TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

ĐỀ CƯƠNG KHOÁ LUẬN TỐT NGHIỆP

DIỄN GIẢI PHƯƠNG PHÁP HỌC MÁY CHO TÁC VỤ PHÂN LỚP ẢNH

(Interpretable Machine Learning in Image Classification Tasks)

1 THÔNG TIN CHUNG

Người hướng dẫn:

– ThS. Nguyễn Trần Duy Minh (Khoa Công nghệ Thông tin)

Nhóm sinh viên thực hiện:

1. Nguyễn Khải Phú (MSSV: 20127062)
2. Thái Cẩm Phong (MSSV: 20127406)

Loại đề tài: Nghiên cứu

Thời gian thực hiện: Từ tháng 1/2024 đến tháng 7/2024

2 NỘI DUNG THỰC HIỆN

2.1 Giới thiệu về đề tài

Phân loại ảnh là một tác vụ cơ bản trong Thị giác Máy tính, mục đích nhằm gán nhãn cho ảnh đầu vào từ một tập các danh mục (còn gọi là lớp) đã định trước. Bài toán phân loại ảnh đã có từ những năm 1980, với những phương pháp giải đầu tiên chủ yếu xoay quanh các chiến lược trích xuất đặc trưng một cách thủ công. Năm 1988, Yann LeCun giới thiệu LeNet-5 [1], một mô hình sử dụng các lớp tích chập cho quá trình trích xuất đặc trưng. Công trình này đã đặt nền móng cho sự phát triển của mạng nơ-ron tích chập (Convolutional Neural Network - CNN). Năm 2012, AlexNet, do Krizhevsky và các cộng sự phát triển [2], giành chiến thắng trong cuộc thi phân loại ảnh ImageNet Large Scale Visual Recognition Challenge (ILSVRC) với độ chính xác vượt trội so với các phương pháp học máy truyền thống, cao hơn đến 40%. Kế thừa thành công của AlexNet, một loạt các kiến trúc CNN mới như VGGNet [3], GoogLeNet [4], ResNet [5] ra đời và được sử dụng rộng rãi trong nhiều bài toán phân loại ảnh khác nhau. Có thể dễ dàng nhận thấy ảnh hưởng to lớn và quan trọng của các mô hình CNN trong mảng Thị giác Máy tính, mở ra nhiều tiềm năng ứng dụng trong thực tế.

ResNet	18-layer	34-layer	50-layer	101-layer	152-layer
FLOPs	1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Bảng 1: Số lượng phép tính toán với số chấm động (FLOPs) tương ứng với từng biến thể của ResNet

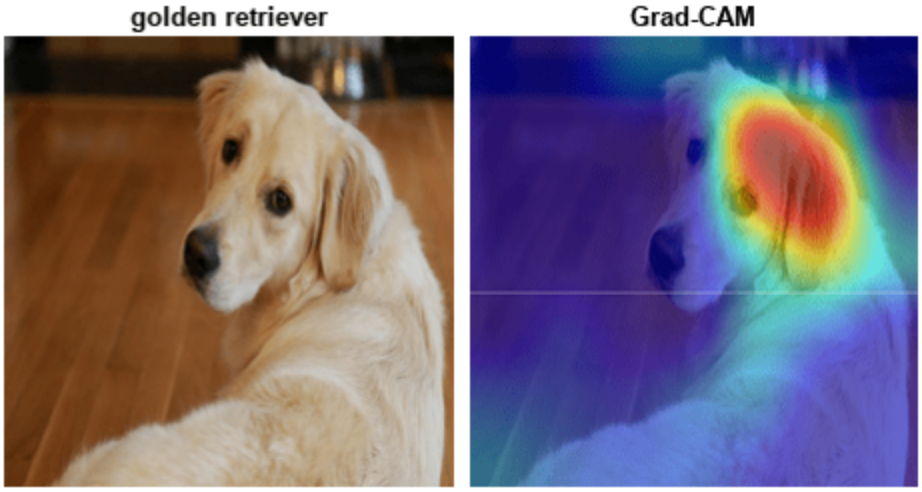
Khác với các phương pháp học máy truyền thống, các mô hình học sâu như CNN có kiến trúc phức tạp với nhiều lớp xử lý và số lượng tham số cực lớn. CNN có hiệu suất ấn tượng trong nhiều bài toán phân loại ảnh khác nhau, nhưng đi kèm với đó là khối lượng tính toán khổng lồ để phân tích ảnh đầu vào qua nhiều giai đoạn trước khi đưa ra dự đoán. Đối với đa phần người sử dụng, kiến thức chuyên môn về Trí tuệ nhân tạo còn hạn chế, sẽ rất khó để diễn giải cho họ hiểu được

