ĐẠI HỌC QUỐC GIA TP HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO MÔN HỌC TRÍ TUỆ NHÂN TẠO NÂNG CAO

Cao học khóa 2022

Đề tài:

Tutorial on Interpretable Machine Learning

Dựa trên bài báo:

MICCAI'18 Tutorial on Interpretable Machine Learning

Wojciech Samek, Fraunhofer Heinrich Hertz Institute
Frederick Klauschen, Charité University Hospital Berlin
Klaus-Robert Müller, Technische Universität Berlin

GIÁO VIÊN HƯỚNG DẪN: GS. TS. Lê Hoài Bắc

TS. Nguyễn Ngọc Thảo

HỌC VIÊN THỰC HIỆN: Võ Hoài Danh – 22C15025

Nguyễn Khắc Duy – 22C15026

Đoàn Minh Hòa – 22C15028

Lê Thị Cẩm Thi – 22C15044

TPHCM - 3/2023

MỤC LỤC

PHÂN CÔNG CÔNG VIỆC

STT	MSSV	Họ và tên	Phân công công việc	Vị trí trang báo cáo thực hiện
1	22C15025	Võ Hoài Danh	- Bản đồ nhiệt cho bằng chứng ung thư - Bản đồ nhiệt cho dấu hiệu phân tử - Kỹ thuật bản đồ nhiệt tìm kiếm độ sai lệch	**Từ:** Trang #33 (5. Case Study: Mô hình học máy) **Đến:** Trang #49 (…..công cụ này trong tương lai).
2	22C15026	Nguyễn Khắc Duy	- Phương pháp diễn giải mô hình. - Mô tả Pixel-wise decomposition . - Mô tả Phương Pháp Layer-wise Relevance Propagation. - Các phương pháp diễn giải.	**Từ:** Trang #2 (Layer-wise Relevance Propagation) **Đến:** Trang #22 (đánh giá phương pháp Layer-wise Relevance Propagation)
3	22C15028	Đoàn Minh Hòa	- Đánh giá và so sánh các phương pháp. - Ứng dụng của LRP.	**Từ:** Trang #22 (3. Đánh giá & So sánh phương pháp) **Đến:** Trang #32 (…..cho bộ phân loại Fish Vector.)
4	22C15044	Lê Thị Cẩm Thi	- Giới thiệu về XAI - Các phương pháp diễn giải.	**Từ:** Trang #1 (1. Giới thiệu và động lực) **Đến:** Trang #12 (……Pixel-wise Decomposition)

TÓM TẮT ĐỒ ÁN

Ngày nay, các thuật toán Học máy (Machine Learning) như Mạng Neural học sâu (Deep Neural Network - DNN) có thể khai thác và xử lý được số lượng lớn dữ liệu từ tập huấn luyện và chuyển đổi chúng thành các dự đoán có tính chính xác cao. Các mô hình DNN đã đạt đến độ chính xác hàng đầu trong các ứng dụng thực tế. Tuy nhiên, các mô hình học máy thường được xem là hộp đen, vì với tính phi tuyến và cấu trúc phức tạp bên trong, khó để hiểu và định lượng được quá trình suy luận của chúng, ví dụ như những gì đã làm cho mô hình học máy huấn luyện đưa ra một quyết định cụ thể cho một điểm dữ liệu. Điều này là một hạn chế lớn đối với các ứng dụng trong đó có tính giải thích của quyết định. Ví dụ trong chuẩn đoán y tế, các dự đoán không chính xác có thể gây ra tử vong, do đó các dự đoán đơn giản của hộp đen không thể tin tưởng hoàn toàn được. Thay vào đó, các dự đoán nên được làm rõ hoặc cho một chuyên gia y tế xác minh.

Gần đây, vấn đề về tính minh bạch đã nhận được nhiều sự chú ý của cộng đồng học máy. Nhiều phương pháp đã được phát triển để hiểu những gì một mô hình học sâu đã học, một số phương pháp tập trung vào việc hình dung các neural hoặc lớp neural cụ thể, các phương pháp khác tập trung vào hình dung tác động của vùng cụ thể trên một hình ảnh đầu vào. Một câu hỏi quan trọng là làm thế nào có thể đo lường một cách khách quan chất lượng của một giải thích cho dự đoán của DNN và làm thế nào để sử dụng các giải thích này để cải thiện mô hình học.

1. GIỚI THIỆU & ĐỘNG LỰC

1.1. Giới thiệu

Trong khi trí tuệ nhân tạo (Artificial Intelligence - AI) thâm nhập ngày càng sâu rộng vào mọi lĩnh vực của đời sống (từ nhận dạng khuôn mặt, ứng dụng đàm thoại, đến xe tự hành hay hệ thống siêu cá nhân hóa,…), thì việc xây dựng và củng cố niềm tin ở AI càng trở nên quan trọng. Tuy nhiên, hầu hết người dùng đều không thể quan sát trực quan hay nhận biết cách thức AI đưa ra quyết định. Phần lớn các thuật toán đang được sử dụng cho học máy cũng gặp hạn chế trong việc thiết lập các lý giải cụ thể, làm ảnh hưởng tiêu cực đến sự tin tưởng của con người đối với các hệ thống trí tuệ nhân tạo. Nhu cầu hiện nay là AI phải vừa có khả năng hoạt động hiệu quả, vừa có thể đưa ra các giải thích minh bạch cho các quyết định của mình. Đây được gọi là Explainable AI (XAI).

Các mô hình Deep Neural Network như một hộp đen, khi ta có một lượng dữ liệu đầu vào rất lớn. Bằng sức mạnh của sự tính toán của máy tính và mô hình DNN chúng ta trích rút ra được các đặc trưng của dữ liệu. Nhưng những đặc trưng này gồm những gì, có đúng chất lượng hay không? Và làm thế nào để chúng ta có thể cải thiện được chất lượng của mô hình máy học?

1.2. Động lực

1.2.1. Xác minh rằng trình phân loại hoạt động như mong đợi

Việc đưa ra quyết định sai có thể gây ra tốn kém và nguy hiểm. Ví dụ:

- Xe tự láy gây tai nạn vì nhận diện sai.
- Hệ thống chuẩn đoán y tế phân loại sai bệnh nhân.

1.2.2. Hiểu điểm yếu và cải thiện bộ phân loại

Đối với hệ thống học máy thông thường không có khả năng diễn giải, thì đầu ra là lỗi rất tổng quát, chúng ta sẽ không thể biết điều gì đã xảy ra trong chiếc hộp đen của mô hình học máy. Do đó mà chúng ta sẽ không biết bắt đầu xem xét từ đâu để cải thiện hệ thống của mình.

Đối với hệ thống học máy có thêm khả năng diễn giải. Thì đầu ra lỗi tổng quát nhưng chúng ta có thể thấy được điều gì đang diễn ra trong chiếc hộp đen của mô hình học máy, bên cạnh đó có thêm kinh nghiệm của con người. Chúng ta có thể tham gia vào hiệu chỉnh vào hệ thống Học máy để cải thiện hệ thống của mình.

1.2.3. Học hỏi những điều mới từ Trí tuệ nhân tạo

Khi trí tuệ nhân tạo càng phát triển, đặc biệt là các lĩnh vực phát triển từ thuật toán học tăng cường (Reinforcement Learning RL) và Giải thuật di truyền (Genetic Algorithm

- GA). Các thuật toán này có thể tìm ra các lời giải mới hơn những lời giải trước đây do con người tạo ra. Và tối ưu hoá các lời giải đã có của con người. Từ đó, còn người có thể học được từ những lời giải do máy tính tìm ra.

1.2.4. Khả năng giải thích trong khoa học:

Tìm hiểu về các cơ chế vật lý / sinh học / hóa học. (ví dụ: tìm gen liên quan đến ung thư, xác định vị trí gắn kết...

Các ví dụ về ứng dụng trí tuệ nhân tạo trong y học:

Năm 2018, các nhà nghiên cứu tại Bệnh viện Đại học quốc gia Seoul (Hàn Quốc) đã phát triển một thuật toán AI gọi là DLAD (Deep Learning based Automatic Detection) để phân rã hình ảnh chụp X-quang ngực cũng như phát hiện sự phát triển bất thường của tế bào (nguyên nhân gây ra bệnh ung thư). Cùng một hình ảnh phim chụp, kết quả đọc của máy tính sẽ được so sánh với kết quả đọc của nhiều bác sỹ khác nhau và thật ngạc nhiên khi những kết luận từ máy tính là vượt trội hơn so với 17/18 các bác sỹ tham gia đọc phim.

Cũng trong năm 2018, thuật toán thứ hai được phát triển bởi các nhà nghiên cứu tại Google AI Healthcare. Họ tạo ra một thuật toán gọi là LYNA (Lymph Node Assistant) giúp phân rã các mẫu bệnh phẩm nhuộm màu để xác định khối ung thư vú di căn từ hạch bạch huyết. Kết quả rất thú vị khi thuật toán này có thể xác định các vùng khả nghi mà mắt thường của con người không thể phân biệt được trong các mẫu sinh thiết được đưa ra. LYNA thử nghiệm trên hai tập dữ liệu và được chứng minh là phân loại chính xác mẫu là ung thư hay không phải ung thư chính xác lên đến 99%. Hơn nữa, thời gian đọc của LYNA nhanh gấp đôi thời gian đọc bởi các bác sỹ thực hành.

AI đang có xu hướng được ứng dụng mạnh mẽ trong y học như: chẩn đoán bệnh, nghiên cứu, phát triển thuốc, tối ưu hóa cho điều trị từng cá nhân, chỉnh sửa gen. Tuy nhiên, dù AI phát triển trong y học đến đâu cũng không thể thay thế hoàn toàn bác sỹ trong quá trình thăm khám và chữa trị, chẳng hạn như AI không thể thực hiện ca phẫu thuật não tự động - nơi mà đôi khi các bác sỹ phẫu thuật phải thay đổi cách tiếp cận của họ ngay khi tổn thương được bộc lộ và nhìn thấy.

1.2.5. Tuân thủ pháp luật:

Pháp luật Việt Nam hiện hành chưa có quy định về trách nhiệm bồi thường thiệt hại liên quan đến AI. Tuy nhiên pháp luật hiện hành có những quy định có thể điều chỉnh được vấn đề bồi thường thiệt hại liên quan đến AI như Luật Chất lượng sản phẩm hàng hóa 2007, sửa đổi bổ sung 2018, hay chế định bồi thường ngoài hợp đồng trong Bộ luật Dân sự 2015. Khi một sản phẩm mang AI vi phạm quy định về chất lượng thì nhà sản xuất có trách nhiệm phải bồi thường.

Điều đó có nghĩa là dù AI có thể đưa ra các dự đoán hoặc quyết định, thì chúng ta cũng cần giữ lại quyết định của con người để phân bổ trách nhiệm.

Do đó, AI cần có khả năng diễn giải, để chúng ta có thể đảm bảo rằng mô hình học sâu mà chúng ta tạo ra làm việc tuân thủ luật đã được đề xuất.

Hiện nay chúng ta có "Artificial Intelligence for Health" - AI4Health là sự hợp tác của ITU và Tổ chức Y tế Thế giới (WHO) nhằm thiết lập một khung đánh giá tiêu chuẩn để đánh giá các phương pháp dựa trên AI đối với các quyết định về sức khỏe, chẩn đoán, phân loại hoặc điều trị.

1.3. Các khía cạnh của khả năng diễn giải

- Dữ liệu: khía cạnh nào, đặc trưng hay thuộc tính nào của dữ liệu phù hợp nhất cho nhiệm vụ..
- Mô hình: chúng ta có thể tách một phần của một danh mục nhất định trong mô hình để xem nó như thế nào.
- Sự dự đoán: Giải thích tại sao một mẫu x nhất định lại được phân loại thành $f(x)$ nhất định.

2. CÁC PHƯƠNG PHÁP DIỄN GIẢI

2.1. Khái niệm

2.1.1. Sự hiểu biết cơ chế - Mechanistic understanding

Hiểu được cơ chế của mạng máy tính dùng để giải quyết một vấn đề hoặc lý giải một hàm chức năng. Tuy nhiên, sự hiểu biết về cơ chế thường áp dụng với các mô hình đơn giản và có thể diễn giải được (ví dụ như Cây quyết định). Hiểu biết về cơ chế có thể diễn giải độc lập với việc nhìn vào dữ liệu vì hiểu biết về cơ chế chỉ phụ thuộc vào kiến trúc của mô hình học máy và diễn ra trước quá trình huấn luyện. Với những mô hình phức tạp với hàng triệu tham số (mạng máy tính với hàng triệu nơ-ron), rất khó để hiểu được cơ chế cụ thể của từng thành phần trong mô hình. Đây cũng là một nhược điểm của hiểu biết cơ chế vì những mô hình đơn giản thường không đủ khả năng để học dữ liệu phức tạp từ thực tế.

2.1.2. Hiểu được về hàm chức năng - Functional understanding

Hiểu được về hàm chức năng nào để hiểu mạng liên kết các biến đầu vào đến đầu ra như thế nào. Các mạng nơ-ron học sâu phức tạp chứa hàng trăm triệu tham số thường rất khó để diễn giải các thành phần cụ thể. Vì vậy, các phương pháp đặc biệt được phát triển để hiểu được về hàm chức năng và diễn giải các quyết định của những mô hình học sâu này. Các phương pháp này diễn ra sau quá trình huấn luyện nhằm mục đích truy ngược và diễn giải lại các mô hình hộp đen đã được sử dụng.

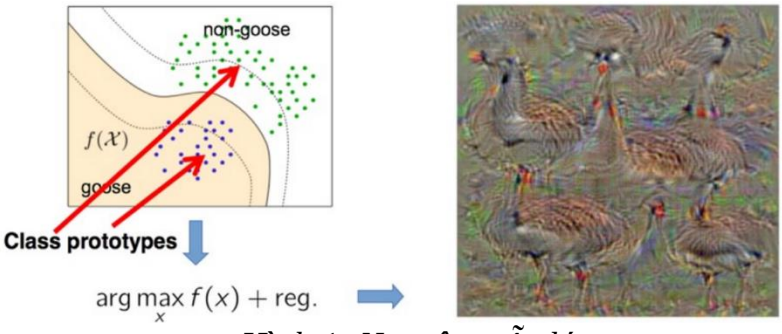
Các phương pháp về hiểu biết về hàm chức năng được phân loại dựa trên độ lân cận (degree of locality). Trong đó, 2 phương pháp chính bao gồm:

- ***Phân tích mô hình:*** Chú trọng về khía cạnh mô hình. Các phương pháp này thường tìm cách diễn giải những biểu diễn nào mà mô hình học sâu đã học được (ví dụ như cách mà mô hình nhận diện được các cạnh, mắt, mũi, tai trong một bức ảnh khuôn mặt) hay diễn giải những khuôn mẫu/ảnh cực đại hóa kích hoạt một nơ ron đặc biệt. Phân tích mô hình có thể biểu diễn toán học dưới dạng xấp xỉ toàn bộ hàm $f(X)$ với X là tập hợp tất các các khả năng của dữ liệu đầu vào (ví dụ như xấp xỉ hàm $f(X)$ để phân loại các cụm dữ liệu phân lớp khác nhau, xấp xỉ ranh giới của bộ phân lớp, …). Ngoài ra, phương pháp này còn diễn giải những đặc trưng đúng với tất cả điểm dữ liệu thuộc cùng một phân lớp.
- ***Phân tích quyết định:*** Chú trọng về khía cạnh dữ liệu đầu vào. Các phương pháp này thường tìm cách giải thích tại sao một dữ liệu đầu vào cụ thể lại được phân loại vào một phân lớp đặc biệt dựa trên dữ liệu (những dữ liệu chủ yếu nào dẫn đến quyết định của mô hình học sâu) và dựa trên các thuộc tính liên quan (ví dụ như điểm ảnh/thuộc tính nào trong ảnh đầu vào đóng góp nhiều vào quyết định cụ thể phân loại tấm ảnh đó). Ngoài ra, phân tích quyết định còn có lợi ích xác nhận lại cách mô hình hoạt động đúng như kỳ vọng cũng như tìm ra các chứng cứ cho quyết định của mô hình học sâu. Vì vậy, phân tích quyết định có vai trò rất quan trọng trong những ứng dụng thực tiễn.

2.2. Phân tích mô hình

2.2.1. Nguyên mẫu lớp - Class prototypes:

Nguyên mẫu lớp là một ví dụ hay một nguyên mẫu điển hình đại diện của một lớp. Ví dụ, trong trường hợp nhận dạng một con ngỗng, một nguyên mẫu lớp cho phân lớp ngỗng có thể là một hình đại diện trung bình của tất cả con ngỗng trong tập dữ liệu, thể hiện hình dáng, màu sắc và các đặc trưng điển hình của một con ngỗng dưới góc nhìn của mạng học sâu. Như vậy, nguyên mẫu lớp cho chúng ta hiểu biết trực quan hơn về cách mô hình đưa ra dự đoán. Cụ thể, phương pháp phân tích mô hình này nhìn vào toàn bộ từng phân lớp (ví dụ như phân lớp ngỗng) và các ranh giới quyết định giữa các phân lớp nhằm phân rã các biểu diễn của nguyên mẫu lớp dưới góc nhìn của mạng nơ-ron học sâu. Nguyên mẫu lớp trả lời câu hỏi "Một chú ngỗng trông thế nào dưới góc nhìn của mạng nơ-ron?"



Hình 1: Nguyên mẫu lớp

Bên cạnh việc xây dựng nguyên mẫu cho mỗi phân lớp, các phương pháp phân rã mô hình còn tìm cách xây dựng những kiểu mẫu tối đa hóa hàm kích hoạt của mỗi nơ-ron riêng biệt ở các lớp khác nhau trong mô hình học sâu. Phân tích mô hình cung cấp một cái nhìn tốt hơn về những biểu diễn bên trong mô hình học sâu.

Một phương pháp khác của diễn giải mô hình rất phát triển trong thời gian gần đây là tối ưu hóa hàm kích hoạt (Activation maximization). Tối ưu hóa hàm kích hoạt là một phương pháp sinh ngược dữ liệu từ mô hình để tìm khuôn mẫu (pattern) tối ưu hóa hàm kích hoạt của một nơ ron cụ thể, ví dụ như khi phân lớp một loài chim cụ thể, một nơ-ron được kích hoạt tối đa dựa trên những loại hình ảnh này thuộc nhãn ngỗng. Ở phương pháp này, ta sẽ tìm:

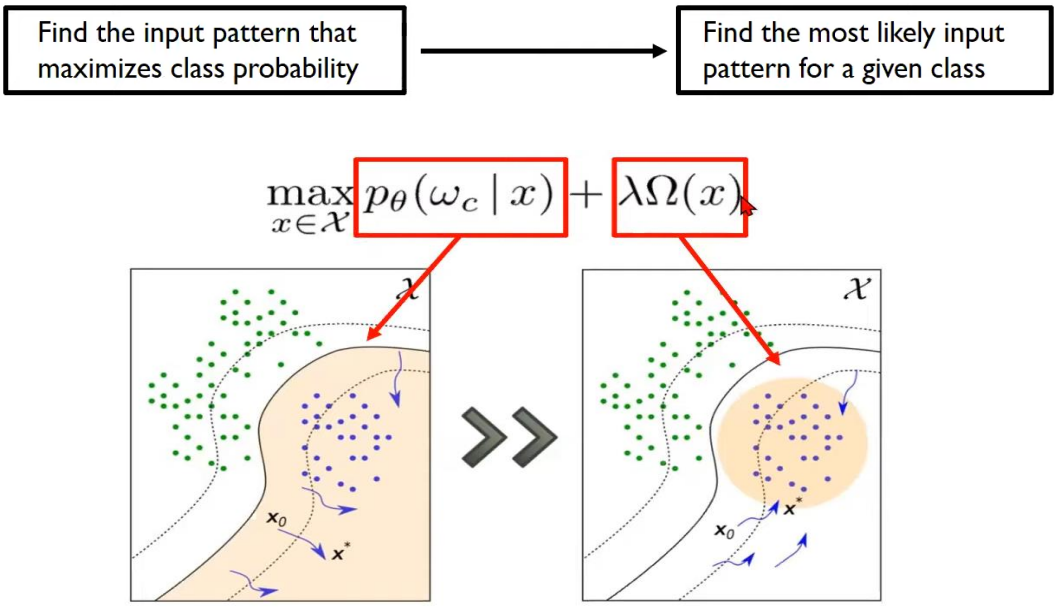
$$\max_{x \in \mathcal{X}} \underbrace{p_\theta(\omega_c \mid x)}_{\text{Class Probability}} + \underbrace{\lambda \Omega(x)}_{\text{Regularization Term}} \quad (1)$$

Giữa tất cả các khả năng của dữ liệu đầu vào, chúng ta sẽ tìm một dữ liệu x sao cho cực đại hóa xác suất sinh ra các trọng số nhất định của một phân lớp cụ thể (cho trước các tham số theta). Xác suất cần được cực đại hóa này còn gọi là xác suất hậu nghiệm cho phân lớp (class posterior probability). Thành phần điều chuẩn (regularization term) ràng buộc việc cực đại hóa sao cho không chọn mọi dữ liệu đầu vào, mà chỉ chọn những dữ liệu thật sự có nghĩa với con người (ví dụ như hình ảnh gần giống con ngỗng cho phân lớp nhãn ngỗng). Nếu không có thành phần điều chuẩn, chúng ta sẽ nhận được những hình ảnh không có nghĩa dưới góc nhìn con người (không thể diễn giải được). Ngược lại, nếu có thành phần điều chỉnh sẽ giúp kết quả đầu ra gần với dữ liệu đầu vào thực sự và có thể diễn giải được (ví dụ như tìm được hình ảnh x có nghĩa và cực đại hóa xác suất hậu nghiệm cho phân lớp con ngỗng). Qua quá trình huấn luyện để cực đại hóa hàm mục tiêu, kết quả ở những vòng lặp đầu rất hỗn độn, không rõ nét ở tất cả các phân lớp và dần hội tụ về những hình ảnh rõ nét hơn. Tuy nhiên, các kết quả hội tụ này vẫn sẽ chưa đạt chất lượng quá tốt cũng như có ích cho việc diễn giải vì chúng vẫn chỉ là những hình ảnh che phủ (overlay) cho tất cả các khả năng của hàm kích hoạt được sử dụng để phân biệt các phân lớp với nhau. Tương tự,

việc biểu diễn giữa các nguyên mẫu lớp (ví dụ như ngỗng so với đà điểu) không thực sự khả dụng từ bài báo Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps (Nguyen et al. 2016). Vì vậy, chúng ta cần phải xác định lại bài toán tối ưu hóa bao gồm cả thành phần điều chuẩn.

