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

PHONG TỤC VÀ CON NGƯỜI TÂY NGUYÊN TRONG TRUYỆN NGẮN VŨ HẠNH

Bùi Ngọc Anh Thư[1*]

[1]Khoa Khoa học Xã hội và Nhân văn, Đại học Cần Thơ

**Email: anhthu12719@gmail.com*

Ngày nhận bài: 27/12/2022 *Ngày nhận bài sửa sau phản biện: 06/03/2023* *Ngày chấp nhận đăng: 14/3/2023*

TÓM TẮT

Vũ Hạnh là một trong những tên tuổi nổi bật của văn học đô thị miền Nam 1954 – 1975. Ông là cây bút hiếm hoi có nhiều tác phẩm đậm tinh thần dân tộc được xuất bản công khai giữa lòng đô thị Sài Gòn vào những tháng năm chiến tranh khốc liệt nhất. Tác giả viết nhiều về phẩm chất nên có của người cầm bút, về hiện trạng rối ren của xã hội, về giá trị truyền thống dân tộc, nhất là văn hóa của các dân tộc ít người ở Tây Nguyên. Bằng việc đưa văn hóa Tây Nguyên vào trong các sáng tác, nhà văn đã thổi vào các truyện ngắn của mình một làn gió mới. Từ góc độ văn hóa, đặc biệt ở hai phương diện phong tục và con người, bài viết sẽ soi chiếu, ghi nhận và đánh giá giá trị của những tác phẩm mang đậm màu sắc văn hóa Tây Nguyên dưới ngòi bút Vũ Hạnh. Qua đó, không chỉ thể hiện được mối liên hệ, gắn bó không thể tách rời giữa văn hóa và văn học mà còn khẳng định được giá trị của những yếu tố văn hóa, đặc biệt là văn hóa Tây Nguyên đậm đà bản sắc giữa bối cảnh rối ren, phức tạp của xã hội miền Nam Việt Nam trong những tháng năm đất nước bị chia cắt.

Từ khóa: *con người Tây Nguyên, phong tục Tây Nguyên, văn hóa Tây Nguyên, văn học đô thị, Vũ Hạnh.*

CUSTOMS AND PEOPLE OF CENTRAL HIGHLANDS IN VU HANH'S SHORT STORIES

ABSTRACT

One of the well-known authors in Southern urban literature between 1954 and 1975 was Vu Hanh, who was a unique writer having a number of his publicly released works showing a strong sense of national pride during the fiercest war years. He wrote extensively on topics such as what makes a good writer, society's problems, national traditions, and particularly the Central Highlands' ethnic minorities' cultures. By incorporating elements of the Central Highlands culture into his writings, the writer has breathed new life into his short stories. The article will reflect, acknowledge, and evaluate the value of works presenting the Central Highlands' cultural traits under Vu Hanh's pen from a cultural perspective, especially in terms of customs and people. By doing so, it affirms the importance of cultural factors, particularly the Central Highlands culture, which is infused with the identity between culture and literature. It can also demonstrate the close relationship and attachment between culture and literature during the period of the country's division, when South Vietnamese society was in a muddled and complex context.

Keywords: *culture of the Central Highlands, customs of the Central Highlands, people of the Central Highlands, urban literature, Vu Hanh.*

1. ĐẶT VẤN ĐỀ

Với văn học đô thị miền Nam 1954 – 1975 nói riêng, văn học Việt Nam nói chung, Vũ Hạnh không phải là một cái tên quá xa lạ. Nhà văn tên thật là Nguyễn Đức Dũng, quê ở huyện Thăng Bình, tỉnh Quảng Nam, một số bút danh khác như Hoàng Thanh Kỳ, Cô Phương Thảo, Minh Hữu, Nguyên Phủ. Ông là một trong những cây bút hiếm hoi có thể vượt qua chính sách *"hốt – cắt – đục"* của chính quyền để hoạt động công khai giữa lòng Sài Gòn trong những năm tháng chiến tranh khốc liệt nhất. Dù từng bị bắt giam đến năm lần, Vũ Hạnh vẫn bền bỉ hoạt động đơn tuyến và cho ra đời những tác phẩm tràn đầy nhiệt huyết, tinh thần dân tộc và lòng yêu nước.

