trên nhiều tập dữ liệu kiểm chuẩn, mặc dù nó có ít tham số và chi phí tính toán ít hơn đáng kể khi so sánh với các mô hình mạng khác. Kiến trúc này giúp *gradient* biến thiên tốt và sử dụng lại các đặc trưng một cách hiệu quả. DenseNet đã chứng tỏ được hiệu suất cao trong các thí nghiệm về phân loại ảnh sâu bệnh [25].

3. EfficientNet

EfficientNet [20] đã đề xuất một phương pháp mở rộng phức hợp mới, sử dụng một hệ số kết hợp để mở rộng đồng nhất chiều rộng, chiều sâu và độ phân giải của kiến trúc mạng. Các tác giả của EfficientNet cho rằng để đạt được độ chính xác tốt hơn, mạng có thể được mở rộng bằng cách tăng số lớp hoặc làm cho mỗi lớp rộng hơn, hoặc ảnh đầu vào có độ phân giải cao hơn hoặc kết hợp tất cả các yếu tố này. Neural Architecture Search được sử dụng để mở rộng kiến trúc mạng cơ sở EfficientNet-B0. Qua đó EfficientNet-B0 đã được mở rộng từ EfficientNet-B1 đến EfficientNet-B7. Kết quả cho thấy các EfficientNet đạt được hiệu quả và độ chính xác cao hơn các DCNN khác. Hơn nữa, EfficienNet đã giảm đáng kể số lượng các tham số huấn luyện mạng. [20] cũng đã chứng minh tính hiệu quả của phương pháp này khi mở rộng áp dụng cho MobileNet và ResNet. EfficientNet-B7 đã đạt được độ chính xác tốt nhất top-1 84,3% trên ImageNet, trong khi kích cỡ mạng nhỏ hơn 8,4 lần và nhưng nhanh hơn 6,1 lần so với các DCNN tốt nhất hiện nay. EfficientNet cũng đạt được độ chính xác tốt nhất trên CIFAR-100 (91,7%), Flowers (98,8%). Trong bài báo này, chúng tôi nghiên cứu một phương pháp tinh chỉnh thích hợp nhằm cải thiện hiệu suất phân loại của EfficientNet trên ba tập dữ liệu ảnh sâu bệnh như trình bày trong Phần A, đặc biệt là trên tập dữ liệu quy mô lớn IP102.

4. Trực quan hóa đặc trưng ảnh với Class Activation Map

Mặc dù phương pháp học sâu đã chứng tỏ độ chính xác cao trong phân loại ảnh, nhưng một trong những vấn đề của phương pháp này là khả năng diễn giải mô hình, thành phần cốt lõi để thông hiểu và gỡ lỗi. Mô hình học sâu thường được xem là "hộp đen" và vì vậy chúng ta không có thông tin về 1) Mạng đang "nhìn" ở đâu trong ảnh đầu vào, 2) Các nơron nào được kích hoạt để chuyển tiếp trong quá trình suy luận hoặc dự đoán, 3) Làm thế nào mạng đưa ra quyết định cuối cùng. Do đó, để đảm bảo mô hình mạng hoạt động đúng đắn, chúng ta cần có công cụ để gỡ lỗi và kiểm tra mô hình mạng đang "nhìn" và "kích hoạt" các vị trí nào trong ảnh đầu vào. Vì vậy, Zhou và cộng sự [27] đã đề xuất một kỹ thuật được gọi là Class Activation Mapping (CAM), cho phép các DCNN thực hiện định vị đối tượng mà không cần sử dụng bất kỳ một *bounding box* nào. Sử dụng CAM, chúng ta có thể kiểm tra một cách trực quan vị trí mạng đang "nhìn", xác định chính xác đối tượng trong hình ảnh và các kích hoạt xung quanh nó. Trong bài báo này, chúng tôi trình bày các hình ảnh trực quan CAM của ResNet-50, DenseNet-121, EfficientNet-B4 và mô hình tinh chỉnh của nó trên các ảnh sâu bệnh. Hình 3 cho thấy EfficienNet-B4 có thể "nhìn" các đối tượng sâu bệnh tốt hơn. Các hình ảnh mẫu này được chọn ngẫu nhiên từ tập dữ liệu kiểm thử của IP102. Hơn nữa, EfficientNet-B4 với chiến lược tinh chỉnh của chúng tôi (cột cuối cùng ở hình 3) đã tập trung vào các vùng có liên quan của đối tượng sâu bệnh với nhiều chi tiết nét, trong khi các DCNN khác "nhìn" thiếu chi tiết hoặc không thể "nhìn" được tất cả các đối tượng trong ảnh.