cách thức hoạt động của các mô hình học sâu. Hệ quả tất yếu của việc này là quá trình phát triển các kỹ thuật phân loại ảnh ứng dụng Trí tuệ nhân tạo bị hạn chế, do những lo ngại về tính đúng đắn và tính minh bạch từ dự đoán đầu ra của các mô hình. Điều này đặc biệt đúng với các ứng dụng công nghệ thuộc các mảng yêu cầu độ chính xác cao như quốc phòng - an ninh, các hệ thống y tế hay xe tự hành. Đây là những lĩnh vực mà chỉ cần sai sót nhỏ trong tính toán cũng có thể gây ra hậu quả nghiêm trọng liên quan đến tính mạng con người, nên người sử dụng đặc biệt quan tâm đến cơ chế hoạt động, tính toán bên trong mô hình. Việc giải đáp những mối lo ngại trên nhằm đạt được sự tin tưởng của người sử dụng là nhu cầu cấp thiết, trong bối cảnh mà chúng ta đang tìm cách khai thác những tiềm năng to lớn của Trí tuệ nhân tạo để phục vụ con người trong nhiều lĩnh vực khác nhau. Trước tình trạng thực tế như trên, diễn giải các phương pháp học máy (Interpretable Machine Learning) trở thành chủ đề nghiên cứu nhận được sự quan tâm của nhiều chuyên gia trong ngành. Xét cụ thể với các mô hình CNN trong bài toán phân loại ảnh, một số phương pháp diễn giải tiêu biểu được nghiên cứu trong những năm gần đây có thể kể đến như sau:

- **Biểu đồ nhiệt** là một cách giải thích rất hiệu quả trong thực tế. Đối với mỗi ảnh đầu vào, ta có thể tính toán và chồng một biểu đồ nhiệt lên trên, thể hiện các khu vực trên ảnh đầu vào có ảnh hưởng lớn đến dự đoán đầu ra của mô hình. Từ đó, người sử dụng có thể dễ dàng biết được mô hình CNN chú trọng vào những đặc điểm nào trên ảnh khi phân lớp.
- Một loại phương pháp khác cũng nhận được nhiều sự chú ý là các **kỹ thuật lan truyền ngược**. Các nơ-ron ở lớp cuối cùng sẽ được chấm điểm dựa trên đóng góp của chúng đối với đầu ra, sau đó các số điểm được lan truyền ngược qua tất cả các lớp của mô hình, về tới hình ảnh đầu vào. Các khu vực khác nhau trong ảnh sẽ nhận được những số điểm khác nhau, đại diện cho đóng góp của chúng với kết quả phân lớp.

- Một hướng tiếp cận độc đáo khác xuất hiện gần đây là **tính xấp xỉ** các mô hình CNN phức tạp bằng cách xây dựng một mô hình đơn giản hơn, đúng với một khu vực lân cận xung quanh điểm dữ liệu cần giải thích. Mô hình đơn giản có thể là những mô hình học máy với thuật toán dễ hiểu như thuật toán hồi quy tuyến tính.

Chủ đề diễn giải phương pháp học máy hiện nay vẫn được xem là tương đối mới, nên các phương pháp diễn giải như trên nhìn chung vẫn còn nhiều nhược điểm. Nội dung của bài chủ yếu sẽ chú trọng vào các phương pháp **Biểu đồ nhiệt** do đây là phương pháp giải thích trực quan nhất, đơn giản cả trong quá trình phát triển và trong lúc sử dụng, rất phù hợp với nhu cầu thực tế.



Hình 1: Minh họa lời giải thích dạng biểu đồ nhiệt [6]

2.2 Mục tiêu đề tài

- Diễn giải mô hình học máy trong tác vụ phân lớp ảnh. Tìm hiểu xem các mô hình tập trung vào những vị trí, đặc điểm nào của ảnh đầu vào để tiến hành phân lớp
- Tìm ra hướng tiếp cận mới để cải thiện thời gian thực thi nhưng không mất đi độ chính xác của các thuật toán diễn giải phương pháp học máy.

- Thực nghiệm, kiểm tra thuật toán đề xuất, so sánh kết quả đầu ra với các phương pháp khác.

2.3 Đối tượng nghiên cứu

- Đối tượng nghiên cứu là các kỹ thuật diễn giải phương pháp học máy được dùng phổ biến trong tác vụ phân lớp ảnh.
- Khảo sát, phân tích ưu nhược điểm của các nghiên cứu về diễn giải phương pháp học máy trong tác vụ phân lớp ảnh, từ đó xây dựng một phương pháp mới nhanh hơn và hiệu quả hơn cho bài toán diễn giải phương pháp học máy nêu trên.