Hình 2 : Biểu diễn các nguyên mẫu lớp Ngỗng và Đà điểu

Tổng kết lại, lợi ích của cực đại hóa hàm kích hoạt trong việc sinh dữ liệu là phương pháp xây dựng được những khuôn mẫu tiêu biểu (ví dụ như chân chim, mỏ chim) cho từng phân lớp cụ thể với mục tiêu là những vật thể nền, không liên quan đến phân lớp đó sẽ không hiện diện trong bức ảnh đầu ra. Tuy nhiên nhược điểm của phương pháp này là kết quả đầu ra không thực sự giống hay tương tự những khuôn mẫu liên quan đến từng phân lớp cụ thể cũng như sẽ làm giảm chất lượng của việc diễn giải cho từng phân lớp cụ thể (hình ảnh có thể có nghĩa với máy móc nhưng không có nghĩa dưới góc nhìn con người).



Hình 3 : Minh hoạ cho công thức Activation maximization

Để giải quyết vấn đề này, chúng ta sẽ cố gắng ép dữ liệu sinh ra x' để giống với dữ liệu đầu vào hơn. Chi tiết hơn, thay vì tìm khuôn mẫu dữ liệu cực đại hóa xác suất phân lớp (tìm kiếm trên cả những khuôn mẫu không giống và không xuất hiện trong tập huấn luyện), chúng ta sẽ đi tìm khuôn mẫu dữ liệu giống nhất cho một phân lớp cụ thể (tìm kiếm cực đại trên những khuôn mẫu dữ liệu giống). Điều này có nghĩa là chúng ta không tìm kiếm trên những bức ảnh được xây dựng nhân tạo trên một tập các điểm ảnh có thể mà chúng ta thực sự tìm kiếm một bức ảnh thực sự từ thực tế (hình ảnh một con ngỗng, một chữ viết tay,...). Thành phần điều chuẩn chính là yếu tố chủ chốt để chúng ta thực hiện mục đích này. Hình thể hiện rõ ý tưởng này bằng việc so sánh việc tìm kiếm khuôn mẫu đầu vào trên toàn bộ không gian X các khả năng có thể của dữ liệu x sao cho cực đại hóa ranh giới của bộ phân lớp và việc chỉ tìm kiếm x^* giữa không gian các dữ liệu đầu vào thực tế để đi tìm khuôn mẫu dữ liệu giống nhất cho một phân lớp cụ thể (ràng buộc bằng thành phần điều chuẩn).

Cực đại hóa hàm kích hoạt bao gồm 2 phương pháp phổ biến:

- Cực đại với thành phần điều chuẩn nguyên mẫu (original constraint/expert).
- Cực đại trong không gian mã các khả năng của dữ liệu đầu vào (Code Space).

Activation Maximization with Expert

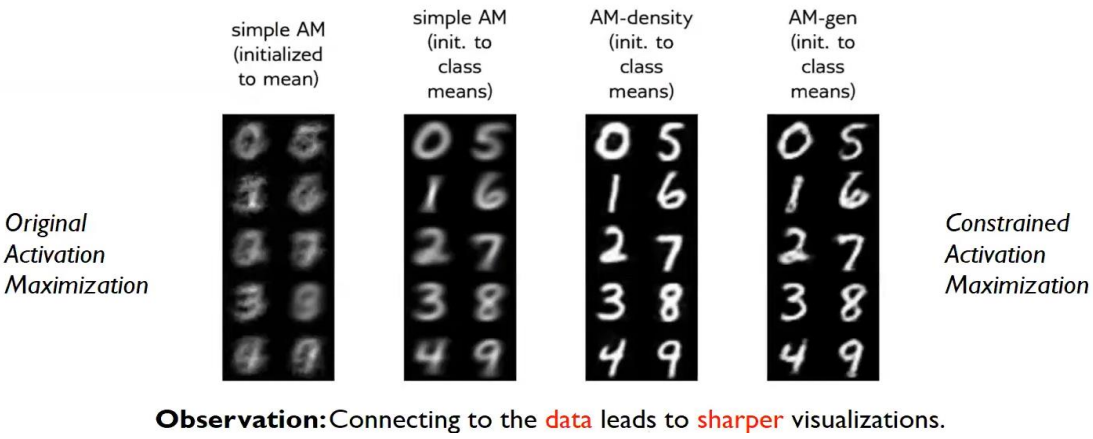
$$p(\boldsymbol{x}|\omega_c) \propto \underbrace{p(\omega_c|\boldsymbol{x})}_{\text{original}} \cdot p(\boldsymbol{x}) \qquad (2)$$

Activation Maximization in Code Space

$$\max_{\boldsymbol{z}\in\mathcal{Z}} p(\omega_c|\underbrace{g(\boldsymbol{z})}_{x}) + \lambda\|\boldsymbol{z}\|^2 \qquad \boldsymbol{x}^\star = g(\boldsymbol{z}^\star)$$

Hai phương pháp này yêu cầu một mô hình không dán nhãn của dữ liệu như:

- Mô hình mật độ $p(x)$ thể hiện những vùng mà dữ liệu tập trung nhất và phân bố mật độ đặc trưng (feature density distribution) để thực hiện lấy mẫu trong không gian các khả năng của dữ liệu đầu vào.
- Một hàm sinh $g(Z)$ không cần thể hiện đầy đủ phân bố dữ liệu nhưng có khả năng lấy mẫu từ phân bố đó.



Hình 4 : Các hình thức Activation maximization

Từ hình (2), (3), (4), chúng ta có thể thấy việc bổ sung thành phần điều chuẩn để ràng buộc bằng mô hình mật độ (cột hình) và mô hình hàm sinh trong không gian mã (cột hình, hình 4) đều cho kết quả rõ nét, trực quan hơn. Từ đó, chúng ta có thể kết luận rằng việc kết nối tối đa hóa hàm kích hoạt với phân bố dữ liệu sẽ sinh ra những hình ảnh thực tế hơn, dễ diễn giải hơn.

Hình 5 : Kết nối tối đa hóa hàm kích hoạt với phân bố dữ liệu sẽ sinh ra những hình ảnh thực tế hơn

Ngoài ứng dụng giải thích phân loại hình ảnh, phân rã mô hình còn có thể giúp tìm hình ảnh trực quan của một mẫu với những tính chất cho trước. Ví dụ, việc tìm nguyên mẫu từ việc phân rã mô hình giúp ta có thể trả lời câu hỏi "Một nguyên tử với tính chất XYZ trông như thế nào?"

Hình 6 : Một nguyên tử với tính chất XYZ trông như thế nào.

Tuy nhiên, chúng ta có những hạn chế khi tìm cách diễn giải toàn cục. Xem một số ảnh chụp xe moto (hình bên dưới), ta có thể thấy các hình ảnh này rất đa dạng từ mẫu mã xe moto, màu sắc đến góc nhìn (nhìn từ vị trí lái, nhìn từ dưới lên, nhìn trực diện trước xe,...). Như vậy, mô hình nào sẽ là nguyên mẫu tốt nhất để diễn giải phân lớp "xe moto"?. Việc tổng hợp tất cả các khía cạnh vào một hình nguyên mẫu duy nhất có thể rất khó khăn như ví dụ về xe moto trên. Một cách diễn giải sẽ càng tốt khi độ đa dạng góc nhìn tăng lên.

Hình 7 : Ảnh chụp xe moto.

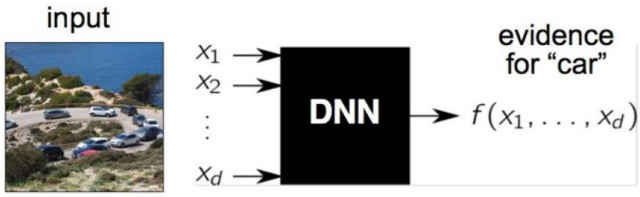
Bên cạnh tìm nguyên mẫu chung cho một lớp, ta cũng cần chú ý đến lời giải thích dưới góc độ cá thể. Ví dụ, đi tìm nguyên mẫu trả lời cho câu hỏi "Một chiếc xe moto điển hình trông như thế nào?", trong khi giải thích dưới góc độ cá thể trả lời câu hỏi "Tại sao ảnh này lại được phân loại là một chiếc xe moto?". Việc giải thích cá nhân này rất thích hợp cho những bài toán tập trung giải quyết vấn đề cho từng trường hợp cá nhân cụ thể, không giống nhau. Ví dụ điển hình là Y học chính xác (Precision Medicine) hay còn gọi là Y học cá nhân hoá (Personalized Medicine), lĩnh vực hướng đến chẩn đoán và điều trị đúng phương pháp cho đúng người bệnh tại đúng thời điểm với đúng liều lượng. Mỗi trường hợp bệnh nhân có thể trạng khác nhau và thay đổi theo từng thời điểm nên cần có lời giải thích riêng để đảm bảo tính chính xác và độ tin cậy. Có thể nói thêm trong lĩnh vực y tế, góc nhìn quần thể cũng rất quan trọng, ví dụ như chúng ta cần biết triệu chứng nào là phổ biến nhất của từng loại bệnh. Cả hai khía cạnh cá thể và quần thể đều quan trọng phụ thuộc vai trò bạn là ai (Cục quản lý Thực phẩm và dược phẩm Hoa Kỳ, bác sĩ, bệnh nhân).

Hiểu biết về hàm chức năng từ cấp độ phân rã mô hình (bao gồm trực quan các filter, cực đại hóa hàm kích hoạt), sang cấp độ phân rã tập trung vào phân bố của dữ liệu đầu vào (Restricted Boltzmann Machine RBM - mạng nơ-ron nhân tạo sinh ngẫu nhiên có thể học phân phối xác suất của một tập input, Deep Grid Net DGN,...) đến phân rã quyết định của mô hình trên dữ liệu thực tế (phân rã độ nhạy và phân tách).

2.2.2. Phương pháp phân rã độ nhạy - Sensitivity Analysis:

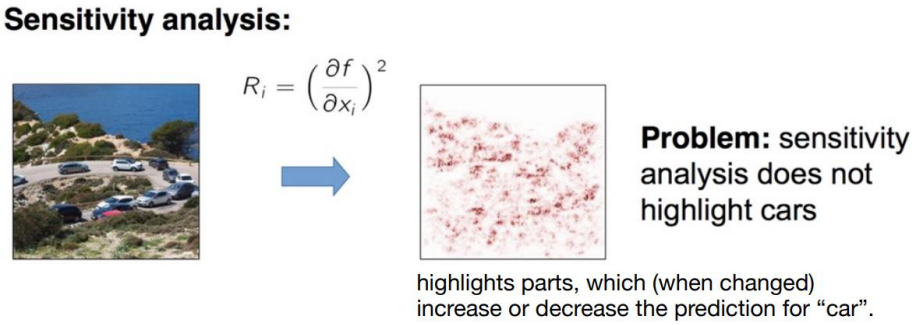
Phân tích độ nhạy bao gồm việc thay đổi dữ liệu đầu vào để đánh giá tác động riêng lẻ của từng biến đối với dữ liệu đầu ra và cung cấp thông tin về các thông tin về các tác động khác của biến được thử nghiệm.

Ví dụ, cho một ảnh, $x_1, x_2, \ldots, x_d$ là các giá trị pixel làm dữ liệu input của mô hình mạng nơ-ron học sâu và output là hàm $f(x_1, x_2, \ldots, x_d)$ chỉ ra có đối tượng xe trong ảnh.



Hình 8 : Input ảnh xe vào mạng DNN.

Độ liên quan (relevance) của đặc trưng đầu vào *i* được tính bởi bình phương của đạo hàm riêng tại điểm đó chỉ ra những vùng highlight (ảnh bên dưới) - những vùng ảnh hưởng đến việc dự đoán chiếc xe. Tuy nhiên, vấn đề là những vùng được highlight bao gồm vị trí của những chiếc xe và những vùng khác ngoài chiếc xe. Như vậy, phân rã độ nhạy chỉ giải thích được mức độ biến động của hàm chức năng chứ không học được chính xác hàm chức năng.



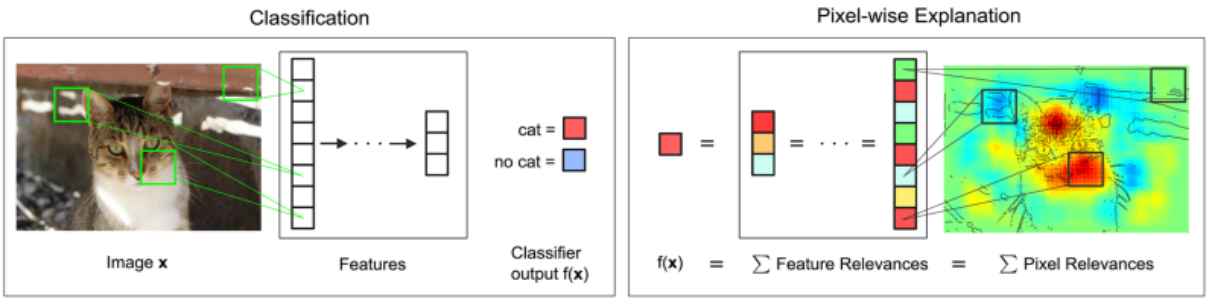
Hình 9 :Highlight vùng ảnh ảnh hưởng đến dự đoán xe.

2.2.3. Khái niệm Phân rã từng pixel - Pixel-wise Decomposition

Ý tưởng chung của phân rã từng pixel là hiểu được đóng góp của một pixel đơn lẻ trong hình ảnh x đến dự đoán $f(x)$ được thực hiện bởi một bộ phân loại f trong một nhiệm vụ phân loại hình ảnh. Chúng ta muốn tìm hiểu rằng, trong mỗi hình ảnh, x những pixel đóng góp trong mức độ nào đến kết quả phân loại một mẫu là đúng (positive), hay sai (negative). Chúng ta sẽ biểu thị mức độ này theo một đơn vị đo lường, giả định rằng bộ phân loại có đầu ra giá trị thực, được ngưỡng tại số không. Giả thiết ánh xạ f: $R^V -> R^V$ sao cho $f(x) > 0$ biểu thị sự hiện diện của cấu trúc đã học. Đầu ra xác suất có thể được xử lý mà không mất tính tổng quát bằng cách trừ đi 0,5. Từ đó ta có thể tìm hiểu đóng góp của mỗi pixel đầu vào $x(d)$ của một hình ảnh đầu vào x đến một dự đoán cụ thể $f(x)$. Ràng buộc quan trọng đặc thù cho việc phân loại đó là tìm những đóng góp khác nhau liên quan đến những trạng thái không chắc chắn nhất đối với việc phân loại, được biểu thị bởi tập các điểm gốc $f(x_0) = 0$. Một cách giải quyết khả thi là phân tách dự đoán $f(x)$ thành một tổng các thành phần của các chiều đầu vào riêng lẻ x_d tương ứng với các pixel:

$$f(x) \approx \sum_{d=1}^{V} R_d \qquad (3)$$

Sự diễn giải được định tính bằng cách: $R_d < 0$ đóng góp bằng chứng ngược lại với sự hiện diện của một cấu trúc cần được phân loại (ví dụ như một con mèo) trong khi : $R_d > 0$ đóng góp bằng chứng cho sự hiện diện của nó. Đối với bước tiếp theo của việc trực quan hóa, các giá trị kết quả quan trọng R_d cho mỗi pixel đầu vào x_d có thể được ánh xạ vào không gian màu và được trực quan hóa theo cách đó như một bản đồ nhiệt (heatmap) tiêu chuẩn. Một ràng buộc cơ bản là các dấu của R_d phải tuân theo phân rã định tính phía trên, tức là các giá trị dương phải thể hiện đóng góp quan trọng, các giá trị âm đại diện cho đóng góp đi ngược lại với quyết định của bộ phân lớp. Hình 10 mô tả ý tưởng chính của phương pháp này.



Hình 10 :mô tả ý tưởng của Phân rã từng pixel

2.2.4. Lan truyền mức độ liên quan theo lớp - Layer-wise Relevance Propagation

Layer-wise Relevance Propagation - LRP là một khái niệm tổng quát để đạt được phân rã từng pixel (Pixel-wise Decomposition) như trong phương trình (4). LRP có thể được hiểu như một ý tưởng chung được định nghĩa bởi một tập các ràng buộc. Bất kỳ giải pháp nào thỏa mãn các ràng buộc này sẽ được xem xét là tuân theo khái niệm LRP.

LRP dạng tổng quát giả định rằng bộ phân loại có thể được phân rã thành nhiều tầng tính toán. Các tầng này có thể là các thành phần của quá trình trích xuất đặc trưng từ hình ảnh (feature extraction) hoặc là các thành phần của thuật toán phân loại chạy trên các đặc trưng tính toán (computed features). Điều này là khả thi cho các đặc trưng Bag of Words với SVM phi tuyến cũng như cho các mạng neuron.

Lớp đầu tiên là đầu vào (các pixel của hình ảnh), lớp cuối cùng là đầu ra dự đoán có giá trị thực của bộ phân loại f. Lớp thứ l được mô hình hoá dưới dạng một vector $z = (z_d^{(l)})_{d=1}^{V(l)}$ với chiều $V(l)$. LRP giả định rằng chúng ta có một điểm thưởng Liên quan $R_d^{(l+1)}$ (Relevance score) cho mỗi chiều z_d^{l+1} của vector z ở tầng $l+1$. Ý tưởng là tìm điểm số Liên quan $R_d^{(l)}$ cho mỗi chiều$z_d^{(l)}$ của vector z tại tầng kế tiếp l (tầng gần với tầng đầu vào hơn), sao cho phương trình sau đây được thỏa mãn:

$$f(x) = \cdots = \sum_{d \in l+1} R_d^{(l+1)} = \sum_{d \in l} R_d^{(l)} = \cdots = \sum_{d} R_d^{(1)} \quad (4)$$

Lặp lại tuần tự phương trình (4) từ tầng cuối cùng là đầu ra của bộ phân loại $f(x)$ đến tầng đầu vào x bao gồm các pixel hình ảnh sẽ dẫn đến phương trình mong muốn (3). Sự tương quan cho tầng đầu vào sẽ phục vụ như việc phân rã thành tổng mong muốn trong phương trình (3). Chúng ta sẽ thêm các ràng buộc khác hơn nữa cho các phương trình (3) và (4) vì một phân rã thành tổng thỏa mãn phương trình (4) đơn thuần không phải là duy nhất, cũng không đảm bảo rằng nó cho ra một giải thích có ý nghĩa về dự đoán của bộ phân loại.

Chúng ta đưa ra đây một phản ví dụ đơn giản. Giả sử chúng ta có một tầng duy nhất. Đầu vào là $x \in R^V$. Chúng ta sử dụng một bộ phân loại tuyến tính với một ánh xạ không gian đặc trưng ϕ_d tùy ý và phụ thuộc vào chiều và một bias b:

$$f(x) = b + \sum_{d} \alpha_d \phi_d(x_d) \quad (5)$$

Độ liên quan cho tầng thứ hai có thể được định nghĩa một cách đơn giản như là $R_1^{(2)} = f(x)$. Tiếp theo, một công thức truyền tính liên quan theo tầng có thể được định nghĩa bằng cách xác định độ liên quan R^l cho đầu vào x như sau:

$$R_d^{(1)} = \begin{cases} f(x)\dfrac{|\alpha_d\phi_d(x_d)|}{\sum_d|\alpha_d\phi_d(x_d)|} & \text{if } \sum_d|\alpha_d\phi_d(x_d)| \neq 0 \\ \dfrac{b}{V} & \text{if } \sum_d|\alpha_d\phi_d(x_d)| = 0 \end{cases} \quad (6)$$

Cài đặt này rõ ràng đáp ứng các phương trình (3) và (4), tuy nhiên, tính liên quan $R_{(x_d)}^{(l)}$ của tất cả các chiều đầu vào có cùng dấu với dự đoán $f(x)$. Về mặt giải thích phân rã từng pixel, tất cả các đầu vào thể hiện sự hiện diện của một cấu trúc nếu $f(x) > 0$ và trỏ về sự vắng mặt của một cấu trúc nếu $f(x) < 0$. Điều này đối với nhiều bài toán phân loại không phải là một giải thích thực tế.

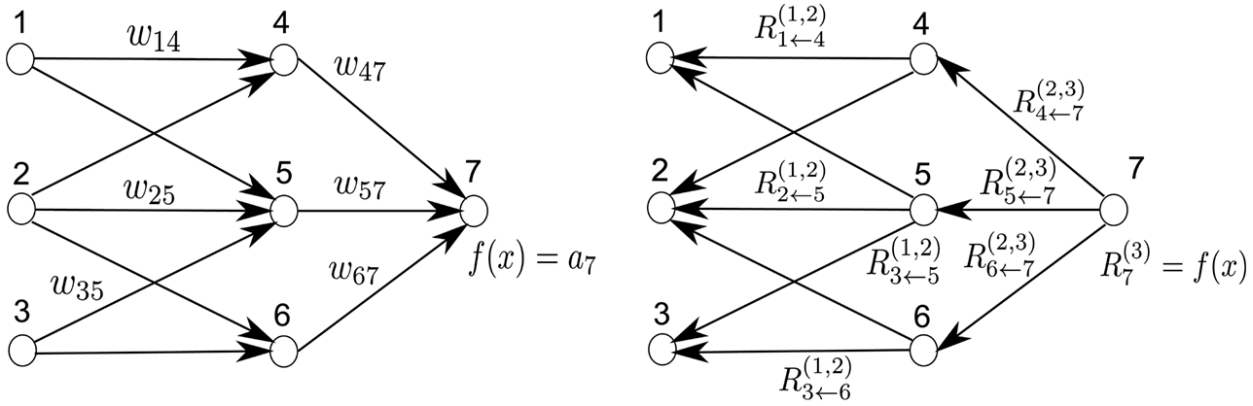
LRP có thể định nghĩa một cách tốt hơn như sau:

$$R_d^{(1)} = \frac{b}{V} + \alpha_d\phi_d(x_d) \quad (7)$$

Độ liên quan của một chiều đặc trưng x_d phụ thuộc vào dấu của thành phần trong phương trình (7). Điều này đối với nhiều bài toán phân loại là một giải thích khả thi hơn. Ví dụ này cho thấy rằng LRP (truyền tính liên quan theo tầng) có thể xử lý các phi tuyến tính như ánh xạ không gian đặc trưng ϕ_d đến một mức độ nào đó và cách mà một mẫu (sample) của LRP đáp ứng Công thức (5) có thể trông như thế nào trong thực tế. Một điều đáng chú ý rằng chúng ta không cần giả định chính quy nào về ánh xạ không gian đặc trưng ϕ_d, chúng có thể thậm chí không liên tục hoặc không đo được theo đo đạc Lebesgue. Công thức (2) có thể được diễn giải như một định luật bảo toàn cho độ đo tính liên quan R giữa các lớp của xử lí các đặc trưng (feature processing)

Ví dụ trên cũng cung cấp cho chúng ta một trực giác về ý nghĩa của độ liên quan R với việc đóng góp cục bộ (local contribution) vào hàm dự đoán $f(x)$. Theo cách hiểu này, độ liên quan của lớp đầu ra chính là dự đoán $f(x)$ đó. Ví dụ đầu tiên cho thấy những gì có thể được mong đợi trong trường hợp phân rã tuyến tính. Trường hợp tuyến tính không phải là điều mới lạ trong học máy, tuy nhiên chúng cung cấp cho chúng ta một trực giác ban đầu.