Hình 3. Class Activation Map (CAM) [27] của mô hình DCNN với các phương pháp khác nhau - Phương pháp của chúng tôi cho phép mô hình (cột cuối cùng) tập trung vào các vùng có liên quan của đối tượng sâu bệnh với mức độ chi tiết nét tốt hơn

C. Học chuyển tiếp

Học chuyển tiếp (transfer learning) là một kỹ thuật sử dụng lại kiến thức đã học làm điểm khởi đầu khi giải quyết các bài toán tương tự khác [28]. Mục đích là để có đường cong học tập với điểm xuất phát cao hơn, độ dốc cao, tiệm cận tốt hơn, qua đó giúp giảm thời gian huấn luyện mô hình mạng. Các trọng số của mạng trước đó được sử dụng làm giá trị khởi đầu để huấn luyện mạng mới. Trong bài báo này, chúng tôi sử dụng các bộ lọc của mạng có độ phân biệt tốt đã được huấn luyện trên ImageNet. Điều quan trọng khi học chuyển tiếp là cần xác định phương pháp tinh chỉnh phù hợp. Nhìn chung có hai phương pháp học chuyển tiếp trong mạng học sâu:

Học chuyển tiếp thông qua trích xuất đặc trưng ảnh: Khi trích xuất đặc trưng ảnh, mạng đã huấn luyện được xem như bộ trích xuất đặc trưng độc lập, nó cho phép ảnh đầu vào truyền và dừng lại ở một lớp xác định nào đó, khi đó đầu ra của lớp này được xem như đặc trưng ảnh kết quả. Trong bài báo này, chúng tôi sử dụng ResNet-50, Densnet-121, EfficientNet-B4 đã được huấn luyện để trích xuất các đặc trưng ảnh, sau đó sử dụng phương pháp học tăng cường để huấn luyện bộ phân loại trên tập đặc trưng ảnh vừa trích xuất. Chúng tôi huấn luyện mô hình mạng trên các tập dữ liệu được chia nhỏ gọi là các lô (batch), quy trình huấn luyện bao gồm các bước sau: 1) Tải từng lô dữ liệu vào bộ nhớ, 2) Huấn luyện mô hình từ các lô dữ liệu này, 3) Lặp lại các bước này cho đến khi mạng đạt được hội tụ cần thiết.

Học chuyển tiếp thông qua tinh chỉnh: Kỹ thuật này chỉnh sửa kiến trúc mạng bằng cách loại bỏ các đầu ra lớp FC, tạo ra các đầu ra mới và sau đó bắt đầu quá trình huấn luyện mạng. Phương pháp này gồm các bước sau: 1) Tải mô hình mạng đã được huấn luyện, chúng tôi sử dụng mạng đã huấn luyện trên ImageNet làm mạng cơ sở; 2) Xây dựng một kiến trúc mạng mới bằng cách sao chép kiến trúc và tham số của mạng cũ ngoại trừ lớp đầu ra; 3) Thêm một lớp đầu ra mới có kích thước đầu ra bằng với số lớp tập dữ liệu đích; 4) Đóng băng các lớp tích chập trước đó của mạng để đảm bảo các đặc trưng nổi bật của mô hình mạng cũ trước đó không bị phá hủy; 5) Bắt đầu huấn luyện mô hình mạng mới trên tập dữ liệu mới, nhưng chỉ đào tạo các đầu ra của lớp FC - bước này được gọi là *warmup*; 6) Rã băng một số lớp tích chập và thực hiện lại quá trình huấn luyện nhưng với tỷ lệ học nhỏ hơn.

D. Hàm Softmax và Power mean support vector machine

1. Hàm Softmax

Đối với các bài toán phân loại sử dụng phương pháp học sâu, hàm *softmax* (hoặc mã hóa từ 1 đến K) thường xuất hiện ở lớp đầu ra của mạng. Ví dụ, nếu bộ phân loại có 10 lớp thì hàm *softmax* sẽ có 10 nút được ký hiệu là p_i, với

