`TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**TRẦN QUỐC VINH– 52000823**

**NGUYỄN QUANG PHƯƠNG NAM – 52000823**

**TÌM HIỂU KỸ THUẬT XAI (EXPLAINABLE AI) ÁP DỤNG TRÊN KIỂU DỮ LIỆU ẢNH**

**DỰ ÁN CÔNG NGHỆ THÔNG TIN**

**HỆ THỐNG THÔNG TIN**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2025**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**TRẦN QUỐC VINH– 52000823**

**NGUYỄN QUANG PHƯƠNG NAM – 52000783**

**TÌM HIỂU KỸ THUẬT XAI (EXPLAINABLE AI) ÁP DỤNG TRÊN KIỂU DỮ LIỆU ẢNH**

**DỰ ÁN CÔNG NGHỆ THÔNG TIN**

**HỆ THỐNG THÔNG TIN**

Người hướng dẫn

**Phạm Văn Huy**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2025**

**LỜI CẢM ƠN**

Chúng em chân thành biết ơn Khoa Công Nghệ Thông Tin, Trường Đại học Tôn Đức Thắng đã tạo điều kiện học tập thuận lợi cho em trong suốt thời gian qua. Đồng thời, em muốn bày tỏ lòng biết ơn sâu sắc đến thầy Phạm Văn Huy, người đã dành thời gian và tâm huyết để giảng dạy và hướng dẫn em trong suốt quãng thời gian này. Thầy đã truyền đạt cho em những kiến thức quý báu, giúp em phát triển không chỉ về mặt học thuật mà còn về mặt kỹ năng và tư duy.

Em xin bày tỏ lòng biết ơn sâu sắc và trân trọng đối với thầy.

*TP. Hồ Chí Minh, ngày 10 tháng 02 năm 2025*

*Vinh*

*Trần Quốc Vinh*

*Nam*

*Nguyễn Quang Phương Nam*

**CÔNG TRÌNH ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của thầy Phạm Văn Huy. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong Dự án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung Dự án của mình**. Trường Đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 10 tháng 02 năm 2025*

*Vinh*

*Trần Quốc Vinh*

*Nam*

*Nguyễn Quang Phương Nam*

**TÌM HIỂU KỸ THUẬT XAI (EXPLAINABLE AI) ÁP DỤNG TRÊN KIỂU DỮ LIỆU ẢNH**

**TÓM TẮT**

**MỤC LỤC**

[DANH MỤC HÌNH VẼ vii](#_Toc190184886)

[DANH MỤC BẢNG BIỂU viii](#_Toc190184887)

[DANH MỤC CÁC CHỮ VIẾT TẮT ix](#_Toc190184888)

[CHƯƠNG 1. MỞ ĐẦU VÀ TỔNG QUAN ĐỀ TÀI 1](#_Toc190184889)

[1.1 Lý do chọn đề tài 1](#_Toc190184890)

[1.2 Mục tiêu thực hiện đề tài 2](#_Toc190184891)

[1.3 Tầm quan trọng của XAI, các thách thức khi áp dụng 3](#_Toc190184892)

[1.3.1 Explainable AI (XAI) là gì và tại sao quan trọng? 3](#_Toc190184893)

[1.3.2 Các thách thức của XAI 3](#_Toc190184894)

[1.3.3 Ứng dụng thực tế của XAI 4](#_Toc190184895)

[1.3.4 Hướng phát triển của XAI 6](#_Toc190184896)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 7](#_Toc190184897)

[2.1 Tổng quan về “Explainability” 7](#_Toc190184898)

[2.2 Như thế nào là “Explainability” 10](#_Toc190184899)

[2.3 So sánh giữa “Explainability” và “Interpretability” 12](#_Toc190184900)

[2.3.1 Interpretable Model (Mô hình có thể diễn giải trực tiếp) 13](#_Toc190184901)

[2.3.2 Explainable Model (Mô hình cần được giải thích) 14](#_Toc190184902)

[2.3.3 So sánh giữa Interpretable Model và Explainable Model 16](#_Toc190184903)

[2.3.4 Explainability với từng loại người dùng 17](#_Toc190184904)

[2.4 Các phương pháp giải thích trong Explainable AI (XAI) 19](#_Toc190184905)

[2.4.1 Global Explanation (GE) – Giải thích toàn cục 19](#_Toc190184906)

[2.4.2 Local Explanation (LE) – Giải thích cục bộ 20](#_Toc190184907)

[2.4.3 Cohort Explanation (CE) – Giải thích theo nhóm dữ liệu 21](#_Toc190184908)

[2.4.4 Attribution Explanation (AE) – Giải thích dựa trên độ quan trọng của feature 22](#_Toc190184909)

[2.4.5 Example-Based Explanations (EBE) – Giải thích dựa trên ví dụ 23](#_Toc190184910)

[2.4.6 Counterfactual Explanations – Giải thích phản thực tế 24](#_Toc190184911)

[2.5 Các Kỹ Thuật Model-Agnostic XAI 25](#_Toc190184912)

[2.5.1 Dữ Liệu Dạng Bảng 25](#_Toc190184913)

[2.5.2 Dữ Liệu Dạng Ảnh 29](#_Toc190184914)

[2.6 Framework 38](#_Toc190184915)

[2.6.1 PyTorch 38](#_Toc190184916)

[2.6.2 PySpark 39](#_Toc190184917)

[CHƯƠNG 3. MÔ HÌNH ĐỀ XUẤT VỚI MẠNG NƠ-RON 40](#_Toc190184918)

[3.1 Neural Graph Collaborative Filtering (NGCF) 40](#_Toc190184919)

[3.2 LightGCN 40](#_Toc190184920)

[3.3 Weighted Addition Input Matrix Graph Convolutional Network (WiGCN) 40](#_Toc190184921)

[CHƯƠNG 4. THỰC NGHIỆM 42](#_Toc190184922)

[4.1 Dữ liệu thực nghiệm 42](#_Toc190184923)

[4.2 Cài đặt thực nghiệm 42](#_Toc190184924)

[4.2.1 Các tham số quan trọng ảnh hưởng đến quá trình huấn luyện mô hình 42](#_Toc190184925)

[4.2.2 Các bước huấn luyện mô hình 45](#_Toc190184926)

[4.3 Kết quả huấn luyện và demo hệ thống khuyến nghị. 49](#_Toc190184927)

[4.3.1 Kết quả sau khi huấn luyện với các tham số phù hợp 49](#_Toc190184928)