Ví dụ thứ hai sẽ trực quan hơn về mặt hình ảnh và không tuyến tính. Bên trái của Hình 11 cho thấy một bộ phân loại hình mạng thần kinh với các nơ-ron và trọng số w_{ij} trên các kết nối giữa các nơ-ron. Mỗi nơ-ron i có đầu ra ai từ một hàm kích hoạt.



Hình 11 :Mô tả ý tưởng lan truyền RLP

Lớp trên cùng bao gồm một nơ-ron đầu ra, được đánh chỉ mục bởi số 7. Đối với mỗi nơ-ron i, chúng ta muốn tính toán độ liên quan R_i. Chúng ta khởi tạo độ liên quan của lớp trên cùng là giá trị của hàm, do đó $R^{(3)}_7 = f(x)$. Việc lan truyền độ liên quan theo lớp trong Công thức (4) yêu cầu đảm bảo được phương trình sau:

$$R_7^{(3)} = R_4^{(2)} + R_5^{(2)} + R_6^{(2)}$$

$$R_4^{(2)} + R_5^{(2)} + R_6^{(2)} = R_1^{(1)} + R_2^{(1)} + R_3^{(1)}$$

Chúng ta sẽ đưa ra hai giả định cho ví dụ này. Thứ nhất, chúng ta diễn tả độ liên quan theo từng lớp dưới dạng thông tinh giữa các nơron i và j có thể được gửi qua mỗi kết nối (connection). Tuy nhiên, các thông tin được điều hướng từ một nơron đến các nơ-ron đầu vào của nó (khác với điều xảy ra trong quá trình dự đoán) như được thể hiện trong bên phải của Hình 11. Thứ hai, chúng ta xác định độ liên quan của bất kỳ nơ-ron nào ngoại trừ nơ-ron 7 là tổng của các thông tin đầu vào:

$$R_i^{(l)} = \sum_{k:\, i \text{ is input for neuron } k} R_{i\leftarrow k}^{(l,l+1)} \tag{8}$$

Ví dụ ta có $R^{(1)}_3 = R^{(1,2)}_{3<-5} + R^{(1,2)}_{3<-6}$ Lưu ý rằng nơ-ron 7 không có bất kỳ thông tin đầu vào nào. Thay vào đó, độ liên quan của nó được xác định là $R^{(3)}_7 = f(x)$. Trong Công thức (8) và các phần tiếp theo, các thuật ngữ đầu vào và nguồn có nghĩa là đầu vào cho một nơ-ron khác theo hướng được xác định trong quá trình phân loại, không phải trong quá

trình tính toán LRP. Ví dụ, trong Hình 11, các nơ-ron 1 và 2 là đầu vào và nguồn cho nơ-ron 4, trong khi nơ-ron 6 là điểm cuối cho nơ-ron 2 và 3. Cho hai giả định được mã hóa trong Công thức (8), việc lan truyền độ liên quan theo lớp bằng Công thức (2) có thể được đáp ứng bởi điều kiện đủ sau đây:

$$R_7^{(3)} = R_{4\leftarrow 7}^{(2,3)} + R_{5\leftarrow 7}^{(2,3)} + R_{6\leftarrow 7}^{(2,3)} \tag{9}$$

$$R_4^{(2)} = R_{1\leftarrow 4}^{(1,2)} + R_{2\leftarrow 4}^{(1,2)} \tag{10}$$

$$R_5^{(2)} = R_{1\leftarrow 5}^{(1,2)} + R_{2\leftarrow 5}^{(1,2)} + R_{3\leftarrow 5}^{(1,2)} \tag{11}$$

$$R_6^{(2)} = R_{2\leftarrow 6}^{(1,2)} + R_{3\leftarrow 6}^{(1,2)} \tag{12}$$

Tổng quát, các điều kiện có thể được biểu diễn:

$$R_k^{(l+1)} = \sum_{i:\ i \text{ is input for neuron } k} R_{i\leftarrow k}^{(l,l+1)}$$

Sự khác biệt giữa điều kiện (13) và định nghĩa (8) là trong điều kiện (13), tổng các nguồn (đầu vào) ở lớp l cho một nơ-ron cố định k ở lớp $l+1$, trong khi trong định nghĩa (8), tổng chạy qua các điểm cuối ở lớp l+1 cho một nơ-ron cố định i ở một lớp l. Khi sử dụng Công thức (8) để xác định độ liên quan của một nơ-ron từ các thông tin của nó, thì điều kiện (13) là điều kiện đủ để đảm bảo rằng Công thức (4) đúng. Tổng của vế trái của Công thức (13) là:

$$\begin{aligned} \sum_k R_k^{(l+1)} &= \sum_k \sum_{i:\ i \text{ is input for neuron } k} R_{i\leftarrow k}^{(l,l+1)} \\ &= \sum_i \sum_{k:\ i \text{ is input for neuron } k} R_{i\leftarrow k}^{(l,l+1)} \\ &\overset{eq.(8)}{=} \sum_i R_i^{(l)} \end{aligned} \tag{13}$$

Một cách giải thích cho điều kiện (13) là thông tin $R^{(l,l+1)}{}_{i<-k}$ được sử dụng để phân phối độ liên quan $R^{(l+1)}{}_k$ của một nơ-ron k vào các nơ-ron đầu vào của nó ở lớp l. Chúng ta

sẽ dựa trên khái niệm này và các điều kiện chặt chẽ hơn về bảo toàn độ liên quan được định nghĩa bởi định nghĩa (8) và điều kiện (13). Chúng ta đặt Công thức (8) và (13) là các ràng buộc chính xác nhất định nghĩa quá trình truyền đạt độ liên quan theo lớp. Một giải pháp theo khái niệm này yêu cầu định nghĩa các thông tin $R^{(l,l+1)}{}_{i<-k}$ theo các phương trình này.

Chúng ta có thể rút ra công thức tường minh cho việc truyền đạt độ liên quan theo lớp cho ví dụ trên bằng cách định nghĩa các thông tin $R^{(l,l+1)}{}_{i<-k}$. Quá trình truyền đạt độ liên quan theo lớp phải thể hiện được các thông tin được lan truyền trong quá trình phân loại. Chúng ta biết rằng trong quá trình phân loại, một nơ-ron i đưa ra đầu vào là $a_i w_{ik}$ cho nơ-ron k sao cho i có một kết nối thắng đến k. Do đó, chúng ta có thể viết lại các phía bên trái của các phương trình (9) và (10) sao cho chúng trùng khớp với cấu trúc các phía bên phải của các phương trình đó bằng cách sau đây:

$$R_7^{(3)} = R_7^{(3)} \frac{a_4 w_{47}}{\sum_{i=4,5,6} a_i w_{i7}} + R_7^{(3)} \frac{a_5 w_{57}}{\sum_{i=4,5,6} a_i w_{i7}} + R_7^{(3)} \frac{a_6 w_{67}}{\sum_{i=4,5,6} a_i w_{i7}} \tag{14}$$

$$R_4^{(2)} = R_4^{(2)} \frac{a_1 w_{14}}{\sum_{i=1,2} a_i w_{i4}} + R_4^{(2)} \frac{a_2 w_{24}}{\sum_{i=1,2} a_i w_{i4}} \tag{15}$$

Sự giống nhau giữa vế phải của các phương trình (9) và (10) với vế phải của (14) và (15) có thể được biểu diễn chung như sau:

$$R_{i \leftarrow k}^{(l,l+1)} = R_k^{(l+1)} \frac{a_i w_{ik}}{\sum_h a_h w_{hk}} \tag{16}$$

Phương pháp mô tả ở (16) cho các thành phần thông tin $R^{(l,l+1)}{}_{i<-k}$ vẫn cần được điều chỉnh để có thể sử dụng ngay cả khi mẫu số trở thành 0. Tuy nhiên, ví dụ được đưa ra trong phương trình (16) cho thấy một ý tưởng về việc một thông tin (message) $R^{(l,l+1)}{}_{i<-k}$ có thể là gì, tức là tính độ liên quan của một neuron đầu ra (sink neuron) $R^{(l+1)}{}_k$ đã được tính toán trước đó, được trọng số theo tỷ lệ với đầu vào của neuron *i* từ tầng trước đó *l*. Định nghĩa này có ý nghĩa tương tự khi chúng ta sử dụng các kiến trúc phân loại khác nhau và thay thế khái niệm của một neuron bằng một chiều của một vector đặc trưng tại một tầng được chỉ định.

Công thức (16) có một tính chất khác là: Dấu của độ liên quan được gửi bởi tin nhắn $R^{(l,l+1)}{}_{i<-k}$ sẽ bị đảo ngược nếu đóng góp của một thần kinh $a_i w_{ik}$ có dấu khác với tổng đóng góp từ tất cả các nơ ron đầu vào, nghĩa là nếu nơ ron bắt tín hiệu (kích hoạt) ngược lại xu hướng chung với các nơ ron tầng trên mà nó thừa kế một phần của độ liên quan. Tính chất này tương tự với ánh xạ tuyến tính trong Công thức (5): một nơ ron đầu vào có thể thừa kế

độ liên quan dương hoặc âm (positive, negative relevance) tùy thuộc vào dấu đầu vào của nơ ron đó và điều này khác với Công thức (4)

Một tính chất khác là công thức về phân phối của độ liên quan có thể áp dụng cho các kích hoạt neuron phi tuyến tính hoặc thậm chí không khả vi hoặc không liên tục a_k. Một thuật toán sẽ khởi tạo với các độ liên quan $R^{(l+1)}$ của lớp $l + 1$ đã được tính toán trước. Sau đó, các thông tin $R^{(l,l+1)}{}_{i<-k}$ sẽ được tính toán cho tất cả các phần tử k từ lớp l + 1 và các phần tử i từ lớp liền trước l - sao cho phương trình (13) được thỏa mãn. Sau đó, định nghĩa (8) sẽ được sử dụng để xác định các độ liên quan R (l) cho tất cả các phần tử của lớp l.

Về cơ bản, tính chất bảo toàn độ liên quan có thể được bổ sung bởi các ràng buộc khác để giảm thiểu tập hợp các giải pháp khả thi. Ví dụ, chúng ta có thể giới hạn các thông tin độ liên quan $R^{(l,l+1)}{}_{i<-k}$ là kết quả từ việc phân bố lại độ liên quan trên các nút thuộc tầng thấp một cách đồng bộ với đóng góp của các nút thuộc tầng thấp đối với lớp trên trong quá trình lan truyền thuận. Nếu một nút i có một giá trị kích hoạt có trọng số lớn hơn $z_{ik} = a_i w_{ik}$ thì nút đó cũng nên nhận một phần lớn hơn trong điểm đo độ liên quan (relevance score) $R^{(l+1)}{}_k$ của nút k (một cách định tính). Đặc biệt, đối với tất cả các nút k thỏa mãn R_k, $\sum i\, z_{ik} > 0$, ta có thể định nghĩa ràng buộc $0 < z_{ik} < z_{i'k} => R^{(l,l+1)}{}_{i<-k} <= R^{(l,l+1)}{}_{i'<-k}$. Tuy nhiên, chúng ta không thể đánh giá gì thêmi về các giá trị độ liên quan chính xác này ngoài việc chúng thỏa mãn với ràng buộc đã nêu ở trên.

Tổng kết lại, LRP đã giới thiệu sự lan truyền độ liên quan theo lớp trong mạng chuyển tiếp dạng feed-forward. Trong định nghĩa đã đề xuất, tổng độ đo liên quan được ràng buộc để được bảo toàn từ một lớp sang lớp khác và độ liên quan của mỗi nút phải bằng tổng của tất cả các thông tin độ liên quan vào nút đó và cũng bằng tổng của tất cả các thông tin độ liên quan đi ra khỏi nút đó. Một chú ý quan trọng nữa là định nghĩa LRP không được đưa ra dưới dạng một thuật toán hoặc một giải pháp cho việc cực tiểu hóa một hàm mục tiêu. Thay vào đó, LRP được đưa ra dưới dạng một tập các ràng buộc mà giải pháp nên thỏa mãn. Vì thế, những thuật toán khác nhau với những giải pháp khác nhau có thể thỏa mãn những ràng buộc này chúng ta cũng thảo luận về một phương pháp dựa trên phân rã Taylor để đưa ra một xấp xỉ của việc truyền đạt ý nghĩa theo lớp. chúng ta sẽ cho thấy rằng đối với một loạt các kiến trúc phân loại phi tuyến, truyền đạt ý nghĩa theo lớp có thể được thực hiện mà không cần sử dụng xấp xỉ bằng cách sử dụng phương pháp mở rộng Taylor.

3. ĐÁNH GIÁ & SO SÁNH CÁC PHƯƠNG PHÁP

3.1. Khái niệm về các phương pháp

Trong bài viết này, chúng tôi sẽ so sánh LRP với 02 phương pháp khác đó là Sensitivity và Deconvolution. Đầu tiên, chúng ta cần hiểu về các phương pháp và mục đích của từng phương pháp:

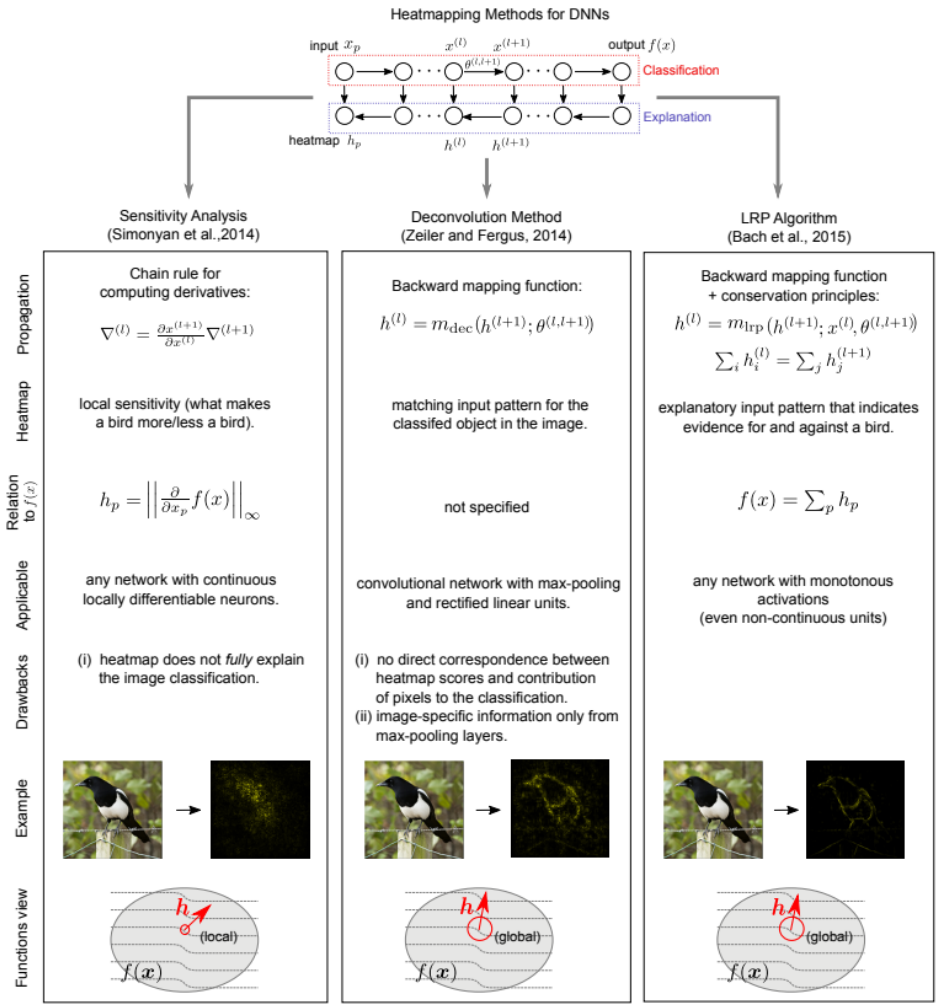
- Phương pháp LRP được sử dụng để xác định mức độ quan trọng của các đặc trưng đầu vào và các neural trong các mô hình học sâu. Nó giải thích quyết định của mô hình bằng cách phân phối giá trị đầu ra của mô hình trở lại các đầu vào ban đầu. LRP có thể cung cấp các giải thích khá trực quan và thường được sử dụng để hiểu cách mà mô hình đưa ra quyết định.
- Sensitivity Analysis là một phương pháp để đánh giá tầm quan trọng của đặc trưng đầu vào đối với đầu ra của mô hình dự đoán. Nó cho phép ta đánh giá tác động của các thay đổi trong giả định, thông số đầu vào hoặc điều kiện của một mô hình hoặc phương pháp tính toán đến kết quả đầu ra của nó. Cơ chế hoạt động của Sensitivity dựa trên tính toán đạo hàm riêng của hàm mục tiêu theo từng biến đầu vào.
- Phương pháp Deconvolution là một phương pháp khá phức tạp và có hiệu quả để tìm hiểu các đặc trưng của mô hình. Nó giải thích các quyết định bằng cách xác định các đặc trưng bên trong các lớp ẩn của mô hình. Deconvolution khôi phục các đặc trưng ban đầu của các đầu vào bằng cách phân tích các neural trong các lớp ẩn của mô hình.

3.2. Đánh giá các phương pháp

Trong phần này, tác giả sử dụng 03 phương pháp để giải thích về các mô hình DNN bao gồm LRP, SA và Deconvolution và trực quan hóa dưới dạng bản đồ nhiệt (heatmap) để hiểu rõ hơn về khả năng của từng phương pháp. Mặc dù được đánh giá dựa vào khả năng quan sát bản đồ nhiệt nhưng đây cũng chỉ là một cách chủ quan bởi con người thực hiện, chúng ta cần một độ đo mang tính khách quan hơn để đánh giá về 03 phương pháp nói trên.

Trong bài báo "Evaluating the visualization of what a DNN has learned" có giới thiệu một phương pháp dựa trên nhiễu loạn vùng (region perturbation) để đánh giá chất lượng của bản đồ nhiệt (heatmap) được tính toán từ DNN, phương pháp này thay đổi các vùng cụ thể của bản đồ nhiệt (heatmap) và quan sát sự thay đổi trong quá trình phân loại hoặc nhận dạng của DNN. Trong bài báo chỉ ra làm thế nào để đánh giá khách quan chất lượng của bản đồ nhiệt (heatmap), giới thiệu một framework để đánh giá bản đồ nhiệt (heatmap) giúp mở rộng ảnh nhị phân sang ảnh màu. So sánh ba phương pháp tính toán bản đồ nhiệt khác nhau trên 3 tập dữ liệu lớn và LRP là phương pháp hiệu quả nhất so với hai phương pháp còn lại.

Giải thích một cách đơn giản cho phương pháp tính toán heatmap, mỗi pixel trên ảnh được gán một giá trị tương ứng, giá trị này được tính bằng một hàm số H(x,f,p) với x là pixel đầu vào, f là hàm số phân loại và p là vị trí của pixel. Heatmap này thường được sử dụng để giải thích quyết định phân loại của một mô hình học sâu.



Hình 12 :So sánh heatmap cho LRP, Deconvolution & SA

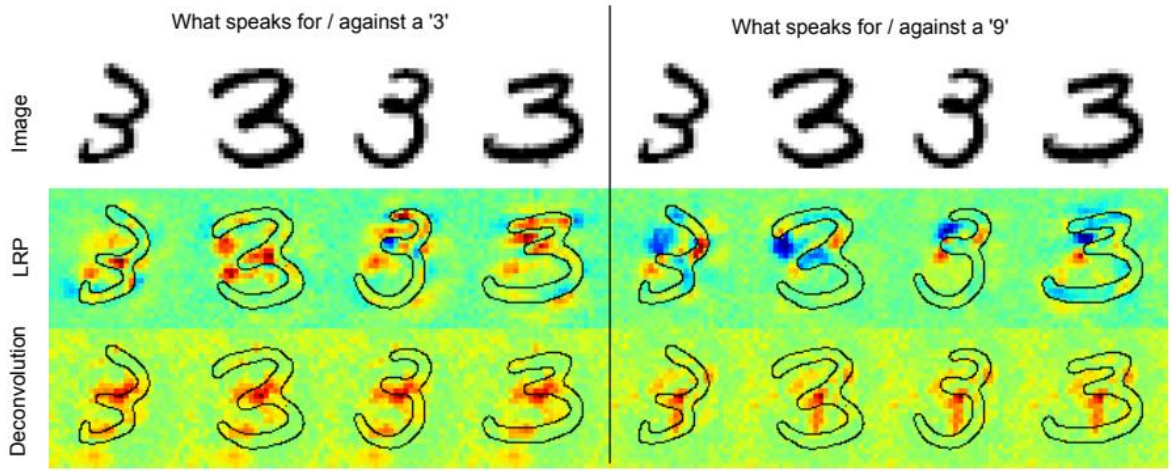
Đoạn văn mô tả ba phương pháp tính toán heatmap khác nhau được sử dụng trong bài báo. Bài báo này muốn tìm hiểu xem một mô hình mạng neural sử dụng để phân loại ảnh có thể được giải thích như thế nào, tức là đối với mỗi ảnh được phân loại, ta có thể biết được đặc điểm nào trong ảnh đã giúp cho mô hình ra quyết định đó.

