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

PHÂN LOẠI ẢNH ĐA NHÃN VỚI ĐỐI TƯỢNG MỚI TỪ TẬP DỮ LIỆU ĐƠN NHÃN DỰA TRÊN MÔ HÌNH CONFORMER MẶT NẠ

Nghiêm Văn Triệu[1*], Ngô Quốc Tạo[2]

[1]*Tổng công ty Viễn thông Mobifone*

[2]*Viện Công nghệ thông tin – Viện Hàn lâm Khoa học và Công nghệ Việt Nam*

**Email: nghiemvantrieu@gmail.com*

Ngày nhận bài: 23/02/2023 *Ngày nhận bài sửa sau phản biện: 25/03/2023* *Ngày chấp nhận đăng: 28/03/2023*

TÓM TẮT

Mô hình Convolutional Neural Network và gần đây là Transformer đã chứng minh hiệu quả trong phân loại ảnh đơn nhãn dựa trên các tập dữ liệu đơn nhãn. Khi mở rộng ra bài toán phân loại ảnh đa nhãn, một rào cản lớn là không đủ các tập dữ liệu đa nhãn cho huấn luyện mô hình. Kết hợp trực tiếp tập ảnh đa nhãn và đơn nhãn (cho đối tượng mới) chưa mang lại kết quả phân loại đa nhãn. Trong bài báo này, chúng tôi đề xuất mô hình Conformer và phương pháp mặt nạ tựa BERT cho phân loại ảnh đa nhãn dựa trên tập dữ liệu đơn nhãn ImageNet và tập dữ liệu đa nhãn Coco. ImageNet được sử dụng để huấn luyện nhận dạng đối tượng "chính" trong ảnh (đối tượng ImageNet) và Coco để nhận dạng các đối tượng "phụ" khác trong ảnh. Kết hợp một lượng nhỏ dữ liệu ngữ cảnh đa nhãn là sự "lai ghép" đối tượng từ Coco và ImageNet để kết nối các tập dữ liệu khác nhau, mô hình đề xuất có thể nhận dạng đối tượng "chính" trong ảnh và các đối tượng thông thường khác. Ngoài ra, mô hình có thể áp dụng cho gán lại đa nhãn tập dữ liệu ImageNet với thông tin ngữ cảnh đặc trưng.

Từ khóa: *gán lại đa nhãn tập ImageNet, mô hình Conformer, phân loại ảnh đa nhãn, tập dữ liệu đơn nhãn, tập dữ liệu ImageNet.*

MULTI-LABEL IMAGE CLASSIFICATION WITH NOVEL OBJECT FROM SINGLE-LABEL DATASET BY MASK CONFORMER MODEL

ABSTRACT

On the basis of single-label datasets, the Convolutional Neural Network (CNN) and, more recently, the Transformer model, have shown to be successful at classifying single-label images. The lack of multi-label datasets for model training is a significant obstacle when it comes to the problem of multi-label image classification. In this paper, we propose a Conformer model and a BERT-like mask method for multi-label image classification based on the ImageNet single-label dataset and Coco multi-label dataset. ImageNet is used to train the "main" object in the image (ImageNet object) and Coco to recognize "secondary" objects in the image. The proposed model can identify the "main" object and other common objects in images when combined with a small amount of multi-label context data, which is a "hybrid" of objects from Coco and ImageNet to connect different datasets. In addition, the model can be applied to a multi-label reassignment of the ImageNet dataset with specific context information.

Keywords: *Conformer model, ImageNet dataset, multi-label image classification, re-label ImageNet, single-label dataset.*

1. ĐẶT VẤN ĐỀ

Những năm gần đây, thị giác máy tính – Computer vision (CV) đã đạt được những bước tiến lớn nhờ tiến bộ của công nghệ học sâu và các tập dữ liệu lớn. Một số mô hình được huấn luyện trên lượng dữ liệu lớn gán nhãn đã đạt và thậm chí vượt qua khả năng của con người trong một số nhiệm vụ cụ thể, chẳng hạn như phân loại ảnh đơn nhãn.

Phân loại ảnh đơn nhãn nhằm mục tiêu gán một nhãn cho ảnh từ tập dữ liệu ảnh đơn nhãn. Trong thời gian dài, mô hình Convolutional Neural Network (CNN) tỏ ra khá hiệu quả trong phân loại ảnh đơn nhãn. Những năm gần đây, với thành công của mô hình Transformer trong lĩnh vực xử lý ngôn ngữ tự nhiên – Natural language processing (NLP), nhiều nghiên cứu đã áp dụng mô hình Transformer trong CV và mang lại kết quả cạnh tranh với mô hình CNN, chẳng hạn mô hình Vision Transformer (ViT) của Google (Dosovitskiy và cs., 2021), Object Detection Transformer (DETR) của Facebook (Carion và cs., 2020). Dữ liệu cho huấn luyện phân loại ảnh đơn nhãn khá phong phú, chẳng hạn ImageNet (Deng và cs., 2009) với 21K lớp đối tượng, hoặc dễ dàng tìm kiếm trên internet theo đối tượng mong muốn.