2.4 Phạm vi của đề tài

- Đề tài chủ yếu khám phá các phương pháp diễn giải mô hình CNN khi được áp dụng vào tác vụ phân lớp ảnh.
- Tập trung vào các biến thể của thuật toán CAM, sẽ được giải thích chi tiết trong phần 2.5.
- Dữ liệu là hình ảnh của đa dạng các đối tượng khác nhau như đồ dùng hàng ngày, phương tiện giao thông, cảnh quan thiên nhiên, động vật,...

2.5 Cách tiếp cận dự kiến

Trong những năm gần đây, đã có nhiều kĩ thuật diễn giải mô hình học máy được nghiên cứu và phát triển. Phần dưới đây sẽ trình bày sơ bộ các phương pháp diễn giải tiêu biểu.

LIME (Local Interpretable Model-agnostic Explanations) [7] là kỹ thuật giải thích dự đoán của mô hình học máy ở phạm vi cục bộ, bằng cách xây dựng và huấn luyện một mô hình phụ với bộ dữ liệu nằm ở lân cận điểm dữ liệu cần được giải thích. Mô hình phụ được chọn là những mô hình đơn giản, con người có thể hiểu

được, ví dụ như hồi quy tuyến tính. LIME có thể giải thích được các mô hình hộp đen (black box models) phức tạp mặc dù không có thông tin gì về kiến trúc của mô hình. Tuy nhiên, LIME không hoạt động tốt khi giải thích các điểm dữ liệu quá khác biệt so với dữ liệu huấn luyện, tức nằm ngoài phạm vi cục bộ.

CAM (Class Activation Map) [8], đề xuất bởi Zhou và các cộng sự, là một bước đột phá đối với quá trình phát triển các phương pháp diễn giải bằng biểu đồ nhiệt. Bằng cách thay đổi mô hình CNN gốc cần được giải thích để thêm một vài lớp cần thiết vào trong mạng nơ-ron và huấn luyện lại, CAM có thể thể hiện các khu vực trong ảnh đầu vào có đóng góp nhiều nhất vào dự đoán của một lớp nào đó. Tuy nhiên, việc yêu cầu thay đổi kiến trúc mô hình cũng chính là nhược điểm lớn nhất của CAM, do người phát triển cần phải huấn luyện lại mô hình, đòi hỏi thêm nhiều tài nguyên và thời gian tính toán hơn.

Dựa trên CAM, Selvaraju cùng các cộng sự đã phát triển **Grad-CAM** [9], tận dụng các giá trị đạo hàm (gradient) cùng với các bản đồ đặc trưng sinh ra từ lớp tích chập cuối cùng để tạo ra các diễn giải ở dạng biểu đồ nhiệt, trong đó nhấn mạnh những khu vực quan trọng trong ảnh đầu vào đối với dự đoán của mô hình. Do các tham số của Grad-CAM đều đã được tính toán sẵn trong quá trình hoạt động của mô hình, phương pháp này không yêu cầu bất kỳ thay đổi nào với kiến trúc mô hình gốc, đồng thời tốc độ xử lý cũng nhanh hơn CAM rất nhiều. Mặc dù vậy, hiện tại vẫn chưa có nghiên cứu nào thực sự chứng minh được độ chính xác của các phương pháp CAM dựa trên đạo hàm như Grad-CAM, nên vẫn còn đó những câu hỏi chưa được giải đáp về tính minh bạch.

Gần đây, nhóm các tác giả đại diện bởi Zheng đã giới thiệu **Shap-CAM** [10], tận dụng những điểm mạnh của giá trị Shapley (Shapley values) [11] để tính toán mức độ quan trọng của các điểm ảnh, thay vì sử dụng đạo hàm như Grad-CAM. Xuất phát từ Lý thuyết Trò chơi, giá trị Shapley được nghiên cứu và khai thác trong lĩnh vực học máy và trí tuệ nhân tạo vì khả năng diễn giải được hoạt động của các mô hình học sâu một cách trực quan và hiệu quả. Cụ thể, giá trị Shapley

xem xét mối quan hệ giữa các điểm ảnh khác nhau, từ đó tính toán ra mức độ đóng góp trung bình và lấy chỉ số này đại diện cho độ quan trọng của điểm ảnh. Khác với các phương pháp CAM dựa trên đạo hàm, lý thuyết về giá trị Shapley đã được nghiên cứu và phát triển trong một khoảng thời gian dài (được công bố năm 1957 [11]), nên các phương pháp diễn giải dựa trên giá trị Shapley được đánh giá là minh bạch và đáng tin cậy. Nhược điểm chính của phương pháp này có độ phức tạp cao $O(2^n)$, yêu cầu rất nhiều tài nguyên và thời gian để tính toán chính xác giá trị Shapley.

