TẠP CHÍ KHOA HỌC ĐẠI HỌC HẠ LONG
Scientific Journal of Ha Long University
http://uhl.edu.vn/

ỨNG DỤNG TRÍ TUỆ NHÂN TẠO THÚC ĐẨY CHUYỂN ĐỔI SỐ TRONG LĨNH VỰC KIỂM LÂM

Phạm Thế Anh[1*], Trịnh Thị Anh Loan[1], Nguyễn Tuấn Anh[2]

[1]Trường Đại học Hồng Đức, tỉnh Thanh Hóa
[2]Chi cục Kiểm lâm Thanh Hóa, tỉnh Thanh Hóa

** Email: phamtheanh@hdu.edu.vn*

Ngày nhận bài: 05/9/2022 *Ngày nhận bài sửa sau phản biện: 10/11/2022* *Ngày chấp nhận đăng: 14/11/2022*

TÓM TẮT

Chuyển đổi số đã và đang tác động mạnh mẽ đến mọi lĩnh vực và đóng vai trò quan trọng thúc đẩy phát triển nhanh và bền vững kinh tế – xã hội. Nhằm đẩy mạnh chuyển đổi số trong lĩnh vực kiểm lâm, bài báo này nghiên cứu ứng dụng công nghệ trí tuệ nhân tạo trên thiết bị di động (mobile app) để giải quyết hiệu quả bài toán nhận dạng các loài động thực vật quý hiếm phục vụ công tác nghiệp vụ của ngành kiểm lâm. Bài báo sử dụng mô hình mạng nơron nhân chập MobileNetV3 thông qua kỹ thuật học chuyển tiếp (transfer learning) để tối ưu thời gian xử lý và nâng cao độ chính xác nhận dạng. Ngoài ra, bài báo cũng tìm hiểu, nghiên cứu các kỹ thuật tăng cường dữ liệu (data augmentation) hiện đại và làm trơn nhãn (label smoothing) để nâng cao hiệu năng của mô hình khi đưa vào sử dụng trong thực tế. Kết quả nhận dạng cho thấy hệ thống hoạt động khá hiệu quả trên các thiết bị di động (Android và iOS), đồng thời cho độ chính xác nhận dạng khá cao.

Từ khóa: *cutout, deep learning, mixup, MobileNetV3.*

AN APPLICATION OF ARTIFICIAL INTELLIGENCE FOR BOOSTING DIGITAL TRANSFORMATION IN THE FIELD OF FOREST MANAGEMENT

ABSTRACT

Digital transformation has been strongly affecting many factors of different fields and is a crucial tool to enable the fast and sustainable development of economy and the modern society. This paper focuses on studying and applying artificial intelligence, specifically its sub-domain in deep learning, to create a case study of digital transformation in the area of forest management with a particular emphasis on solving the problem of animal and plant recognition. Specifically, the paper proposes using the MobileNetV3 as the backbone network because of its advantages in efficiency and accuracy. Following that, optimized learning techniques such as soft labeling, data augmentation, and transfer learning were used to improve generality and performance. Experimental results showed that the model performs well in terms of both recognition accuracy and inference time in comparison with other methods. Finally, we have developed an application on mobile platforms (iOS and Android) and the deployment test showed promising performance.

Keywords: *cutout, deep learning, mixup, MobileNetV3.*

1. ĐẶT VẤN ĐỀ

Trong những năm gần đây, thế giới đã chứng kiến sự phát triển mạnh mẽ của công nghệ thông tin (CNTT), tập trung chủ yếu vào các công nghệ lõi của cuộc cách mạng công nghiệp lần thứ tư (CMCN 4.0) như: trí tuệ nhân tạo (artificial intelligence hay AI), mạng Internet vạn vật (IoT), chuỗi khối (block chain), thực tế ảo (VR), v.v.. Đặc biệt, trí tuệ nhân tạo (AI) đã nổi lên như một xu thế phát triển tất yếu của xã hội hiện đại, quyết định đến sự thành công hay thất bại của các doanh nghiệp, là yếu tố quan trọng để định hướng sự phát triển của doanh nghiệp trong việc cung cấp các dịch vụ, sản phẩm chất lượng cao cho khách hàng. Trên phương diện vĩ mô, AI đóng góp rất nhiều vào sự thúc đẩy và tăng trưởng kinh tế – xã hội, giúp giải quyết nhiều bài toán nan giải trong nhiều lĩnh vực chủ chốt và quan trọng như quân sự, y tế, nông nghiệp, kiểm lâm, giáo dục, v.v..

Những tác động và ảnh hưởng của AI trong phát triển kinh tế – xã hội đã và đang diễn ra trên quy mô toàn cầu, ở khắp các quốc gia trên thế giới. Tuy nhiên, sự thành công của AI phụ thuộc nhiều vào điều kiện cụ thể của từng lĩnh vực ứng dụng, điều kiện kinh tế của từng vùng miền, sự phát triển của cơ sở hạ tầng, sự hỗ trợ của các doanh nghiệp viễn thông và công nghệ, đặc biệt là định hướng chiến lược phát triển của các nhà lãnh đạo. Trong bài viết này, tác giả nghiên cứu ứng dụng công nghệ AI để giải quyết bài toán đặc thù của ngành kiểm lâm đó là xây dựng ứng dụng hỗ trợ nhận dạng các loài động thực vật quý hiếm, nguy cấp nhằm đóng góp vào quá trình bảo vệ và bảo tồn thiên nhiên cũng như góp phần thúc đẩy chuyển đổi số trong lĩnh vực này.