Ba phương pháp tính toán heatmap được mô tả như sau:

- **Sensitivity heatmaps**: Phương pháp này sử dụng đạo hàm riêng để tính toán heatmap. Heatmap được tính bằng cách đo lường sự thay đổi của đầu ra khi thay đổi giá trị của từng pixel trong ảnh. Các pixel có giá trị relevance score lớn có nghĩa là những pixel quan trọng trong phương pháp tính toán heatmap.
- **Deconvolution method**: Phương pháp này sử dụng 2 mạng CNN khác nhau, một mạng CNN đầu vào để phân loại ảnh, sau đó lấy mạng CNN đầu vào làm đầu vào cho mạng CNN thứ hai. Mạng CNN thứ hai được xây dựng sao cho có khả năng "hoàn tác" các thao tác được thực hiện bởi mạng CNN thứ nhất. Hạn chế của phương pháp này đó là một số negative evidence sẽ bị loại bỏ đi và không chuẩn

hóa điểm số trong quá trình lan truyền ngược, mối quan hệ giữa điểm số heatmap và đầu ra phân loại f(x) sẽ không rõ ràng.

- **Layer-wise Relevance Propagation (LRP)**: Phương pháp này phân tích đầu ra của mạng phân loại f(x) thành các pixel có liên quan bằng cách tuân thủ nguyên lý bảo toàn theo từng lớp, tức là các positive evidence và negative evidence không bị mất đi trong quá trình lan truyền ngược. Thuật toán này không sử dụng đạo hàm và do đó có thể áp dụng cho các kiến trúc mạng khác nhau. LRP giải thích toàn cục quyết định phân loại và điểm số heatmap có một diễn giải rõ ràng là positive evidence và negative evidence.



Hình 13 : Hình minh hoạ bản đồ nhiệt cho LRP và Deconvolution

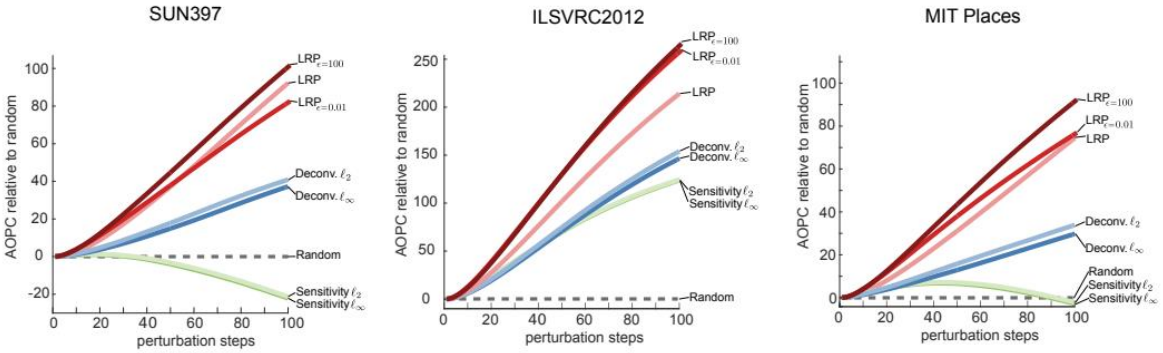
Trong hình minh họa, chúng ta có một mạng neural không có lớp pooling và hai hình ảnh: một được phân loại là số "3" và một được phân loại là số "9". Hình được chia làm hai phần trên dưới, có hai heatmap, một dùng phương pháp LRP và một dùng phương pháp deconvolution. Ở phía trái, chúng ta thấy heatmap LRP và deconvolution cho hình ảnh được phân loại là số "3". Cả hai heatmap đều chỉ ra những đặc trưng quan trọng của hình ảnh đó, nhưng heatmap LRP cung cấp cả positive evidence và negative evidence. Trong khi đó, deconvolution heatmap không cung cấp được negative evidence. Ở phía phải, chúng ta thấy heatmap LRP và deconvolution cho hình ảnh được phân loại là số "9". Heatmap LRP cung cấp cả positive evidence và negative evidence, trong khi deconvolution heatmap chỉ cung cấp positive evidence.

Vì trong phương pháp Deconvolution các negative evidence đã bị loại bỏ trong quá trình backpropagation qua lớp ReLU. Trong khi đó, heatmap LRP có thể cung cấp được cả hai loại evidence và được tùy chỉnh cho từng hình ảnh cụ thể.

3.3. Đánh giá trên tập dữ liệu lớn

Hình 14 : Các tập dữ liệu ảnh có số lượng dữ liệu lớn

Trong phần này, tác giả mô tả phương pháp đánh giá heatmap được đề xuất để so sánh chất lượng của các heatmap được tính bằng 03 phương pháp LRP, Deconvolution và SA. Các heatmap được tính toán cho các mô hình phân loại cho các tập dữ liệu lớn MIT Places và ImageNet, các thí nghiệm được thực hiện với một lượng ảnh lớn và được lặp lại nhiều lần để đảm bảo tính chính xác và độ tin cậy của kết quả.

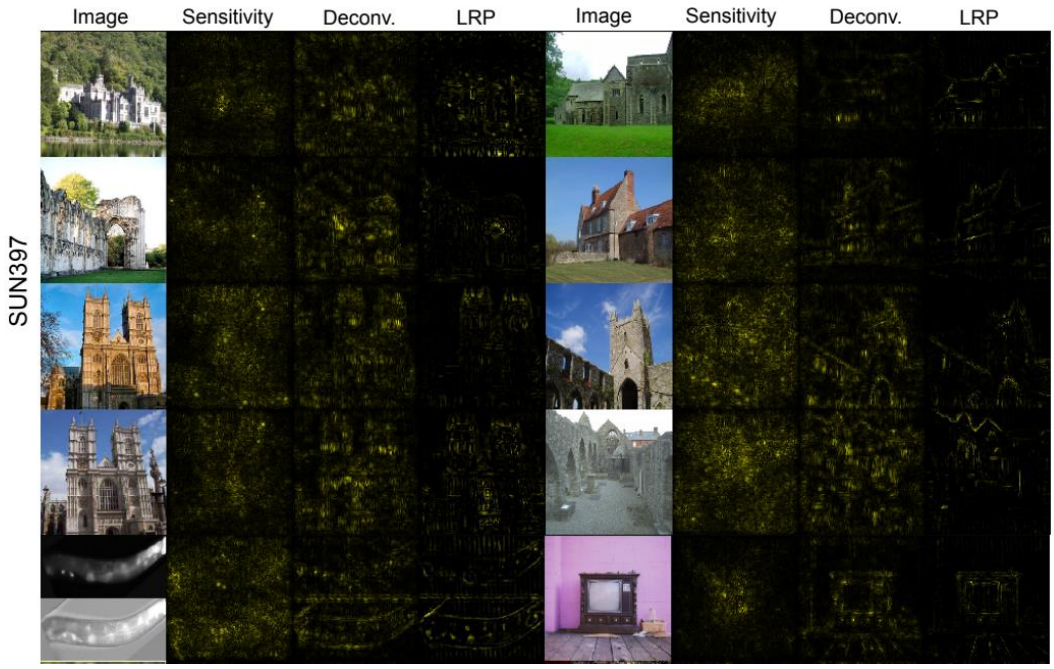


Hình 15 : So sánh chất lượng của bản đồ nhiệt bằng AOPC

Hình ảnh trên đề cập đến việc so sánh chất lượng của các bản đồ nhiệt được tạo ra bằng ba thuật toán đã nêu ở trên. Ngoài ra, cũng so sánh kết quả đó với việc tạo ra các bản đồ nhiệt ngẫu nhiên. Để so sánh, tác giả sử dụng đường cong AOPC để đo lường hiệu quả của các bản đồ nhiệt. Từ hình vẽ, ta có thể thấy rằng các bản đồ nhiệt được tính bằng thuật toán LRP có giá trị AOPC lớn nhất, tức là chúng có khả năng xác định các pixel liên quan đến tác vụ phân loại tốt hơn so với bản đồ nhiệt được tạo ra bằng phương pháp SA hoặc phương pháp Deconvolution. Tuy nhiên, phương pháp Deconvolution vẫn có thể cạnh tranh được với LRP và vượt trội hơn so với bản đồ nhiệt ngẫu nhiên. Điều này được giải thích bởi việc LRP phân biệt giữa các positive evidence và negative eviden, và chuẩn hóa điểm số một cách đúng đắn, do đó nó cung cấp các bản đồ nhiệt ít nhiễu hơn so với phương pháp deconvolution. Trong khi đó, phương pháp SA hướng tới một vấn đề khác và cung cấp các giải thích về quyết định của bộ phân loại dựa trên các thông tin cục bộ trong hình ảnh, nhưng nó có thể bỏ qua các đặc trưng toàn cục của một lớp. Theo như nghiên cứu, cũng chỉ ra rằng bản đồ nhiệt được tính trên tập dữ liệu ILSVRC2012 có chất lượng tốt hơn so với các bản đồ nhiệt khác, vì hình ảnh trong tập này chứa nhiều đối tượng và ít cảnh vật lộn xộn hơn so với các tập dữ liệu SUN397 và MIT Places. Các tập dữ liệu này có chất

lượng kém hơn vì chúng chứa nhiều thông tin liên quan đến nền tảng hơn so với đối tượng cần phân loại.

Theo nghiên cứu, tác giả đã so sánh ba phương pháp tính toán độ quan trọng của các đặc trưng là LRP, deconvolution và sensitivity trên ba tập dữ liệu khác nhau. Kết quả cho thấy phương pháp LRP là phương pháp hiệu quả nhất và cho kết quả giải thích tốt nhất về quyết định của mô hình. Ngoài ra, các heatmap được tính bằng phương pháp LRP cũng có độ phức tạp thấp nhất và kích thước file nhỏ nhất. Tác giả cũng đã đề cập đến việc đánh giá tính phức tạp của các heatmap, bằng cách so sánh độ thưa và độ ngẫu nhiên của các giải thích. Các heatmap tốt là những heatmap chỉ tập trung vào các khu vực quan trọng và không chứa quá nhiều thông tin vô ích hay nhiễu. Kết quả cho thấy heatmap được tính bằng phương pháp LRP có độ phức tạp thấp nhất và dễ nén nhất, còn deconvolution và sensitivity thì có độ phức tạp cao hơn.



Hình 16 : Một số hình ảnh heatmap từ tập dữ liệu lớn khi áp dụng 3 phương pháp

Tóm lại trong nội dung về đánh giá khả năng diễn giải, tác giả đã tập trung vào việc hiểu và làm rõ quá trình ra quyết định của một mô hình mạng neural sâu thông qua khái niệm heatmap. Heatmap là một khái niệm cho phép ta xác định đóng góp của từng pixel trong ảnh đầu vào đến kết quả dự đoán của mạng neural sâu. Tác giả đã đề xuất một phương pháp đánh giá chất lượng heatmap bằng cách sử dụng chiến lược region perturbation. Bằng cách lật ngược các pixel quan trọng nhất đầu tiên, phương pháp này đo lường hiệu suất giảm của mạng neural sâu. Kết quả này cho ta một chỉ số đánh giá chất lượng heatmap. Tác giả cũng đã chứng minh rằng heatmap được tính bằng phương pháp LRP

có độ chính xác và ít nhiễu hơn so với phương pháp tính sensitivity map và heatmap bằng phương pháp deconvolution. Tác giả cũng đã chỉ ra rằng heatmap có thể được sử dụng để đánh giá hiệu suất của mạng neural sâu. Một heatmap tốt không chỉ giúp hiểu rõ hơn về mạng neural sâu mà còn có thể được sử dụng để ưu tiên các vùng ảnh quan trọng hơn trong quá trình phân tích ảnh.

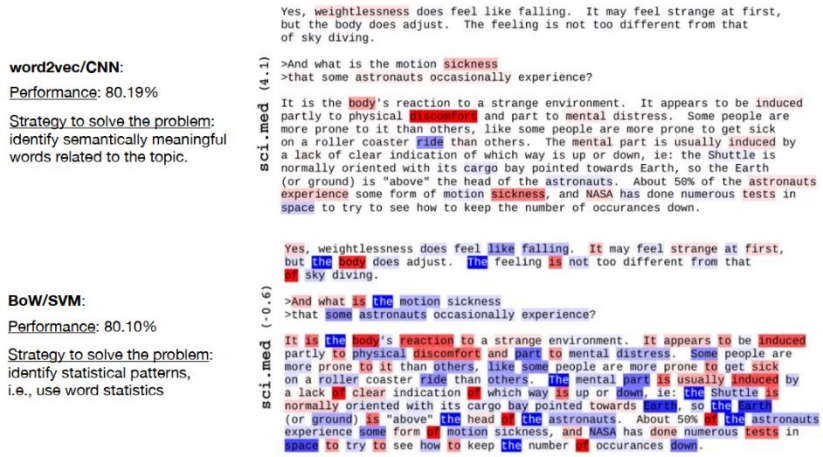
4. ỨNG DỤNG CỦA LRP

4.1. So sánh các mô hình (Compare Models)

Một trong những ứng dụng của LRP trong lĩnh vực xử lý về ngôn ngữ tự nhiên NLP, nội dung của văn bản có thể khác nhau tùy thuộc vào ngữ nghĩa, văn phong và chủ đề mà văn bản đó hướng đến. Các mô hình học máy được huấn luyện để tự động ánh xạ các văn bản đến các khái niệm trừu tượng trên, bên cạnh việc dự đoán chính xác về thể loại văn bản, chúng ta cũng cần phải hiểu về cách thức và lý do quá trình dự đoán thực hiện và đó cũng là một trong những ứng dụng của LRP trong lĩnh vực về xử lý ngôn ngữ tự nhiên.

LRP có thể được sử dụng để xác định các điểm tương đồng và sự khác biệt của chúng. Bằng cách trực quan hóa các đặc trưng quan trọng có ảnh hưởng đến kết quả đầu ra của mô hình. Giúp chúng ta có thể hiểu rõ hơn về cách mà mô hình xử lý thông tin và đưa ra quyết định. Thông tin này có thể được sử dụng để xác định các vùng mà mô hình đồng ý hoặc không đồng ý cho kết quả dự đoán. Qua đó, người dùng có thể so sánh hiệu quả các mô hình khác nhau và chọn ra mô hình có độ chính xác và hiệu quả tốt nhất cho nhu cầu của họ.

Ví dụ như chúng ta có hai mô hình được đào tạo để phân loại ảnh của động vật gồm mô hình mạng neural truyền thẳng và CNN. Bằng cách áp dụng LRP cho cả hai mô hình và trực quan hóa các đặc trưng đóng góp cho kết quả đầu ra, so sánh cách hai mô hình xử lý dữ liệu hình ảnh. Chúng ta có thể thấy mạng neural truyền thẳng tập trung nhiều hơn vào các đặc trưng đơn giản như màu sắc và hình dạng, trong khi CNN sẽ tập trung vào các đặc trưng phức tạp hơn như kết cấu của hình, hoa văn.

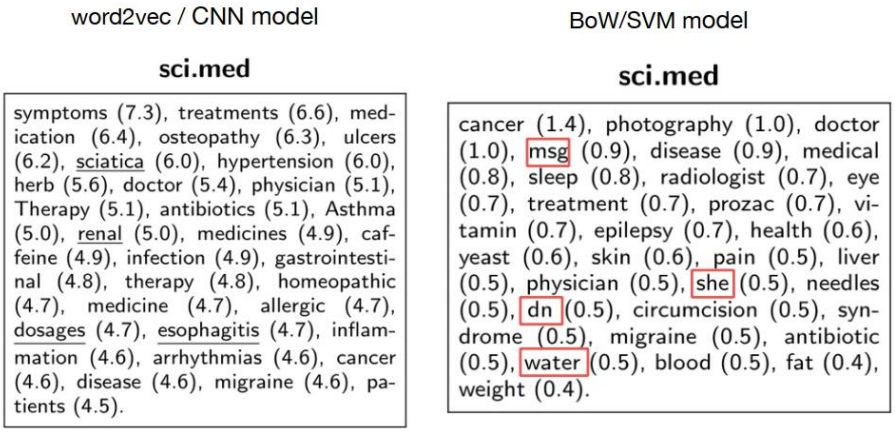


Hình 17 : Phân loại văn bản bằng hai mô hình CNN & BoW/SVM

Ví dụ ở đây tác giả đào tạo hai mô hình trong bài toán phân loại văn bản, bao gồm mạng CNN và bộ phân loại BoW/SVM, sau đó áp dụng LRP để tính toán chỉ số cho biết các đóng góp từ các đặc trưng bao nhiêu vào quyết định phân loại tổng thể.

- Word2vec/CNN là một kiến trúc mạng neural tích chập được huấn luyện trên tập dữ liệu 20Newsgroup và đạt được độ chính xác 80.19%, chiến lược giải quyết vấn đề của mô hình này là xác định các từ ngữ mang ý nghĩa liên quan đến chủ đề (Ở đây là chủ đề về y tế). LRP có thể giúp xác định các từ nào trong câu được coi là quan trọng nhất để mô hình đưa ra dự đoán về chủ đề của văn bản. Kết quả LRP được biểu thị dưới dạng heatmap, trong đó các từ có độ quan trọng cao được tô đậm so với các từ có độ quan trọng thấp.
- BoW/SVM là mô hình sử dụng phương pháp Bag of Words kết hợp với mô hình SVM, cũng được huấn luyện trên tập dữ liệu 20Newsgroup và đạt được độ chính xác 80.10%, chiến lược của mô hình này xác định các mô hình thống kê của các từ trong văn bản. LRP giúp xác định những từ nào trong văn bản được coi là quan trọng đối với kết quả phân loại của mô hình. Kết quả của LRP có thể được biểu thị dưới dạng danh sách các từ theo thứ tự độ quan trọng giảm dần, ở đây các từ màu đỏ sẽ quan trọng và các từ màu xanh thì không quan trọng.

Kết luận, mặc dù hai mô hình hoạt động tương tự nhau về độ chính xác phân loại, nhưng mô hình CNN thể hiện mức độ giải thích cao hơn, khiến nó dễ hiểu hơn đối với chúng ta và có khả năng hữu ích hơn cho các ứng dụng khác.



Hình 18 : Relevance Score của 02 mô hình

4.2. Định lượng ngữ cảnh (Quantify Context Use)

Ngoài lĩnh vực về ngôn ngữ tự nhiên, LRP còn ứng dụng rất mạnh trong lĩnh vực về xử lý hình ảnh. Như trong bài báo "Analyzing Classifer: Fisher Vectors and Deep Neural Networks", trình bày về hai phương pháp phổ biến trong lĩnh vực phân loại hình ảnh, đó là Fisher Vectors (FV) và Deep Neural Networks (DNNs), và nhấn mạnh rằng cả hai đều được xem là "black box" vì cách để đưa ra kết quả dự đoán rất khó hiểu và khó diễn giải. LRP giúp hiểu rõ hơn cách thức lý giải bên trong các mô hình phân loại phi tuyến phức tạp như mô hình Bag of Feature hoặc DNNs.

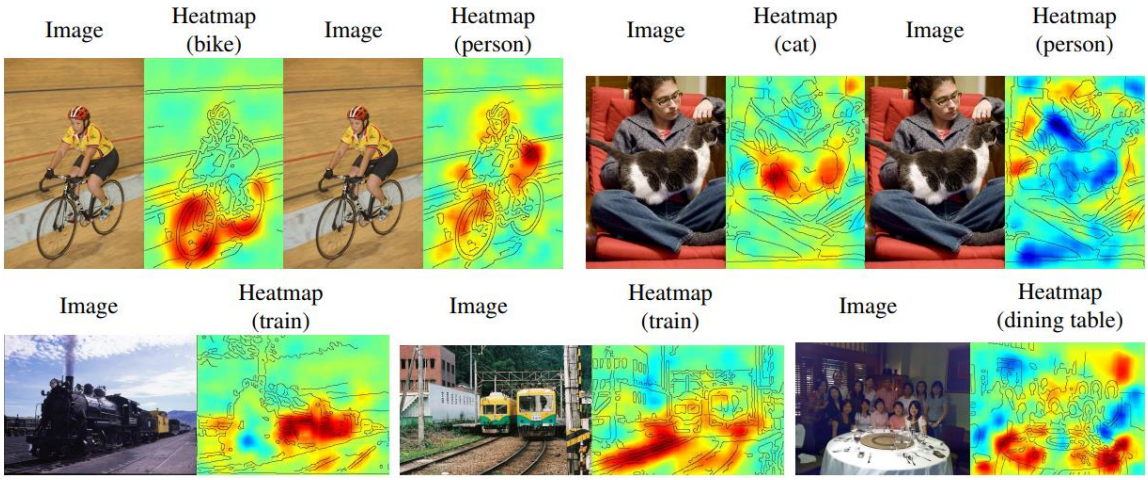
Trong bài viết, áp dụng LRP cho FV Classifier và sử dụng như một công cụ phân tích để đo lường tầm quan trọng của ngữ cảnh đối với phân loại ảnh, đây là một trong những ứng dụng quan trọng của LRP được nêu ra và chúng ta sẽ tìm hiểu chi tiết về nội dung này. Ngoài ra, còn so sánh định tính giữa DNNs và FV Classifiers về các vùng quan trọng và phát hiện các lỗi và độ thiên vị trong dữ liệu, nội dung này sẽ giải thích chi tiết ở phần tiếp theo. Tất cả thí nghiệm đều được thực hiện trên bộ dữ liệu PASCAL VOC 2007 và ILSVRC 2012.

Trong nghiên cứu này, các đóng góp vào các nội dung như định nghĩa các chỉ số để đo lường mức độ phụ thuộc vào ngữ cảnh (context dependence) trong việc dự đoán một hình ảnh. Áp dụng chỉ số phụ thuộc vào ngữ cảnh cho các bộ phân loại dựa trên FV và DNNs trên bộ dữ liệu PASCAL VOC 2007, chứng minh phương pháp này có khả năng phát hiện các trường hợp mạnh về phụ thuộc ngữ cảnh và thiên hướng trong dữ liệu huấn luyện ngay cả khi không sử dụng thông tin bouding box.