Nhận xét về Vũ Hạnh, Trần Hữu Tá cho rằng: *"Vũ Hạnh không dừng lâu ở một đề tài. Ông viết truyện đường rừng (Lòng suối, Mối thù của Khoan Ray, Cây đàn trong núi, Lửa rừng...), truyện về nông thôn (Miếng thịt vịt), truyện về tầng lớp dưới đáy của xã hội thị thành (Người chồng thời đại, Mụ Tư Cò...)"* (Trần Hữu Tá (Nghiên cứu, sưu tầm, tuyển chọn), 2000). Thời kỳ đầu, nhiều người từng biết đến một Vũ Hạnh đầy thâm trầm, sâu sắc với lối văn *"ở ẩn"* trong những *Bút máu*, *Chất ngọc*... nhưng ít ai biết rằng, vẫn còn có một Vũ Hạnh rất mực tài tình, uyên bác khi viết về mảnh đất Tây Nguyên. Dấu ấn văn hóa Tây Nguyên được nhà văn thể hiện ở nhiều truyện ngắn trên tạp chí *Bách Khoa* (một tờ bán nguyệt san gồm 426 số xuất bản công khai ở Sài Gòn từ năm 1956 – 1975). Vũ Hạnh có đến 6/19 truyện ngắn viết về Tây Nguyên từng được đăng trên tạp chí Bách Khoa bên cạnh các trang viết về hoàn cảnh của những người giáo khổ trường tư, những người dân nghèo thành thị cơ cực, những phẩm hạnh và nhiệm vụ của người nghệ sĩ chân chính... Việc lựa chọn mảnh đất Tây Nguyên để khai thác đã tạo nên một chuỗi các *"truyện đường rừng"* đầy lí thú, độc đáo và giàu ý nghĩa đối với truyện ngắn Vũ Hạnh nói riêng, truyện ngắn trong văn học đô thị miền Nam nói chung. Giữa bối cảnh hết sức phức tạp của xã hội miền Nam những năm 1954 – 1975, những câu chuyện giản dị, mộc mạc về Tây Nguyên của Vũ Hạnh như một khúc ca về truyền thống, về nguồn cội, vang xa, vọng sâu vào lòng người, khơi gợi trong lòng những người dân đô thị tình yêu dân tộc và ý thức về cội nguồn khi đang đứng trước nguy cơ bị "bật gốc" bởi sự đổ xô dồn dập, ồ ạt của các lý thuyết phục vụ cho chính sách nô dịch của thực dân đế quốc.

Trước năm 1975, *"văn học Tây Nguyên vẫn như một mảnh đất hoang sơ, đây đó có những "sự sống" chưa thực mạnh mẽ."* (Đỗ Thị Thu Huyền, 2020). Các tác phẩm truyện ghi nhận một nỗ lực tuyệt vời trong việc phục dựng lại một văn hóa Tây Nguyên riêng biệt, giàu bản sắc. Trong truyện đường rừng của Vũ Hạnh trên tạp chí Bách Khoa, văn hóa Tây Nguyên được thể hiện qua nhiều phương diện, trong đó, phải kể đến những phong tục tập quán và hình ảnh con người nơi đây. Sự am tường về mảnh đất Tây Nguyên được nhà văn phát huy triệt để, kết hợp với lối kể chuyện khúc chiết, đầy kịch tính đã tạo nên những tác phẩm vừa hấp dẫn, vừa giàu ý nghĩa.

2. PHƯƠNG PHÁP NGHIÊN CỨU

Để thực hiện bài viết, chúng tôi sử dụng một số phương pháp nghiên cứu sau đây:

Phương pháp phân tích, tổng hợp: tùy vào từng tiêu chí mà lựa chọn phân tích các tác phẩm phù hợp, xác đáng. Trong từng tác phẩm, lựa chọn phân tích các chi tiết, hình ảnh thể hiện được phong tục và con người Tây Nguyên. Từ đó, có thể đúc kết thành những đánh giá, nhận xét khái quát thông qua việc tổng hợp các phép phân tích đã thực hiện.