$i = 1, \cdots, 10$. p_i là một phân phối xác suất rời rạc, vì vậy $\sum_i^{10} p_i = 1$. Gọi $\boldsymbol{h}$ là các nút kích hoạt của lớp áp chót, $\boldsymbol{W}$ là trọng số nối lớp áp chót với lớp *softmax*, tổng giá trị đầu vào đối với lớp *softmax* của nút thứ i là:

$$a_i = \sum_k h_k W_{ki} \tag{1}$$

Khi đó ta có:

$$p_i = \frac{exp(a_i)}{\sum_j^{10} exp(a_j)} \tag{2}$$

Giá trị dự đoán của lớp thứ i được cho bởi công thức:

$$\hat{\imath} = \arg\max_i p_i = \arg\max_i a_i \tag{3}$$

2. Power mean support vector machine

Xét bài toán phân loại nhị phân tuyến tính trên tập dữ liệu huấn luyện $\mathbb{T} = \{(x_i, y_i)\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \{+1, -1\}$ với n là số lượng mẫu huấn luyện. Thuật toán SVM [29] sẽ tìm một siêu phẳng có lề lớn nhất để phân tách các điểm dữ liệu của hai lớp $\mathrm{y} = +1$ và $\mathrm{y} = -1$. Nó thực hiện đồng thời cực đại hóa lề giữa hai siêu phẳng hỗ trợ của hai lớp và tối thiểu hóa mức độ lỗi. Điều này được thực hiện bằng cách giải bài toán tối ưu ở dạng đối ngẫu (4).

$$\min_{\alpha \in \mathbb{R}^n} \mathrm{f}(\alpha) = \frac{1}{2}\alpha^{\mathrm{T}} \mathrm{Q} \alpha - \mathrm{e}^{\mathrm{T}}\alpha \qquad \text{s.t.} \begin{cases} \mathrm{y}^{\mathrm{T}}\alpha = 0 \\ 0 \le \alpha_{\mathrm{i}} \le \mathrm{C}, \forall \mathrm{i} = 1, 2, \cdots, \mathrm{n} \end{cases} \tag{4}$$

với $e = [1, \cdots, 1]_n^T$, C là hằng số dùng để điều chỉnh độ rộng lề và mức độ lỗi, $\alpha = (\alpha_1, \cdots, \alpha_n)$ là các nhân tử Lagrange, Q là ma trận đối xứng $n \times n$ với $\mathrm{Q}_{ij} = y_i y_j K\langle x_i, x_j \rangle$, và $K\langle x_i, x_j \rangle$ là hàm *kernel*.

Các vector hỗ trợ (SV) ($\alpha_i > 0$) được xác định sau khi giải bài toán tối ưu (4), sau đó siêu phẳng và độ lệch b sẽ được xác định thông qua các vector hỗ trợ này. Khi đó phân lớp của dữ liệu mới x được xác định bởi (5):

$$sign\left(\sum_{i=1}^{\#SV} y_i \alpha_i K\langle x_i, x_j \rangle - b\right) \tag{5}$$

Để chuyển đổi từ bộ phân lớp tuyến tính sang phi tuyến tính, ta cần thay K bởi một *kernel* phi tuyến. Hàm *kernel* được sử dụng phổ biến là hàm *polynomial* bậc d, *Radial Basis Function* hoặc *sigmoid*. PmSVM được đề xuất bởi Wu [21] thay thế hàm *kernel* $K\langle x_i, x_j \rangle$ trong (4) và (5) bởi *Power mean kernel* $M_p\langle x_i, x_j \rangle$ (x_i và $x_j \in \mathbb{R}_+^d$), được trình bày trong (6), là dạng tổng quát của *additive kernel* (χ^2 kernel, histogram intersection kernel hoặc Hellinger"s kernel).

$$M_p\langle x_i, x_j \rangle = \sum_{z=1}^{d} \left(x_{i,z}^p + x_{j,z}^p\right)^{\frac{1}{p}} \tag{6}$$

với $p \in \mathbb{R}$ là một hằng số. χ^2 kernel ($p = -1$): $M_{-1}(x, y) = K_{\chi^2}(x, y) = \frac{2xy}{x+y}$, Histogram intersection kernel ($p = -\infty$): $M_{-\infty} = K_{HI}(x, y) = min(x, y)$, Hellinger kernel ($p = 0$): $M_0(x, y) = \sqrt{xy}$.

PmSVM sử dụng thuật toán tối ưu *Coordinate descent* để huấn luyện mô hình. Hơn nữa, *gradient* được ước lượng xấp xỉ bởi *polynomial regression* với chi phí tính toán thấp [21]. Vì vậy, PmSVM rất hiệu quả trong huấn luyện và kiểm thử mô hình. Các tác giả của [22], [30] đã thay thế hàm *softmax* của DCNN bởi SVM. Trong khi hàm *softmax* cực tiểu hóa *cross-entropy* hoặc cực đại hóa *log-likelihood*, thì SVM chỉ tìm siêu phẳng lớn nhất giữa các điểm dữ liệu của các lớp khác nhau. Thí nghiệm cho thấy việc thay thế hàm *softmax* bởi SVM có kết quả tốt hơn trong nhiều trường hợp. Trong bài báo này, chúng tôi đề xuất mô hình kết hợp giữa DCNN và SVM để phân loại ảnh sâu bệnh. Cụ thể, chúng tôi đánh giá bộ phân lớp phi tuyến PmSVM khi phân loại đặc trưng ảnh được trích xuất từ mô hình EfficientNet.