Công thức tính tỷ lệ positive relevance có bouding box:

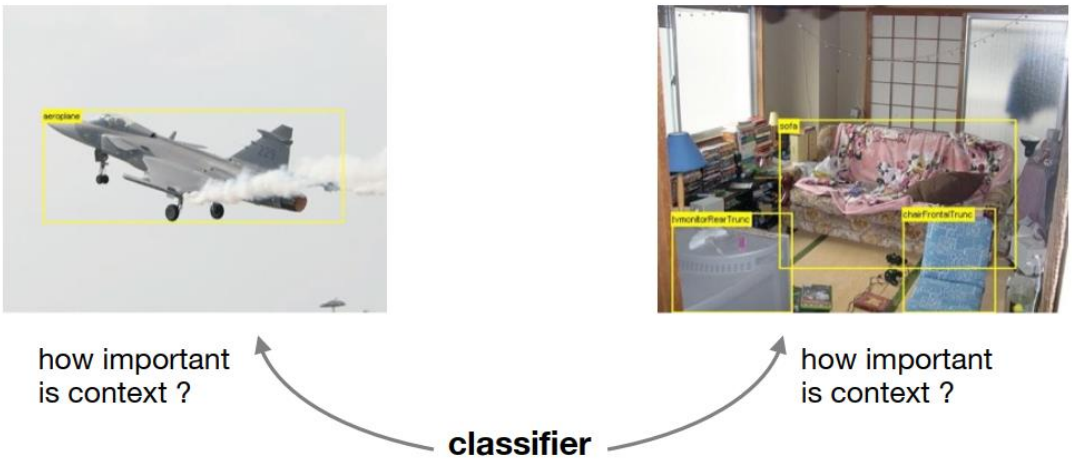
$$\mu = \frac{\frac{1}{|P_{\text{out}}|} \sum_{q \in P_{\text{out}}} R_q^{(1)}}{\frac{1}{|P_{\text{in}}|} \sum_{p \in P_{\text{in}}} R_p^{(1)}} \qquad (17)$$

Sử dụng LRP để đo lường về ngữ cảnh, phân bổ các positive relevance trên heatmap có thể được sử dụng để đánh giá tính quan trọng của bối cảnh với một phân loại ảnh cụ thể. Nếu các thông số bounding box có sẵn (như trong bộ dữ liệu PASCAL VOC), chúng ta có thể tính toán độ đo tỷ lệ tương quan bên trong và ngoài. Tỷ lệ positive relevance cao cho thấy mô hình phân loại sử dụng nhiều bối cảnh để hỗ trợ ra quyết định, còn tỷ lệ thấp cho thấy mô hình phân loại tập trung vào đối tượng để hỗ trợ quyết định. Lưu ý rằng độ đo này không thể chính xác 100% trong các trường hợp, ví dụ máy bay trong quá trình cất cánh cũng sẽ bao phủ lượng lớn nền của ảnh.



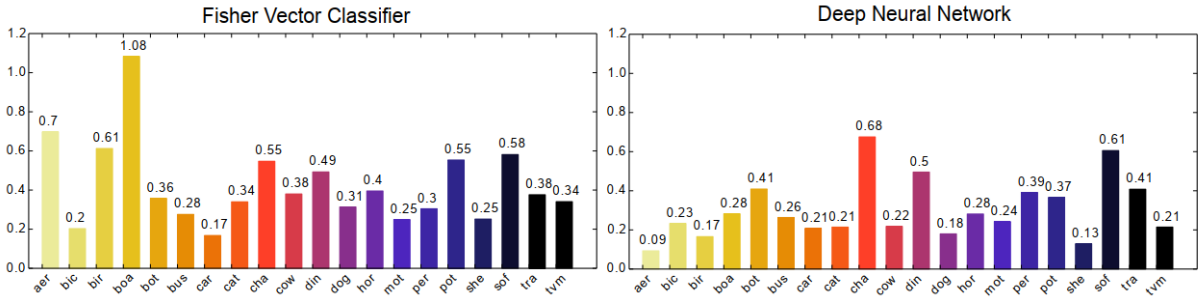
Hình 19: Heatmap trên bộ dữ liệu Pascal VOC

Hình trên mô tả các heatmap tính toán trên các hình ảnh thử nghiệm mẫu của bộ dữ liệu Pascal VOC, xem xét các điểm dự đoán cho một lớp cụ thể. Ví dụ: Như hình đầu tiên, FV nhận thấy rằng "bánh xe" quan trọng cho lớp "xe đạp" , "đường ray" là đối tượng quan trọng cho lớp "tàu hỏa" và "đồ dùng bàn ăn" là quan trọng đối với lớp "bàn ăn". Điều này cho thấy rằng phần quan trọng để đưa ra kết quả không nhất thiết phải nằm trên đối tượng trung tâm, ngược lại có bối cảnh có thể là phần tác động nhiều đến kết quả đưa ra dự đoán.



Quantify Context Use là một trong những ứng dụng của LRP. Trong nhiều trường hợp, ý nghĩa một thông tin cụ thể sẽ phụ thuộc vào ngữ cảnh mà nó xuất hiện. Ví dụ, trong xử lý ngôn ngữ tự nhiên, ý nghĩa của một từ có thể phụ thuộc vào các từ đứng trước hoặc đứng sau nó. Trong xử lý hình ảnh, ý nghĩa của một đặc trưng có thể phụ thuộc vào các đặc trưng khác có trong ảnh. LRP có thể được sử dụng để định lượng mức độ đóng góp của một đặc trưng hoặc một phần thông tin cụ thể vào đầu ra của mô hình, dựa trên ngữ cảnh mà nó xuất hiện.

Giả sử chúng ta có một mô hình được đào tạo để phân loại hình ảnh của các đối tượng và chúng ta muốn hiểu bối cảnh của hình ảnh ảnh hưởng đến dự đoán của mô hình như thế nào. Bằng cách áp dụng LRP cho từng hình ảnh riêng lẻ và phân tích kết quả trực quan hóa, chúng ta có thể xác định được đặc trưng nào của hình ảnh là quan trọng nhất để đưa ra dự đoán chính xác và cách các đặc trưng tương tác với nhau để tạo ra ý nghĩa cụ thể theo ngữ cảnh.

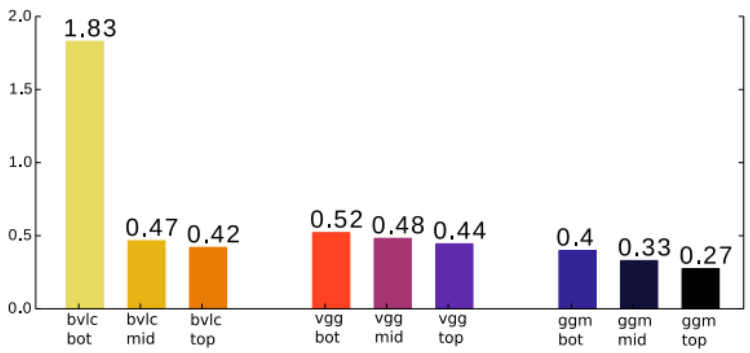


Hình 20: Đánh giá chất lượng Heatmap trên Fisher vector & DNN

Trong phần này, tác giả đã tiến hành đánh giá chất lượng của heatmap được tính bằng Fisher vector cho từng lớp và mô hình, bằng cách đo tỷ lệ giữa relevance outside-inside. Relevance outside-inside được tính bằng công thức số (17) và được sử dụng để đo lường mức độ mà mô hình sử dụng ngữ cảnh (outside) hay đối tượng thực tế (inside) để quyết định lớp của ảnh.

Kết quả được thể hiện trong hình trên, cho thấy mô hình Fisher vector sử dụng nhiều ngữ cảnh hơn so với mô hình Deep Neural Networks (DNN), và mức độ sử dụng ngữ cảnh còn phụ thuộc vào từng lớp. Các lớp như "boat" và "airplane" có xu hướng sử dụng nhiều ngữ cảnh trong mô hình Fisher vector do các yếu tố ngữ cảnh như mặt nước và bầu trời có tương quan mạnh với các đối tượng cần phân loại. Các lớp khác như "bicycle", "car", "motorbike" và "sheep" không cần sử dụng nhiều ngữ cảnh để phân loại.

Đối với mô hình DNN, các lớp như "aeroplane", "bird", "sheep", "dog", "car", "cat" và "tvmonitor" không sử dụng nhiều ngữ cảnh để phân loại và cho kết quả tốt hơn so với mô hình Fisher vector.



Hình 21: Kết quả của 3 mô hình BVLC CaffeNet, VGG CNN & GoogleNet

Đây là kết quả đánh giá tính quan trọng của ngữ cảnh trong việc dự đoán hình ảnh trên tập dữ liệu ImageNet 2012 của ba mô hình được đào tạo trước là BVLC CaffeNet, VGG CNN S và GoogleNet. Dữ liệu được chia thành 333 lớp có độ chính xác dự đoán thấp nhất và cao nhất trên mỗi mô hình. Kết quả cho thấy GoogleNet sử dụng ít ngữ cảnh hơn so với hai mô hình còn lại.

Đối với sự khác biệt giữa kết quả trên ImageNet và PASCAL VOC 2007, nguyên nhân có thể là do các bounding box trên ImageNet bao phủ ít hơn vật thể hơn so với các bounding box trên PASCAL VOC.

4.3. Phát hiện tham số & cải thiện mô hình (Detect Bias & Improve Model)

Trong bài báo "Understanding and Comparing Deep Neural Networks for Age and Gender Classification" đề cập đến việc sử dụng DNN để nhận dạng tuổi và giới tính của khuôn mặt trong hình ảnh. Tuy nhiên, đánh giá cho thấy các mô hình này được áp dụng không hiệu quả, không cung cấp thông tin về các đặc trưng khuôn mặt được sử dụng để dự đoán và làm thế nào các đặc trưng này phụ thuộc vào việc tiền xử lý ảnh, khởi tạo mô hình và lựa chọn kiến trúc. Nghiên cứu này so sánh 4 cấu trúc mạng neural phổ biến, nghiên cứu tác động của việc huấn luyện, đánh giá tính ổn định của phương pháp tiền xử lý thông qua việc thay đổi giữa tập kiểm tra và trực quan hóa chiến lược dự đoán bằng việc áp dụng LRP.

Age classification

	A	C	G	V
[i]	51.4 87.0	52.1 87.9	54.3 89.1	–
[r]	51.9 87.4	52.3 88.9	53.3 89.9	–
[m]	53.6 88.4	54.3 89.7	56.2 90.7	–
[i,n]	–	51.6 87.4	56.2 90.9	53.6 88.2
[r,n]	–	52.1 87.0	57.4 91.9	–
[m,n]	–	52.8 88.3	58.5 92.6	56.5 90.0
[i,w]	–	–	–	59.7 94.2
[r,w]	–	–	–	–
[m,w]	–	–	–	62.8 95.8

Gender classification

	A	C	G	V
[i]	88.1	87.4	87.9	–
[r]	88.3	87.8	88.9	–
[m]	89.0	88.8	89.7	–
[i,n]	–	89.9	**91.0**	**92.0**
[r,n]	–	90.6	**91.6**	–
[m,n]	–	90.6	**91.7**	**92.6**
[i,w]	–	–	–	90.5
[r,w]	–	–	–	–
[m,w]	–	–	–	**92.2**

A = AdienceNet
C = CaffeNet
G = GoogleNet
V = VGG-16

[i] = in-place face alignment
[r] = rotation based alignment
[m] = mixing aligned images for training
[n] = initialization on Imagenet
[w] = initialization on IMDB-WIKI

(Lapuschkin et al., 2017)

Hình 22: Bốn kiến trúc mạng neural phổ biến trong bài toán nhận diện tuổi và giới tính

Trong nghiên cứu này, tác giả tiến hành so sánh hiệu suất của bốn kiến trúc mạng neural phổ biến khi áp dụng cho việc nhận dạng giới tính và độ tuổi trên hình ảnh khuôn mặt con người. Nghiên cứu tập trung vào việc khám phá ảnh hưởng của việc tiền xử lý ảnh, khởi tạo mô hình và sự khác biệt giữa các phương pháp căn chỉnh ảnh trên kết quả dự đoán của mô hình. Để đánh giá hiệu suất của các mô hình huấn luyện, tác giả sử dụng kỹ thuật đánh giá oversampling và sử dụng trung bình dự đoán từ mười cách cắt (bốn góc và một giữa, cộng với phiên bản đối xứng) cho mỗi mẫu. Kết quả cho việc dự đoán độ tuổi và giới tính được liệt kê trong các bảng trên. Các cột của cả hai bảng tương ứng với các mô hình đã mô tả; AdienceNet, CaffeNet, Googlenet và VGG-16.

Tác giả cũng báo cáo kết quả độ chính xác 1-off - tức là độ chính xác đạt được khi dự đoán ít nhất một nhãn tuổi kề cận với nhãn chính xác - cho nhiệm vụ dự đoán độ tuổi. Các bảng cũng liệt kê kết quả thực hiện các bước tiền xử lý khác nhau trên dữ liệu, bao gồm căn chỉnh ảnh và khởi tạo trọng số sử dụng hai tập dữ liệu khác nhau: ImageNet và IMDB-WIKI. Kết quả của các thử nghiệm cho thấy việc khởi tạo tham số phù hợp

dẫn đến việc nhận thức toàn diện của đầu vào, đồng thời bù đắp các biểu diễn dữ liệu nhân tạo. Kết hợp các bước tiền xử lý đơn giản, tác giả đạt được hiệu suất tốt nhất đối với việc nhận dạng giới tính. Cuối cùng, tác giả nhận xét rằng có xu hướng chung trong lựa chọn kiến trúc, cách tổ chức dữ liệu và tiền xử lý và khởi tạo mô hình khi thực hiện các thử nghiệm.

Figure 4. Heatmaps for **GoogleNet** models and **gender** recognition. Input images are shown above heatmaps for a DNN pretrained on Imagenet, which are shown above heatmaps for a DNN initialized randomly. The finetuned model predicts based on an ensemble of facial features, whereas the model starting with random weights has overfit on an isolated set of features characteristic to the target classes.

Figure 5. Heatmaps for **VGG-16** and **age** prediction. Input images are shown above heatmaps for a DNN pretrained on IMDB-WIKI, which are shown above heatmaps for a DNN pretrained on ImageNet.

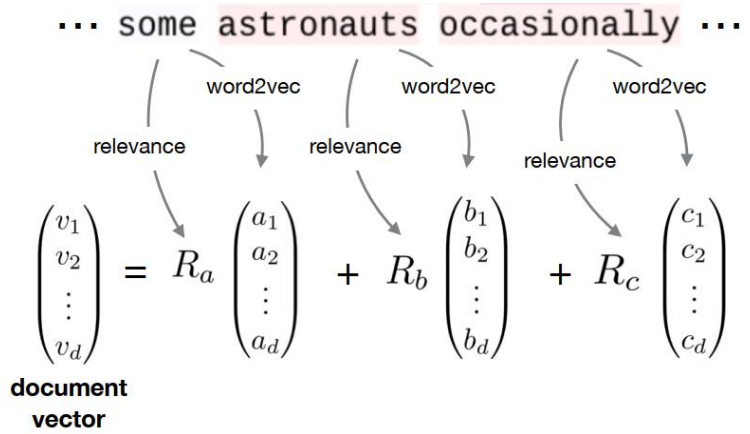
Hình 23 : Heatmap của mô hình GoogleNet và VGG-16 cho bài toán giới tính và tuổi

Tác giả đã thực hiện một nghiên cứu về việc sử dụng các mô hình GoogleNet, VGG16 và CaffeNet để dự đoán giới tính và độ tuổi dựa trên ảnh khuôn mặt. Kết quả cho thấy, trong các cài đặt khác nhau, kiến trúc GoogleNet và VGG16 đều vượt trội hơn so với kiến trúc CaffeNet. Đối với bài toán dự đoán giới tính, các mô hình VGG-16 tốt nhất vượt trội hơn so với các mô hình GoogleNet tốt nhất. Tác giả đã thực hiện các thí nghiệm với việc fine-tuning trên trọng số đã được ImageNet huấn luyện trước và phát hiện ra rằng GoogleNet phản hồi tốt hơn so với CaffeNet khi được khởi tạo với các trọng số tương ứng. Điều này có thể do chất lượng của các tham số khởi đầu: trong khi GoogleNet đạt được lỗi top-5 là 6,6% trên ImageNet, CaffeNet chỉ đạt được 19,6%. Sử dụng các trọng số huấn luyện trước từ ImageNet hoặc IMDB-WIKI dẫn đến sự cải thiện đáng kể về hiệu suất của các mô hình VGG-16 khi so sánh với các mô hình khởi tạo ngẫu nhiên. Tác giả cũng phát hiện ra rằng việc khởi tạo trọng số thích hợp có tác động tích cực đến kết quả dự đoán. Mô hình sử dụng trọng số từ ImageNet hoặc IMDB-WIKI tập trung nhiều hơn vào các đặc điểm khuôn mặt chính, trong khi mô hình khởi tạo ngẫu nhiên chọn ra các đặc điểm riêng lẻ trong quá trình huấn luyện. Đối với bài toán dự đoán độ tuổi, mô hình VGG-16 khởi tạo trọng số từ ImageNet có kết quả tốt hơn so với mô hình sử dụng trọng số từ IMDB-WIKI.

Trong tổng quát, kết quả của tác giả cho thấy rằng khởi tạo trọng số và xử lý dữ liệu đầu vào đóng vai trò quan trọng trong hiệu suất của các mô hình.

4.4. Học cách biểu diễn (Learn new representation)

Trong bài báo "Explaining Predictions of Non-linear Classifier in NLP", tác giả tập trung vào việc giải thích cách mà mô hình phân loại phi tuyến trong NLP đưa ra các dự đoán. Bài báo sử dụng hai mô hình để đánh giá mô hình phân loại là LRP và SA, để giải thích cho các dự đoán của mô hình, tác giả cũng đề xuất các phương pháp tính trọng số khác nhau cho các embeddings từ để tạo ra biểu diễn mới dựa trên các điểm quan trọng của từ liên quan đến lớp mục tiêu được phân loại bởi mô hình. Bài báo đưa ra một số ví dụ về cách sử dụng các mô hình đánh giá giá trị quan trọng của từ và tính trọng số khác nhau để giải thích các dự đoán của mô hình. Ứng dụng của LRP ở đây chính là các biểu diễn mới được tạo ra bằng cách sử dụng các embeddings từ của tài liệu và các giá trị quan trọng của từ được ước tính bởi mô hình LRP và SA để tính toán trọng số cho các embeddings đó. Kết quả là việc biểu diễn mới này giúp cho việc giải thích mô hình dự đoán của mô hình phân loại phi tuyến tính trong xử lý ngôn ngữ tự nhiên trở nên dễ hiểu hơn.



Hình 24: LRP tính toán biểu diễn mới trong bài toán NLP

Một ứng dụng quan trọng của LRP là học cách biểu diễn mới từ các mô hình học sâu. Tương tự như các ứng dụng khác thì sử dụng LRP để xác định các đặc trưng quan trọng đầu vào đóng góp cho kết quả dự đoán đầu ra của mô hình, các đặc trưng này sau đó có thể được sử dụng để làm biểu diễn mới cho mô hình.

Ví dụ, trong bài toán phân loại hình ảnh, LRP có thể được sử dụng để xác định các vùng quan trọng của ảnh góp phần vào quyết định phân loại. Những vùng này sau đó có thể được trích xuất và sử dụng như các biểu diễn mới cho ảnh. Bằng cách này, các biểu diễn mới có thể chứa các thông tin quan trọng và phân biệt nhất từ ảnh đầu vào, và có thể được sử dụng cho các bài toán như truy xuất hình ảnh (image retrieval) hoặc phát hiện đối tượng (object detection).

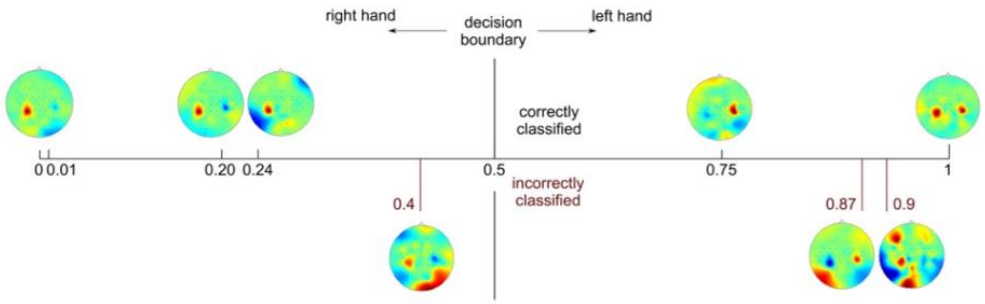
Ngoài ra, LRP có thể được áp dụng cho bài toán xử lý ngôn ngữ tự nhiên (NLP) và nhận dạng giọng nói. Trong NLP, LRP có thể được sử dụng để phát hiện các từ hoặc cụm từ quan trọng nhất góp phần vào classification decision hoặc sentiment analysis. Những từ hoặc cụm từ quan trọng này có thể được sử dụng làm cách trình bày mới cho dữ liệu văn

bản. Nhìn chung, việc học các biểu diễn mới bằng LRP có thể giúp cải thiện hiệu suất và khả năng diễn giải của các mô hình học sâu. Bằng cách xác định các tính năng đầu vào quan trọng nhất và sử dụng chúng làm biểu diễn mới, các mô hình có thể nắm bắt tốt hơn các mẫu và mối quan hệ cơ bản trong dữ liệu, dẫn đến hiệu suất tốt hơn đối với các quá trình huấn luyện sau.

4.5. Giải thích dữ liệu khoa học (Interpreting Scientific Data)

Giải thích dữ liệu khoa học cũng là một ứng dụng quan trọng trong LRP, trong đó LRP được sử dụng để xác định các đặc trưng và mối quan hệ trong dữ liệu khoa học. Dữ liệu khoa học thường chứa các mối quan hệ phức tạp, gây khó khăn trong quá trình tìm hiểu. LRP được sử dụng để tạo ra các biểu diễn của dữ liệu có thể hiểu được, làm nổi bật các đặc trưng và mối quan hệ quan trọng, hỗ trợ việc diễn giải và hiểu dữ liệu.

Một ứng dụng về LRP được áp dụng cho dữ liệu khoa học là trong nghiên cứu khoa học thần kinh. Các nhà thần kinh học sử dụng LRP để xác định các đặc trưng quan trọng và mối quan hệ trong dữ liệu hình ảnh não, chẳng hạn như dữ liệu fMRI (functional magnetic resonable imaging) và EGG (electroencephalography). LRP có thể giúp xác định vùng của não hoạt động tích cực trong các tác vụ cụ thể và mạng não nào tham gia vào các quá trình nhận thức.