Trong thực tế, dữ liệu ảnh, kể cả từ tập dữ liệu đơn nhãn, thường có nhiều hơn một đối tượng trong đó. Chẳng hạn, đối tượng *Accordion* trong ImageNet thường kèm theo các đối tượng *person*, *chair*… trong ảnh. Do đó, bài toán phân loại ảnh đa nhãn mang lại nhiều thông tin giá trị hơn, có thể áp dụng tốt hơn cho nhiều bài toán khác nhau, như trong nhận dạng các đối tượng cho mô tả ảnh đối tượng mới. Nhận dạng các đối tượng trong ảnh, đặc biệt các đối tượng mới, là bước đầu tiên rất quan trọng, quyết định nhiều đến chất lượng của mô tả ảnh đối tượng mới.

Thành công của các mô hình phân loại ảnh đơn nhãn tạo nguồn cảm hứng cho phân loại ảnh đa nhãn. Tuy nhiên, không đơn giản chỉ là chuyển đổi từ mô hình phân loại ảnh đơn nhãn sang phân loại đa nhãn. Bởi trong phân loại đơn nhãn, các đối tượng "phụ" thường không được chú ý, trong khi phân loại ảnh đa nhãn, sự quan tâm chia đều cho các đối tượng và có tình trạng trùng lặp/che khuất của các đối tượng trong ảnh.

Thêm vào đó, dữ liệu ảnh đa nhãn thường không đầy đủ, không đa dạng và tốn rất nhiều tài nguyên, công sức cho gán lại đa nhãn, hoặc tìm kiếm trên internet. Chẳng hạn: Coco dataset (https://cocodataset.org) là tập dữ liệu đa nhãn, tập trung vào 80 lớp đối tượng; Open image (Kuznetsova và cs., 2020) là tập dữ liệu đa nhãn với 600 lớp đối tượng, nhưng chỉ tập nhỏ dữ liệu trong đó được gán đa nhãn. Hơn nữa, việc sử dụng tập dữ liệu đa nhãn sẵn có thường hạn chế sự phong phú của các đối tượng nhận dạng. Việc thu thập và gán đa nhãn cho tập dữ liệu lớn vẫn là một thách thức. Dữ liệu ảnh đa nhãn có thể thu thập từ internet bằng việc kết hợp nhiều từ khóa, tuy nhiên kết quả tìm kiếm trả về nhiều ảnh không phù hợp với nội dung nên cần phải có sự rà soát thủ công. Gán đa nhãn toàn bộ tập dữ liệu lớn là một công việc tẻ nhạt, tốn thời gian, công sức và dễ bị lỗi và yêu cầu sự tham gia của chuyên gia trong một số lĩnh vực, chẳng hạn trong lĩnh vực y khoa.

Để giải quyết tình trạng không đầy đủ, phong phú dữ liệu đa nhãn, một giải pháp đề xuất là sử dụng tập dữ liệu đơn nhãn cho huấn luyện phân loại ảnh đa nhãn. Sử dụng đơn thuần tập dữ liệu đơn nhãn cho huấn luyện phân loại ảnh đa nhãn không đem lại kết quả mà cần phải có các dữ liệu bổ sung hoặc các kỹ thuật chuyên sâu khác nữa. Đã có một số nghiên cứu trên thế giới theo hướng này mang lại kết quả khả quan. Nghiên cứu của tác giả Sangdoo Yun và cộng sự thuộc Phòng thí nghiệm NAVER AI (Hàn Quốc) (Yun và cs., 2021) đề xuất gán lại đa nhãn cho tập dữ liệu ImageNet. Theo đó, tác giả sử dụng phương pháp "random crop and resize" – chọn ngẫu nhiên một vùng ảnh để hy vọng nhận được đối tượng, sau đó qua mô hình nhận dạng để phát hiện đối tượng trong ảnh. Từ đó, nghiên cứu đã thực hiện gán lại đa nhãn cho ImageNet với độ chính xác lên đến 80%. Tác giả Baoyuan Wu và cộng sự (Wu và cs., 2019) xây dựng kho dữ liệu đa nhãn Tencent multi-label Images dựa trên các tập dữ liệu ImageNet và Open Image, bằng cách sử dụng