Việt Nam là một quốc gia có nhiều loài động thực vật quý hiếm nguy cấp trong sách đỏ cần được bảo vệ. Tuy nhiên, hiện nay, một số loài động thực vật quý hiếm trước đây được ghi nhận tại các địa bàn trên phạm vi cả nước đã tuyệt chủng tại địa phương do tình trạng khai thác trái phép, săn bắn, bẫy bắt quá mức, mất sinh cảnh sống làm suy giảm số lượng loài, cá thể trên địa bàn địa phương (Thanh Hóa, Ninh Bình, Nghệ An). Trong những năm gần đây, được sự quan tâm của Chính phủ và các ngành liên quan đã có rất nhiều chương trình, dự án về bảo vệ, bảo tồn và phát triển các loài động, thực vật được triển khai trên địa bàn tỉnh Thanh Hóa, đặc biệt tại các khu rừng đặc dụng, là khu vực có các sinh cảnh sống chủ yếu của các loài động, thực vật nguy cấp, quý hiếm, trong đó có các loài được ưu tiên bảo vệ. Do đó, việc xây dựng và ứng dụng phần mềm nhận dạng nhanh các loài động, thực vật rừng nguy cấp, quý, hiếm có ý nghĩa rất quan trọng trong lĩnh vực kiểm lâm nhằm bảo vệ các loài động thực vật quý hiếm.

Trên cơ sở phân tích, đánh giá các phần mềm tra cứu và nhận dạng động thực vật quý hiếm (Miao và cs., 2019; Willi và cs., 2019), chúng tôi nhận thấy rằng các phần mềm kể trên có nhược điểm là đơn điệu và kém hiệu quả (chỉ hỗ trợ theo từng nhóm đối tượng động vật, hoặc thực vật, yêu cầu kiến thức chuyên môn, v.v.). Những nhược điểm này gây trở ngại lớn cho các nhà khoa học, sinh thái học, cơ quan chức năng và đặc biệt là người dân có thể tham gia vào quá trình theo dõi động, thực vật hoang dã trong một môi trường mở mà không đòi hỏi nhiều kiến thức chuyên ngành. Ứng dụng các thành tựu gần đây của cuộc cách mạng công nghiệp 4.0 (đặc biệt là các công nghệ trí tuệ nhân tạo, thị giác máy), trong bài báo này chúng tôi xây dựng hệ thống tra cứu, nhận diện một số loài động, thực vật quý hiếm, cần bảo tồn đảm bảo độ chính xác cao và thời gian nhận dạng, có thể hoạt động hiệu quả trên các thiết bị di động và không cần kết nối internet. Ngoài ra, hệ thống có thể được sử dụng như một công cụ tin cậy và hiệu quả để hỗ trợ đội ngũ cán bộ kiểm lâm thực hiện các nghiệp vụ bảo vệ rừng cũng như động, thực vật quý hiếm. Các đóng góp chính của bài báo gồm: (i) xây dựng hệ thống mạng nơron học sâu dựa vào mạng MobileNetV3 bằng cách áp dụng kỹ thuật học chuyển tiếp; (ii) áp dụng các kỹ thuật tối ưu về tăng cường dữ liệu và làm trơn nhãn để cải thiện hiệu quả huấn luyện mô hình; (iii) xây dựng ứng dụng trên môi trường di động tích hợp mô hình nhận dạng đã huấn luyện nhằm hỗ trợ người dùng tra cứu, nhận dạng nhanh các loài động, thực vật quý hiếm.

2. PHƯƠNG PHÁP NGHIÊN CỨU

Bài báo sử dụng hai phương pháp nghiên cứu chính là: phương pháp phân tích và tổng hợp lý thuyết và phương pháp nghiên cứu thực nghiệm. Cụ thể, chúng tôi áp dụng và triển khai quy trình nghiên cứu như sau:

– Tìm hiểu tổng quan các công nghệ, phương pháp xây dựng mạng nơron học sâu đã có và phân tích ưu nhược điểm của các giải pháp đã tồn tại.

– Đề xuất các giải pháp, cải tiến mới và thiết kế các thuật giải.

– Cài đặt và đánh giá/so sánh hiệu năng của các giải pháp đề xuất với các giải pháp khác.

– Sử dụng các cơ sở dữ liệu chuẩn (được cung cấp bởi các cộng đồng nhà khoa học cùng chuyên ngành) và phương pháp/quy trình đánh giá chuẩn để phân tích và so sánh tính hiệu quả của các giải pháp đề xuất.

Để giải quyết các bài toán đặt ra ở trên, cần thiết phải kết hợp các phương pháp nghiên cứu lý thuyết và nghiên cứu thực nghiệm. Phương pháp nghiên cứu lý thuyết được vận dụng để hình thành các ý tưởng chính, hoàn thiện ý tưởng, xây dựng bản mẫu/quy trình nghiên cứu hay mô hình khái niệm, thiết kế các kiến trúc mạng tích chập học sâu, phân tích và đánh giá ưu nhược điểm của các thành phần mạng về phương diện độ phức tạp tính toán cũng như tính năng dự đoán của mạng. Phương pháp nghiên cứu thực nghiệm sau đó được áp dụng để xây dựng các thử nghiệm (experiments), lựa chọn các tập dữ liệu phục vụ đánh giá kiểm thử (benchmark datasets), lựa chọn giao thức (evaluation protocol) và tiêu chuẩn (metrics, criteria) để đánh giá thử nghiệm.