Phương pháp nghiên cứu liên ngành: vận dụng một số nội dung của lĩnh vực văn hóa, đặc biệt là văn hóa Tây Nguyên để đối chiếu và lý giải cho ý nghĩa của các chi tiết, hình ảnh xuất hiện trong các tác phẩm.

Phương pháp hệ thống: đặt đối tượng nghiên cứu vào hệ thống văn học đô thị miền Nam 1954 – 1975 để đánh giá vai trò, tầm ảnh hưởng của truyện ngắn Vũ Hạnh. Đồng thời đặt mảng *"truyện đường rừng"* vào hệ thống truyện ngắn của Vũ Hạnh để thấy giá trị của các truyện ngắn sáng tác theo đề tài này.

3. NỘI DUNG NGHIÊN CỨU

3.1. Một số phong tục của người dân Tây Nguyên trong truyện ngắn Vũ Hạnh

3.1.1. Phong tục Cuôi ba dùm

Phong tục Cuôi ba dùm được giới thiệu trong truyện ngắn cùng tên của Vũ Hạnh. Cuôi ba dùm có nghĩa là *"ngủ thân ái"* với khách lạ, một tục lâu đời của người Thượng ở miền Đông Đường (Vũ Hạnh, 1960). Tục lệ ấy là một hành động trao gửi yêu thương, đem tấm lòng của bộ lạc để chở che, an ủi giấc ngủ cho những kẻ tha phương giữa núi rừng rộng lớn. Thông qua một đêm *"cuôi ba dùm"* giữa nhân vật tôi và Y Sao, nhà văn không chỉ khắc họa được một nét đẹp văn hóa lâu đời của đồng bào Thượng mà còn ca ngợi sự bao dung, hiếu khách của những người con núi rừng. Trước hành động thiếu kiềm chế của tôi, Y Sao vừa nhẹ nhàng vừa dứt khoát chống trả. Khí khái dõng dạc, cương quyết của Y Sao đã minh chứng cho ý thức giữ gìn, bảo vệ phong tục nghìn đời của dân tộc, không để cho nó méo mó, lệch lạc đi. Với đồng bào Thượng, sức mạnh tinh thần nằm ở chỗ tin tưởng vào phong tục: *"Đồng bào Thượng bám vào phong tục như dây leo bám lấy thân cành, như hổ dữ bám vào hang đá"* (Vũ Hạnh, 1960). Thế nên, việc gìn giữ phong tục đối với họ không chỉ là một cách để dung dưỡng giá trị truyền thống mà còn là một hành động mang ý nghĩa trang trọng, linh thiêng. Đó là mạch nguồn sự sống, là vốn liếng của dân tộc để tồn tại giữa chốn rừng núi hoang vu.

Một đêm *"cuôi ba dùm"* ngắn ngủi nhưng đã để lại cho nhân vật tôi nhiều ấn tượng sâu sắc về khí khái của đồng bào, về vẻ đẹp của phong tục truyền thống. Nhất là trong bối cảnh xã hội đương thời, khi các giá trị truyền thống đứng trước nguy cơ bị lung lạc bởi thế lực ngoại xâm, thì việc quý trọng và bảo vệ phong tục của đồng bào Thượng đã tạo thành một lời nhắc nhớ mạnh mẽ cho tinh thần dân tộc, cho ý thức công dân với cộng đồng.

3.1.2. Phong tục đón khách, cưới hỏi

Nếu trong *Cuôi ba dùm*, người đọc được chứng kiến sự hiếu khách, thân ái của người dân Tây Nguyên đối với khách ở xa thì với *Mùa xuân trên đỉnh non cao* (Vũ Hạnh, 1962a), ta lại càng bất ngờ hơn trước cách từ chối đón khách khéo léo của người dân sơn dã. Không phải lúc nào, buôn làng cũng sẵn lòng để thiết đãi khách ở xa. Khi vào giữa vụ mùa, đồng bào Thượng thường từ chối cho người lạ vào buôn bằng cách gài lá trước cổng vào. Buôn đã có dấu cấm kỵ, tuyệt nhiên không một vị khách nào có thể vào được, nếu làm trái sẽ phải chịu trừng phạt nặng nề. Sự tôn sùng phong tục một lần nữa được khắc họa sâu sắc thông qua thái độ cương quyết của người dân núi rừng. Chỉ khi nào lá cây được gỡ khỏi cổng, đồng bào mới sẵn lòng đón tiếp những người khách vào buôn. Tuy vậy, một khi đã đón khách, đồng bào Thượng du lúc nào cũng thiết đãi thân tình, bày đủ mọi món ngon, rượu quý, ca hát nhảy múa thâu đêm.