[4.3.2 Demo hệ thống Lưu trữ thông tin người dùng và doanh nghiệp vào Database 50](#_Toc190184929)

[CHƯƠNG 5. KẾT LUẬN 58](#_Toc190184930)

[5.1 Kết luận 58](#_Toc190184931)

[5.2 Hướng phát triển 59](#_Toc190184932)

[TÀI LIỆU THAM KHẢO 60](#_Toc190184933)

# DANH MỤC HÌNH VẼ

[Hình 1 Câu hỏi mang tính chất “Explainability” 8](#_Toc190184934)

[Hình 2 Cấu trúc cơ bản của một hệ thống 9](#_Toc190184935)

[Hình 3 Các kỹ thuật giải thích mô hình (XAI) 10](#_Toc190184936)

[Hình 4 Thế nào là Explainability 12](#_Toc190184937)

[Hình 5 Interpretable Model 14](#_Toc190184938)

[Hình 6 Explainable model 15](#_Toc190184939)

[Hình 7 So sánh giữa Interpretable Model và Explainable Model 16](#_Toc190184940)

[Hình 8 Các phương pháp giải thích Explanation 19](#_Toc190184941)

[Hình 9 Global Explanation (GE) 20](#_Toc190184942)

[Hình 10 Local Explanation (LE) 21](#_Toc190184943)

[Hình 11 Cohort Explanation (CE) 22](#_Toc190184944)

[Hình 12 Attribution Explanation (AE) 23](#_Toc190184945)

[Hình 13 Example-Based Explanations (EBE) 24](#_Toc190184946)

[Hình 14 Counterfactual Explanations 25](#_Toc190184947)

[Hình 15 Premodeling Explainability (EDA) 25](#_Toc190184948)

[Hình 16 Ví dụ dự đoán giá nhà của PDPs 28](#_Toc190184949)

[Hình 17 Sơ đồ hoạt động của CAM 31](#_Toc190184950)

[Hình 18 Sơ đồ hoạt động của Grab-CAM 32](#_Toc190184951)

[Hình 19 Phân đoạn thành các siêu pixel 34](#_Toc190184952)

[Hình 20 Giá trị trung bình của các giá trị điểm ảnh riêng lẻ 36](#_Toc190184953)

[Hình 21 Kết quả của thuật toán LIME 37](#_Toc190184954)

[Hình 22 Thư mục chứa dữ liệu cho quá trình huấn luyện 42](#_Toc190184955)

[Hình 23 Hàm mất mát BPR (Bayesian Personalized Ranking) 48](#_Toc190184956)

[Hình 24 Kết quả sau khi huấn luyện với các tham số phù hợp 49](#_Toc190184957)

[Hình 25 Dữ liệu của người dùng trong database 50](#_Toc190184958)

[Hình 26 Dữ liệu của doanh nghiệp trong database 50](#_Toc190184959)

[Hình 27 Hàm load mô hình đã huấn luyện và sử dụng để dự đoán 51](#_Toc190184960)

[Hình 28 Kết quả khi sử dụng mô hình dự đoán 52](#_Toc190184961)

[Hình 29 Giao diện login cho người quản lý 53](#_Toc190184962)

[Hình 30 Giao diện trang chủ khi đăng nhập thành công 53](#_Toc190184963)

[Hình 31 Giao diện danh sách người dùng của Yelp 2025 54](#_Toc190184964)

[Hình 32 Giao diện danh sách doanh nghiệp của Yelp 2025 54](#_Toc190184965)

[Hình 33 Giao diện chi tiết thông tin của doanh nghiệp 55](#_Toc190184966)

[Hình 34 Giao diện của hệ thống khuyến nghị 55](#_Toc190184967)

[Hình 35 Giao diện khi nhấn send và gửi api về mô hình khuyến nghị 56](#_Toc190184968)

[Hình 36 Kết quả đề xuất các doanh nghiệp phù hợp cho người dùng. 56](#_Toc190184969)

[Hình 37 Thông tin doanh nghiệp sau khi nhấn vào tên trong danh sách đề xuất 57](#_Toc190184970)

[Hình 38 Giao diện khi truy cập đường dẫn nhưng chưa đăng nhập 57](#_Toc190184971)

# DANH MỤC BẢNG BIỂU

[Bảng 1 So sánh 2 kỹ thuật (EDA) và (PDPs) 29](#_Toc190184972)

# DANH MỤC CÁC CHỮ VIẾT TẮT

XAI Explainable Artificial Intelligence (Trí tuệ nhân tạo có thể giải thích)

GE Global Explanation (Giải thích tổng quát)

LE Local Explanation (Giải thích cục bộ)

CE Cohort Explanation (Giải thích theo nhóm dữ liệu)

AE Attribution Explanation (Giải thích theo quy kết)

EBE Example-Based Explanation (Giải thích dựa trên ví dụ)

PDP Partial Dependence Plot (Biểu đồ phụ thuộc cục bộ)

EDA Exploratory Data Analysis (Phân tích dữ liệu khám phá)

GAP Global Average Pooling (Tổng hợp trung bình toàn cục)

CAM Class Activation Map (Bản đồ kích hoạt lớp)

Grad-CAM Gradient-weighted Class Activation Mapping

CNN Convolutional Neural Network (Mạng nơ-ron tích chập)

KNN K-Nearest Neighbors (Thuật toán K láng giềng gần nhất)

LIME: Local Interpretable Model-Agnostic Explanations (Giải thích mô hình cục bộ và không phụ thuộc vào mô hình)

SVM Support Vector Machine (Máy vector hỗ trợ)

RF Random Forest (Rừng ngẫu nhiên)

DNN Deep Neural Network (Mạng nơ-ron sâu)

API Application Programming Interface (Giao diện lập trình ứng dụng)

# MỞ ĐẦU VÀ TỔNG QUAN ĐỀ TÀI

## Lý do chọn đề tài

Trong thời đại dữ liệu lớn và trí tuệ nhân tạo phát triển mạnh mẽ, các mô hình học sâu (Deep Learning) đã trở thành công cụ quan trọng trong nhiều lĩnh vực, từ nhận diện hình ảnh, xử lý ngôn ngữ tự nhiên đến hệ thống tự động hóa. Tuy nhiên, một trong những thách thức lớn nhất của các mô hình này là tính "hộp đen" (Black Box), khi mà quá trình ra quyết định của mô hình không thể giải thích rõ ràng. Điều này đặt ra nhiều vấn đề về sự minh bạch, khả năng kiểm tra và tính trách nhiệm, đặc biệt trong các lĩnh vực quan trọng như y tế, tài chính và phương tiện tự hành.