Nội dung và mục đích chính của đề tài này là kết hợp các phương pháp CAM để cải thiện tốc độ tính toán của Shap-CAM, từ đó có thể tạo ra những biểu đồ diễn giải minh bạch, đáng tin cậy với thời gian tính toán ngắn hơn.

2.6 Kết quả dự kiến của đề tài

- Chứng minh tính khả thi của kỹ thuật được đề xuất thông qua các phương pháp thực nghiệm để so sánh tốc độ và độ chính xác với các thuật toán trước đây.
- Xuất bản một bài báo khoa học cho hội nghị/tạp chí trong nước hoặc quốc tế trình bày về phương pháp được đề xuất.
- Giá trị thực tiễn: cải tiến về tốc độ thực thi trong việc diễn giải phương pháp học máy trong tác vụ phân lớp ảnh

2.7 Kế hoạch thực hiện

Thời gian	**Nội dung**
01 tháng	Khảo sát tổng quan về bài toán diễn giải phương pháp học máy trong tác vụ phân lớp ảnh.
01 tháng	Tìm ra hướng tiếp cận sơ bộ cho bài toán đã đề ra.
02 tháng	Cài đặt môi trường và tiến hành thực nghiệm. Tinh chỉnh lại phương pháp để nâng cao hiệu quả.
02 tháng	Hoàn chỉnh thuật toán để xuất. Viết 01 bài báo khoa học cho hội nghị - tạp chí chuyên ngành trong hay ngoài nước.
01 tháng	Viết khóa luận tốt nghiệp. Báo cáo khóa luận tốt nghiệp.

Bảng 2: Kế hoạch thời gian cho khóa luận tốt nghiệp

Thời gian	**Thành viên**	**Nội dung**
01 tháng	Nguyễn Khải Phú	Khảo sát các bài toán diễn giải phương pháp học máy.
	Thái Cẩm Phong	Tổng quan các mô hình học máy cho tác vụ phân lớp ảnh.
01 tháng	Nguyễn Khải Phú	Đưa ra những thiếu sót của các phương pháp diễn giải trước đây để cải tiến. Lập kế hoạch triển khai thực nghiệm.
	Thái Cẩm Phong	Tìm hiểu cách triển khai thuật toán của phương pháp.
02 tháng	Nguyễn Khải Phú	Tiến hành cài đặt môi trường và thực nghiệm.
	Thái Cẩm Phong	
02 tháng	Nguyễn Khải Phú	Tìm hiểu về các cách đánh giá độ hiệu quả và viết báo cáo.
	Thái Cẩm Phong	Hoàn chỉnh thuật toán đề xuất.
01 tháng	Nguyễn Khải Phú	Hoàn chỉnh khóa luận tốt nghiệp.
	Thái Cẩm Phong	

Bảng 3: Kế hoạch phân chia công việc cho khóa luận tốt nghiệp

Tài liệu

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, pp. 84–90, May 2017.

[3] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Apr. 2015. arXiv:1409.1556 [cs].

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," Sept. 2014. arXiv:1409.4842 [cs].

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 2015. arXiv:1512.03385 [cs].

[6] MathWorks, "Grad-cam reveals the why behind deep learning decisions." `https://www.mathworks.com/help/deeplearning/ug/gradcam-explains-why.html`.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," Aug. 2016. arXiv:1602.04938 [cs, stat].

[8] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," Dec. 2015. arXiv:1512.04150 [cs].

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *Int J Comput Vis*, vol. 128, pp. 336–359, Feb. 2020. arXiv:1610.02391 [cs].

[10] Q. Zheng, Z. Wang, J. Zhou, and J. Lu, "Shap-CAM: Visual Explanations for Convolutional Neural Networks based on Shapley Value," Aug. 2022. arXiv:2208.03608 [cs].

[11] L. S. Shapley, "17. A Value for n-Person Games," in *Contributions to the Theory of Games (AM-28), Volume II* (H. W. Kuhn and A. W. Tucker, eds.), pp. 307–318, Princeton University Press, Dec. 1953.

XÁC NHẬN
CỦA NGƯỜI HƯỚNG DẪN
(Ký và ghi rõ họ tên)

TP. Hồ Chí Minh, ngày... tháng... năm...
NHÓM SINH VIÊN THỰC HIỆN
(Ký và ghi rõ họ tên)