Tương tự như nhiều dân tộc khác, người Thượng xem cưới hỏi là một việc vô cùng hệ trọng. Để bày tỏ tình cảm, người con gái Thượng cũng ít khi vòng vo, rào trước đón sau mà trực tiếp thổ lộ nỗi lòng mình. Tính cách thẳng thắn, bộc trực và sự chân thành của Xiu Peng trong *Mùa xuân trên đỉnh non cao* khiến Dụng không khỏi sững sờ, bối rối. Trước chàng trai miền xuôi mà mình cảm mến, cô sơn nữ sẵn sàng đề nghị: *"Tôi cưới anh nhé!... Tôi ưng anh mà... Nghèo thì Xiu Peng nuôi. Xiu Peng có rẫy, có trâu, có gà."* (Vũ Hạnh, 1962a). Vốn theo chế độ mẫu hệ, trong phong tục của người Thượng, phụ nữ sẽ là người hỏi cưới đàn ông. Tình yêu và hôn nhân của họ cũng rất mộc mạc, không quá câu nệ lễ nghi hay so đo vật chất. Người phụ nữ thường giữ vai trò chủ động. Họ lựa chọn bạn đời bằng những tiêu chí rất giản đơn: *"Nhiều cô sẵn sàng yêu ta, chết sống với ta chỉ vì ta biết điều, hiền hậu, chứ không phải vì ta có nhiều bạc tiền, địa vị, bằng cấp hay là một thứ nhãn hiệu nào khác."* (Vũ Hạnh, 1962a). Vẻ đẹp truyền thống cùng sự thô mộc của con người giữa chốn hoang vu đã nhiều lần làm thổn thức trái tim của những kẻ tha phương. Vẻ đẹp của phong tục không chỉ làm nên bản sắc riêng biệt cho dân tộc mà còn là mạch nguồn dung dưỡng tâm hồn, làm nền tảng sức sống trường tồn của con người giữa núi rừng hoang vu.

Điều hấp dẫn những người con miền xuôi, nơi phố thị ồn ào như tôi trong *Cuôi ba dùm* hay Dụng, An, Khánh, Hiệp trong *Mùa xuân trên đỉnh non cao* không chỉ là phong cảnh hùng vĩ, nên thơ, không chỉ là những bản làng mờ trong sương cùng những cô sơn nữ xinh đẹp, dịu dàng mà còn là những phong tục giản dị, đơn sơ và rất giàu bản sắc. Sự tôn thờ tục lệ cùng ý thức giữ gìn truyền thống nghìn đời của những con người núi rừng luôn bị đánh giá là lạc hậu, quê mùa đã khiến cho những thanh niên phố thị cứ ngỡ mình tiến bộ, văn minh phải nhìn nhận lại, đánh giá lại cái gọi là nguồn cội, bản sắc. Song song đó, sự khác biệt nơi núi rừng đã làm dấy lên ước muốn thay đổi trong lòng những thanh niên như Dụng, An, Hiệp, Khánh. Họ vừa ngưỡng mộ cuộc sống bình yên nơi thôn dã, vừa mong muốn thay đổi cuộc sống hiện tại nơi phố thị của mình. Họ không chấp nhận sống một cuộc đời nhàn hạ trong khi còn biết bao sự đổi dời, xáo trộn ngoài kia. *"Làm sao có thể sống như cây cỏ không hề băn khoăn vì mục đích gì?"* (Vũ Hạnh, 1962a), câu hỏi ấy như một hồi chuông thức tỉnh và thúc giục con người buộc phải hành động để cải thiện hiện tại và thay đổi tương lai của chính mình. Kết thúc *Mùa xuân trên đỉnh non cao*, các nhân vật đã cùng đi đến một quyết định tất yếu và cần thiết để hiện thực hóa mong ước có một "mùa xuân" đích thực cho chính mình: *"... mùa xuân của ta không thể tìm thấy trên đỉnh non cao. Mùa xuân thật sự của ta ở trong cuộc sống này đây và dưới đồng bằng..."* (Vũ Hạnh, 1962a). Có thể thấy rằng, các phong tục văn hóa Tây Nguyên mà tác giả đan cài trong các tác phẩm không chỉ dừng lại ở việc quảng bá, tăng tính hấp dẫn cho câu chuyện mà còn đóng vai trò thúc đẩy tình cảm, giúp con người nhìn nhận lại những giá trị chân chính của dân tộc, của cuộc sống giữa một thời đất nước chia cắt đau thương.