Các kết quả thử nghiệm trong nhiều tình huống có thể mâu thuẫn với mô hình khái niệm và các thuật giải đã đề xuất, xây dựng. Trong trường hợp đó, chúng tôi sẽ tiến hành xem xét, đánh giá lại các thuật toán, phát hiện các điểm chưa hoàn thiện, phát triển và tinh chỉnh thuật toán nếu cần thiết. Trong lĩnh vực thị giác máy và máy học, phương pháp thử nghiệm còn được vận dụng rất nhiều để đánh giá sự ảnh hưởng và tác động của các tham số liên quan trong thuật toán đề xuất. Một hệ thống thị giác máy bền vững phải ít lệ thuộc vào sự thay đổi của các tham số hệ thống, có khả năng tổng quát hóa cao, bền vững với các loại nhiễu, sự thay đổi và sự đa dạng của dữ liệu.

2.1. Tổng quan tình hình nghiên cứu

Sự phát triển của các mạng nơron nhân chập CNN (convolutional neural network) đã được ứng dụng để giải quyết nhiều bài toán khó trong lĩnh vực thị giác máy tính như nhận dạng khuôn mặt, dò tìm đối tượng, xử lý tiếng nói, v.v.. Tuy nhiên, các mạng CNN thường có nhược điểm về độ phức tạp tính toán. Một trong những giải pháp tiềm năng là sử dụng các mạng xấp xỉ mạng CNN hay còn gọi là mạng nhân chập khả tách (separable convolution). Ý tưởng sử dụng các phép chập phân tách lần đầu tiên được giới thiệu trong (Sifre & Mallat, 2014) và sau đó đã được ứng dụng trong (Howard và cs., 2017, 2019; Sandler và cs., 2018) để phát hiện và phân loại đối tượng. Trong bài báo này, phép toán tích chập thông thường được phân tích thành hai phép tích chập đơn giản hơn: tích chập theo chiều sâu (depthwise convolutions) và sau đó là tích chập điểm (pointwise convolutions). Phép chập theo chiều sâu chia một dữ liệu (tensor) đầu vào có dạng D×D×M thành M thành phần (mỗi thành phần có kích thước D×D×1). Mỗi thành phần này, sau đó được nhân chập với một bộ lọc có kích nhỏ (thường là 3×3×1), tạo ra M bản đồ đặc trưng (feature maps) có kích thước D×D×1.

Bảng 1. So sánh giữa nhân chập truyền thống và nhân chập phân tách

Lớp tích chập chuẩn	Tích chập khả tách
Input: D×D×M → Output: D×D×N Thuật toán: áp dụng bộ lọc tích chập trên tín hiệu đầu vào để tạo ra tín hiệu có kích thước đầu ra. Cụ thể, bộ lọc tích chập sẽ có kích thước: 3×3×M×N (giả sử stride = 1 và spatial filter size: 3×3).	Input: D×D×M → Output: D×D×N Thuật toán: Thực hiện 2 vòng tích chập sau: i) Depthwise convolution – Chia tín hiệu đầu vào thành M feature maps có kích thước: D×D×1 – Sử dụng M bộ lọc có kích thước: 3×3×1 (channel 1)để tạo ra M feature maps có kích thước D×D×1 – Ghép (concatenating) các feature maps ở trên thành một tensor có dạng: D×D×M. ii) Pointwise convolution: – Áp dụng bộ lọc tích chập có kích thước: 1×1×M×N lên đầu ra của bước 1 để tạo ra tín hiệu cuối cùng: D×D×N
Độ phức tạp tính toán: D×D×M ×N×3×3	**Độ phức tạp tính toán**: – Depthwise convolution: D×D×M×3×3 – Pointwise convolution: D×D×M×N Tổng: D×D×M×(9+N) → giảm độ phức tạp từ 8 – 9 lần với kích thước mặt nạ lọc là 3×3.
Độ chính xác: cao (do không có sự làm tròn)	**Độ chính xác**: thấp do sử dụng dạng xấp xỉ của phép tích chập chuẩn.
Số lượng tham số: D×D×M×N×3×3	**Số lượng tham số:** D×D×M×(9+N)
Overfitting: cao (do nhiều tham số)	**Overfitting**: thấp (do ít tham số hơn)

Do các phép nhân chập theo chiều sâu hoạt động riêng biệt trên các kênh đầu vào và do đó cần kết hợp các đầu ra để khai thác tốt hơn mối tương quan trực quan của các đặc trưng. Phép tích chập điểm thực hiện công việc này bằng cách trước tiên ghép các feature maps này thành một tensor mới có dạng D×D×M, sau đó áp dụng phép nhân chập với bộ lọc có kích thước 1×1×M×N, tạo ra đầu ra cuối cùng là D×D×N (tức là N kênh đầu ra). Về mặt lý thuyết, phép tích chập phân tách giảm độ phức tạp tính toán từ 8 đến 9 lần so với bộ lọc thông thường khi sử dụng kích thước bộ lọc 3×3 nhưng độ chính xác cũng bị giảm đi một tỉ lệ nhỏ. Bảng 1 so sánh hiệu quả của lớp nhân chập khả tách và lớp nhân chập chuẩn.