III. KẾT QUẢ THỰC NGHIỆM

A. Cấu hình thí nghiệm và huấn luyện

Tất cả các thí nghiệm được thực hiện trên máy trạm Intel (R) Core (TM) i7-8565U CPU @ 1,80 GHz, RAM 8 GB, hệ điều hành Ubuntu. Các thuật toán được lập trình bằng ngôn ngữ lập trình Python và Keras. Tất cả hình ảnh được chuyển đổi từ RGB sang BGR, sau đó mỗi kênh màu được chuẩn hoá dữ liệu về dạng có trung tâm là 0 (zero-centered) với các thuật toán *batch/layer normalization*. Kế tiếp, các ảnh được thay đổi kích thước cho tương thích với đầu vào của từng kiến trúc mạng. Cụ thể, kích thước ảnh được thay đổi là 224 × 224 điểm ảnh cho ResNet-50, DenseNet-121 và EfficientNet-B0, 240 × 240 với EfficientNet-B1, 260 × 260 với EfficientNet-B2, 300 × 300 với EfficientNet-B1, và 380 × 380 với EfficientNet-B4. Các mô hình mạng được huấn luyện bằng phương pháp học chuyển tiếp như được trình bày trong Phần II.C. Trọng số đã huấn luyện trên ImageNet được sử dụng để rút ngắn thời

gian huấn luyện các mô hình mạng mới. Các lớp FC cuối cùng của các mô hình mạng có 1.000 đầu ra được điều chỉnh thành 24, 40 và 102 đầu ra để phù hợp với các tập dữ liệu Xie24 (24 lớp), D0 (40 lớp) và IP102 (102 lớp). Các thông số như số lớp tích chập được đóng băng, số lớp FC, tỷ lệ *dropout*, thuật toán tối ưu, tỷ lệ học và số lần lặp được xác định bằng phương pháp *thử-và-sai*. Để đánh giá hiệu suất trích xuất đặc trưng ảnh, số lượng các lớp FC được giữ nguyên như nhau trong tất cả các mô hình mạng. Các mạng được tinh chỉnh sử dụng cùng một thuật toán tối ưu hóa như trên ImageNet, đó là hàm *categorical cross-entropy* và thuật toán tối ưu *Adam*. Mỗi mô hình mạng được huấn luyện qua hai giai đoạn: 1) Giai đoạn đầu tiên là *warmup* mạng, trong đó chỉ các đầu ra của lớp FC được huấn luyện, tất cả các lớp trước đó được đóng băng, tỷ lệ học là 0,01, số lần lặp là 10 cho các mô hình mạng; 2) Ở giai đoạn hai, tùy theo kiến trúc mạng, một số lớp tích chập được rã băng và huấn luyện lần thứ hai với tỷ lệ học nhỏ hơn là 0.0001, số lần lặp là 15 trên D0, Xie24 và 30 trên IP102. Để tránh *overfitting* khi huấn luyện, kỹ thuật tăng cường dữ liệu được áp dụng trên các tập ảnh là xoay, lật, phóng to và phản chiếu. Trong bước phân loại ảnh, chúng tôi đã huấn luyện các mô hình SVM với các *kernel* tuyến tính và phi tuyến tính trên các đặc trưng ảnh được trích xuất bởi DCNN. Tất cả các mô hình phân loại nhiều lớp SVM đều được thực hiện với chiến lược *one-versus-all*. Chúng tôi đã so sánh PmSVM với LIBLINEAR [31], LIBSVM [34]. Trong đó LIBLINEAR được cấu hình với các tham số mặc định và $C = 1$, LIBSVM với hàm nhân RBF, PmSVM được huấn luyện với các tham số như trong [21] là $p = -1$ (tương ứng χ^2 kernel) và $C = 0.01$.