3.1.3. Phong tục thách đấu

Để phân định thắng thua hoặc chứng minh trong sạch, người dân Tây Nguyên thường dùng việc thách đấu với nhau bằng nhiều hình thức, dưới sự chứng kiến của chúa làng cùng buôn dân. Dù thắng hay thua thì kết quả vẫn được xem là một sự an bày của thần linh, không thể chối cãi. Trong *Lòng suối* (Vũ Hạnh, 1962b), để chứng minh bản thân không phải là loài ma dai như lời buộc tội của Kha Roát, A Đun và Kha Roát đã chấp nhận thử thách nhúng tay vào chì nóng theo lời của ông Ha Râm. Việc thách đấu không chỉ để khẳng định sức mạnh của con người mà còn được xem là một hình thức để dò xét ý nghĩ của thần linh. Nhà văn đã khéo léo thể hiện tư duy hồn nhiên, chất phác của đồng bào dân tộc thiểu số thông qua niềm tin vào thần linh, ma quỷ. Họ tin tưởng tuyệt đối vào sự tồn tại và nhất mực tôn trọng, tuân theo phán quyết của thần linh.

Hơn thế nữa, tục lệ người Thượng từ lâu đã quy định nếu thực sự có tội, người bị buộc tội sẽ bị người buộc tội giết chết. Dựa vào điểm này, A Đun đã thách đấu hòng tiêu diệt tên Kha Roát ác ôn. Cuộc thi đấu được diễn ra bằng hình thức bơi đua. Dưới sự giúp đỡ của Y Rít, A Đun đã tiêu diệt được Kha Roát nhưng bản thân cũng phải bỏ mạng dưới dòng suối chảy xiết. Qua đó, có thể thấy, phong tục thách đấu đôi khi mang lại những hiệu quả nhất định nhưng cũng có khi gây tổn hại đến những người vô tội khi họ phải trả giá bằng mạng sống của chính mình.

Không chỉ tập trung thể hiện, ca ngợi những tục lệ đẹp đẽ, đáng quý của đồng bào Tây Nguyên, Vũ Hạnh còn chỉ ra những hạn chế, tiêu cực của những hủ tục mà đồng bào nơi đây vẫn một mực tin tưởng. Có thể thấy, bằng sự am tường sâu sắc cùng một giọng văn phóng khoáng, lôi cuốn, Vũ Hạnh đã tái hiện một bức tranh phong tục Tây Nguyên với đủ đầy màu sắc đối nghịch, đan xen giữa sáng và tối, giữa tốt và xấu, giữa nên lưu giữ và đáng bài trừ.

3.2. Con người Tây Nguyên trong truyện ngắn Vũ Hạnh

3.2.1. Con người dũng cảm, tài năng

Nổi lên trong những trang viết về Tây Nguyên của Vũ Hạnh là hình ảnh của những chàng trai khỏe khoắn, đầy tài năng và rất mực gan dạ, dũng cảm. Dường như ở bất kì câu chuyện nào, nhà văn cũng đều đặt con người vào thế đối diện với thử thách. Và cũng từ những hoàn cảnh cam go, thử thách ấy, tài

năng và lòng dũng cảm của những người con núi rừng có cơ hội được bộc lộ rõ nét.