Tăng cường dữ liệu (data augmentation) là một kỹ thuật tối ưu hiệu năng huấn luyện mạng CNN trong trường hợp dữ liệu huấn luyện không đủ về mặt số lượng hoặc không đa dạng về nội dung, ngữ cảnh hoặc điều kiện môi trường thu nhận ảnh bị giới hạn. Khi đó, các mô hình mạng CNN sau khi huấn luyện sẽ có xu hướng gặp phải vấn đề khả năng mở rộng hay tổng quát hóa (generlization) hoặc gặp phải vấn đề học quá nhớ (overfitting). Kết quả là mô hình CNN có thể cho kết quả khá tốt trong tập dữ liệu huấn luyện nhưng thường hoạt động không hiệu quả trong các ngữ cảnh thực tế. Các kỹ thuật tăng cường dữ liệu truyền thống bao gồm: biến đổi ngẫu nhiên giá trị điểm ảnh, thay đổi độ bảo hòa, độ sáng, độ tương phản; biến đổi ngẫu nhiên hình học của ảnh như xoay, dịch chuyển, cắt ảnh. Trong thời gian gần đây, nhiều kỹ thuật tăng cường dữ liệu hiện đại đã được đề xuất và chứng tỏ hiệu năng vượt trội khi huấn luyện các mô hình CNN, bao gồm: CutOut (DeVries & Taylor, 2017), MixUp (Zhang và cs., 2018), CutMix (Yun và cs., 2019).

CutOut (DeVries & Taylor, 2017) là một kỹ thuật tăng cường dữ liệu bằng cách xóa bỏ một phần nội dung của bức ảnh bởi một mặt nạ hình vuông theo cách ngẫu nhiên. Cụ thể, kích thước và vị trí của vùng bị xóa là được

xác định một cách ngẫu nhiên. Toàn bộ nội dung bên trong mặt nạ xóa sẽ được gán màu đen. Mục tiêu của CutOut nhằm tạo ra các dạng dữ liệu để huấn luyện mô hình mạng trở nên bền vững với các ngữ cảnh đối tượng bị che một phần trong thực tế.

Kỹ thuật MixUp được đề xuất trong (Zhang và cs., 2018) dùng để trộn dữ liệu và nhãn của hai ảnh với nhau theo công thức sau:

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j$$

$$\tilde{y} = \lambda y_i + (1-\lambda)y_j$$

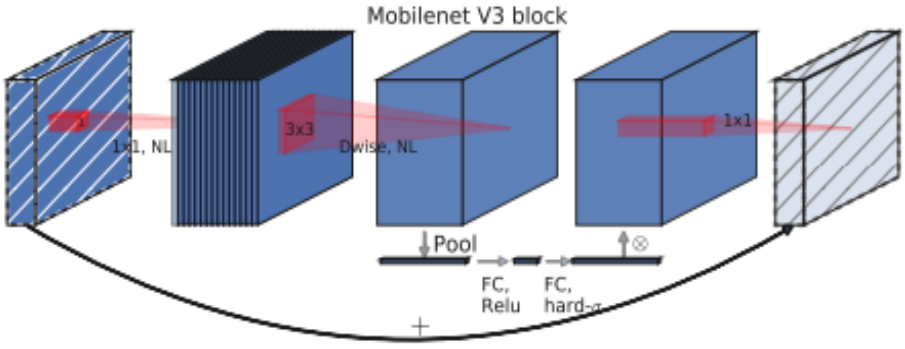
Trong đó: λ có giá trị có phạm vi [0, 1] và được lấy mẫu từ bản phân phối Beta, x_i và x_j là các ảnh đầu vào, y_i và y_j là các nhãn được mã hóa theo dạng chuỗi one-hot. Các mạng nhân chập thường yếu trong việc học các mẫu có nhãn không chính xác (không đầy đủ). Do vậy, kỹ thuật này nhằm trộn nhãn của các mẫu lại để giúp hệ thống học bền vững hơn với các mẫu khó (hard samples).

CutMix (Yun và cs., 2019) dùng để giải quyết vấn đề mất thông tin và kém hiệu quả trong các chiến lược DropOut (loại bỏ một phần các tham số để tránh hiện tượng overfitting). Thay vì loại bỏ các điểm ảnh và gán chúng bằng giá trị màu đen (đôi khi là các giá trị nhiễu của hàm Gaussian), CutMix sẽ thay thế vùng bị xóa bằng một vùng ảnh từ từ một hình ảnh khác. Đồng thời, nhãn của dữ liệu tương ứng cũng được biến đổi theo một hàm tổ hợp tuyến tính với trọng số dựa trên số lượng các điểm ảnh bị thay thế.

2.2. Lựa chọn mô hình mạng nhân chập học sâu

Các mô hình mạng nơron nhân chập có khả năng học tự động các đặc trưng của đối tượng để thực hiện các nhiệm vụ như phân lớp hoặc dò tìm đối tượng với hiệu năng tương đương hoặc vượt con người trong một số ngữ cảnh. Tuy nhiên, các mô hình này thường yêu cầu một lượng lớn dữ liệu để huấn luyện và học. Trong nhiều ngữ cảnh, chúng ta không thể thu thập được lượng dữ liệu cần thiết để huấn luyện mô hình và do vậy sẽ đối mặt với các vấn đề về Underfitting (học chưa đủ), dẫn đến hiệu năng mạng hạn chế. Một giải pháp tuyệt vời cho vấn đề trên đó là học chuyển tiếp (transfer learning). Về cơ bản, kỹ thuật này liên quan đến việc sử dụng một mạng CNN đã được huấn luyện (Pre-trained model) để giải quyết bài toán khác có liên quan.