cây phân cấp ngữ nghĩa và đồng xuất hiện giữa các lớp đối tượng. Đây là kho dữ liệu ảnh khá lớn với khoảng 18 triệu ảnh cho 11K lớp đối tượng. Trong nghiên cứu (Huang và cs., 2020), tác giả xây dựng đồ thị mô tả các lớp đối tượng dựa trên tập dữ liệu Coco và ImageNet, tính toán độ tương đồng dựa trên lưới từ và sử dụng ngưỡng xác định cạnh giữa các node trên đồ thị. Sau đó, sử dụng mạng Relational GraphConvolutional Network (GCN) cho huấn luyện phân loại ảnh đa nhãn đối tượng mới trong ảnh. Trong nghiên cứu (Wei và cs., 2016), tác giả sử dụng mô hình phát hiện đối tượng BING, để phát hiện các đối tượng trong ảnh, được coi như là tập các ứng viên đối tượng trong ảnh, sau đó sử dụng mô hình CNN chia sẻ kết nối các ứng viên trong một tổng thể bởi lớp max pooling cho kết quả cuối cùng là dự đoán đa nhãn cho ảnh. Trong nghiên cứu (Verelst và cs., 2023), tác giả sử dụng phương pháp cắt ảnh ngẫu nhiên và chuyển đổi kích thước ảnh làm dữ liệu bổ sung, dùng hàm mất mát "spatial consistency loss" cho huấn luyện phân loại ảnh đa nhãn và giảm bớt nhiễu phân loại do cắt ảnh ngẫu nhiên gây không đồng bộ nhãn huấn luyện và ảnh đầu vào.

Nhìn chung, các nghiên cứu thường sử dụng hệ phát hiện đối tượng để trích xuất các đối tượng làm dữ liệu đầu vào cho huấn luyện mô hình nên kết quả nhận dạng phụ thuộc khá nhiều vào độ chính xác và độ đa dạng của hệ phát hiện đối tượng. Việc đồng thời sử dụng kết quả thứ cấp làm đầu vào cũng ảnh hưởng phần nào đến hiệu năng mô hình. Việc xây dựng quan hệ giữa các đối tượng dựa trên lưới từ/đồ thị phụ thuộc ngữ cảnh ngôn ngữ nhiều hơn là ngữ cảnh hình ảnh.

Trong quá trình thử nghiệm thực tế, chúng tôi nhận thấy rằng, việc sử dụng trực tiếp ImageNet và Coco không mang lại kết quả cho phân loại ảnh đa nhãn với đối tượng mới. Do vậy, trong bài báo này, chúng tôi đề xuất phương pháp "lai ghép" đối tượng ảnh và sử dụng phương pháp mặt nạ tựa BERT trong mô hình Conformer để đạt được kết quả phân loại ảnh đa nhãn với đối tượng mới từ tập ImageNet. Cụ thể, chúng tôi sử dụng tập dữ liệu đơn nhãn ImageNet và một lượng rất nhỏ dữ liệu được chú thích đa nhãn từ chính ImageNet, kết hợp tập dữ liệu đa nhãn Coco để làm dữ liệu huấn luyện. Bằng phương pháp mặt nạ hóa trong mô hình Conformer, chúng tôi đã thử nghiệm thành công mô hình phân loại ảnh đa nhãn dựa trên tập dữ liệu ảnh đơn nhãn ImageNet. Có thể mở rộng áp dụng cho các tập dữ liệu đơn nhãn khác hoặc các lớp đối tượng tự tìm kiếm trên internet, đồng thời có thể áp dụng cho gán lại đa nhãn tập dữ liệu ảnh ImageNet.

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Phương pháp trích xuất lai ghép các đối tượng

Coco là tập dữ liệu ảnh sử dụng cho phát hiện đối tượng, phân đoạn và chú thích ảnh quy mô lớn với 328K ảnh cho 80 lớp đối tượng khác nhau. Các đối tượng xuất hiện với tần suất khá lớn, chẳng hạn đối tượng "dog" có tần suất là 18.000 lần.

ImageNet là tập dữ liệu ảnh nổi tiếng trong CV nói chung và phân loại ảnh nói riêng. Đây là tập dữ liệu lớn gồm 14 triệu ảnh được chú thích theo phân cấp mạng từ với 21K lớp đối tượng. Tập con thường được sử dụng của ImageNet bao gồm 1.000 lớp đối tượng với 1.281.167 ảnh cho huấn luyện, 50.000 cho kiểm thử và 100.000 cho thử nghiệm.

Dưới góc độ phân loại ảnh, Coco là tập đa nhãn có thể sử dụng cho phân loại ảnh đa nhãn, ImageNet là tập dữ liệu đơn nhãn chủ yếu sử dụng cho phân loại ảnh đơn nhãn. Kết hợp trực tiếp hai tập dữ liệu này không mang lại kết quả cho phân loại ảnh đa nhãn, do xác suất trong phân loại đơn nhãn cao hơn nhiều so với phân loại ảnh đa nhãn.