Mặc dù các mô hình học sâu mang lại hiệu suất cao trong dự đoán và xử lý dữ liệu phức tạp, nhưng việc thiếu khả năng giải thích khiến chúng trở nên khó kiểm soát, làm giảm sự tin cậy từ người dùng và nhà quản lý. Chính vì vậy, nghiên cứu về việc cải thiện tính minh bạch và khả năng giải thích của các mô hình học sâu là một vấn đề cấp thiết.

Việc đề xuất các phương pháp giúp mô hình trở nên dễ hiểu hơn không chỉ giúp tăng độ tin cậy mà còn hỗ trợ quá trình ra quyết định trong các hệ thống quan trọng. Những nghiên cứu về XAI (Explainable AI - Trí tuệ nhân tạo có thể giải thích) đang dần mở ra hướng đi mới, giúp con người có thể hiểu rõ hơn về cách thức hoạt động của các thuật toán AI, từ đó áp dụng chúng một cách hiệu quả và an toàn hơn.

Chính vì vậy, đề tài này tập trung vào việc nghiên cứu các phương pháp giúp tăng cường tính minh bạch của các mô hình học sâu, từ đó nâng cao hiệu quả ứng dụng AI trong thực tiễn.

## Mục tiêu thực hiện đề tài

Dự án này đặt mục tiêu nghiên cứu và ứng dụng các kỹ thuật XAI (Explainable AI) nhằm cải thiện khả năng giải thích và minh bạch của các mô hình học sâu trên dữ liệu ảnh. Trong bối cảnh trí tuệ nhân tạo ngày càng phát triển và được ứng dụng rộng rãi, việc hiểu rõ cách thức hoạt động của các mô hình học sâu trở nên quan trọng, đặc biệt trong các lĩnh vực yêu cầu tính minh bạch cao như y tế, an ninh và tự động hóa.

Trước tiên, dự án sẽ tìm hiểu tổng quan về XAI, bao gồm các phương pháp phổ biến như LIME (Local Interpretable Model-agnostic Explanations), SHAP (Shapley Additive Explanations) và Grad-CAM (Gradient-weighted Class Activation Mapping). Những phương pháp này giúp làm rõ cách các mô hình học sâu đưa ra quyết định, từ đó hỗ trợ việc đánh giá và kiểm soát quá trình hoạt động của AI.

Tiếp theo, hệ thống sẽ thu thập và tiền xử lý dữ liệu ảnh để chuẩn bị cho quá trình huấn luyện mô hình. Sau đó, một mô hình học sâu, chẳng hạn như CNN (Convolutional Neural Network), sẽ được xây dựng và huấn luyện để thực hiện một nhiệm vụ cụ thể như phân loại hình ảnh hoặc nhận diện đối tượng.

Sau khi mô hình được huấn luyện, các phương pháp XAI sẽ được áp dụng để phân tích và giải thích lý do đằng sau các dự đoán của mô hình. Điều này giúp phát hiện sai sót, giảm thiểu thiên lệch và tăng cường sự tin cậy của hệ thống AI. Ngoài ra, dự án sẽ so sánh và đánh giá hiệu quả của từng phương pháp XAI dựa trên các tiêu chí như mức độ minh bạch, tính trực quan và khả năng ứng dụng trong thực tế.

Cuối cùng, các kết quả nghiên cứu sẽ được tích hợp vào một hệ thống thử nghiệm nhằm minh họa cách XAI có thể hỗ trợ việc giải thích mô hình học sâu trên dữ liệu ảnh. Hy vọng rằng dự án này sẽ đóng góp vào việc nâng cao hiểu biết về Explainable AI, thúc đẩy tính minh bạch trong AI và mở rộng khả năng ứng dụng của học sâu trong nhiều lĩnh vực quan trọng.

## Tầm quan trọng của XAI, các thách thức khi áp dụng

### Explainable AI (XAI) là gì và tại sao quan trọng?

Explainable AI (XAI) là một lĩnh vực trong trí tuệ nhân tạo (AI) nhằm phát triển các mô hình học máy (ML) có khả năng giải thích quyết định của mình theo cách dễ hiểu đối với con người. Cốt lõi của XAI là khả năng minh bạch, giúp người dùng không chỉ nhận được kết quả mà còn hiểu được lý do đằng sau các dự đoán của mô hình. Điều này rất quan trọng vì:

* Giúp cải thiện chất lượng và hiệu suất của mô hình.
* Xây dựng sự tin tưởng của người dùng đối với AI.
* Cho phép phát hiện và điều chỉnh khi mô hình dự đoán không chính xác.
* Hỗ trợ việc kiểm tra, xác minh tính hợp lệ của mô hình trong các lĩnh vực quan trọng như y tế, tài chính và pháp lý.

Trong khi các mô hình AI ngày càng phức tạp và mạnh mẽ, sự thiếu khả năng giải thích có thể dẫn đến những thách thức lớn trong ứng dụng thực tế. Người dùng cần hiểu rõ lý do đằng sau các quyết định của AI, đặc biệt khi các kết quả này có thể ảnh hưởng đến cuộc sống con người.

### Các thách thức của XAI

Mặc dù XAI mang lại nhiều lợi ích, nhưng việc áp dụng vào thực tế vẫn đối mặt với nhiều thách thức lớn. Một số vấn đề quan trọng bao gồm:

* Tính đúng đắn về mặt ngữ nghĩa
  + Lời giải thích cần dễ hiểu đối với con người, không chỉ hợp lý về mặt toán học.
  + Một mô hình có thể cung cấp thông tin chính xác về cách nó đưa ra dự đoán, nhưng nếu lời giải thích không đủ rõ ràng, người dùng vẫn khó hiểu và không thể đưa ra quyết định đúng đắn dựa trên kết quả AI.
* Kết hợp các kỹ thuật giải thích
  + Các phương pháp XAI như LIME, SHAP, Grad-CAM hoạt động theo các nguyên tắc khác nhau.
  + Cần có cách tiếp cận để kết hợp các kỹ thuật này nhằm mang lại sự hiểu biết toàn diện, thay vì gây nhầm lẫn cho người sử dụng.
* Khám phá và thăm dò
  + Người dùng cần công cụ trực quan để kiểm tra và điều chỉnh lời giải thích từ AI.
  + Ví dụ, với dữ liệu hình ảnh, nếu AI sử dụng Gradient tích hợp để làm nổi bật các pixel quan trọng, người dùng cần biết liệu việc thay đổi hoặc loại bỏ các pixel này có ảnh hưởng đến kết quả hay không.
* Hiệu quả tính toán
  + Nhiều phương pháp XAI đòi hỏi lượng tính toán lớn, có thể làm giảm hiệu suất hệ thống.
  + Điều này đặc biệt quan trọng với các mô hình phức tạp hoặc hệ thống xử lý dữ liệu lớn như nhận diện hình ảnh, xử lý ngôn ngữ tự nhiên (NLP) và xe tự hành.
* Tính mạnh mẽ của các kỹ thuật giải thích
  + Cần có các tiêu chí đánh giá mức độ tin cậy của các phương pháp XAI.
  + Một số nghiên cứu đã chỉ ra rằng nhiều kỹ thuật XAI có thể bị tấn công hoặc bị thao túng, làm giảm tính chính xác và độ tin cậy của các lời giải thích.