B. Chỉ số đánh giá mô hình mạng

Phương pháp phổ biến để đánh giá hiệu suất của các hệ thống phân loại đối tượng với nhiều lớp là tính Average Precision (AP) với công thức (7), Average Recall (AR) với công thức (8), Average F1-score (AF1) với công thức (9) và Accuracy (A) với công thức (10). Xét các giá trị trong ma trận nhầm lẫn thu được từ kết quả phân loại, các giá trị đo lường của các công thức từ (7) đến (10) được tính toán bằng cách sử dụng các chỉ số True Positive (TP), True Negative (TN), False Positive (FP) và False Negative (FN). Trong đó, TP là số lượng ảnh được phân loại chính xác trong mỗi lớp, TN là số lượng ảnh được phân loại chính xác vào các lớp khác. FN là số lượng ảnh bị phân loại sai vào các lớp khác của lớp đang xét. FP là số lượng ảnh bị phân loại sai của các lớp khác vào lớp đang xét.

$$Average\ Precision(AP) = \frac{1}{\#classes}\sum_{k=1}^{\#classes} P \qquad (7)$$

$$Precision(P) = \frac{TP}{TP + FP} \qquad (7")$$

$$Average\ Recall(AR) = \frac{1}{\#classes}\sum_{k=1}^{\#classes} R \qquad (8)$$

$$Recall(R) = \frac{TP}{TP + FN} \qquad (8")$$

$$Average\ F1-score(AF1) = \frac{1}{\#classes}\sum_{k=1}^{\#classes} F1 \qquad (9)$$

$$F1-score(F1) = 2 \times \frac{P \times R}{P + R} \qquad (9")$$

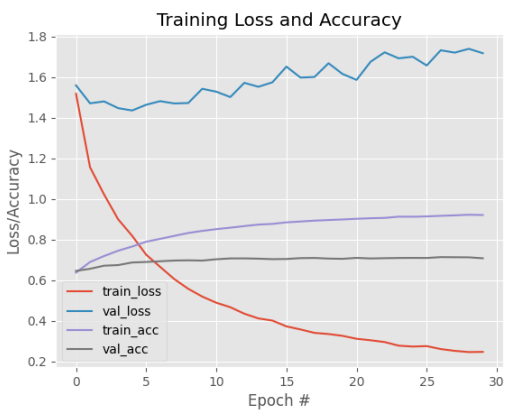
$$Accuracy(A) = \frac{TP + TN}{TP + TN + FP + FN} \qquad (10)$$

C. Kết quả và thảo luận

1. Kết quả

Chúng tôi đã cài đặt ResNet-50, DenseNet-121 và 6 mô hình khác nhau của EfficientNet để đánh giá hiệu suất trên tất cả các tập dữ liệu. Hình 4 và hình 5 là biểu đồ độ chính xác và độ lỗi trong giai đoạn *warmup* và rã băng (unfrozen) của mô hình EfficientNet-B4 trên tập dữ liệu huấn luyện và kiểm thử của IP102. Bảng 1 và bảng 2 tóm tắt hiệu suất của từng mô hình trên tập dữ liệu Xie24 và D0. Bảng 3 tóm tắt hiệu suất của từng mô hình trên tập dữ liệu IP102. Ký hiệu "-" có nghĩa là các tác giả đã không báo cáo kết quả. Các giá trị được đánh dấu in đậm trong bảng 1, bảng 2 và bảng 3 biểu thị cho trường hợp có giá trị tốt nhất theo tiêu chí hiệu suất tương ứng.

Hình 4. Biểu đồ Accuracy/loss của EfficientNet-B4 trên dữ liệu huấn luyện và kiểm thử của IP102 trong giai đoạn warmup



Hình 5. Biểu đồ Accuracy/loss của EfficientNet-B4 trên dữ liệu huấn luyện và kiểm thử của IP102 trong giai đoạn unfrozen

Bảng 1. Độ chính xác phân loại (%) của các phương pháp trên tập dữ liệu ảnh Xie24

Methods	Accuracy	Avg. Precision	Avg. Recall	Avg. F1-score
Thenmozhi el al. [12]	97.47	-	-	-
DenseNet-121	90.00	90.00	90.00	90.00
ResNet-50	98.00	98.00	98.00	98.00
EfficientNet-B0	**99.00**	**99.00**	**99.00**	**99.00**

Bảng 2. Độ chính xác phân loại (%) của các phương pháp trên tập dữ liệu ảnh D0

Methods	Accuracy	Avg. Precision	Avg. Recall	Avg. F1-score
Ayan et al. [19]	98.81	98.88	98.81	98.81
DenseNet-121	85.00	86.00	85.00	85.00
ResNet-50	93.00	94.00	93.00	93.00
EfficientNet-B0	**99.00**	**99.00**	**99.00**	**99.00**