Chúng tôi sử dụng kiến trúc mạng MobileNetV3 (Howard và cs., 2019) làm kiến trúc mạng để huấn luyện mô hình nhận dạng các loài động thực vật quý bởi tính ưu việt của mô hình này cả về tốc độ xử lý và độ chính xác. MobileNetV3 là phiên bản cải tiến của hai mô hình MobileNetV2 và MobileNetV1 với số lượng tham số giảm gần 50%. Điểm cải tiến của MobileNetV3 so với MobileNetV2 đó là việc sử dụng bổ sung cấu trúc Squeeze-and-Excite (Hu và cs., 2018) trong mỗi khối cơ bản của MobileNetV2 để học được nhiều đặc trưng ngữ cảnh hơn (Hình 1). Ngoài ra, mạng MobileNetV3 được xây dựng bằng cách áp dụng kỹ thuật tìm kiếm mạng NAS (Tan và cs., 2019) để tối ưu kiến trúc mạng tổng thể trên cơ sở tối ưu hóa các khối mạng thành phần.



Hình 1. Kiến trúc khối cơ bản trong MobileNetV3 (Howard và cs., 2019)

2.3. Kỹ thuật tăng cường dữ liệu

Để giảm hậu quả của vấn đề học quá nhớ và tăng khả năng tổng quát hóa của mô hình, chúng tôi áp dụng các phép tăng cường dữ liệu sau:

– Các phép biến đổi giá trị điểm ảnh: Để tạo ra sự đa dạng về điều kiện sáng hay màu sắc của ảnh, một hàm biến đổi ngẫu nhiên sẽ được tạo ra để thay đổi giá trị mầu của ảnh đầu vào. Giả sử ảnh đầu vào là ảnh màu trong hệ RGB, khi đó mỗi thành phần mầu sẽ được biến đổi ngẫu nhiên về giá trị điểm ảnh. Các phép biến đổi được sử dụng bao gồm: biến đổi ngẫu nhiên độ tương phản, màu, độ bão hòa, độ sáng, thêm nhiễu (noise).

– Biến đổi hình học của ảnh: Các phép biến đổi hình học được áp dụng ngẫu nhiên lên ảnh đầu vào nhằm tạo ra các đối tượng có sự phong phú về hình dáng mô phỏng phép chiếu phối cảnh trong thực tế. Các phép biến đổi phổ biến gồm xoay ảnh, dịch chuyển, cắt ảnh.

– Các phép biến đổi hiện đại, gồm: CutMix (Yun và cs., 2019), MixUp (Zhang và cs., 2018) (Hình 2).

Hình 2. Kết quả ảnh sau khi áp dụng tăng cường dữ liệu: CutMix (Yun và cs., 2019), MixUp (Zhang và cs., 2018)

2.4. Kỹ thuật làm trơn nhãn (Label Smoothing)

Kỹ thuật làm trơn nhãn được nghiên cứu và ứng dụng gần đây trong các mô hình mạng nơron học sâu (Goodfellow và cs., 2016; Guo và cs., 2017; Müller và cs., 2019; Pereyra và cs., 2022; Szegedy và cs., 2016) nhằm khắc phục các vấn đề về tự tin quá (overconfidence) và nâng cao khả năng tổng quát hóa cho mô hình khi sử dụng thực tế. Giả sử y_{hot} là vector nhãn của các đối tượng ở dạng biểu diễn one–hot (nghĩa là nếu có K lớp đối tượng thì y_{hot} là vector có N phần tử trong đó chỉ số tương ứng với nhãn của đối tượng sẽ có giá trị 1, các vị trí còn lại có giá trị 0). Để biến đổi y_{hot} thành y_{smooth} biểu diễn nhãn mềm (label smoothing), chúng ta áp dụng công thức biến đổi như sau:

$$y_{smooth} = (1 - \alpha) * y_{hot} + \alpha/K$$

Trong đó: α là hệ số mờ và thường được chọn là $\alpha = 0.1$. Kỹ thuật làm trơn nhãn đặc biệt hữu ích cho các bài toán phân lớp đối tượng và mô hình dự đoán sử dụng hàm Softmax để tạo ra chuỗi giá trị biểu diễn xác suất của mỗi lớp đối tượng. Chẳng hạn, nếu $K = 10$ và $\alpha = 0.1$, các vector biểu diễn đối tượng có nhãn ở vị trí thứ 5 sẽ được tạo ra như sau:

$$y_{hot} = [0, 0, 0, 0, 1, 0, 0, 0, 0, 0\ ]$$

$$y_{smooth} = [0.01, 0.01, 0.01, 0.01, 0.91, 0.01,$$
$$0.01, 0.01, 0.01, 0.01, 0.01, 0.01]$$

2.5. Kỹ thuật học chuyển tiếp

Để hạn chế vấn đề học quá nhớ (overfitting) và khai thác hiệu quả các mạng đã được huấn luyện sẵn trên tập dữ liệu có kích thước lớn, chúng tôi áp dụng kỹ thuật học chuyển tiếp (transfer learning) trên mạng MobileNetV3. Cụ thể, chúng tôi sử dụng mạng MobileNetV3 đã được tiền huấn luyện trên tập dữ liệu ImageNet (Russakovsky và cs., 2015) (chứa khoảng 15 triệu ảnh thuộc 1000 lớp đối tượng khác nhau). Giả sử ảnh

đầu vào có kích thước: 224×224×3, quy trình áp dụng kỹ thuật học chuyển tiếp trên mạng MobileNetV3 được thực hiện như sau:

– Pha 1: Trích chọn đặc trưng và tạo một bộ phân lớp mới (classfication head):

+ Khởi tạo mô hình: model = MobileNetV3 (loại trừ tầng phân lớp ở đỉnh mạng) và đóng băng (khóa các tham số học) các tầng của mô hình hiện tại.