### Ứng dụng thực tế của XAI

* XAI có nhiều ứng dụng quan trọng trong thực tế, giúp cải thiện tính minh bạch và đáng tin cậy của AI trong nhiều lĩnh vực khác nhau:
* Y tế (Healthcare)
  + Giải thích các mô hình AI trong chẩn đoán bệnh, phát hiện ung thư, phân tích hình ảnh y khoa.
  + Giúp bác sĩ hiểu rõ hơn về quyết định của mô hình, từ đó có thể tin tưởng hơn vào các hệ thống hỗ trợ ra quyết định.
  + Hỗ trợ cá nhân hóa phác đồ điều trị cho bệnh nhân dựa trên dữ liệu sức khỏe.
* Tài chính (Finance)
  + Hỗ trợ phát hiện gian lận, đánh giá tín dụng và dự báo thị trường.
  + Giúp các ngân hàng và tổ chức tài chính tuân thủ các quy định pháp lý bằng cách cung cấp các lời giải thích rõ ràng cho khách hàng về lý do từ chối khoản vay hoặc quyết định giao dịch.
  + Tăng cường niềm tin của khách hàng vào các hệ thống tài chính sử dụng AI.
* Xe tự hành (Autonomous Vehicles)
  + Giải thích hành vi của hệ thống lái tự động, chẳng hạn như lý do chọn đường đi, tránh va chạm hoặc dừng xe.
  + Tăng cường an toàn và sự tin tưởng của người dùng đối với công nghệ xe tự lái.
  + Hỗ trợ trong việc phát triển các hệ thống điều khiển xe đáng tin cậy hơn.
* An ninh mạng (Cybersecurity)
  + Giải thích các mô hình phát hiện tấn công mạng hoặc phần mềm độc hại.
  + Giúp các chuyên gia bảo mật hiểu rõ hơn về các mối đe dọa và phản ứng kịp thời để ngăn chặn các cuộc tấn công.
  + Nâng cao khả năng bảo vệ hệ thống trước các cuộc tấn công tinh vi.
* Các lĩnh vực liên quan khác (Other Relevant Domains)
  + Ngành pháp lý: Giải thích các mô hình dự đoán khả năng tái phạm hoặc phân tích tài liệu pháp lý, giúp đưa ra quyết định công bằng và minh bạch hơn.
  + Thương mại điện tử: Đề xuất sản phẩm dựa trên các yếu tố dễ hiểu với người dùng, cải thiện trải nghiệm mua sắm và tăng hiệu quả kinh doanh.

### Hướng phát triển của XAI

Mặc dù XAI đang đạt được nhiều tiến bộ, nhưng vẫn chưa có một phương pháp nào được chấp nhận rộng rãi có thể áp dụng trong mọi tình huống. Do đó, việc phát triển các kỹ thuật XAI không chỉ dừng lại ở việc tạo ra những lời giải thích hợp lý, mà còn cần đảm bảo chúng thực sự minh bạch, dễ hiểu và có thể áp dụng vào thực tế. Điều này đặc biệt quan trọng trong các lĩnh vực nhạy cảm như y tế, tài chính và an ninh, nơi mà tính minh bạch và sự đáng tin cậy là yếu tố cốt lõi.

# CƠ SỞ LÝ THUYẾT

## Tổng quan về “Explainability”

“I’m increasingly inclined to think that there should be some regulatory oversight, maybe at the national and international level, just to make sure that we don’t do something very foolish. I mean with artificial intelligence we’re summoning the demon”

**Elon Musk**

Explainability (tạm dịch: khả năng có thể giải thích được) luôn là 1 phần của các model machine learning/deep learning từ lúc chúng được tạo ra. Khi nói về khả năng Explainability, ta nghĩ ngay tới việc làm sao các model có thể giải thích được lý do tại sao 1 dự đoán được đưa ra, feature nào có ảnh hưởng lớn nhất và vô số các câu hỏi khác nữa để trả lời câu hỏi “WHY” được đưa ra bởi end-users. Một số câu hỏi mang tính chất Explainability bao gồm:

- Model của tôi thực chất hoạt động như thế nào?

- Tại sao model của tôi lại đưa ra 1 dự đoán như vậy?

- Đâu là feature quan trọng và ít quan trọng nhất của model?

- Giới hạn và bias của tôi bao gồm những gì?

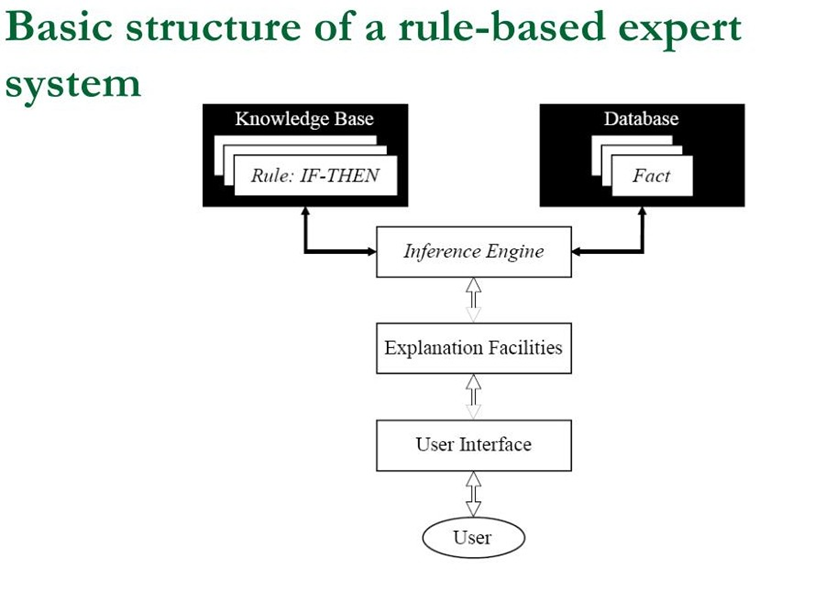
- Làm sao tôi có thể tin tưởng vào quyết định của model

- Liệu model có thể đưa ra lý giải (reasoning) đằng sau đó khộng?