Qua quá trình phân tích và thử nghiệm thực tế, chúng tôi nhận thấy rằng, kết hợp ImageNet với Coco, cộng thêm một lượng tối thiểu ảnh từ ImageNet được chú thích đa nhãn (theo thử nghiệm của chúng tôi là 20 ảnh/ lớp đối tượng) cho phân loại ảnh đa nhãn đem lại kết quả khả quan. Cụ thể: (1) Coco: thực hiện lấy danh sách nhãn các đối tượng từ chú thích và 80 đối tượng, cộng thêm một số lớp đối tượng có tần suất xuất hiện nhiều trong chú thích; (2) ImageNet: là tập đơn

nhãn sẵn có; (3) ImageNet^{++}: chọn 20 ảnh trong mỗi lớp đối tượng của ImageNet và thực hiện chú thích đa nhãn cho các ảnh này. Các ảnh này được chọn sao cho số lượng các đối tượng trong ảnh là đa dạng khác nhau. Khi thực hiện chú thích đa nhãn cho 20 ảnh/lớp đối tượng, ngoài các đối tượng thông dụng xuất hiện trong ảnh cùng với đối tượng chính ImageNet, chẳng hạn *chair, man*… bên cạnh *accordion*, chúng tôi còn thực hiện phiên âm các đối tượng ngữ cảnh trong ảnh, chẳng hạn *grass, field, building, sky*, *street*… Các đối tượng này thường bị bỏ qua trong các tập dữ liệu ảnh khác nhưng lại khá quan trọng trong ImageNet bởi phần lớn ảnh trong đó có kèm theo các đối tượng ngữ cảnh, đồng thời lại rất có giá trị trong nghiệp vụ mô tả ảnh. ImageNet^{++} có thể coi là tập dữ liệu "lai ghép" đối tượng giữa hai tập ImageNet và Coco, nghĩa là các ảnh mà có xuất hiện của cả đối tượng ImageNet và Coco được chú thích đa nhãn.

Ý tưởng trích xuất "lai ghép" các đối tượng trong ảnh dựa trên các tập dữ liệu nêu trên cụ thể như sau: ImageNet được sử dụng để nhận dạng đối tượng "chính" ImageNet trong ảnh; Coco được sử dụng để nhận các đối tượng "phụ" trong ảnh, kể cả đối tượng ngữ cảnh; ImageNet^{++} được sử dụng như một "mồi nhử" cho mục đích cùng với Coco "ép buộc, định hướng" mô hình thực hiện theo hướng phân loại ảnh đa nhãn cho tập ImageNet và gợi ý các đối tượng ngữ cảnh đại diện cho tập dữ liệu ImageNet. Khi được huấn luyện theo phương pháp mặt nạ trên mô hình Confomer có thể thực hiện được phân loại ảnh đa nhãn từ tập dữ liệu đơn nhãn ImageNet.

Hình 1. Nhận dạng các đối tượng trong theo phương pháp lai ghép

Trong Hình 1, ảnh (a), ImageNet được huấn luyện để nhận dạng đối tượng chính **robin**; ảnh (b), Coco được sử dụng để huấn luyện đối tượng phụ, ngữ cảnh ***fence, green tree***; ảnh (c), khi thực hiện dự đoán, mô hình nhận dạng được cả đối tượng chính ImageNet và đối tượng phụ trong ảnh.

2.2. Mô hình Confomer với phương pháp mặt nạ

Trước đây, trong CV, mô hình CNN đã thống trị một thời gian dài và đã chứng minh được hiệu quả trong nhiều nghiệp vụ. Năm 2017, khi xuất hiện mô hình Transfomer trong lĩnh vực xử lý ngôn ngữ tự nhiên cho kết quả vượt trội, nhiều tác giả đã lấy cảm hứng từ mô hình Transfomer trong NLP áp dụng trong lĩnh vực CV và mang lại kết quả rất cạnh tranh với mô hình CNN, chẳng hạn mô hình ViT trong phân loại ảnh của Google, mô hình DETR trong phát hiện đối tượng của Facebook. Đặc biệt trong lĩnh vực chú thích/mô tả ảnh, Transfomer đã dần thay thế mô hình RNN/LSTM và chiếm xu thế chủ đạo trong các nghiên cứu về mô tả ảnh hiện nay.

Mô hình Transfomer khắc phục được các vấn đề của mô hình trước đó về sự phụ thuộc xa giữa các từ trong câu do sự biến mất của đạo hàm (gradient), tốc độ huấn luyện chậm do xử lý tuần tự, đặc biệt là cơ chế "tự chú ý" (Self Attention) và đem lại hiệu quả hơn cho mô hình. Do đó, mô hình Transformer và các biến thể của nó như BERT, GPT-3 đã tạo ra kết quả hiện đại (State of the art – SOTA) cho các tác vụ liên quan đến NLP.