Bảng 3. Độ chính xác phân loại (%) của các phương pháp trên tập dữ liệu ảnh IP102

Methods	Params	Accuracy	Avg. Precision	Avg. Recall	Avg. F1-score
Ayan et al. [19]	-	67.13	67.17	67.13	65.76
FR-ResNets [18]	31M	55.24	-	-	54.18
AlexNet	57M	49.41	-	-	48.22
DenseNet-121	8M	54.59	-	-	52.97
ResNet-50	26M	54.19	54.19	54.19	48.22
ResNet-101	45M	53.07	-	-	52.00
GoogleNet	10M	52.17	-	-	51.24
VGG-16	138M	51.84	-	-	51.20
EfficientNet-B0	5M	67.00	67.00	67.00	67.00
EfficientNet-B1	8M	69.00	69.00	69.00	69.00
EfficientNet-B2	9M	69.00	69.00	69.00	69.00
EfficientNet-B3	12M	70.00	70.00	70.00	70.00
EfficientNet-B4	19M	71.00	71.00	71.00	70.00
EfficientNet-B5	31M	71.17	71.17	71.17	71.00
EfficientNet-B4+LIBSVM	19M	49.85	-	-	-
EfficientNet-B4+LIBLINEAR	19M	69.31	-	-	-
EfficientNet-B4+PmSVM	19M	**71.84**	-	-	-

2. Thảo luận

Kết quả thực nghiệm cho thấy sự kết hợp của EfficientNet và PmSVM đã mang lại hiệu quả phân loại tốt nhất, vì vậy có thể được áp dụng thành công trong phân loại ảnh sâu bệnh. Như đã thấy trong bảng 1, trong tập dữ liệu Xie24, tất cả các mô hình đều cho kết quả với độ chính xác trung bình rất gần nhau. Mô hình EfficientNet-B0 cho độ chính xác tốt nhất với 99% khi so sánh với phương pháp tốt nhất Thenmozhi và cộng sự [12] với 97,47% và các mô hình DCNN khác (DenseNet-121 với 90% và ResNet-50 với 98%). Trong tập dữ liệu D0, EfficientNet-B0 cũng cho độ chính xác cao nhất với 99%, so với phương pháp tốt nhất Ayan và cộng sự [12] với 98,81%, DenseNet-121 với 85% và ResNet-50 với 93%. Trên tập dữ liệu IP102, bảng 3 đã cho thấy Efficientnet-B4 với phương pháp tinh chỉnh của chúng tôi có hiệu quả tốt hơn các phương pháp khác, độ chính xác lên đến 71%. Chúng tôi đã đánh giá EfficientNet-B5 trên IP102, tuy nhiên độ chính xác không được cải thiện đáng kể, trong khi tham số huấn luyện mạng khá lớn (31M). Trong khi đó, sự kết hợp Efficientnet-B4 với PmSVM (EfficientNet-B4+PmSVM) đã cho kết quả vượt trội hơn các phương pháp khác về độ chính xác. Như thể hiện trong bảng 3, nó đạt được độ chính xác cao nhất với 71,84% đối với IP102, cải thiện được 4,71% so với phương pháp của Ayan và cộng sự [19] với 67,13%, FR-ResNets [18] với 55,24% và các mô hình DCNN khác (DenseNet-121 với 54,59%, ResNet-101 với 53,07%). Chúng tôi cũng đã đánh giá sự kết hợp của EfficientNet-B4 với LIBLINEAR và LIBSVM, tuy nhiên độ chính xác chỉ là 69,31% và 49.85% (bảng 3), không tốt hơn hàm *softmax* của EfficientNet-B4 và sự kết hợp của nó với PmSVM. Chú ý rằng, Efficientnet-B4 có ít tham số hơn (19M) khi so sánh với các mô hình DCNN khác và vì vậy nó có chi phí tính toán thấp trong quá trình huấn luyện mạng. Điều này cho thấy rằng phương pháp của chúng tôi có khả năng mở rộng áp dụng trên các tập dữ liệu quy mô lớn.