Hình A diagram of a cat and a cat

AI-generated content may be incorrect. Câu hỏi mang tính chất “Explainability”

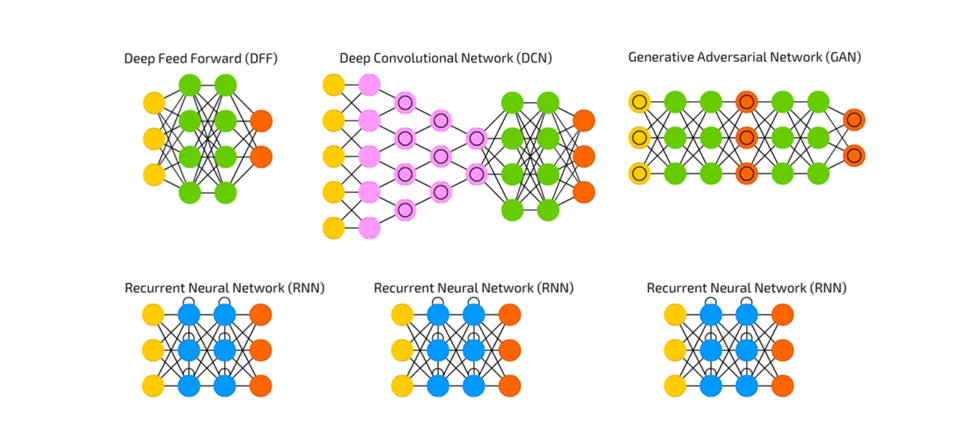
Phía trên chỉ là một số ít câu hỏi mang tính chất Explainability có thể liệt kê ra. Đối với mỗi dạng end-users khác nhau, các câu hỏi còn rất nhiều và đa dạng tùy thuộc vào trách nhiệm và domain của họ. Một trong những phương pháp để hỗ trợ Explainability xuất hiện từ rất lâu đời bao gồm: rule-based chain systems và decision-tree-system (DTS). Đối vối DTS, các prediction của chúng ta được đưa ra dựa theo threshold và một số rules được extract từ model sau khi train. Với rule-based chain systems, ta có thể xem đây là một phiên bản cải tiến hơn của DTS, nó cho phép các rules được flexible hơn và cho phép end-users có thể diễn giải (interpretable) khi đọc.



Hình 2 Cấu trúc cơ bản của một hệ thống

Các kỹ thuật giải thích mô hình (XAI) tiếp tục được phát triển nhằm hỗ trợ các nhà nghiên cứu và người dùng cuối trong việc hiểu và tối ưu hóa các mô hình AI. Tuy nhiên, sự ra đời của mạng nơ-ron nhân tạo (Neural Network) vào đầu những năm 1980 đã tạo ra một bước ngoặt lớn, làm thay đổi hoàn toàn cách tiếp cận trong lĩnh vực AI. Sự phát triển vượt bậc của mạng nơ-ron sâu (Deep Neural Networks - DNNs) đã giúp các mô hình AI đạt được độ chính xác cao trong nhiều bài toán phức tạp, nhưng đồng thời cũng khiến việc giải thích cách chúng đưa ra quyết định trở nên khó khăn hơn.

Bên cạnh đó, các chu kỳ "AI Winter" (thời kỳ suy giảm đầu tư và quan tâm vào AI) trong quá khứ đã làm chậm quá trình nghiên cứu về Explainability. Khi các mô hình AI chưa thực sự mang lại hiệu quả rõ ràng hoặc gặp phải những rào cản kỹ thuật, nguồn lực dành cho việc phát triển các kỹ thuật giải thích bị suy giảm đáng kể. Điều này khiến Explainable AI bị tụt hậu so với sự tiến bộ nhanh chóng của các mô hình học sâu.

Mãi đến những năm gần đây, với sự bùng nổ của AI trong các lĩnh vực quan trọng như y tế, tài chính, an ninh mạng và xe tự hành, nhu cầu về tính minh bạch và giải thích của AI mới thực sự trở thành một vấn đề cấp thiết. Các tổ chức, doanh nghiệp và cơ quan quản lý ngày càng yêu cầu các mô hình AI không chỉ chính xác mà còn phải có khả năng giải thích rõ ràng để đảm bảo tính công bằng, trách nhiệm và khả năng kiểm tra. Điều này đã thúc đẩy sự hồi sinh của Explainable AI, với nhiều phương pháp và công cụ mới được phát triển để giúp con người hiểu rõ hơn về cách AI ra quyết định.

Hình 3 Các kỹ thuật giải thích mô hình (XAI)

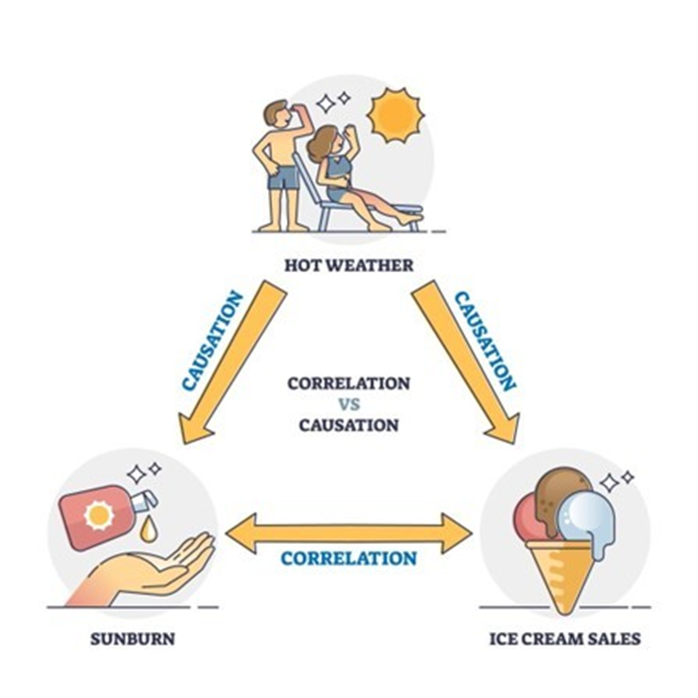
## Như thế nào là “Explainability”