Phần lõi của mô hình Transfomer là cơ chế Scaled Dot-Product Attention, trong CV, cho phép huấn luyện một (một số từ) chú ý đến một vùng ảnh nhất định. Các trọng số của lớp "chú ý" được điều chỉnh trong quá trình huấn luyện dựa theo độ đa dạng của các ảnh đầu vào. Do vậy, mô hình này khá phù hợp cho các nghiệp vụ CV.

Lấy cảm hứng từ kết hợp mô hình CNN với Transfomer và phương pháp "mặt nạ" từ mô hình BERT, chúng tôi đề xuất mô hình Mask Confomer áp dụng cho phân loại ảnh đa nhãn từ tập dữ liệu đơn nhãn và mang lại kết quả khả quan. Mô hình Conformer là sự kết hợp của CNN và Transformer.

Transformer và CNN độc lập đều có những hạn chế nhất định. Mặc dù Transformer có khả năng mô hình hóa bối cảnh toàn cục tầm xa (long-range global context), chúng lại ít có khả năng trích xuất các mẫu tính năng cục bộ chi tiết. CNN thì ngược lại, chúng có khả năng khai thác thông tin cục bộ và được sử dụng làm các khối tính toán trong CV. Chúng học các nhân dựa trên vị trí được chia sẻ thông qua các cửa sổ cục bộ với khả năng dịch chuyển và có khả năng mô hình hóa được các đặc trưng như cạnh hay hình dạng. Một giới hạn nữa của kết nối cục bộ là cần nhiều tham số hoặc lớp mạng để học được các thông tin toàn cục.

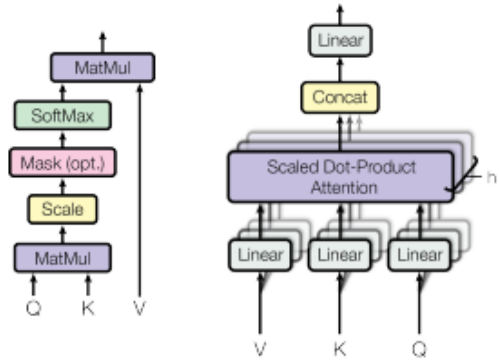
Gần đây, đã có một số nghiên cứu (Gulati và cs., 2020; Wei và cs., 2016) chỉ ra rằng, việc kết hợp tích chập (convolution) và cơ chế "chú ý" giúp cải tiến hiệu năng mô hình hơn là việc sử dụng riêng rẽ từng loại hình do sự kết hợp tạo điều kiện cho việc học được cả các đặc trưng mang tính cục bộ và toàn cục. Mô hình Conformer trong thử nghiệm của chúng tôi bao gồm Conformer encoder và Transformer decoder. Phương pháp huấn luyện được dựa theo mặt nạ hóa tựa BERT.

Scaled Dot-Product Attention là một cơ chế chú ý dựa vào việc nhân ma trận (dot-product) và sau đó nhân tiếp cho một hệ số tỉ lệ (scaling factor), cụ thể ở mô hình Transformer là $\frac{1}{\sqrt{d_k}}$.

$$Attention\,(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

trong đó: Q, K, V tương ứng là ma trận Query, Key và Value, d_k là số chiều của véc tơ key.

Multi-Headed Attention (Hình 2) được sử dụng để tăng khả năng chạy song song với cơ chế chú ý nhiều lần. Các đầu ra chú ý là độc lập, được ghép nối và chuyển đổi tuyến tính đến số chiều mong muốn. Theo trực giác, multi-head attention cho phép tham gia vào những phần khác nhau của chuỗi từ, chẳng hạn phần phụ thuộc dài hạn, phần phụ thuộc ngắn hạn…



Hình2. Scaled Dot-Product Attention (trái) và Multi-head attention (phải) (Vaswani và cs., 2017)

$$MultiHead(Q,K,V) = \left[head_1,\ldots,head_h\right]W_0,$$

trong đó $head_i = Attention(QW^q, KW^k, VW^v)$.

Khối Conformer trong thử nghiệm của chúng tôi bao gồm:

$x = FeedForward(x)$
$x = Self\_attn(x)$
$x = pre\_norm(x)$
$x = ConformerConvModule(x)$
$x = FeedForward(x)$
$x = post\_norm(x)$

$ConformerConvModule$ gồm:

$y = LayerNorm(y)$
$y = Conv1d(y)$
$y = GLU(y)$
$y = DepthWiseConv1d(y)$
$y = BatchNorm1d(y)$
$y = Swish(y)$
$y = Conv1d(y)$
$y = Dropout(y)$

Trong đó:

$FeedForward$: lớp truyền thẳng;

$Self\_attn$: lớp tự chú ý;

$pre\_norm, post\_norm$: lớp chuẩn hóa;