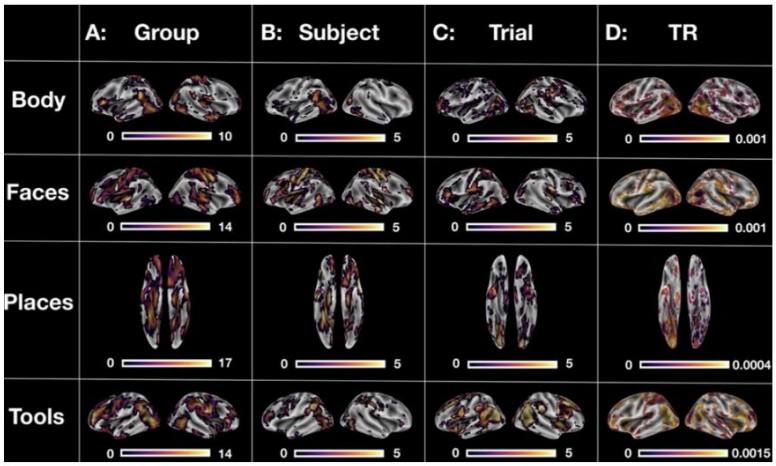


Hình 25: Áp dụng LRP trong bài toán EGG

Một ví dụ khác trong việc nghiên cứu gen, LRP có thể áp dụng để xác định các gen quan trọng và các biến thể di truyền góp phần tạo nên một đặc điểm hoặc bệnh cụ thể. Bằng cách gán các điểm phù hợp cho từng gen hoặc biến thể di truyền, LRP có thể giúp xác định gen nào quan trọng nhất hoặc bệnh cụ thể và cách mà chúng tương tác với nhau.

Ngoài các lĩnh vực trên, LRP có thể được áp dụng cho nhiều lĩnh vực khoa học khác, chẳng hạn như vật lý, hóa học và khoa học môi trường, để hỗ trợ giải thích và hiểu dữ liệu phức tạp.

Nhìn chung, diễn giải dữ liệu khoa học với LRP liên quan đến việc sử dụng LRP để tạo các biểu diễn dữ liệu có thể diễn giải làm nổi bật các đặc trưng và quan hệ quan trọng. Điều này có thể hỗ trợ xác định được các mô hình phức tạp và cơ chế của dữ liệu, giúp hiểu rõ hơn về các hiện tượng khoa học.



Hình 26: Áp dụng bài toán LRP trong bài toán fMRI

Như đã nói ở trên LRP có thể được sử dụng để giải thích dữ liệu fMRI bằng cách xác định các vùng não nào là quan trọng đối với một quá trình hoặc một tác vụ nhận thức cụ thể. Điều này được thực hiện bằng cách áp dụng LRP cho các trọng số của mô hình đã được đào tạo về dữ liệu fMRI. Mô hình học máy được sử dụng để phân loại các trạng thái não hoặc quá trình nhận thức khác nhau dựa trên các mẫu hoạt động của não được đo bằng fMRI.

Khi mô hình đã được đào tạo, LRP có thể áp dụng cho các trọng số của mô hình để xác định vùng não nào phù hợp nhất với quyết định của mô hình. Relevance score của LRP có thể sử dụng để tạo bản đồ của các vùng não quan trọng đối với tác vụ hoặc quá trình nhận thức đang được nghiên cứu. Bản đồ này có thể được sử dụng để xác định các mạng lưới và vùng não liên quan đến quá trình hoặc tác vụ nhận thức và đồng thời có thể giúp xác định các giải thuyết và lý thuyết mới về các cơ chế thần kinh cơ bản.

Ví dụ LRP có thể được nghiên cứu các cơ chế thần kinh làm cơ sở cho nhận thức tri giác. Bằng cách áp dụng LRP cho các trong số của mô hình học máy được đào tạo trên dữ liệu fMRI, các nhà nghiên cứu có thể xác định được vùng não nào quan trọng nhất đối với nhận thức thị giác. Các nhà nghiên cứu phát hiện ra rằng vỏ não thị giác chính (Primary visual cortex), chịu trách nhiệm xử lý các thông tin thị giác, là vùng não quan trọng nhất đối với nhận thức thị giác. Họ cũng phát hiện ra các vùng não khác, chẳng hạn như vỏ não đỉnh (parietal cortext) và vỏ não trước trán (prefrontal cortex) có liên quan đến việc xử lý thông tin thị giác ở cấp độ cao hơn, chẳng hạn như sự chú ý và ra quyết định. Tóm lại, LRP là một công cụ mạnh mẽ để diễn giải dữ liệu bằng cách xác định các đặc trưng quan trọng từ dữ liệu đầu vào và biết được độ quan trọng đối với kết quả dự đoán, điều này giúp chúng ta hiểu rõ hơn về những tri thức mới và cách hoạt động như trong bài toán fMRI.

4.6. Hiểu về mô hình và đạt được góc nhìn mới (Understand Model & Obtain new insight)

Hiểu mô hình máy học và thu được các hiểu biết mới với LRP liên quan đến việc sử dụng LRP để phân tích các hoạt động bên trong của mô hình và hiểu rõ cách mô hình

đưa ra dự đoán. Điều này có thể giúp xác định các mẫu và mối quan hệ dữ liệu chưa biết trước đây và có thể hỗ trợ phát triển các giả thuyết và lý thuyết mới.

LRP hoạt động bằng cách đưa ra các relevance score cho mỗi đặc trưng đầu vào hoặc mỗi neural trong mô hình, đại diện cho sự đóng góp của đặc trưng hoặc neural đó vào kết quả dự đoán đầu ra. Bằng cách phân tích các relevance score, có thể hiểu rõ hơn về hoạt động bên trong mô hình và hiểu cách mô hình đi đến một dự đoán cụ thể.

Một ứng dụng của LRP trong việc hiểu rõ được mô hình là xác định các đặc trưng quan trọng nhất đối với quyết định của mô hình. Điều này có thể giúp xác định các đặc trưng nào đang thúc đẩy dự đoán và có thể hỗ trợ giải thích kết quả đầu ra của mô hình. Ví dụ: Trong phân loại hình ảnh, LRP có thể được sử dụng để xác định vùng nào của hình ảnh quan trọng nhất đối với kết quả phân loại cụ thể.

Một ứng dụng khác là xác định sự tương tác giữa các đặc trưng hoặc neural khác nhau trong mô hình. Bằng cách phân tích relevance score của các đăc trưng hoặc neural khác nhau, có thể xác định được tổ hợp các đặc trưng hoặc neural nào là quan trọng nhất đối với một dự đoán cụ thể. Điều này có thể dẫn đến việc phát hiện ra các mối quan hệ và tương tác trong dữ liệu, có thể hỗ trợ các giả thuyết và lý thuyết mới.

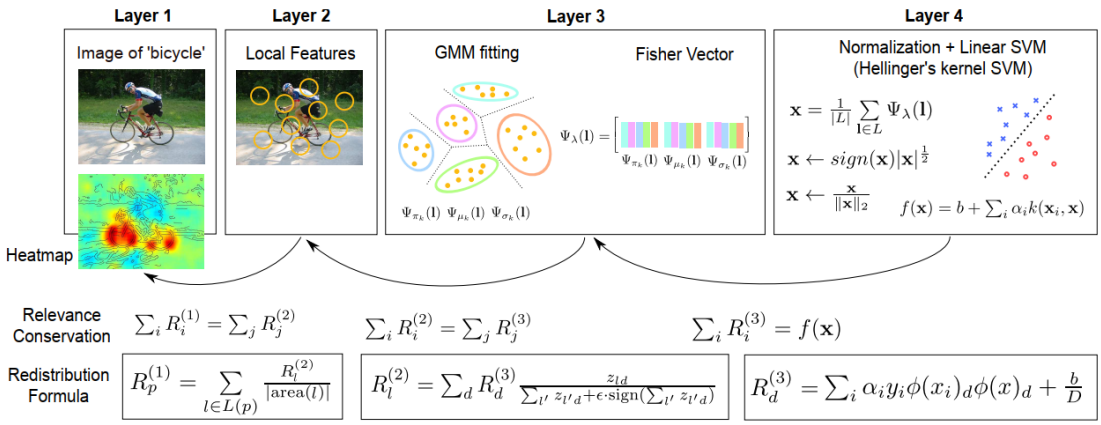


Figure 1. Computing Fisher Vector representation of an image and explaining the classification decision.

Hình 27: Áp dụng LRP cho bộ phân loại Fish Vector

Trong bài báo về "Analyzing Classifiers: Fisher Vectors and Deep Neural Network" mô tả quá trình tính toán Fisher Vector của một hình ảnh và giải thích quyết định phân loại của mô hình. Cụ thể, hình ảnh được chuyển qua bước tiền xử lý để trích xuất đặc trưng và tạo ra vector đặc trưng Fisher. Sau đó, vector đặc trưng được sử dụng để dự đoán lớp của hình ảnh. Để giải thích cho quyết định phân loại này, LRP được áp dụng để tính toán giá trị liên quan cho mỗi đặc trưng và cho biết đặc trưng nào đã được mô hình quan tâm để đưa ra quyết định phân loại. Hình trên, giải thích về việc sử dụng LRP cho bộ phân loại Fisher vector.

5. CASE STUDY: MÔ HÌNH HỌC MÁY CÓ THỂ GIẢI THÍCH ĐƯỢC TRONG MÔ BỆNH HỌC

Các mô hình học máy được ứng dụng rộng rãi trong nhiều lĩnh vực của cuộc sống, ví dụ ứng dụng mô hình máy học trong nhận dạng khuôn mặt người, nhận dạng chữ viết tay, dịch thuật, chuyển đổi âm thanh sang chữ viết, mô tả hình ảnh, sáng tác nhạc, tranh và nhiều ứng dụng khác. Trong y tế, các mô hình máy học cũng đóng vai trò rất quan trọng, mô hình máy học trong mô bệnh học là một ứng dụng cụ thể, nó giúp bác sĩ chuẩn đoán ung thư, tình trạng diễn biến của bệnh…

Bên cạnh những ứng dụng to lớn đó vẫn luôn tồn tại các câu hỏi như – Liệu con người có thể tin vào dự đoán của mô hình học máy? Tại sao mô hình lại đưa ra dự đoán như vậy?...

Các mô hình máy học như là một "Hộp đen" và con người không thể biết được những gì diễn ra bên trong "hộp đen"

Vì vậy các mô hình máy học cần có thể giải thích được là một yêu cầu tất yếu trong thực tế.

Việc giải thích mô hình máy học giúp con người hiểu rõ hơn về dữ liệu, hiểu rõ hơn về nguồn gốc của các dự đoán…

Heatmapping - là một tron những kỹ thuật phổ biến và hiệu quả giúp con người hiểu rõ mô hình học máy. Heatmapping thể hiện trực quan dưới dạng một bản đồ nhiệt, thông qua đó hỗ trợ con người xem xét khu vực ảnh hưởng đến quá trình đưa ra dự đoán của mô hình học máy

5.1. Bản đồ nhiệt cho bằng chứng ung thư

Vấn đề: Mô hình dự đoán đã dựa vào đâu để đưa ra kết luận ung thư, mô hình có nhầm lẫn hay bác sĩ đã bỏ sót chi tiết nào đó?

Bài báo tham khảo chính: "Towards computational fluorescence microscopy: Machine learning-based integrated prediction of morphological and molecular tumor profiles" - Alexander Binder et, al. [1]

Các mô hình máy học được ứng dụng trong công việc dự đoán ung thư. Trước đây con người chỉ tập trung vào độ chính xác của mô hình, hiện nay ngoài độ chính xác, con người cần quan tâm hơn đến việc giải thích các quyết định của mô hình học máy.

Ý tưởng đươc đưa ra cho vấn đề này là lập bản đồ nhiệt thể hiện bằng các điểm ảnh với màu sắc khác nhau nhằm khoanh vùng các pixel liên quan đến dấu hiệu bệnh (ung thư).

Một bộ phân loại giải thích dược gồm có 2 pha cơ bản

Truyền thẳng: Bộ phân loại đưa ra dự đoán từ ảnh đầu vào.

Truyền ngược: Xây dựng bản đồ nhiệt để trực quan hóa, giải thích cho dự đoán của bộ phân loại (trên bản đồ nhiệt các pixel thể hiện màu sắc từ cam đến đỏ đối với các tế bào

ung thư, màu xanh làm cho các tế bào bình thường). Bác sĩ sẽ dựa vào bản đồ nhiệt này để đưa ra kết luận.

Kỹ thuật BoW, bộ phân lớp SVM, và kỹ thuật Layer-wise Relevance Propagation (LRP) được chọn làm công cụ trích xuất đặc trưng, phân lớp và xây dựng bản đồ nhiệt giải thích cho mô hình dự đoán ung thư trong bài báo này.

5.1.1. Bag of Word (BoW)

BoW là một kỹ thuật xây dựng vector đặc trưng phổ biến, hiệu quả phục vụ cho việc dự đoán hồ sơ phân tử, lập bản đồ nhiệt thuộc tính phân tử, lập bản đồ nhiệt trong chuẩn đoán điều trị ung thư, làm đầu vào cho huấn luyện các mô hình máy học…

BoW Hiệu quả tốt trên các tập dữ liệu nhỏ (số lượng mẫu nhỏ hơn 1000).

Ổn định với các thay đổi nhỏ khi tăng cường dữ liệu, lấy mẫu âm tính (Trong mô bệnh học, các mẫu âm tính thường chiếm đa số so với mẫu dương tính, và phần lớn chúng không mang thông tin hữu ích cho quá trình phân lớp, dự đoán. Áp dụng kỹ thuật BoW xây dựng vector đặc trưng sẽ hạn chế ảnh hưởng xấu từ các mẫu âm tính).

Xây dựng vector BoW trải qua các bước:

- Chọn vùng ảnh để tính toán đặc trưng cục bộ.
- Trích xuất các đặc trưng cục bộ trên các vùng đã chọn (Dùng các đặc trưng SIFT, Gradient Norm Quantiles, Color Intensity Quantiles).
- Tạo tập từ điển (*set of visual word*) – bước này chỉ làm duy nhất 1 lần cho việc xử lý dữ liệu phục vụ cho việc training (dùng thuật toán k-means để phân cụm, mỗi cụm tương đương với 1 visual word).
- Ánh xạ các vector đặc trưng cục bộ vào tập từ điển để tạo vector BoW.

Trong bài báo [1], vector đặc trưng BoW được xây dựng trên 3 loại đặc trưng cục bộ là SIFT, Gradient Norm Quantiles (gnq), Color Intensity Quantiles (ciq) .

Local features: SIFT: tính toán đặc trưng SIFT được thực hiện dựa trên việc tìm keypoints (các điểm ảnh đặc trưng không thay đổi với tỉ lệ) sau đó tính toán tập các vector gradients cho các keypoint này. Chi tiết được trình bày trong [2].

Từ 1 vùng ảnh được chọn, trích xuất đặc trưng SIFT [2] (trên các kênh màu khác nhau), mỗi đặc trưng SIFT được biểu diễn bằng 1 vector có 128 chiều.

Local features: Gradien Norm Quantiles: được tính toán trên một vùng ảnh hình tròn có đường kính gấp 6 lần và đồng nhất với đặc trưng SIFT, vòng tròn này được chia làm 2 phần, đường chia là đường trực giao với hướng chính của gradient, mỗi nữa hình tròn được

biểu diễn bằng 1 vector 9 chiều là lượng tử hóa của gradient, tổng cộng có 18 chiều biểu diễn cho một đặc trưng Gradient Norm Quantiles. Ở đây Gradient Norm là tính từ l_2-*norm* của ngõ ra gradient khi áp dụng bộ lọc Sobel (không làm mờ ảnh), sau đó lượng tử vào vector 9 chiều cho mỗi 1/2 hình tròn phạm vi tính đặc trưng.

Local features: Intensity Quantiles: : được tính toán trên một vùng ảnh hình tròn có đường kính gấp 6 lần và đồng nhất với đặc trưng SIFT, vòng tròn này được chia làm 2 phần, đường chia là đường trực giao với hướng chính của gradient, sau đó chia toàn bộ phổ màu sắc của vùng ảnh này thành các phân vị dựa trên cường độ màu sắc của các pixel trong vùng đó. Cụ thể, các giá trị pixel trong vùng được sắp xếp theo thứ tự tăng dần và sau đó chia phổ này thành các phân vị bằng cách chia số lượng pixel cho số lượng phân vị được yêu cầu. Các phân vị được đánh số từ 1 đến số lượng phân vị và cho biết giá trị cường độ màu sắc tương ứng với phân vị đó. Sau đó, đặc trưng Intensity Quantiles được tính toán bằng cách sử dụng các giá trị quantiles như một tập hợp các giá trị đặc trưng, mỗi vector đặc trưng có 9 chiều cho mỗi 1/2 hình tròn.

Để nâng cao hiệu suất, các vector đặc trưng cục bộ (SIFT, Gradient Norm Quantile, Intensity Quantiles) được triển khai trên các kênh màu khác nhau (cụ thể là đỏ và xanh dương) với tỉ lệ (scale) khác nhau và ghép lại [1] tạo thành vector đặc trưng cục bộ có số chiều lên tới (128 + 2.9 + 2.9) x (đỏ, xanh dương) = 328 chiều. Ứng với mỗi giá trị scale (1.5/2.0/2.5) và sự kết hợp các loại đặc trưng cục bộ sẽ cho ra tập túi từ *(set of visual word)* với số chiều khác nhau.

- Đặc trưng SIFT (128 chiều) : Kích thước của túi từ (set of visual word) là 384
- Đặc trưng Gradient Norm Quantiles (18 chiều) : Kích thước của túi từ (set of visual word) là 384
- Đặc trưng Color Intensity Quantiles (18 chiều) : Kích thước của túi từ (set of visual word) là 384

Table 1: Base local feature types when used for a single color channel

type	scale	feature radius	local feature dim. per color channel
gradient norm quantiles (gnq)	1.5/2.0/2.5	9/12/15	18
color intensity quantiles (ciq)	1.5/2.0/2.5	9/12/15	18
SIFT	1.5/2.0/2.5	9/12/15	128
SIFT+gnq	1.5/2.0/2.5	9/12/15	146
SIFT+ciq	1.5/2.0/2.5	9/12/15	146
SIFT+ciq+gnq	1.5/2.0/2.5	9/12/15	164

Từ tập vector đặc trưng cục bộ được rút ra từ tập ảnh huấn luyện, tác giả dùng bộ phân cụm *k-means* để tạo ra tập từ điển (a set of visual word) dùng để xây dựng vector đặc trưng BoW. Một vector BoW là một biểu đồ histrogram biểu diễn sự phân bổ của các vector đặc trưng cục bộ và được chuẩn hóa bằng l_1-*norm*, mỗi ngăn của biểu đồ histrogram này là thể hiện của 1 "*visual word*", số chiều của 1 vector BoW là số lượng "*visual word*".

$$m_d(l) = 1[d == \text{argmin}_{d'} \|l - w_{d'}\|]$$

Figure 1: Hàm ánh xạ đặc trưng cục bộ vào BoW

l: Là vector đặc trưng cục bộ

d: Là phần tử thứ d thuộc vector BoW

m_d *: Là giá trị đánh giá đặc trưng cục bộ l thuộc thành phần d của vector BoW,* $m_d \in \{0,1\}$.

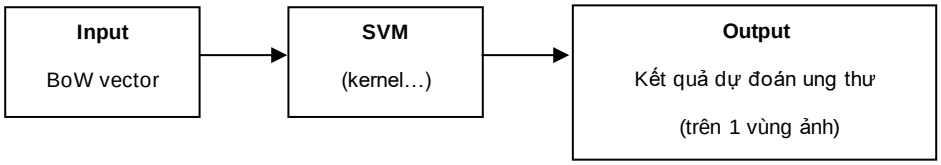
Với cách làm như vậy dù 1 ảnh có xuất hiện bao nhiêu đặc trưng chúng ta cũng có thể ánh xạ vào vector đặc trưng BoW với số chiều nhất định, đại diện chung cho 1 khu vực ảnh được tính toán.

5.1.2. Mô hình học máy dự đoán ung thư (bộ phân lớp SVM).

Xây dựng bộ phân lớp SVM

Tập huấn luyện được xây dựng từ tập hình ảnh H&E stain (Nhộm H&E) đã dược gán nhãn, và trích xuất đặc trưng để làm dữ liệu đầu vào, nhãn được gán cho mỗi vùng ảnh (là 1 ô vuông kích thước tương ứng với khu vực có hoặc không có tế bào ung thư) mang giá trị +1 nếu nó chứa ít nhất một tế bào ung thư và -1 nếu không, vector đặc trưng BoW được tính trên từng vùng ảnh, mỗi ảnh huấn luyện được lấy mẫu trên các ô vuông có kích thước 102x102 (người ta trượt 1 cửa sổ kích thước 102x102 với bước nhảy 34 cho mỗi lần dịch cửa sổ trên ảnh chính). Trong bài báo, tác giả xây dựng vector BoW có 510 chiều (kích thước này là chọn trên cơ sở kinh nghiệm).

Bộ phân lớp SVM được huấn luyện để phân lớp các bản vá hình ảnh (image patch), cụ thể sẽ dự đoán từng vùng ảnh này có phải là biểu hiện của ung thư hay không (điểm số đánh giá là 1 số thực, điểm này kết hợp với ngưỡng sẽ tạo ra bộ phân lớp nhị phân).



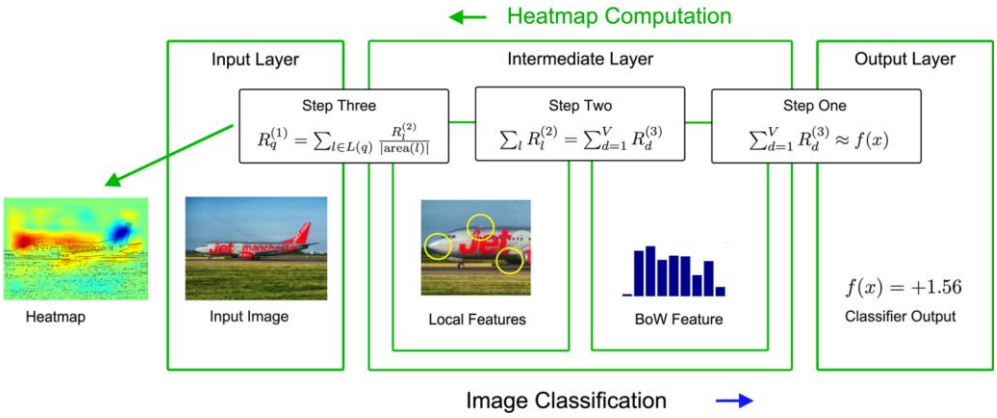
$\chi 2$ – kernel được chọn làm kernel cho bộ phân lớp SVM trong bài báo [1].

$$K(x,z) = \exp(-\sigma \sum_{d|x_d+z_d>0} \frac{(x_d - z_d)^2}{x_d + z_d})$$

Sử dụng *SHOGUN Machine Learning Toolbox* [3] để tạo ra bộ phân lớp SVM với dữ liệu đầu vào là tập vector đặc trưng BoW.