Trong vòng một thập kỷ trở lại đây, các nghiên cứu về Explainability mới thực sự trở nên phổ biến trở lại, phần lớn do những lo ngại ngày càng tăng về AI cũng như nhu cầu từ phía người dùng cuối (end-users) trong việc hiểu rõ tại sao một mô hình lại đưa ra quyết định như vậy. Chỉ trong vài năm qua, nhiều kỹ thuật Explainability đã được đề xuất nhằm giải quyết vấn đề này. Tuy nhiên, tốc độ phát triển nhanh chóng không đồng nghĩa với việc lĩnh vực nghiên cứu này đã hoàn thiện. Ngược lại, sự đa dạng của các khái niệm và phương pháp đôi khi khiến người dùng cuối cảm thấy bối rối. Nhiều định nghĩa vẫn còn mơ hồ và chưa được chuẩn hóa rõ ràng. Một ví dụ tiêu biểu cho điều này là sự khác biệt giữa hai thuật ngữ “Explainability” và “Interpretability” – một vấn đề mà chúng ta sẽ tìm hiểu kỹ hơn ở phần sau.

Khi một mô hình AI thực hiện dự đoán, các phương pháp XAI sẽ cung cấp những thông tin chi tiết (insights) giúp giải thích hành vi của mô hình và lý do đằng sau các dự đoán đó. Hiểu một cách đơn giản, Explainability tập trung vào việc trả lời các câu hỏi mang tính “WHY” – “Tại sao mô hình lại đưa ra kết quả như vậy?”. Chẳng hạn, giả sử chúng ta có một mô hình phát hiện vật thể (object detection) và nó được sử dụng để nhận diện con người trong một bức ảnh. Nếu mô hình có thể giải thích rằng nó xác định đối tượng là con người dựa trên các đặc điểm như có hai chân, hai tay, khuôn mặt và cổ dài, thì đó chính là khả năng Explainability mà chúng ta đang đề cập. Nghiên cứu về Explainability giúp chúng ta hiểu được yếu tố nào ảnh hưởng nhiều nhất đến dự đoán của mô hình, cơ chế hoạt động bên trong của mô hình và cách mà dữ liệu đầu vào tác động đến kết quả đầu ra.

Tuy nhiên, việc đưa ra những lời giải thích đơn giản theo dạng “X dẫn đến Y” vẫn chưa đủ để giúp người dùng hiểu rõ về mô hình. Do đó, chúng ta cần các phương pháp giải thích nâng cao hơn, chẳng hạn như Counterfactual Explanations. Một lời giải thích theo hướng Counterfactual không chỉ trả lời câu hỏi “Những yếu tố nào đã dẫn đến kết quả này?” mà còn giải đáp “Những yếu tố nào, nếu thay đổi, có thể khiến kết quả khác đi?”. Ví dụ, khi một mô hình dự đoán thời tiết hôm nay có mưa hay không, không chỉ cần biết yếu tố nào góp phần gây mưa (như độ ẩm cao, nhiệt độ thấp), mà còn phải xác định những yếu tố nào không ảnh hưởng đến mưa (chẳng hạn như tốc độ gió nhẹ hoặc áp suất khí quyển trung bình).

Trong thực tế, chúng ta thường tìm kiếm các dạng giải thích mang tính nhân quả (causal relationship), vì chúng giúp lý giải trực tiếp về cách một mô hình hoạt động. Nếu có X thì sẽ có Y, còn nếu bỏ X đi thì Y sẽ không xảy ra. Tuy nhiên, việc xác định các mối quan hệ nhân quả trong mô hình AI là một thách thức lớn, bởi vì một sự kiện có thể bị ảnh hưởng bởi nhiều yếu tố khác nhau. Không phải tất cả các yếu tố có tương quan với kết quả đều mang ý nghĩa nhân quả. Một số chỉ đơn thuần là mối tương quan (correlation relationship) mà không thực sự có tác động đến dự đoán. Do đó, một phần quan trọng của Explainability là phân biệt giữa các yếu tố có tác động thực sự và các yếu tố chỉ xuất hiện cùng với kết quả nhưng không phải nguyên nhân của nó.



Hình Thế nào là Explainability

## So sánh giữa “Explainability” và “Interpretability”

Một trong những câu hỏi phổ biến nhất khi nghiên cứu về XAI chính là sự khác biệt giữa hai khái niệm “Interpretability” và “Explainability”. Để phân biệt rõ ràng hai thuật ngữ này, trước hết chúng ta cần hiểu bản chất của từng khái niệm.

### Interpretable Model (Mô hình có thể diễn giải trực tiếp)

Khi một mô hình được coi là interpretable, điều đó có nghĩa là chúng ta có thể hiểu được cách nó đưa ra quyết định mà không cần bất kỳ công cụ hỗ trợ nào. Nói cách khác, người dùng có thể trực tiếp quan sát các tham số hoặc cấu trúc của mô hình và từ đó suy luận được cách nó hoạt động. Những mô hình thuộc nhóm này thường được gọi là "intrinsically interpretable models" – tức là mô hình có tính diễn giải một cách tự nhiên.

**"Interpretable models can be understood by a human without any other aids/techniques."**

Một ví dụ tiêu biểu của mô hình có tính interpretable là **Decision Tree (Cây quyết định - DST)**. Giả sử ta có một mô hình dự đoán nguy cơ vỡ nợ dựa trên hai yếu tố: **tuổi** và **thu nhập**. Khi nhìn vào cấu trúc của cây quyết định, chúng ta có thể ngay lập tức hiểu được quy trình mà mô hình sử dụng để đưa ra dự đoán, mà không cần sự hỗ trợ của bất kỳ công cụ phân tích nào khác.

Tương tự, trong trường hợp của **Hồi quy tuyến tính (Linear Regression)**, ta có thể dễ dàng hiểu được cách dự đoán được thực hiện chỉ bằng cách quan sát phương trình của mô hình. Ví dụ:

**Y = 100\*age + 10\*income + 200**

Dựa vào các hệ số (coefficients) trong phương trình, ta có thể đánh giá độ quan trọng của từng biến đầu vào đối với kết quả dự đoán. Trong ví dụ trên, yếu tố **tuổi** có trọng số cao hơn so với **thu nhập**, nghĩa là nó có ảnh hưởng lớn hơn đến dự đoán cuối cùng.