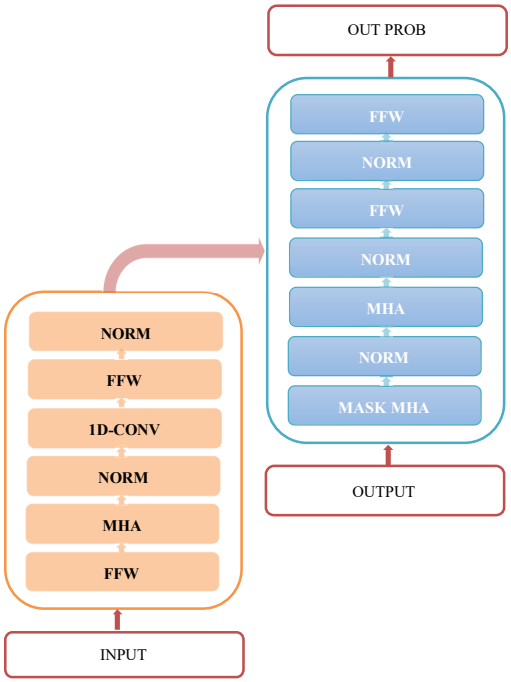
$ConformerConvModule$: lớp tích chập;

$DepthWiseConv1d, Conv1d$: module tích chập 1D;

$GLU, Swish$: hàm kích hoạt;

$Swish$: module "bỏ học".

Mô hình Conformer (Hình 3): mô hình gồm encoder là các lớp conformer block và decoder là các lớp transformer.



Hình 3. Mô hình Conformer cho phân loại ảnh đa nhãn

Pha huấn luyện:

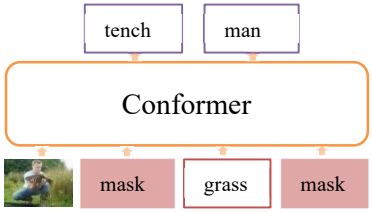
Dữ liệu ảnh đầu vào được chuyển về kích thước 224×224, sau đó qua blackbone là mô hình pretrained – Mobilenet v3 đã được huấn luyện trên 21K lớp đối tượng ImageNet, thu được bản đồ đặc trưng 960×7×7, rồi được làm phẳng thành chuỗi từ có gồm 49 từ với số chiều 960 và sau đó được chuyển thành các từ có số chiều 256.

Nhãn đầu ra là danh sách các đối tượng trong ảnh tương ứng. Với tập dữ liệu ImageNet, nhãn đầu ra gồm một nhãn là đối tượng chính trong ảnh. Tập dữ liệu ImageNet^{++} là nhãn các đối tượng trong ảnh. Với tập dữ liệu Coco, chúng tôi thực hiện theo theo phương án: 51% là nhãn đối tượng thực trong ảnh, 49% thực hiện thay thế đối tượng Coco bởi các đối tượng ImageNet, ví dụ ***dog***, *man, grass* → ***samoyed***, *man, grass*. Xác suất thay thế là 49%, nhỏ hơn xác xuất đối tượng thực 51% (đối tượng Coco) nên kết quả nhận dạng không ảnh hưởng đến đối tượng thực. Việc thay thế này nhằm mục đích tăng cường xác suất liên kết giữa đối tượng chính ImageNet với các đối tượng Coco khác, đồng thời duy trì được xác suất của đối tượng phụ Coco.

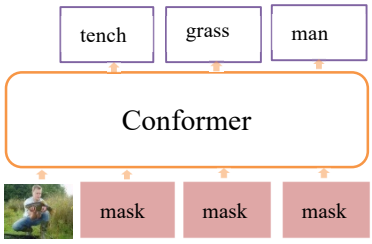
Nhãn dữ liệu đầu ra sau đó được mặt nạ hóa ngẫu nhiên (random mask) theo tỉ lệ: 33% che mặt nạ 1 từ, 33% mặt nạ 2 từ và 34% che mặt nạ toàn bộ các từ. Các nhãn đối tượng được mã hóa thành các thẻ (token) bởi pretrained – BERT và chuyển đổi (embedding) thành số chiều là 256, sau đó ghép với chuỗi từ ảnh, tạo thành chuỗi từ đầu vào cho mô hình.

Mô hình được huấn luyện để dự đoán các từ được thay thế bởi từ mặt nạ đầu vào dựa vào dữ liệu ảnh. Chỉ tính "cross entropy" cho các từ được dự đoán (che mặt nạ).

(a)



(b)



Hình 4. Mô hình Conformer cho phân loại ảnh đa nhãn theo các pha huấn luyện (a) và dự đoán (b)

Pha nhận dạng:

Dữ liệu đầu vào cho nhận dạng là ảnh cần dự đoán, phần ngôn ngữ được che mặt nạ toàn bộ. Do theo phương pháp huấn luyện của mô hình, số lượng đối tượng dự đoán phụ thuộc vào số từ mặt nạ được sử dụng trong dữ liệu đầu vào. Nghĩa là có thể thực hiện điều chỉnh số lượng mặt nạ đầu vào để dự đoán số lượng đối tượng đầu ra trong ảnh. Trong thử nghiệm của chúng tôi, nhận dạng ảnh được thực hiện theo ba bước:

Bước 1: thực hiện nhận dạng với số lượng từ mặt nạ đầu vào là 1. Khi đó, mô hình trở thành phân loại ảnh đơn nhãn và dễ dàng nhận dạng được đối tượng chính trong ảnh với độ chính xác rất cao, chẳng hạn nhận dạng được đối tượng *tench*.