Lòng dũng cảm được xem là một loại vũ khí sắc bén của con người để chống lại những hiểm nguy nơi rừng núi hoang vu, hiểm trở. Đứng trước chúa tể sơn lâm đầy hung tợn, Khoan Ray trong *Mối thù của Khoan Ray* (Vũ Hạnh, 1959b) không hề nao núng, e sợ. Trái lại, chàng trai dũng mãnh vẫn bình tĩnh lao mũi giáo để hạ con vật đang điên tiết. Trước sự chống trả quyết liệt, lồng lộn của con vật to lớn, Khoan Ray vẫn quyết tâm nắm chắc ngọn giáo, lao thẳng vào con vật, chiến đấu và chiến thắng nó để trả thù cho đứa con đáng thương. Sức mạnh của Khoan Ray đã được khẳng định bằng chiến tích lẫy lừng, bằng lòng quả cảm và quyết tâm cao độ. Không chỉ chiến đấu với con vật to lớn, con người Tây Nguyên còn sẵn sàng chiến đấu chống lại những kẻ gian ác, gây ra tai họa cho dân làng. A Đun và Y rít trong *Lòng suối* sẵn sàng hy sinh thân mình để tiêu diệt tên Kha Roát tham lam, hiểm độc. A Đun đã chiến thắng sức nóng của chì sôi bằng đôi bàn tay lao động chai sạn và chiến thắng kẻ ác ôn bằng tài năng cùng sự kiên nghị của mình. Con người quả cảm ấy dù biết có thể gặp phải nguy hiểm, thậm chí phải bỏ mạng giữa lòng suối nhưng vẫn khảng khái thách đấu hòng tiêu diệt kẻ gian ác, trừ hại cho buôn dân.

Giữa một không gian núi rừng rộng lớn, hoang vu, Vũ Hạnh đã khắc họa nên những con người vĩ đại, tràn đầy tài năng, sức mạnh và lòng quả cảm. Vang vọng trong từng câu chữ là tấm lòng cao quý, tài năng lẫm liệt, sức mạnh quật cường của những chàng trai sơn dã. Tất cả làm nên sự đáng trọng, đáng quý của con người Tây Nguyên giản dị mà phi thường.

3.2.2. Con người thủy chung, giàu tình cảm

Nhắc đến người dân Tây Nguyên là nhắc đến những con người đầy tình yêu thương và có lòng thủy chung, sơn sắt. Giữa chốn núi rừng hoang vu, với điều kiện sống vô cùng khó khăn, thiếu thốn nhưng những người dân nơi đây vẫn sẵn sàng sẻ chia, giúp đỡ những kẻ lỡ bước lạc đường. Tấm lòng của Y Sao trong *Cuối ba dùm* hay Xiu Peng trong *Mùa xuân trên đỉnh non cao* là một biểu hiện đầy cảm động cho sự hiếu khách cùng tấm lòng nhân hậu của người dân chốn đại ngàn. Họ không chỉ thiết đãi người xa bằng rượu ngon, thịt quý mà còn gửi trao bằng tất cả tình cảm nhiệt thành, thân ái của mình.

Dù chỉ dựng nên những cốt truyện đơn giản, với cách thể hiện nhân vật chủ yếu thông qua hành động nhưng Vũ Hạnh đã thành công trong việc khắc họa nên những con người Tây Nguyên thủy chung, trọng nghĩa tình. Đằng sau những lời lẽ không mấy hoa mỹ là một trái tim luôn chan chứa biết bao tình cảm chân thành. Tình cảm cha con thiêng liêng, tình nghĩa vợ chồng sâu nặng, tình yêu đôi lứa thiết tha đều được Vũ Hạnh khắc họa vô cùng cảm động. Trong *Mối thù của Khoan Ray*, người đọc không khỏi cảm phục trước một người cha bất chấp mọi hiểm nguy để tìm diệt con ác thú, trả thù cho đứa con bé bỏng của mình. Đến với *Lòng suối*, ta lại có cơ hội chứng kiến tình nghĩa vợ chồng sâu nặng giữa A Đun đối với người vợ Mi Sao. Ngay cả khi Mi Sao chỉ còn lại là một nấm mồ lạnh lẽo, A Đun vẫn quyết lòng chung thủy và ngày đêm thương nhớ nàng. Và càng xúc động hơn trước tình yêu âm thầm mà cháy bỏng của Yan đối với Y Mo trong *Ổ ong rừng* (Vũ Hạnh, 1959a). Chàng trai thật thà Yan chỉ vì một lời nói đùa mà sẵn sàng bất chấp thử thách ngậm cả một ổ ong rừng để có thể cưới được Y Mo làm vợ. Tất cả họ đã cùng tạo nên hình ảnh con người Tây Nguyên tuy giản dị, mộc mạc mà chan chứa bao sức mạnh tiềm tàng cấu thành bởi lòng thủy chung và tình yêu thương bao la rộng lớn.