Nhờ vào tính trực quan và dễ hiểu, các mô hình interpretable thường được ưu tiên trong các lĩnh vực đòi hỏi tính minh bạch cao, chẳng hạn như y tế, tài chính và pháp lý.

A diagram of a flowchart

AI-generated content may be incorrect.

Hình Interpretable Model

### Explainable Model (Mô hình cần được giải thích)

Chúng ta có thể hình dung một mô hình học máy (ML) như một hàm số, trong đó các đặc trưng (features) đóng vai trò là đầu vào và dự đoán là đầu ra. Tuy nhiên, không phải mô hình nào cũng có thể dễ dàng hiểu được bằng cách trực tiếp quan sát nó.

Một Explainable Model là một mô hình có độ phức tạp cao đến mức con người không thể trực tiếp hiểu được cách nó đưa ra quyết định. Những mô hình này còn được gọi là black-box models (hộp đen), vì quá trình dự đoán của chúng không rõ ràng với người dùng. Để hiểu và giải thích các mô hình này, chúng ta cần áp dụng các phương pháp hoặc kỹ thuật chuyên biệt.

**"Explainable models require additional techniques to be understood by humans."**

**Ví dụ về Explainable Model:**

Một ví dụ điển hình của mô hình cần được giải thích là **Random Forest (RF)**. RF được xây dựng từ nhiều **Decision Trees (Cây quyết định - DST)**, trong đó mỗi cây đóng góp vào quá trình đưa ra dự đoán cuối cùng. Để hiểu cách RF hoạt động, ta cần hiểu toàn bộ tập hợp các cây quyết định bên trong nó. Tuy nhiên, ngay cả khi chỉ có một số lượng nhỏ DST, việc này vẫn gần như không thể thực hiện đối với con người.

Mọi thứ trở nên phức tạp hơn khi xét đến các thuật toán thuộc nhóm **Neural Networks (Mạng nơ-ron nhân tạo)**. Ví dụ điển hình là **AlexNet**, một mạng nơ-ron tích chập (CNN) được sử dụng để nhận dạng hình ảnh. Mô hình này chứa đến **62.378.344 tham số**, khiến cho việc trực tiếp lý giải cách nó đưa ra quyết định gần như không thể đối với con người.

Vì vậy, trong trường hợp của các **black-box models**, việc sử dụng các kỹ thuật Explainable AI (XAI) là vô cùng cần thiết để giúp chúng ta hiểu rõ hơn về cách thức hoạt động của chúng, từ đó đảm bảo tính minh bạch và đáng tin cậy trong quá trình áp dụng AI vào thực tế.

A diagram of different types of blocks

AI-generated content may be incorrect.

Hình Explainable model

### So sánh giữa Interpretable Model và Explainable Model

Để hiểu rõ sự khác biệt giữa hai loại mô hình này, ta có thể xem xét sự khác nhau về kiến trúc và cách tiếp cận.

Interpretable Models có thiết kế đơn giản và dễ hiểu, giúp người dùng có thể lý giải trực tiếp cách chúng hoạt động mà không cần công cụ hỗ trợ. Tuy nhiên, do tính minh bạch dựa vào chính kiến trúc của mô hình, khả năng giải thích của chúng chỉ giới hạn trong phạm vi model cụ thể.

A blue line with black text

AI-generated content may be incorrect.Explainable Models, hay còn gọi là các mô hình "hộp đen" (black-box models), có kiến trúc phức tạp khiến việc hiểu và lý giải kết quả trở nên khó khăn. Để giải thích cách các mô hình này đưa ra dự đoán, ta cần đến các phương pháp của Explainable AI (XAI). Một ưu điểm của XAI là tính Model-agnostic, tức là nó có thể áp dụng để giải thích nhiều loại mô hình khác nhau mà không phụ thuộc vào kiến trúc cụ thể.

Hình So sánh giữa Interpretable Model và Explainable Model

### Explainability với từng loại người dùng

Explainability là một khái niệm rộng, và cách hiểu cũng như nhu cầu sử dụng của nó phụ thuộc vào loại người dùng cụ thể. Một Data Scientist phát triển mô hình học sâu sẽ cần một cách lý giải rất khác so với một người dùng không chuyên về công nghệ.

Hiện tại, phần lớn các phương pháp XAI tập trung vào hỗ trợ technical users những người có nền tảng công nghệ và trực tiếp làm việc với mô hình. Trong khi đó, đối với non-technical users, việc giải thích AI vẫn là một thách thức lớn, cần thêm nhiều nghiên cứu và phát triển để giúp họ hiểu được cách AI đưa ra quyết định.

Dưới đây là ba nhóm người dùng chính có nhu cầu về Explainability và cách họ sử dụng nó:

**Practitioners – Data Scientists và ML Engineers**

* Data Scientists sử dụng Explainability để phân tích và tinh chỉnh mô hình trong quá trình phát triển. Họ quan tâm đến cách tối ưu hóa hiệu suất, điều chỉnh kiến trúc mô hình, cải thiện dữ liệu huấn luyện và tối ưu hóa schema của tập dữ liệu. Mục tiêu chính của họ là giảm loss và kiểm tra độ chính xác của mô hình trước khi triển khai.
* ML Engineers quan tâm nhiều hơn đến hiệu suất của data pipeline. Họ cần hiểu tính toán nào tốn nhiều tài nguyên nhất hoặc dữ liệu nào có chi phí thu thập cao nhất. Nhìn chung, Practitioners tập trung vào việc xây dựng, phát triển và tối ưu hóa mô hình, do đó họ cần Explainability để hiểu rõ cách hệ thống vận hành.