IV. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Phân loại sâu bệnh tự động bằng ảnh là một nhiệm vụ quan trọng trong nông nghiệp thông minh. Trong bài báo này, chúng tôi đã đề xuất một phương pháp mới dựa trên EfficientNet và sự kết hợp của nó với PmSVM để phân loại ảnh sâu bệnh. Cụ thể, chúng tôi đã đánh giá 6 kiến trúc mạng khác nhau của EfficientNet. Các thí nghiệm cho thấy sự kết hợp của EfficientNet-B4 với PmSVM đã đạt được hiệu suất tốt nhất với chi phí tính toán thấp. Nó đã thiết lập một hiệu suất phân loại mới tốt nhất trên tập dữ liệu ảnh sâu bệnh quy mô lớn IP102 với độ chính xác lên đến 71,84%. Tiếp theo chúng tôi sẽ nghiên cứu các kiến trúc mạng và các phương pháp trích xuất đặc trưng ảnh sâu bệnh hiệu quả hơn. Ngoài ra, trong giai đoạn phân loại ảnh, khi tập dữ liệu đặc trưng ảnh vượt quá giới hạn bộ nhớ chính máy tính thì

PmSVM sẽ gặp vấn đề. Chúng tôi sẽ nghiên cứu áp dụng cách tiếp cận như trong [32]: Một phương pháp học tăng cường cho PmSVM. Cách tiếp cận này tránh tải toàn bộ dữ liệu vào bộ nhớ bằng cách chia dữ liệu thành các khối nhỏ và lưu trữ trong các tập tin riêng biệt, sau đó lần lượt từng khối dữ liệu được tải vào bộ nhớ để huấn luyện mô hình.

V. LỜI CẢM ƠN

Công trình nghiên cứu này đã nhận được sự tài trợ từ Quỹ nghiên cứu Exploration Grants of National Geographic Society (NGS-KOR-59552T-19), Microsoft AI for Earth; Sự hỗ trợ kỹ thuật từ Trường Đại học An Giang, Đại học Quốc gia Thành phố Hồ Chí Minh, Việt Nam.

TÀI LIỆU THAM KHẢO

1 Stork, N.: “How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth?”, Annual Review of Entomology, 63, 2018.

2 Ngugi, L. C., Abelwahab, M., and Abo-Zahhad, M.: “Recent advances in image processing techniques for automated leaf pest and disease recognition - A review”, Information Processing in Agriculture, 8, (1), pp. 27-51, 2021.

3 Lowe, D. G.: “Distinctive image features from scale-invariant keypoints”, 2004, 60, (2), pp. 91-110.

4 Dalal, N., Triggs, B., Dalal, N., and Triggs, B.: “Histograms of Oriented Gradients for Human Detection”, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 886-893, 2005.

5 Samanta, R. K., Samanta, R. K., and Ghosh, I.: “Tea Insect Pests Classification Based on Artificial Neural Networks Artificial Intelligence in Agriculture: A Literature Survey View”, 2, 2012.

6 Manoja, M., Rajalakshmi, M. J., and Student, P. G.: “Early Detection of Pests on Leaves Using Support Vector Machine”, 2, pp. 187-194, 2014.

7 Rani, R. U., Amsini, P.: “Pest identification in leaf images using SVM classifier”, 6, (1), pp. 248-260, 2016.

8 Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A.: “Going deeper with convolutions”, in Editor (Ed.): “Book Going deeper with convolutions” (edn.), pp. 1-9, 2015.

9 He, K., Zhang, X., Ren, S., and Sun, J.: “Deep residual learning for image recognition”, in Editor (Ed.)^(Eds.): “Book Deep residual learning for image recognition” (edn.), pp. 770-778, 2016.

10 Kamilaris, A., and Prenafeta-Boldú, F. X.: “Deep learning in agriculture: A survey”, Computers and Electronics in Agriculture, 147, (February), pp. 70-90, 2018.

11 Xia, D., Chen, P., Wang, B., Zhang, J., and Xie, C.: “Insect detection and classification based on an improved convolutional neural network”, Sensors (Switzerland), 18, (12), pp. 1-12, 2018.

12 Thenmozhi, K., and Srinivasulu Reddy, U.: “Crop pest classification based on deep convolutional neural network and transfer learning”, Computers and Electronics in Agriculture, 164, 2019.

13 Xie, C., Zhang, J., Li, R., Li, J., Hong, P., Xia, J., Chen, P. J. C., and Agriculture, E.i.: “Automatic classification for field crop insects via multiple-task sparse representation and multiple-kernel learning”, 119, pp. 123-132, 2015.

14 Xie, C., Wang, R., Zhang, J., Chen, P., Dong, W., Li, R., Chen, T., and Chen, H.: “Multi-level learning features for automatic classification of field crop pests”, Computers and Electronics in Agriculture, 152, pp. 233-241, 2018.

15 Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., and Fei-Fei, L.: “ImageNet Large Scale Visual Recognition Challenge”, International Journal of Computer Vision, 115, (3), pp. 211-252, 2015.