+ Thêm một tầng nhân chập CNN với các tham số: *filters* = 2*NUM_CLASSES, *kernel_size* = 3, *strides* = 2. Trong đó: NUM_CLASSES là tổng số lớp đối tượng cần nhận dạng, *filters* là số bộ lọc, *kernel_size* là kích thước bộ lọc và *strides* là bước nhảy của bộ lọc.

+ Thêm một tầng nhân chập CNN với các tham số: *filters* = NUM_CLASSES, *kernel_size* = 3, *strides* = 2.

+ Thêm một tầng Pooling để thu được vector chứa NUM_CLASSES giá trị tương ứng với giá trị xác suất của mỗi lớp đối tượng bằng cách áp dụng hàm GlobalAveragePooling2D của TensorFlow[1].

+ Áp dụng kỹ thuật làm trơn nhãn (smooth labeling) để tăng tính tổng quát hóa của mô hình học và hạn chế vấn đề mô hình tự tin quá khi đưa ra các giá trị xác suất biểu diễn các lớp đối tượng dự đoán.

+ Huấn luyện mô hình (chính xác là các tầng mới thêm) trên tập dữ liệu động thực vật quý hiếm với các tham số: base_learning_rate = 0.001, epochs = 25.

Trong pha 1 trình bày ở trên, để chuyển tiếp mô hình ban đầu dùng để nhận dạng 1000 đối tượng thành mô hình mới chỉ nhận dạng số lớp đối tượng là NUM_CLASSES (NUM_CLASSES = 56 trong bài báo này), chúng tôi đã loại bỏ hoàn toàn tầng dự đoán của mô hình gốc (dùng để dự đoán 1000 nhãn) và thay bằng 3 tầng mới gồm: 2 tầng nhân chập và một tầng Pooling để phục vụ bài toán mới là nhận dạng NUM_CLASSES lớp đối tượng.

– Pha 2: Làm mịn (fine-tuning):

Sau khi kết thúc pha 1, chúng ta tiếp tục áp dụng pha 2 để thực hiện làm mịn mô hình bằng cách huấn luyện lại một số tầng ở phía cuối mạng bằng cách sử dụng tham số học (learning_rate) khá nhỏ để đảm bảo không biến đổi nhiều các trọng số đã học được từ mô hình cơ sở. Các kết quả thực nghiệm gần đây[2] chỉ ra việc áp dụng pha 2 sẽ giúp cải thiện đáng kể độ chính xác và hiệu năng của mô hình. Trong phần thử nghiệm, chúng tôi cũng sẽ đánh giá hiệu quả của việc áp dụng bước làm mịn này.

Các bước chính của pha 2 bao gồm như sau:

+ Mở khóa các tầng thứ L của mô hình mạng. Trong thực nghiệm, chúng tôi chọn $L = 200$ dựa vào kết quả thực nghiệm trên cơ sở xem xét tổng số tầng của mạng cơ sở là khoảng 356 tầng.

+ Huấn luyện mô hình trên tập dữ liệu động, thực vật quý hiếm với các tham số: learning_rate = 0.0001, epochs = 25.

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Xây dựng dữ liệu huấn luyện

Dữ liệu huấn luyện được chúng tôi thu thập là dữ liệu ảnh của 35 loài động vật và 21 loài thực vật với tổng số ảnh gần 25000 (NUM_CLASSES = 56). Mục đích của việc thu thập ảnh là để huấn luyện các mạng CNN phân lớp phục vụ nhận dạng danh tính của mỗi loài. Do vậy, trong mỗi ảnh, chúng tôi thu thập ảnh chứa một loài với nhiều vị trí khác nhau (như thân, đuôi, đầu với động vật hoặc gốc, rễ, vân lá, thân cây với thực vật). Ngoài ra, các ảnh được chụp ở các góc độ khác nhau, điều kiện ánh sáng khác nhau và tại nhiều thời điểm trong năm để có thể bao quát được các giai đoạn phát triển của loài (ví dụ, các loài lan khi chưa nở hoa và sau khi nở hoa). Quá trình tiền xử lý dữ liệu gồm các bước như sau:

– Chuẩn hóa kích thước ảnh về: 224×224×3.

[1]https://www.tensorflow.org/api_docs/python/tf/keras/layers/GlobalAveragePooling2D

[2] https://keras.io/guides/transfer_learning/

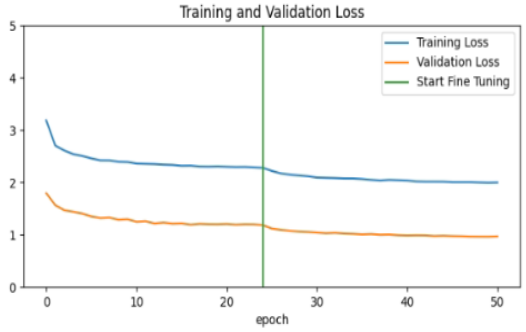
– Chuyển đổi định dạng ảnh về chuẩn JPEG (đuôi: *.jpg)

– Quá trình gán nhãn được thực hiện bởi các chuyên gia của ngành kiểm lâm (Chi cục Kiểm lâm Thanh Hóa).

Dữ liệu sau tiền xử lý được chia thành ba tập: Training, Validation và Testing với tỉ lệ 80%, 5% và 15%.

3.2. Kết quả huấn luyện

Kết quả huấn luyện mô hình (pha 1 và pha 2) của hàm mục tiêu (Loss Functions) được trình bày trực quan trên Hình 3.