4. KẾT LUẬN

Hơn 80 năm cầm bút, nhà văn Vũ Hạnh đã trở thành "một tên tuổi lớn, một tấm gương về nhân cách sống, về tình yêu dành cho văn học với ý nghĩa cao đẹp nhất của người cầm bút" (Ban Tuyên giáo Thành ủy TPHCM, 2021). Riêng việc đưa văn hóa Tây Nguyên vào trang viết có thể xem là một hướng đi mới mẻ trong văn học đô thị miền Nam giai đoạn 1954 – 1975. Khi bầu không khí văn học đô thị đã đặc quánh bởi những tác phẩm ngoại lai cổ vũ lối sống hưởng thụ, buông thả hay những sáng tác "minh họa" mang tính hiệu

triệu, đậm chất chính trị thì "truyện đường rừng" của Vũ Hạnh đã đưa người đọc đến với một không gian mới với tất cả sự thô mộc, hồn nhiên của thiên nhiên và con người.

Với những câu chuyện giản dị, mộc mạc, Vũ Hạnh liên tiếp mở ra trước mắt người đọc một bức tranh thú vị, đầy màu sắc về phong tục và con người Tây Nguyên. Vũ Hạnh không hẳn là cây bút xuất sắc nhất khi viết về Tây Nguyên, nhưng thông qua những câu chuyện đăng trên tạp chí Bách Khoa, nhà văn đã mang đến cho người đọc những tri nhận mới mẻ về văn hóa ở mảnh đất xa xôi, huyền bí này. Giữa chốn đại ngàn hoang sơ, rộng lớn, con người không hề nhỏ bé, lầm lũi mà tỏa sáng rạng rỡ với những vẻ đẹp truyền thống đáng quý, đáng trọng. Những bài học về nghĩa thủy chung, tình dân tộc, về giá trị của bản sắc truyền thống cùng những phẩm chất đáng quý cũng từ đó mà âm thầm thấm nhuần vào trong tâm khảm của mỗi con người.

TÀI LIỆU THAM KHẢO

Ban Tuyên giáo Thành ủy TPHCM. (2021). *Nhà văn Vũ Hạnh – Biểu tượng đẹp của tinh thần văn hóa dân tộc* [Văn hóa]. Trang tin điện tử Đảng bộ Thành phố Hồ Chí Minh. http://www.thanhuytphcm.vn-/tin-tuc/nha-van-vu-hanh-bieu-tuong-dep-cua-tinh-than-van-hoa-dan-toc-1491882534

Đỗ Thị Thu Huyền. (2020). *Vài nét về văn học hiện đại các dân tộc thiểu số Tây Nguyên* [Bình luận văn nghệ]. Văn nghệ quân đội. http://vannghequandoi.com.vn-/binh-luan-van-nghe/vai-net-ve-van-hoc-hien-dai-cac-dan-toc-thieu-so-tay-nguyen-_11391.html

Trần Hữu Tá (Nghiên cứu, sưu tầm, tuyển chọn). (2000). *Nhìn lại một chặng đường văn học*. Nxb Thành phố Hồ Chí Minh.

Vũ Hạnh. (1959a). Ổ ong rừng. *Tạp chí Bách khoa, 62,* 98–103.

Vũ Hạnh. (1959b). Mối thù của Khoan Ray. *Tạp chí Bách khoa, 55,* 84–97.

Vũ Hạnh. (1960). Cuôi ba dùm. *Tạp chí Bách khoa, 73,* 221–231.

Vũ Hạnh. (1962a). Mùa xuân trên đỉnh non cao. *Tạp chí Bách khoa, 122,* 142–166.