5.1.3. Xây dựng bản đồ nhiệt (Heatmapping) cho mô hình dự đoán ung thư [4].

Bản đồ nhiệt được tạo bằng kỹ thuật Lan truyền ngược qua 03 bước:



Lan truyền ngược (Backpropagate) từ ngõ ra của bộ phân lớp *f(x)* đến ngõ vào của kernel $x_{(d)}$

$x_{(d)}$ là thành phần thứ *d* của vector BoW, kernel sử dụng cho phần này là χ2-kernel.

Kết quả đầu ra *f(x)* của bộ phân loại SVM có thể xem là tổng có trọng số của khoảng cách giữa điểm dữ liệu *x* đến từng điểm hỗ trợ trong SVM.

$$f(x) = b + \sum_i a_i y_i k(z_i, x)$$

$$\text{goal: } f(x) \approx \sum_d R_d^{(3)}(x)$$

a_i: là tham số của mô hình SVM ($(a_i * y_i)$ là trọng số theo cách giải thích trên).
b : là độ lệch của mô hình SVM.
y_i: là nhãn của điểm dữ liệu z_i trong tập các điểm hỗ trợ.
x: là vector BoW thể hiện của 1 input.
z_i: là điểm dữ liệu thứ i của tập các điểm hỗ trợ.
k: là hàm kernel (có thể là HIK-kernel, χ2-kernel).
R_d: là mức đóng góp của thành phần thứ d của x vào kết quả dự đoán.
d: là số chiều của 1 input.

Trường hợp kernel là χ2-kernel

$$k(z, x) = \exp(-\gamma \sum_{d: z_{(d)} + x_{(d)} > 0} \frac{(z_{(d)} - x_{(d)})^2}{z_{(d)} + x_{(d)}}) \quad [1]$$

Áp dụng khai triển Taylor tại x_0 sao cho $f(x_0) = 0$

[1] *Phương trình biểu diễn χ2-kernel*

$$f(x) \approx 0 + \sum_d (x_{(d)} - x_{0,(d)}) \sum_i a_i y_i \frac{\partial k(z_i, x_0)}{\partial x_{0,(d)}} ?$$

$$f(x) \approx 0 + \sum_d (x_{(d)} - x_{0,(d)}) \sum_i a_i y_i \frac{\partial k(z_i, x_0)}{\partial x_{(d)}}$$

Khi đó giá trị của $R_d^{(3)}$ được tính theo công thức:

$$R_d^{(3)}(x) = (x_{(d)} - x_{0,(d)}) \sum_i a_i y_i \frac{\partial k(z_i, x_0)}{\partial x_{(d)}}$$

Thực hiện lan truyền ngược từ ngõ vào $x_{(d)}$ của kernel đến đặc trưng cục bộ.

Giá trị $R_d^{(2)}$ được tính theo công thức.

$$R^{(2)}(l) = \sum_d R_d^{(3)} \frac{m_d(l)}{\sum_{l'} m_d(l')}$$

Ý nghĩa: Phân phối toàn bộ giá trị $R_d^{(3)}$ đến từng đặc trưng cục bộ khi giá trị m_d tương ứng là 1.

Trường hợp tổng quát có xét tới $d | \sum_l m_d(l) = 0$, $R_d^{(2)}$ tính theo công thức sau.

$$R^{(2)}(l) = \sum_{d \notin Z(x)} R_d^{(3)} \frac{m_d(l)}{\sum_{l'} m_d(l')} + \sum_{d \in Z(x)} R_d^{(3)} \frac{1}{\sum_{l'} 1}$$

Trong đó: Z(x) = { $d | \sum_l m_d(l) = 0$}.

Thực hiện lan truyền ngược từ đặc trưng cục bộ lên các pixel, tạo heatmapping.

Ý tưởng: Chỉ phân bổ giá trị $R_l^{(2)}$ của các vector đặc trưng cục bộ đến các pixel nằm trong chính vector đặc trưng cục bộ ấy (các pixel đó gọi là pixel hỗ trợ, ký hiệu q).

$$LF(q) = \{ l \mid q \in supp(l) \}$$

$$R^{(1)}(q) = \sum_{l \in LF(q)} \frac{R^{(2)}(l)}{|supp(l)|}$$

LF(q): Là tập các vector đặc trưng cục bộ có chứa pixel q
|supp(l)| : Là tổng số lượng pixel có trong đặc trưng cục bộ l
$R^{(1)}(q)$: Là giá trị đóng góp của pixel q cho f(x)

5.2. Bản đồ nhiệt cho bằng chứng dấu hiệu phân tử

Bằng chứng dấu hiệu phân tử là chứng về việc một gen hay một protein được biểu hiện (expressed) trong một mẫu tế bào hoặc mẫu mô cụ thể, các thông tin này rất quan trọng trong việc nghiên cứu các quá trình sinh học cơ bản, bao gồm sự phát triển và bệnh lý. Nó cũng đóng vai trò quan trọng trong việc phát hiện và chẩn đoán các bệnh lý liên quan đến gene và protein.

5.2.1. Mô hình dự đoán hồ sơ phân tử (Molecular profile prediction).

Cấu trúc tập dữ liệu huấn luyện:

Copy number variations (CNV)[2]	554 mẫu
Gene expression (RNASEQ)[3]	563 mẫu
DNA methylation data (METH)[4]	400 mẫu
Protein profiles from reverse-phase protein arrays (PROT)[5]	565 mẫu
Somatic mutations (SOM)[6]	10 mẫu

Dữ liệu của các phép đo protein/gene được lượng tử hóa và so sánh với ngưỡng để gán nhãn. Nếu giá trị đo > ngưỡng thì nhãn là +1, ngược lại nhãn là -1. Riêng đối với đột biến Soma (SOM) nhãn là +1 cho trường hợp có đột biến, -1 cho trường hợp bình thường.

Xây dựng vector BoW:

Tương tự như việc huấn luyện cho bộ phân loại ung thư. Mô hình dự đoán hồ sơ phân tử cũng sử dụng phương pháp vector đặc trưng BoW cho công đoạn trích xuất đặc trưng và bộ phân lớp SVM để dự đoán hồ sơ phân tử.

---

[2] *CNV là sự thay đổi số lượng bản sao của một đoạn DNA so với số lượng bản sao mặc định trong các tế bào của một cá nhân. CNV có thể dẫn đến sự thay đổi trong sản phẩm gen (protein) được tạo ra từ gen đó, gây ra sự thay đổi trong cấu trúc và chức năng của protein, và có thể góp phần vào phát triển bệnh lý.*

[3] *Gene expression (RNASEQ) là quá trình chuyển đổi thông tin di truyền từ DNA sang RNA và protein trong tế bào. Gene expression quyết định các tính chất và chức năng của tế bào và là quá trình cơ bản để các tế bào có thể phát triển và hoạt động.*

[4] *DNA methylation data là dữ liệu được tạo ra từ các phép đo định lượng mức độ methylation trên các vị trí CpG trên toàn bộ hoặc một phần của bộ gen trong một mẫu tế bào hoặc mẫu tế bào khác nhau. METH data cung cấp thông tin về mức độ methylation tại các vị trí CpG trên toàn bộ hoặc một phần của bộ gen. Dữ liệu METH có thể được sử dụng để phân tích sự thay đổi methylation liên quan đến bệnh lý, phát hiện các biomarker và tìm kiếm các gene liên quan đến các quá trình sinh lý và bệnh lý*

[5] *PROT là một công cụ phân tích protein được sử dụng để xác định mức độ biểu hiện của các protein trong một mẫu, PROT được sử dụng để giúp xác định các biến số sinh học quan trọng trong các quá trình bệnh lý, bao gồm ung thư, theo dõi tác động của các thuốc hoặc các liệu pháp điều trị khác lên các mạng protein.*

[6] *Đột biến soma (SOM) là một loại đột biến trong tế bào thần kinh gây ra bởi sự thay đổi trong một hoặc nhiều gen của tế bào soma, Sự đột biến trong gen của tế bào soma có thể dẫn đến sự thay đổi trong chức năng của tế bào thần kinh, gây ra sự tăng trưởng không kiểm soát, phân chia tế bào bất thường, và phát triển khối u.*

Trích xuất đặc trưng cục bộ (local feature): Một hình ảnh được chia thành các ô (image tiles) có kích thước 201 × 201. Ô được tạo ra bằng cách trượt 1 cửa sổ dọc theo toàn bộ ảnh (với bước trượt là 201/3 = 67). Đặc trưng BoW trên mỗi ô này. Một điểm khác biệt so với đặc trưng BoW ở bộ phân loại ung thư , ở phần này tác giả bài báo chỉ sử dụng SIFT để trích xuất đặc trưng cục bộ, vì các đặc trưng cục bộ khác không hoạt động tốt với nhiệm vụ dự đoán cho đột biến Soma (SOM). Đặc trưng SIFT được trích xuất trên 2 kênh màu đỏ và xanh lam, tổng số chiều là 256.

Tạo túi từ điển (Bag of Word): Dùng thuật toán *K-means* phân cụm các vector đặc trưng cục bộ thành 510 cụm, tạo ra túi từ với 510 từ (tương ứng với 510 chiều của vector BoW). Số cụm được chọn dựa vào thực nghiệm và kinh nghiệm.

Tạo vector đặc trưng BoW: Ảnh sau khi được trích xuất đặc trưng cục bộ, các vector đặc trưng cục bộ này được ánh xạ vào vector BoW theo phương pháp "*rank-weighted soft*". Sau đó chuẩn hóa bằng l_1*-norm.*

$$m_d(l) = \begin{cases} 2^{-RK_d(l)} & \text{if } RK_d(l) \leq 4 \\ 0 & \text{else.} \end{cases}$$

l: Là vector đặc trưng cục bộ
d: Là phần tử thứ d thuộc vector BoW
md : Là giá trị đánh giá đặc trưng cục bộ l thuộc thành phần d của vector BoW, md $\in [0,1]$.

Huấn luyện mô hình:

Bộ phân loại SVM được sử dụng cho mục đích dự đoán hồ sơ phân tử. Sử dụng *SHOGUN Machine Learning Toolbox* [3] để tạo ra bộ phân lớp SVM với dữ liệu đầu vào là tập vector đặc trưng BoW đã được gán nhãn.

5.2.2. Bản đồ nhiệt cho mô hình dự đoán hồ sơ phân tử.

Thực hiện xây dựng bản đồ nhiệt cho mô hình dự đoán hồ sơ phân tử theo cách tương tự như phần 1.3 của báo cáo.

Heatmap thể hiện các giá trị mức độ biểu hiện của protein được mã hóa bằng màu sắc (từ xanh lục sang đỏ).

Bản đồ nhiệt được tạo bằng kỹ thuật Lan truyền ngược qua 03 bước:

Lan truyền ngược (Backpropagate) từ ngõ ra của bộ phân lớp *f(x)* đến ngõ vào của kernel $x_{(d)}$

$x_{(d)}$ là thành phần thứ *d* của vector BoW, kernel được chọn cho phần này là HIK-kernel[7].

Kết quả đầu ra *f(x)* của bộ phân loại SVM có thể xem là tổng có trọng số của khoảng cách giữa điểm dữ liệu *x* đến từng điểm hỗ trợ trong SVM.

$$f(x) = b + \sum_i a_i y_i k(z_i, x)$$

$$\text{goal: } f(x) \approx \sum_d R_d^{(3)}(x)$$

[7] *Histogram intersection kernel*

a_i: là tham số của mô hình SVM ($(a_i * y_i)$ là trọng số theo cách giải thích trên).
b : là độ lệch của mô hình SVM.
y_i: là nhãn của điểm dữ liệu z_i trong tập các điểm hổ trợ.
x: là vector BoW thể hiện của 1 input.
z_i: là điểm dữ liệu thứ i của tập các điểm hổ trợ.
k: là hàm kernel (có thể là HIK-kernel, $\chi 2$-kernel).
R_d: là mức đóng góp của thành phần thứ d của x vào kết quả dự đoán.
d: là số chiều của 1 input.

Trường hợp kernel là HIK-kernel

$$k(z,x) = \sum_d \min(z_{(d)}, x_{(d)})$$

Khi đó $R_d^{(3)}$ được tính theo công thức:

$$f(x) = b + \sum_i a_i y_i k(z_i, x)$$

$$= b + \sum_d \sum_i a_i y_i k_d(z_{(d)}, x_{(d)})$$

$$R_d^{(3)}(x) = \frac{b}{D} + \sum_i a_i y_i k_d(z_{(d)}, x_{(d)})$$

D: Là số chiều của vector BoW

b: Là bias

z: Là vector hổ trợ

x: Là vector input BoW

Thực hiện lan truyền ngược từ ngõ vào $x_{(d)}$ của kernel đến đặc trưng cục bộ.

Giá trị $R_d^{(2)}$ được tính theo công thức.

$$R^{(2)}(l) = \sum_d R_d^{(3)} \frac{m_d(l)}{\sum_{l'} m_d(l')}$$

Ý nghĩa: Phân phối toàn bộ giá trị $R_d^{(3)}$ đến từng đặc trưng cục bộ khi giá trị m_d tương ứng > 0 .

Trường hợp tổng quát có xét tới $d | \sum_l m_d(l) = 0$, $R_d^{(2)}$ tính theo công thức sau.

$$R^{(2)}(l) = \sum_{d \notin Z(x)} R_d^{(3)} \frac{m_d(l)}{\sum_{l'} m_d(l')} + \sum_{d \in Z(x)} R_d^{(3)} \frac{1}{\sum_{l'} 1}$$

Trong đó: Z(x) = { $d | \sum_l m_d(l) = 0$}.

Thực hiện lan truyền ngược từ đặc trưng cục bộ lên các pixel, tạo heatmapping.

Ý tưởng: Chỉ phân bổ giá trị $R_l^{(2)}$ của các vector đặc trưng cục bộ đến các pixel nằm trong chính vector đặc trưng cục bộ ấy (các pixel đó gọi là pixel hổ trợ, ký hiệu q).

$$LF(q) = \{l \mid q \in supp(l)\}$$

$$R^{(1)}(q) = \sum_{l \in LF(q)} \frac{R^{(2)}(l)}{|supp(l)|}$$

LF(q): Là tập các vector đặc trưng cục bộ có chứa pixel q

|supp(l)| : Là tổng số lượng pixel có trong đặc trưng cục bộ l

R(1)(q): Là giá trị đóng góp của pixel q cho f(x)

Hình ảnh bản đồ nhiệt cho hồ sơ phân tử:

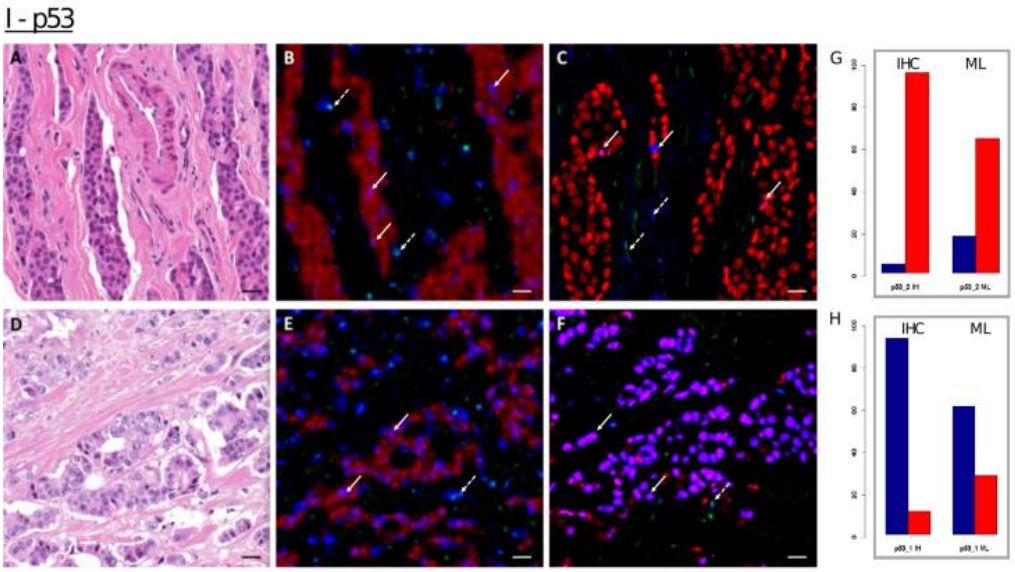


Figure 2: Hình ảnh H&E gốc được sử dụng để dự đoán khối u, p53 biểu hiện thấp, phân tán (A) và p-53 khối u biểu hiện cao (D), dự đoán tính toán tương ứng (B,E) và nhuộm IHC[8] (C,F)

Protein p53 là một protein ức chế khối u, có khả năng kiểm soát sự phân chia tế bào, sửa chữa DNA hư hỏng. Khi xảy ra đột biến trong gene p53, protein p53 sẽ bị mất chức năng, gây ra sự tăng trưởng tế bào bất thường và dẫn đến sự phát triển của khối u ác tính. Hình ảnh trên cho thấy tương quan giữa Heatmap cung cấp bằng chứng phân tử của p53 (ảnh B,E) và phát họa của p53 bằng phương pháp IHC. Heatmap đã thể hiện được khả năng cung cấp bằng chứng cho dấu hiệu phân tử (mặc dù chất lượng không cao bằng phương pháp IHC), nhưng bù lại việc xây dựng Heamap là nhanh hơn và có sự hỗ trợ mạnh mẽ của máy tính, chi phí cũng rẻ hơn so với phương pháp IHC.

5.2.3. Kính hiển vi huỳnh quan điện toán (Computational fluorescence microscopy)

Bằng cách kết hợp nhiều bản đồ nhiệt (Heatmap cho bằng chứng ung thư, Heatmap cho bằng chứng phân tử…). Bài báo đề xuất một phương pháp xây dựng hình ảnh huỳnh quang (*fluorescence*[9]**)** gọi là *"Computational fluorescence microscopy"*.

[8] IHC (Immunohistochemistry) phác họa p53 là một phương pháp được sử dụng trong bệnh học và nghiên cứu để phát hiện sự biểu hiện của protein p53 trong mẫu mô.

[9] Fluorescence trong mô bệnh học là một phương pháp sử dụng các chất sắc thể fluorochrome để đánh dấu các tế bào hoặc các thành phần khác trong mẫu mô bệnh học để phát hiện và phân tích chúng bằng các kỹ thuật hình ảnh hoặc phân tích dữ liệu.

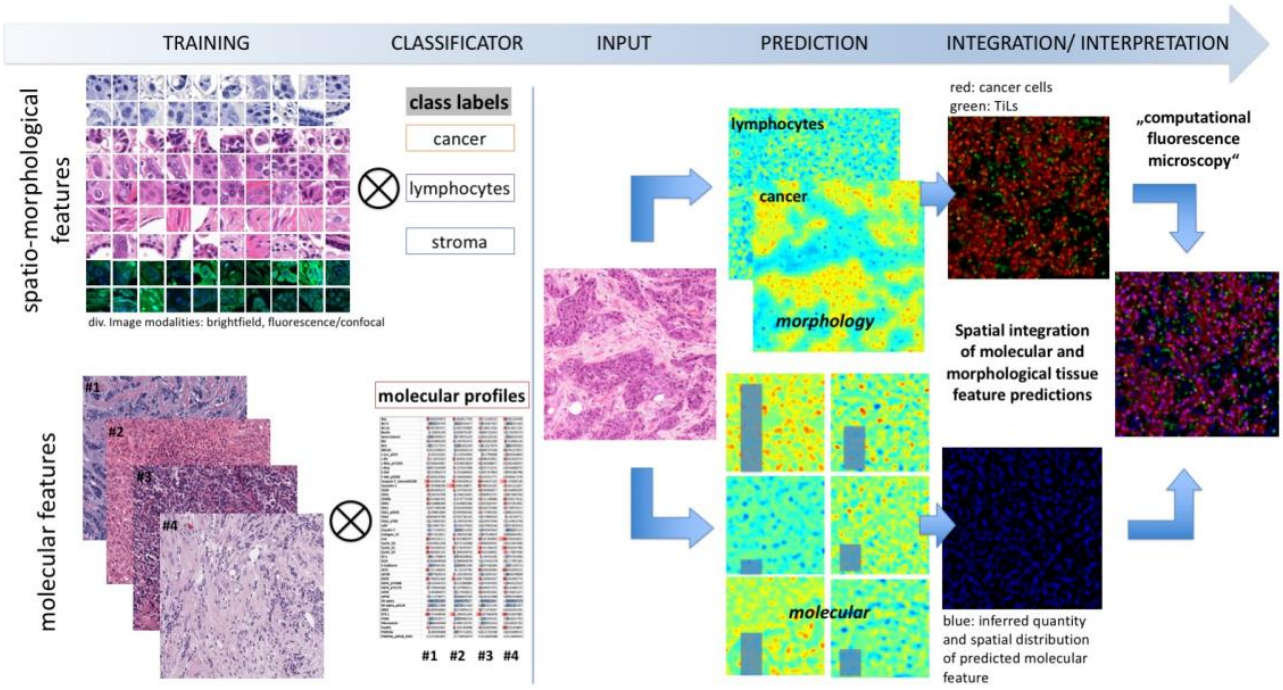


Figure 3: Workflow of machine learning-based integrated prediction of morphological and molecular tumor profiles (Computational fluorescence microscopy)

5.3. Kỹ thuật bản đồ nhiệt tìm kiếm độ lệch (SAI LỆCH) trong mạng học sâu

Vấn đề: Làm cách nào phát hiện ra sai lệch của mô hình dự đoán và khởi nguồn của sai lệch?

Bài báo tham khảo chính: "Resolving challenges in deep learning-based analyses of histopathological images using explanation methods" - Miriam Hägele et, al. [5]

Chẩn đoán mô bệnh học được thực hiện bởi các chuyên gia được đào tạo bằng cách đánh giá trực quan các đặc tính của mô ở cấp độ vi mô. Tuy nhiên, số lượng yêu cầu chẩn đoán ngày càng tăng và nhu cầu đánh giá định lượng bổ sung về các đặc tính của mô đòi hỏi các phương pháp phân tích hình ảnh điện toán hỗ trợ. Các kỹ thuật học sâu được áp dụng vào bài toán này do hiệu suất và độ chính xác cao của nó. Ví dụ đối với nhiệm vụ cụ thể là phân loại tổn thương da, mạng lưới thần kinh tích chập cho thấy hiệu suất ngang với các chuyên gia y tế [5].