Hình 3. Kết quả hàm mục tiêu trên tập Training và Validation

Quan sát Hình 3, chúng ta thấy giá trị hàm mục tiêu trên tập huấn luyện cao hơn hẳn so với trên tập Validation. Điều này khá ngạc nghiên nhưng được giải thích bởi tác dụng của các hàm tăng cường dữ liệu nâng cao CutMix và MixUp. Trong quá trình huấn luyện, các ảnh trong tập Training được áp dụng các phép tăng cường dữ liệu nhưng điều này không đúng trên tập Validation. Do vậy, các ảnh huấn luyện sau tăng cường có xu hướng trở thành các mẫu khó (hard samples) và cưỡng ép mô hình phải học các mẫu khó này để có thể trở nên thông minh hơn, bền vững hơn khi hoạt động thực tế.

Quan sát Hình 3, chúng ta cũng phát hiện một điểm nổi bật đó là hiệu quả của áp dụng hai pha học chuyển tiếp là khá rõ ràng. Khi kết thúc pha 1, đồ thị hàm mục tiêu dừng lại ở điểm 1.2 (tập Validation). Sau khi áp dụng pha 2, đồ thị hàm mục tiêu tiếp tục giảm sâu chứng tỏ quá trình học diễn ra rất hiệu quả và tối ưu.

3.3. Kết quả thực nghiệm trên tập Testing

Để đánh giá kết quả thực nghiệm trên tập Testing, chúng tôi áp dụng độ đo chuẩn là độ chính xác (Accuracy) được tính như sau:

Accuracy(i) = Tổng số ảnh nhận dạng đúng của lớp đối tượng (i) / Tổng số ảnh của lớp đối tượng (i); Trong đó: một ảnh được coi là nhận dạng đúng nếu nó chứa đối tượng X và hệ thống trả về kết quả dự đoán là X với xác suất đủ lớn. Đầu ra của mô hình nhận dạng là nhãn dự đoán của đối tượng và xác suất dự đoán đối tượng đó. Giá trị xác suất nằm trong đoạn [0, 1]. Vì vậy, chúng tôi quy định hệ thống nhận dạng đúng nếu xác suất nhận dạng phải đủ lớn (hay lớn hơn một ngưỡng cho trước). Trong các kết quả thử nghiệm sau, chúng tôi đặt ngưỡng nhận dạng được là 0.5 (hay 50%).

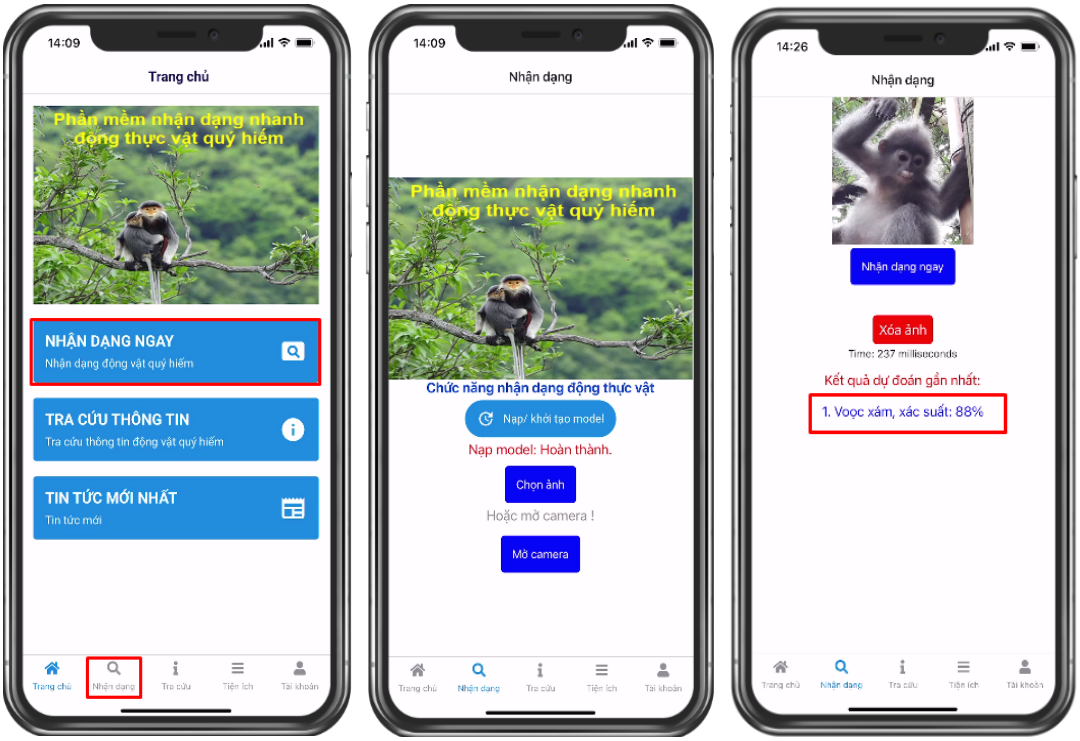
Kết quả thực nghiệm trên hệ thống cho thấy độ chính xác trung bình của nhận dạng 56 loài động, thực vật là 82%. Nếu không sử dụng ngưỡng nhận dạng (ngưỡng nhận dạng = 0), độ chính xác trung bình là 96.1%. Bảng 2 so sánh kết quả nhận dạng của một số hệ thống tiêu biểu cho bài toán nhận dạng động, thực vật quý hiếm khi không dùng ngưỡng nhận dạng.