16 Wu, X., Zhan, C., Lai, Y. K., Cheng, M.M., and Yang, J.: “IP102: A large-scale benchmark dataset for insect pest recognition”, Proceedings of the IEEE Computer Society Conference on CVPR, 2019-June, pp. 8779-8788, 2019.

17 He, K., Zhang, X., Ren, S., and Sun, J.: “Deep Residual Learning for Image Recognition”, in Editor: “Book Deep Residual Learning for Image Recognition” (edn.), pp. 770-778, 2016.

18 Ren, F., Liu, W., and Wu, G.: “Feature reuse residual networks for insect pest recognition”, IEEE Access, 7, pp. 122758-122768, 2019.

19 Ayan, E., Erbay, H., and Varçın, F.: “Crop pest classification with a genetic algorithm-based weighted ensemble of deep convolutional neural networks”, Computers and Electronics in Agriculture, 179, 2020.

20 Tan, M., and Le, Q. V.: “EfficientNet: Rethinking model scaling for convolutional neural networks”, 36th International Conference on Machine Learning, ICML 2019, 2019-June, pp. 10691-10700, 2019.

21 Wu, J.: “Power mean SVM for large scale visual classification”, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2344-2351, 2012.

22 Tang, Y.: “Deep Learning using Linear Support Vector Machines”, 2013.

23 LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P.: “Gradient-based learning applied to document recognition”, Proceedings of the IEEE, 86, (11), pp. 2278-2323, 1998.

24 Simard, P. Y., Steinkraus, D., and Platt, J. C.: “Best practices for convolutional neural networks applied to visual document analysis”, ICDAR, 2003-January, (May 2014), pp. 958-963, 2003.

25 Shi, Z., Dang, H., Liu, Z., and Zhou, X.: “Detection and identification of stored-grain insects using deep learning: A more effective neural network”, IEEE Access, 8, pp. 163703-163714, 2020.

26 Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K.Q.: “Densely connected convolutional networks”, Proceedings - 30th IEEE Conference on CVPR, 2017-January, pp. 2261-2269, 2017.

27 Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A.: “Learning Deep Features for Discriminative Localization”, Proceedings of the IEEE Computer Society Conference on CVPR, pp. 2921-2929, 2016.

28 Yosinski, J., Clune, J., Bengio, Y., and Lipson, H.: "How transferable are features in deep neural networks?", Advances in Neural Information Processing Systems, 4, (January), pp. 3320-3328, 2014.

29 Vapnik, V.: "The nature of statistical learning theory" (Springer science & business media, 2013), 2013.

30 Petrovska, B., Atanasova-Pacemska, T., Corizzo, R., Mignone, P., Lameski, P., Zdravevski, E.: "Aerial scene classification through fine-tuning with adaptive learning rates and label smoothing", Applied Sciences, pp. 1-25.

31 Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., and Lin, C. J.: "LIBLINEAR: A library for large linear classification", Journal of Machine Learning Research, 9, (2008), pp. 1871-1874, 2008.

32 Doan, T. N., Do, T. N., and Poulet, F.: "Parallel incremental power mean SVM for the classification of large-scale image datasets", International Journal of Multimedia Information Retrieval, 3, (2), pp. 89-96, 2014.

PEST INSECT CLASSIFICATION USING EFFICIENTNET AND POWER MEAN SVM

Doan Thanh Nghi

ABSTRACT: *Insects play an essential part in agricultural productivity. Early identification and precise classification of crop pests are extremely beneficial for Integrated Pest Management, which improves the quality and production of agricultural goods. As a result, it is critical to study an effective automatic classification system of pest insect based on images. Recently, most research have focused on utilizing deep learning convolutional neural network models to recognize insect images, with the "softmax" function used to predict the classification label and minimize "cross-entropy loss". In this paper, we present a new approach for pest insect image classification that combines EfficientNet and Power Mean SVM classifier. In which, the EfficientNet network model is fine-tuned and retrained on the new data set. This resulting network model is then utilized to extract image features from the insect datasets. We replace the "softmax" function with a nonlinear classifier Power Mean SVM in the image classification step. Therefore, the learning process will focus on minimizing "margin-based loss" instead of "cross-entropy loss". Our method has been evaluated on several benchmark pest insect datasets. The experiments have shown that our method outperforms several state-of-the-art algorithms for pest insect image classification. Specifically, the accuracy of EfficientNet and Power Mean SVM is 99% on Xie24, 99% on D0 and 71.84% on IP102. To the best of our knowledge, this is the best performance classification result on these data sets.*