Bước 2: thực hiện nhận dạng ảnh với số lượng từ mặt nạ đầu vào mong muốn, thông thường là 3. Do nhận dạng là đa nhãn, nên có một xác suất nào đó trong kết quả dự đoán không xuất hiện đối tượng chính và bị nhận dạng sang đối tượng giống loài tương ứng, ví dụ *tench* bị nhận dạng thành *fish*.

Bước 3: thực hiện thay thế đối tượng chính được nhận dạng từ bước 1 cho đối tượng giống loài được nhận dạng ở bước 2, chẳng hạn *fish* → *tench*. Kết quả đầu ra là các nhãn đối tượng bao gồm cả đối tượng chính trong ImageNet và đối tượng khác được nhận dạng trong ảnh.

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Tập dữ liệu huấn luyện

Trong các thử nghiệm, chúng tôi sử dụng tập dữ liệu ImageNet và Coco cho huấn luyện mô tả ảnh. Do hạn chế về nguồn lực và tài nguyên, chúng tôi chọn ngẫu nhiên một tập lớp đối tượng từ ImageNet 1K cho các thử nghiệm mà không làm mất tính tổng quát của phương pháp hay hiệu năng của mô hình thực hiện. Từ tập dữ liệu này, chúng tôi chọn mỗi lớp 20 ảnh đại diện để thực hiện gán đa nhãn cho ảnh. Đối với tập dữ liệu Coco, nhãn các đối tượng được lấy từ danh sách đối tượng và chú thích tương ứng. Chúng tôi cố gắng lấy danh sách đối tượng nhiều nhất có thể với mỗi ảnh để tăng khả năng nhận dạng nhiều nhãn và bù đắp cho số lượng nhãn đối tượng từ tập ImageNet.

3.2. Mô hình

Trong bài báo này, chúng tôi sử dụng mô hình Conformer với encoder gồm 1 layer conformer, decoder gồm 1 layer multi-head attention với số lượng tham số huấn luyện là 7.689.088. Mạng cơ sở được sử dụng là Mobilenet v3 đã được huấn luyện trên 21K lớp đối tượng ImageNet với số lượng tham số sử dụng là 2.971.952.

Mô hình được huấn luyện trên Colab với GPU Tesla T4 16GB. Thời gian huấn luyện là 4 giờ cho 50 epoch. Thời gian nhận dạng ảnh là 0,5 s/ảnh.

3.3. Kết quả nhận dạng

Do ImageNet là tập dữ liệu đơn nhãn, không có tập dữ liệu chuẩn cho đánh giá mô hình đa nhãn. Hơn nữa, mô hình được thử nghiệm để nhận dạng cả đối tượng ngữ cảnh nên khó so sánh với các mô hình dựa trên tập dữ liệu khác. Do vậy, để đánh giá kết quả thử nghiệm mô hình, chúng tôi thu thập 165 ảnh cho 33 lớp đối tượng ImageNet và thực hiện đánh giá theo các chỉ số tương tự mô hình NOC (Venugopalan và cs., 2017) như dưới đây:

(1) $S1$: tỉ lệ phần trăm lớp đối tượng ImageNet được nhận dạng khi có ít nhất một kết quả nhận dạng của lớp đối tượng (X) trên tổng các lớp đối tượng được dự đoán (C). $S1 = X/C * 100$. Chỉ số này cho biết tỉ lệ lớp đối tượng mới được xuất hiện trong kết quả dự đoán trên tổng số các lớp đối tượng.

(2) $S2$: tỉ lệ phần trăm các đối tượng ImageNet được nhận dạng (Y) trên tổng các kết quả dự đoán (S): $S2 = Y/S * 100$. Chỉ số này cho biết tỉ lệ đối tượng mới được xuất hiện trong các dự đoán.

(3) $S3$: độ chính xác của nhận dạng là tỉ lệ kết quả nhận dạng đúng (Z) trên tổng các dự đoán (S): $S3 = Z/S * 100$. Tỉ lệ này cho biết độ chính xác của dự đoán từ mô hình.

(4) $S4$: điểm số $F1$ được tính toán dựa trên precision và recall của đối tượng mới ImageNet trong các kết quả dự đoán. $F1 = 2 * Precision * Recall / (Precision + Recall)$. Chỉ số này là số dung hòa giữa Recall và Precision để có căn cứ đánh giá, lựa chọn mô hình.