Bảng 2. So sánh độ chính xác của một số hệ thống khác nhau

Hệ thống	Độ chính xác	Ghi chú
MobileNetV2 (Howard và cs., 2017)	91.1%	Áp dụng transfer learning trên tập dữ liệu chung
InceptionV3 (Szegedy và cs., 2016)	95.0%	
Hệ thống đề xuất	96.1%	

Chúng tôi cũng triển khai mô hình nhận dạng trên môi trường thiết bị di động (Android và iOS) bằng cách chuyển đổi mô hình nhận dạng sử dụng công cụ

TensorFlowJS[3]. Mục đích của thử nghiệm này nhằm đánh giá thời gian xử lý trên các thiết bị di động của mô hình. Kết quả thực nghiệm trên các máy Andoird (LG G8 Thin Q) và iOS (iPhone XR) cho thấy để nhận dạng một ảnh có kích thước bất kỳ thì thời gian trung bình khoảng 350 ms trong đó khoảng 50% thời gian được sử dụng để thực hiện phép chuẩn hóa kích thước ảnh về 224×224. Như vậy, mô hình chỉ cần khoảng 175 ms để hoàn thành việc dự đoán nhãn của ảnh. Đây là tốc độ xử lý rất hiệu quả, phù hợp với yêu cầu thời gian thực khi sử dụng trong thực tế.



Hình 4. Giao diện một số chức năng chính của ứng dụng trên iOS

4. KẾT LUẬN

Trong bài báo này, chúng tôi xây dựng hệ thống nhận dạng động, thực vật quý hiếm ứng dụng mạng CNN học sâu kết hợp nhiều kỹ thuật học tối ưu bao gồm: học chuyển tiếp, làm trơn nhãn và tăng cường dữ liệu. Hệ thống sử dụng kiến trúc mạng MobileNetV3 và sử dụng mô hình đã được tiền huấn luyện (pre-trained) trên tập dữ liệu ImageNet. Chúng tôi áp dụng các kỹ thuật học tối ưu để tiếp tục cải tiến mô hình cho bài toán nhận dạng động thực vật quý hiếm. Kết quả huấn luyện cho thấy hệ thống có tính tổng quát cao, có độ chính xác cao và đặc biệt thời gian xử lý nhanh. Hệ thống đã được triển khai xây dựng thành các ứng dụng đặc thù trên iOS và Android phục vụ công tác nghiệp vụ của ngành kiểm lâm. Trong các nghiên cứu tiếp theo, chúng tôi tiếp tục tăng cường dữ liệu và hoàn thiện ứng dụng để có thể triển khai rộng cho các đơn vị kiểm lâm phạm vi toàn quốc.

LỜI CẢM ƠN

Bài báo này được tài trợ bởi đề tài khoa học công nghệ cấp tỉnh "Xây dựng phần mềm nhận dạng nhanh một số loài động, thực vật nguy cấp, quý hiếm phục vụ công tác quản lý, bảo vệ rừng và bảo tồn đa dạng sinh học trên địa bàn tỉnh Thanh Hóa", 2020 – 2022.

[3] https://www.tensorflow.org/js

TÀI LIỆU THAM KHẢO

DeVries, T., & Taylor, G. W. (2017). *Improved Regularization of Convolutional Neural Networks with Cutout* (arXiv:1708.04552). arXiv.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning* (Illustrated edition). The MIT Press.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. *Proceedings of the 34th International Conference on Machine Learning*, 1321–1330.

Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L.-C., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Adam, H., & Le, Q. (2019). *Searching for MobileNetV3*, 1314–1324. https://doi.org-/10.1109/ICCV.2019.00140

Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications* (arXiv:1704.04861). arXiv.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-Excitation Networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7132–7141. https://doi.-org/10.1109/CVPR.2018.00745

Miao, Z., Gaynor, K. M., Wang, J., Liu, Z., Muellerklein, O., Norouzzadeh, M. S., McInturff, A., Bowie, R. C. K., Nathan, R., Yu, S. X., & Getz, W. M. (2019). Insights and approaches using deep learning to classify wildlife. *Scientific Reports*, *9*(1), Art. 1. https://doi.org-/10.1038/s41598-019-44565-w

Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help? *Advances in Neural Information Processing Systems*, *32*.

Pereyra, G., Tucker, G., Chorowski, J., Kaiser, L., & Hinton, G. (2022). *Regularizing Neural Networks by Penalizing Confident Output Distributions*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007-/s11263-015-0816-y

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). *MobileNetV2: Inverted Residuals and Linear Bottlenecks*, 4510–4520. https://doi.-org/10.1109/CVPR.2018.00474

Sifre, L., & Mallat, S. (2014). *Rigid-Motion Scattering for Texture Classification* [PhD Thesis, arXiv]. http://arxiv.org/-abs/1403.1687

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). *Rethinking the Inception Architecture for Computer Vision*, 2818–2826. https://doi.-org/10.1109/CVPR.2016.308

Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., & Le, Q. V. (2019). MnasNet: Platform-Aware Neural Architecture Search for Mobile. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2815–2823. https://doi.org/10.1109/CVP-R.2019.00293

Willi, M., Pitman, R. T., Cardoso, A. W., Locke, C., Swanson, A., Boyer, A., Veldthuis, M., & Fortson, L. (2019). Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution*, *10*(1), 80–91. https://doi.org/-10.1111/2041-210X.13099

Yun, S., Han, D., Chun, S., Oh, S. J., Yoo, Y., & Choe, J. (2019). *CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features*, 6022–6031. https://doi.org/10.1109/ICCV.2019.00612

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). *Mixup: Beyond Empirical Risk Minimization*. International Conference on Learning Representations 2018, Vancouver Convention Center, Vancouver, BC, Canada.