Vũ Hạnh. (1962b). Lòng suối. *Tạp chí Bách khoa, 132,* 83–91.

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.

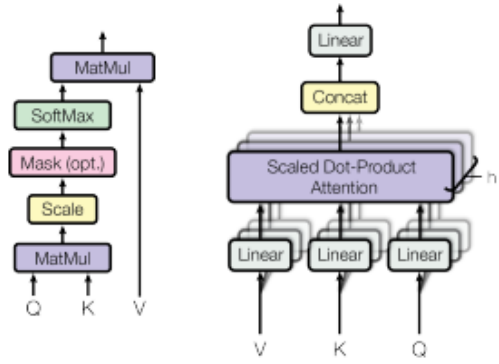
Gần đây, đã có một số nghiên cứu (Gulati và cs., 2020; Wei và cs., 2016) chỉ ra rằng, việc kết hợp tích chập (convolution) và cơ chế "chú ý" giúp cải tiến hiệu năng mô hình hơn là việc sử dụng riêng rẽ từng loại hình do sự kết hợp tạo điều kiện cho việc học được cả các đặc trưng mang tính cục bộ và toàn 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;

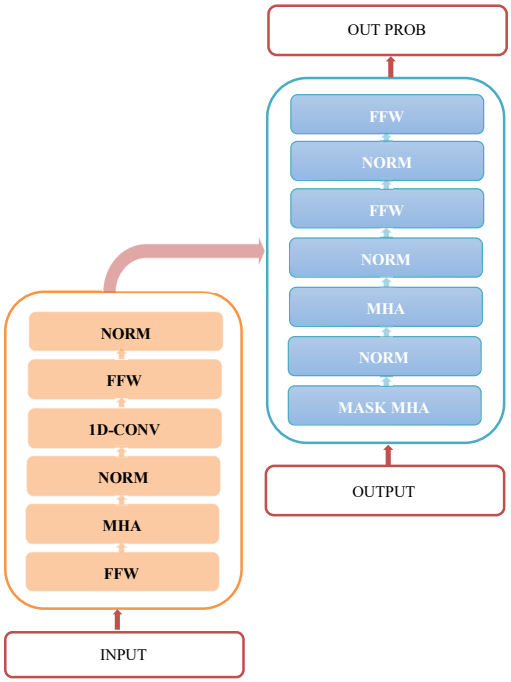
$ConformerConvModule$: lớp tích chập;

$DepthWiseConv1d, Conv1d$: 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:

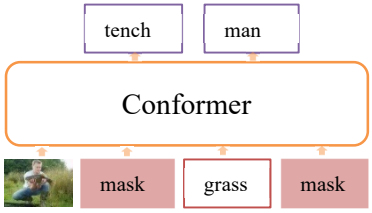
Dữ liệu ảnh đầu vào được chuyển về kích thước 224×224, sau đó qua blackbone là mô hình pretrained – Mobilenet v3 đã được huấn luyện trên 21K lớp đối tượng ImageNet, thu được bản đồ đặc trưng 960×7×7, rồi được làm phẳng thành chuỗi từ có gồm 49 từ với số chiều 960 và sau đó được chuyển thành các từ có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

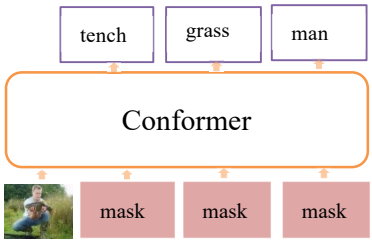
Nhãn dữ liệu đầu ra sau đó được mặt nạ hóa ngẫu nhiên (random mask) theo tỉ lệ: 33% che mặt nạ 1 từ, 33% mặt nạ 2 từ và 34% che mặt nạ toàn bộ các từ. Các nhãn đối tượng được mã hóa thành các thẻ (token) bởi pretrained – BERT và chuyển đổi (embedding) thành số chiều là 256, sau đó ghép với chuỗi từ ảnh,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

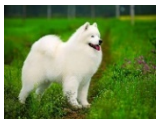

man, **bison**, grass field, road

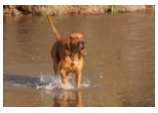

songbird, tree, leaves


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.