Nhưng trong khi chất lượng hiệu suất dự đoán là cần thiết, thì việc làm cho quá trình phân loại minh bạch, giải thích được là rất quan trọng, đặc biệt đối với ứng dụng y tế.

5.3.1. Một số đặc điểm của mạng học sâu.

Mạng học sâu đang được ứng dụng rất mạnh mẽ trong những năm gần đây do nhưng ưu thế của nó so với các phương pháp truyền thống. Với mạng học sâu con người không cần

dạy cho mạng phương pháp trích xuất đặc trưng, con người chỉ cần cung cấp dữ liệu huấn luyện đủ và phù hợp, mạng học sâu sẽ tự tìm ra cách để trích xuất đặc trưng của dữ liệu. Do đó đặc điểm chung của mạng học sâu là:

- Cần một lượng dữ liệu lớn cho huấn luyện mạng.
- Có rất nhiều tham số "hyperparameters" cần phải cài đặc phù hợp (batch size, minibatch structure, learning rate, learning rate decay, optimizer…).
- Dữ liệu huấn luyện nếu không đủ buộc phải sử dụng các kỹ thuật "Data Augmentation" để phát sinh thêm dữ liệu cho bộ huấn luyện.
- Dữ liệu bị lệch "biases" hay còn gọi là mất cân bằng dữ liệu, nghĩa là phân phối giữa các lớp trong dữ liệu có sự chên lệch quá lớn làm giảm độ chính xác của các mô hình máy học. Vấn đề này là không dễ giải quyết và xuất hiện nhiều trong các tập dữ liệu về y tế.

Ví dụ: Một mô hình mạng nơ ron có chức năng dự đoán tế bào ung thư.

- Nếu tập huấn luyện bị lệch (các mẫu "dương tính" quá ít so với "mẫu âm tính" (rất thường xuyên xuất hiện trong dữ liệu về y tế) thì mô hình sẽ dự đoán không hiệu quả khi đưa vào sử dụng.
- Mô hình học từ dữ liệu bao gồm các thông tin không cần thiết, điều này làm cho mô hình có tính khái quát cao quá mức và giảm hiệu suất dự đoán, mô hình không phân biệt được sự khác nhau giữa các mẫu "dương tính" và "âm tính".
- Dữ liệu huấn luyện bị mất cân bằng do quá trình gắn nhãn và chúng ta không phát hiện ra, dẫn đến mạng hoạt động không tốt.

Trong chuẩn đoán ung thư (trong ví dụ là ung thư vú) bộ phân lớp đưa ra dự đoán "dương tính" và ung thư ở mức độ 3 (grade-3), vấn đề là bác sĩ lâm sàng không thể tin hoàn toàn vào dự đoán, mô hình phân loại cũng ko giải thích được lý do cho dự đoán này. Hoặc trong trường hợp bộ phân loại nhầm lẫn, hoặc bác sĩ đã bỏ sót các chi tiết nào đó.

Kỹ thuật bản đồ nhiệt sẽ hỗ trợ bác sĩ đưa ra chuẩn đoán của riêng mình, bản đồ nhiệt khoanh vùng các khu vực có liên quan đến dự đoán của mô hình hổ trợ bác sĩ kiểm tra, xác nhận tính chính xác của mô hình dự đoán.

5.3.2. Độ lệch (sai lệch, thiên lệch, thiên vị) trong các mô hình học sâu.

Mạng học sâu có đóng góp quan trọng trong việc chuẩn đoán bệnh ý vì độ chính xác cao của nó tuy nhiên trong lĩnh vực y tế yêu cầu giải thích và hiểu biết sâu sắc để hiểu rõ hơn ngoài đánh giá hiệu suất của mô hình dự đoán. Đối với các mô hình dự đoán từ dữ liệu hình ảnh mô tế bào thường tồn tại một số loại sai lệch (biases).

- Sai lệch của toàn bộ dữ liệu (Dataset bias)
- Sai lệnh ngẫu nhiên có tương quan với nhãn phân lớp (Class correlated bias)
- Sai lệch lấy mẫu (Sample bias)

Các vấn đề của tập dữ liệu huấn luyện gây ảnh hưởng đến hiệu suất của mô hình học sâu khi được học từ các tập dữ liệu bị sai lệch.

- Dữ liệu bị gán nhãn không thống nhất do yếu tố chủ quan của những chuyên gia, người quan sát.
- Phương sai giữa các tập dữ liệu khác nhau.
- Sự mất cân bằng giữa các lớp trong cùng một tập dữ liệu.
- Dữ liệu quá ít.

Xây dựng bản đồ nhiệt là một cách tiếp cận linh hoạt và hiệu quả giúp chúng ta phát hiện và loại bỏ các tác động từ biases, giúp ta khái quát, hiểu rõ hơn về tập dữ liệu, chúng ta có thể thấy rõ tác dụng này qua các minh chứng từ bài báo [6].

5.3.3. Xây dựng bản đồ nhiệt cho tìm kiếm độ lệch của tập dữ liệu.

Sau khi huấn luyện mạng học sâu (ví dụ: mạng học sâu cho nhiệm vụ phân loại nhị phân). Heatmap được xây dựng dựa trên kỹ thuật LRP, các điểm liên quan *(relevance score)* từ đầu ra của mạng được phân phối lại trên từng pixel của ảnh đầu vào dựa theo đóng góp của từng pixel vào kết quả dự đoán, các giá trị này tập trung quanh giá trị 0, sau đó chúng được chuẩn hóa về phạm vi [-1,1] biểu diễn bằng dải màu tương ứng từ xanh lam với giá trị -1,đỏ với giá trị +1.

Trong bài báo [6], tác giả dùng phương pháp LRP với quy tắc *LRP-ε* cho lớp FC (fully connector) để lan truyền ngược từ output

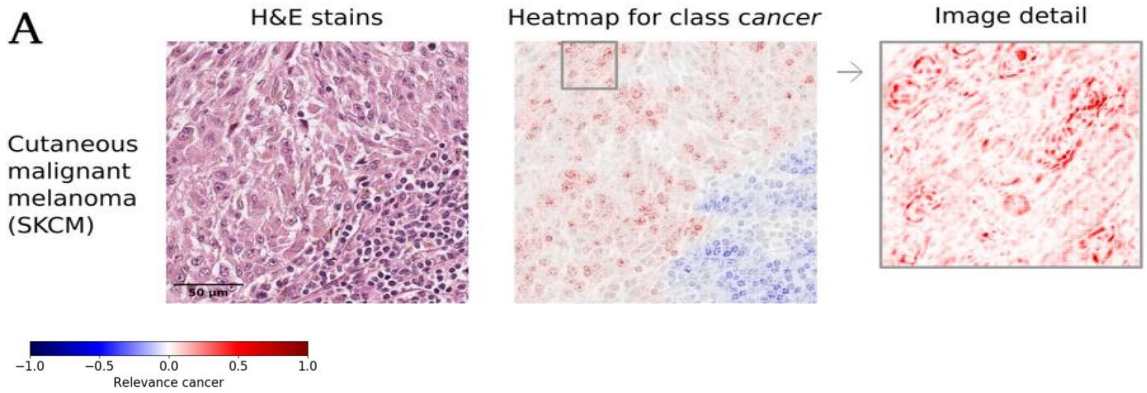
$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_{i'} Z_{i'j} + \varepsilon \cdot \text{sign}\left(\sum_{i'} Z_{i'j}\right)} R_j^{(l+1)}$$

Đối với các lớp chập (Convolution layers) tiếp theo được áp dụng quy tắc *LRP-αβ* (α = 1 và β = 0):

$$R_i^{(l)} = \sum_j \left(\alpha \cdot \frac{z_{ij}^+}{\sum_{i'} z_{i'j}^+} + \beta \cdot \frac{z_{ij}^-}{\sum_{i'} z_{i'j}^-} \right) R_j^{(l+1)}$$

Vì Input của mạng học sâu có kích thước cố định và chỉ là kích thước của một phần nhỏ của ảnh y tế (ảnh với kích thước thật các bác sĩ hay quan sát, ảnh này có kích thước lớn hơn nhiều so với ảnh đưa vào mạng học sâu) nên bản đồ nhiệt thực tế phải tổng hợp từ nhiều bản đồ nhiệt nhỏ qua các lần dự đoán của mô hình.

Verifying learned features: Bản đồ nhiệt trực quan hóa các dấu hiệu ung thư – giúp kiểm tra, xác minh lại chuẩn đoán của mô hình.



Data sampling strategies: Bản đồ nhiệt phát hiện ảnh hưởng của tỉ lệ lấy mẫu - Giúp đưa ra quyết định điều chỉnh tỉ lệ lấy mẫu phù hợp với mục đích bài toán.

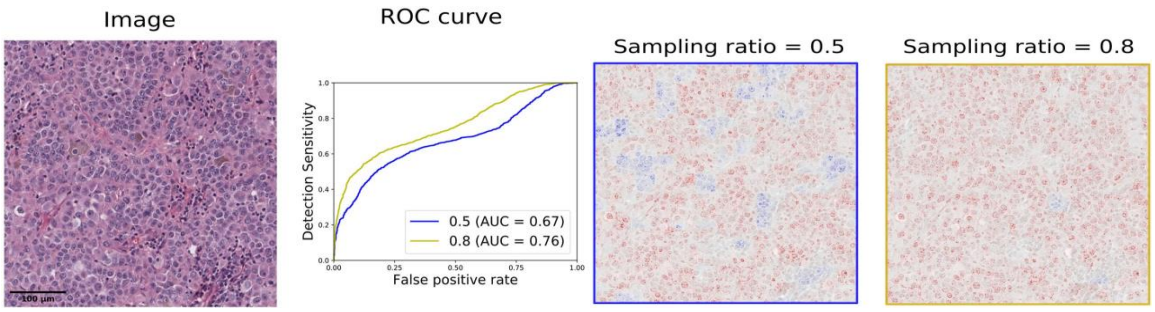


Figure 4: Điều tra ảnh hưởng của tỷ lệ lấy mẫu lớp cố định trong các 'mini-batches' đối với bản đồ nhiệt giải thích

Đường cong ROC[10] ứng với Sampling ratio = 0.8 (lấy mẫu quá mức có lợi cho lớp ung thư) ở trên đường cong ROC ứng với Sampling ratio = 0.5 (lấy mẫu cân bằng) và bản đồ nhiệt ở mỗi trường hợp, cho thấy kỹ thuật bản đồ nhiệt đã phát hiện ra ảnh hưởng của tỉ lệ lấy mẫu.

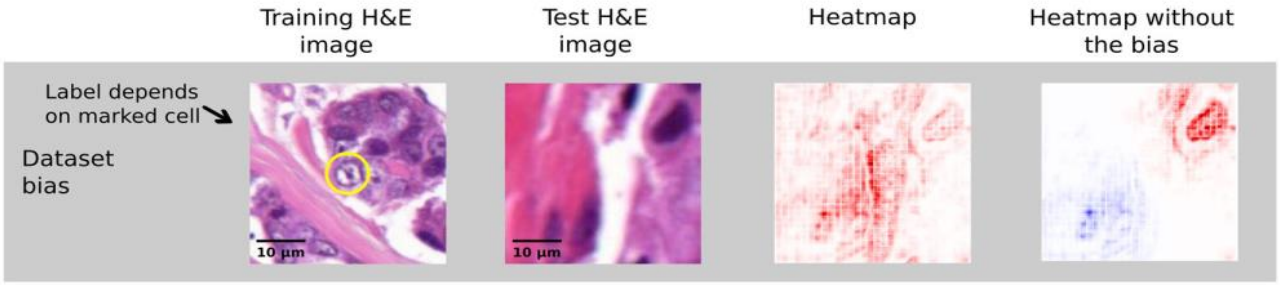
Dataset bias:

[10] *Đường cong ROC biểu thị sự phân tách của các điểm dữ liệu thuộc hai nhóm, được biểu diễn trên một biểu đồ hai chiều với trục x là tỷ lệ dự đoán sai (false positive rate) và trục y là tỷ lệ dự đoán đúng (true positive rate).*

Là loại bias liên quan đến sự khác biệt giữa các tập dữ liệu. Nó xảy ra khi một mô hình học máy được huấn luyện trên một tập dữ liệu không đại diện cho dữ liệu trong thực tế, tức là các đặc trưng (feature) của tập dữ liệu huấn luyện khác với các đặc trưng của dữ liệu thực tế mà mô hình sẽ được sử dụng. Điều này dẫn đến sự khác biệt giữa độ chính xác dự đoán của mô hình trên dữ liệu huấn luyện và trên dữ liệu thực tế. Dataset bias còn có thể xảy ra khi một số dữ liệu bị thiếu hoặc không có trong tập huấn luyện.

Trong hình là kết quả dự đoán của 1 mô hình nhưng được huấn luyện trên 2 tập data khác nhau. Một tập huấn luyện mà nhãn dương tính chỉ phụ thuộc vào nhãn ô trung tâm (Dataset bias do quá trình gán nhãn, dữ liệu huấn luyện không thể hiện đủ các trường hợp trong thực tế) và một tập huấn luyện mà nhãn dương tính được gán tại nơi có dấu hiệu ung thư.

Bản đồ nhiệt thể hiện rõ được vấn đề “dataset bias”



Mô hình huấn luyện trên tập dataset bị bias đôi khi không phát hiện được dấu hiệu ung thư ở các vị trí nằm ngoài ô trung tâm, mặc dù mô hình hoạt động rất tốt trên cả tập train và tập test (với cùng một kiểu gán nhãn).

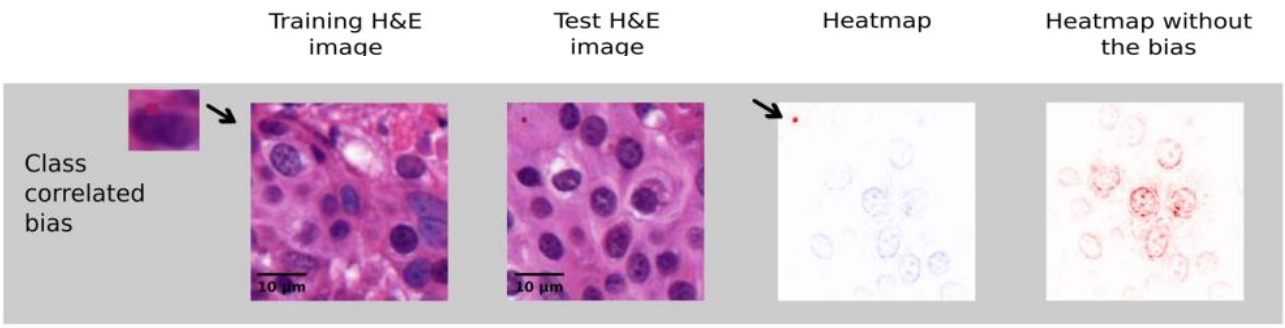
Class-correlated bias:

Là loại bias liên quan đến sự khác biệt giữa các lớp dữ liệu. Nó xảy ra khi một mô hình học máy được huấn luyện trên một tập dữ liệu không cân bằng về các lớp, tức là tỷ lệ các lớp trong tập dữ liệu không đồng đều. Khi đó, mô hình sẽ có xu hướng học tốt các lớp thiểu số hơn là các lớp đa số, gây ra sai lệch trong dự đoán cho các lớp đa số. Class-correlated bias còn có thể xảy ra khi một số đặc trưng (feature) của dữ liệu liên quan đến lớp dữ liệu.

Những tác động ngoài môi trường, những sai sót kỹ thuật trong việc thu thập dữ liệu, hoặc thậm chí là những lỗi trong quá trình xử lý dữ liệu tác động lên dữ liệu làm cho dữ liệu bị bias.

Trong thực tế các tế bào ung thư và tế bào bình thường sẽ có sự khác biệt chút ít về kích thước, hình dạng và màu sắc (ví dụ tế bào ung thư thường lớn, màu sắc hơi khác và hình dạng không đều so với tế bào bình thường). Nếu mô hình phân loại được đào tạo trên tập mẫu bao gồm nhiều ảnh tế bào bình thường nhưng chỉ có số lượng nhỏ tế bào ung thư thì mô hình sẽ không phân biệt tốt các tế bào này khi áp dụng trên một ảnh mới.

Để mô phỏng tác động của "Class-correlated bias", trong bài báo, tác giả tô 1 vùng nhỏ (5px*5px) màu trùng với màu ảnh được nhộm H&E vào khu vực chứa dấu hiệu ung thư của một số ảnh. Việc này làm cho mô hình đưa ra dự đoán sai. Thể hiện rất rõ trong hình bênh dưới.



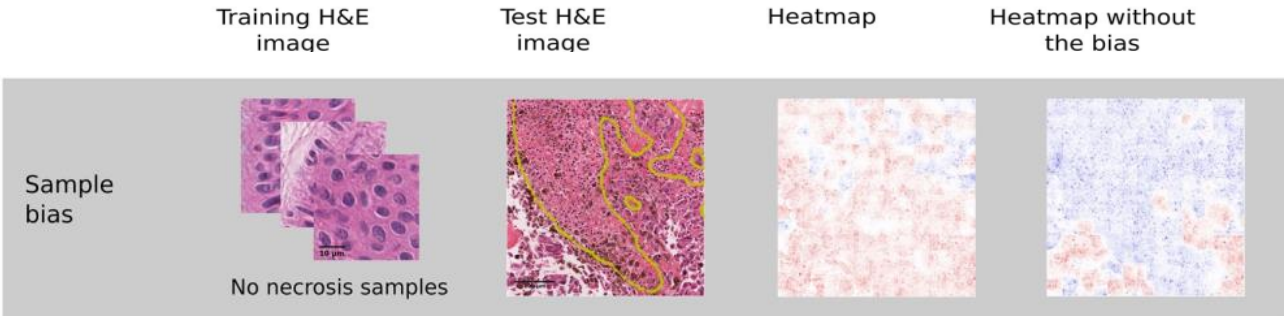
Sample bias:

Là vấn đề về tập huấn luyện không chứa đầy đủ các tính chất đặc trưng của đối tượng nghiên cứu, do quá trình lấy mẫu sinh ra, ví dụ lấy mẫu thiếu, lượt bỏ các thành phần không liên qua tới mục đích bài toán một cách thiếu xem xét.

Sample bias có thể xảy ra khi lựa chọn mẫu không đại diện cho quần thể, thiếu dữ liệu cho một số đối tượng trong quần thể, mẫu được thu thập bằng cách không ngẫu nhiên hoặc không tuân theo quy trình khoa học, lựa chọn mẫu quá nhỏ.

Để mô tả lỗi trên, các ảnh H&E huấn luyện cho bộ phân loại ung thư được cắt bỏ phần hình ảnh mô bị hoại tử (phần này không phải là ung thư, không liên quan đến tế bào ung thư). Kết quả sau huấn luyện thì mô hình nhầm lẫn các tế bào hoại tử là tế bào ung thư.

Bản đồ nhiệt thể hiện rất rõ và trực quan vấn đề Sample bias.



5.4. Kết luận

Heatmap là một công cụ hữu ích hỗ trợ cho các mô hình dự đoán mô bệnh học, heatmap cho phép trực quan hóa các thông tin mà mô hình máy học dựa vào để cho ra kết quả dự đoán. Qua các minh chứng bằng thực nghiệm heatmap đã chứng tỏ được vai trò của mình trong việc giải thích kết quả dự đoán của một mô hình máy học, phát hiện ra sai lệch của dữ

liệu và còn nhiều chức năng khác cần được chúng ta khai phá và ứng dụng công cụ này trong tương lai.

Tài liệu tham khảo

[1] Klaus-Robert Müller, “Explaining and Interpreting Deep Neural Networks”, Tu-Berlin, CoSIPICDL2017.

[2] Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K. R. (2019). Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, 193-209.

[3] Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W (2015) On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLoS ONE 10(7): e0130140. https://doi.org/10.1371/journal.pone.0130140.

[4] W Samek, A Binder, G Montavon, S Lapuschkin, KR Müller. Evaluating the visualization of what a Deep Neural Network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660-2673, 2017.

[5] L Arras, F Horn, G Montavon, KR Müller, W Samek. Explaining Predictions of Non-Linear Classifiers in NLP. *Workshop on Representation Learning for NLP,* Association for Computational Linguistics, 1-7, 2016.

[6] L Arras, F Horn, G Montavon, KR Müller, W Samek. "What is Relevant in a Text Document?": An Interpretable Machine Learning Approach. *PLOS ONE, 12(8):e0181142*, 2017.

[7] S Lapuschkin, A Binder, G Montavon, KR Müller, Wojciech Samek. Analyzing Classifiers: Fisher Vectors and Deep Neural Networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2912-20, 2016.

[8] S Lapuschkin, A Binder, KR Müller, W Samek. Understanding and Comparing Deep Neural Networks for Age and Gender Classification. *IIEEE International Conference on Computer Vision Workshops (ICCVW)*, 1629-38, 2017.

[9] I Sturm, S Lapuschkin, W Samek, KR Müller. Interpretable Deep Neural Networks for Single-Trial EEG Classification. *Journal of Neuroscience Methods*, 274:141–145, 2016.

[10] B. Alexander, "Towards computational fluorescence microscopy: Machine learning-based integrated prediction of morphological and molecular tumor profiles," *arxiv,* p. 58, 2018.

[11] D. Lowe, "Distinctive image features from scale invariant keypoints," *International Journal of Computer Vision,* 2004.

[12] Sonnenburg, Ratsch, Henschel, Widmer, Behr, Zien, Bona, Binder, Gehl and Franc, "The SHOGUN Machine Learning Toolbox," *Journal of Machine Learning Research,* 2010.

[13] B. Sebastian, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLOS One,* p. 46, 10 July 2015.

[14] Esteva, A. et al, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature,* 2017.

[15] M. Hägele, P. Seegerer, S. Lapuschkin, M. Bockmayr, W. Samek, F. Klauschen, K.-R. Müller and A. Binder, "Resolving challenges in deep learning-based analyses of histopathological images using explanation methods".

[16] J. Uijlings, A. Smeulders and R. Scha, "The Visual Extent of an Object Suppose We Know the Object Locations," *International Journal of Computer Vision,* p. 18, 2012.