Kết quả thử nghiệm như trong Bảng 1 và Bảng 2 dưới đây. Kết quả dự đoán một số mẫu phân loại ảnh đa nhãn được trình bày tại Hình 5.

Bảng 1. Điểm số đánh giá cho 33 lớp đối tượng ImageNet

S1 (%)	S2 (%)	S3 (%)	S4 (%)
100	86,67	77,57	86,60

Bảng 2. Điểm số đánh giá cho 19 lớp đối tượng ImageNet có cây phân cấp giống loài

S1 (%)	S2 (%)	S3 (%)	S4 (%)
100	93,68	78,95	86,81


accordion, woman, bench, grass field


water buffalo, man, grass field

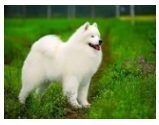

man, **bison**, grass field, road

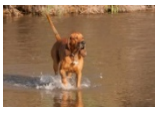

songbird, tree, leaves


goldfinch, tree


robin, grass


cello, woman, chair, room


samoyed, grass, field


redbone, water


cocker spaniel, car


siamese cat, bed


tiger cat, bed, blanket


academic gown, man, tree


lampshade, table, flowers


kimono, woman, tree

Hình 5. Một số kết quả phân loại ảnh đa nhãn với đối tượng mới từ tập dữ liệu ImageNet

Bảng 1 là chỉ số đánh giá cho 33 lớp đối tượng mới ImageNet được lựa chọn. Bảng 2 là chỉ số đánh giá 19 lớp đối tượng có xuất hiện cây phân cấp giống loài, ví dụ *dog* → *samoyed.*

Theo kết quả dự đoán, mô hình đã nhận dạng được đa nhãn: đối tượng mới trong ảnh và các đối tượng khác. Xác suất nhận dạng đối tượng mới khá cao dựa theo kỹ thuật mặt nạ và điều chỉnh số lượng từ mặt nạ đầu vào trong pha dự đoán.

4. KẾT LUẬN

Trong bài báo này, chúng tôi xây dựng và thử nghiệm mô hình Conformer theo phương pháp mặt nạ cho huấn luyện phân loại ảnh đa nhãn với đối tượng mới từ tập dữ liệu đơn nhãn ImageNet. Mô hình thực hiện khá đơn giản với tập dữ liệu đơn nhãn rất phong phú, có thể dễ dàng thu thập được từ internet. Do vậy có thể mở rộng cho tập các đối tượng bất kỳ theo mục đích sử dụng. Đồng thời, có thể áp dụng cho gán lại đa nhãn tập dữ liệu ImageNet với các đối tượng ngữ cảnh phù hợp.

LỜI CẢM ƠN

Chúng tôi xin chân thành cảm ơn nhiệm vụ cao cấp "Hỗ trợ hoạt động nghiên cứu khoa học cho nghiên cứu viên cao cấp năm 2023", mã số NVCC02.01/23-23 đã hỗ trợ trong quá trình thực hiện nghiên cứu này.

TÀI LIỆU THAM KHẢO

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-End Object Detection with Transformers. Trong A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (B.t.v), *Computer Vision – ECCV 2020* (tr 213–229). Springer International Publishing.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* (arXiv:2010.11929). arXiv.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). Conformer: Convolution-augmented Transformer for Speech Recognition. *Interspeech 2020*, 5036–5040.

Huang, H., Chen, Y., Tang, W., Zheng, W., Chen, Q.-G., Hu, Y., & Yu, P. (2020). *Multi-label Zero-shot Classification by Learning to Transfer from External Knowledge* (arXiv:2007.15610). arXiv.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., & Ferrari, V. (2020). The Open Images Dataset V4. *International Journal of Computer Vision*, *128*(7), 1956–1981.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*.

Venugopalan, S., Hendricks, L. A., Rohrbach, M., Mooney, R., Darrell, T., & Saenko, K. (2017). Captioning Images with Diverse Objects. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1170–1178.

Verelst, T., Rubenstein, P. K., Eichner, M., Tuytelaars, T., & Berman, M. (2023). Spatial Consistency Loss for Training Multi-Label Classifiers from Single-Label Annotations. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 3868–3878.

Wei, Y., Xia, W., Huang, J., Ni, B., Dong, J., Zhao, Y., & Yan, S. (2016). CNN: Single-label to Multi-label. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(9), 1901–1907.

Wu, B., Chen, W., Fan, Y., Zhang, Y., Hou, J., Liu, J., & Zhang, T. (2019). Tencent ML-Images: A Large-Scale Multi-Label Image Database for Visual Representation Learning. *IEEE Access*, *7*, 172683–172693.

Yun, S., Oh, S. J., Heo, B., Han, D., Choe, J., & Chun, S. (2021). Re-labeling ImageNet: From Single to Multi-Labels, from Global to Localized Labels. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2340–2350.