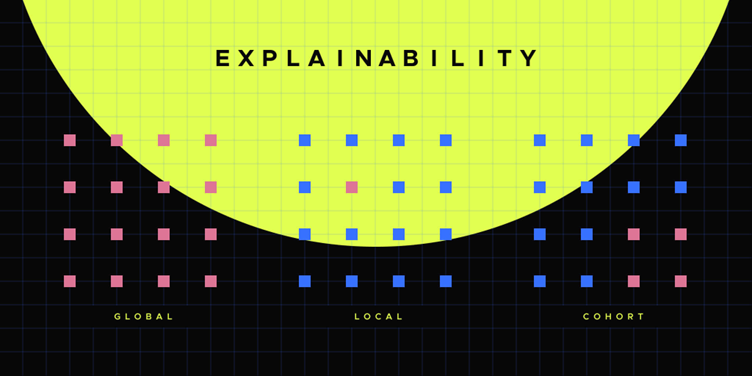
**Observers – Business Stakeholders và Regulators**

* Đây là nhóm người không trực tiếp tham gia vào quá trình phát triển mô hình, nhưng có vai trò quan trọng trong khía cạnh kinh doanh hoặc quản lý.
* Stakeholders (các bên liên quan trong doanh nghiệp) có thể bao gồm nhà đầu tư, ban giám đốc hoặc những người chịu trách nhiệm thương mại hóa mô hình. Họ cần các giải thích đơn giản, dễ hiểu, nhấn mạnh vào tác động kinh doanh của mô hình thay vì các chi tiết kỹ thuật.
* Regulators (cơ quan quản lý bên ngoài) quan tâm đến tính đạo đức, tính công bằng, độ thiên lệch (bias), tính minh bạch và khả năng kiểm toán của mô hình. Họ đánh giá mô hình không chỉ dựa trên hiệu suất mà còn trên các tiêu chí liên quan đến pháp lý và xã hội.

**End Users – Domain Experts và Affected Users**

* Domain Experts là những chuyên gia trong lĩnh vực mà AI được áp dụng, ví dụ như bác sĩ sử dụng AI để chẩn đoán bệnh. Họ có chuyên môn trong lĩnh vực của mình nhưng không nhất thiết phải hiểu sâu về AI. Họ cần Explainability để tin tưởng vào kết quả mà AI đưa ra và hỗ trợ quyết định của họ.
* Affected Users là những người chịu tác động trực tiếp từ dự đoán của AI, chẳng hạn như sinh viên bị từ chối khoản vay do mô hình dự đoán khả năng vỡ nợ cao. Những người dùng này cần đảm bảo rằng mô hình đưa ra quyết định một cách công bằng, minh bạch và không thiên lệch.

## Các phương pháp giải thích trong Explainable AI (XAI)



Hình 8 Các phương pháp giải thích Explanation

Trong XAI, có nhiều phương pháp giải thích khác nhau để giúp người dùng hiểu được cách mô hình đưa ra dự đoán. Những phương pháp này có thể chia thành ba nhóm chính:

* **Global Explanation (GE)** – Giải thích toàn cục
* **Local Explanation (LE)** – Giải thích cục bộ
* **Cohort Explanation (CE)** – Giải thích theo nhóm dữ liệu

Ngoài ra, còn có một số phương pháp khác như **Attribution Explanation (AE)**, **Example-Based Explanations (EBE)** và **Counterfactual Explanations (CE)**.

### Global Explanation (GE) – Giải thích toàn cục

* GE giúp ta hiểu cách mô hình hoạt động trên toàn bộ tập dữ liệu. Điều này có nghĩa là ta có thể xác định được trọng số và mức độ ảnh hưởng của từng feature lên tổng thể dự đoán của mô hình.
* Từ khóa quan trọng: "toàn bộ" – GE không phân tích một dự đoán cụ thể mà đánh giá mô hình trên phạm vi rộng.

Ứng dụng:

* Các phòng ban như marketing có thể sử dụng GE để xác định feature quan trọng nhất nhằm phân khúc khách hàng hiệu quả.
* Nhà quản lý có thể điều chỉnh chiến lược kinh doanh dựa trên insight từ GE.

A screenshot of a graph

AI-generated content may be incorrect.

Hình Global Explanation (GE)

### Local Explanation (LE) – Giải thích cục bộ

* Ngược lại với GE, LE tập trung vào **một dự đoán cụ thể** và tìm hiểu xem feature nào ảnh hưởng mạnh nhất đến quyết định của mô hình tại thời điểm đó.

**Ví dụ:**

* Một khách hàng bị từ chối khoản vay và muốn biết lý do. LE có thể giúp xác định rằng yếu tố “thu nhập thấp” hoặc “lịch sử tín dụng xấu” là nguyên nhân chính.

**Ứng dụng:**

* Dùng để debug mô hình trong môi trường production.
* Giải thích lý do tại sao một dự đoán cụ thể được đưa ra, giúp tăng tính minh bạch với người dùng cuối.

A graph with red lines and numbers

AI-generated content may be incorrect.

Hình Local Explanation (LE)

### Cohort Explanation (CE) – Giải thích theo nhóm dữ liệu

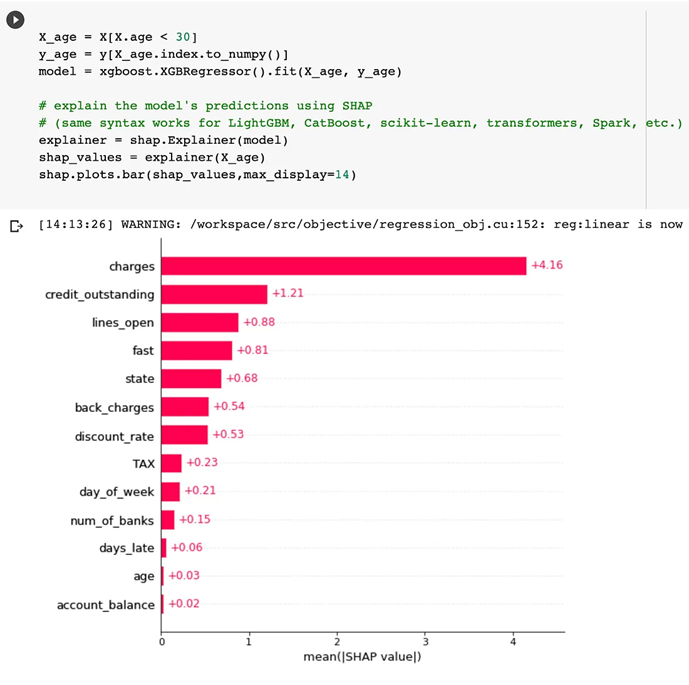
CE không xem xét toàn bộ dữ liệu như GE hay một điểm dữ liệu cụ thể như LE, mà tập trung vào **một tập con (subset) của dữ liệu**.

**Ví dụ:**

* Nếu ta muốn phân tích những khách hàng **dưới 30 tuổi**, CE sẽ giúp tìm ra feature nào quan trọng nhất đối với nhóm này.

**Ứng dụng:**

* Kiểm tra mô hình hoạt động tốt hay kém trên từng nhóm khách hàng.
* Phát hiện bias hoặc các hạn chế của mô hình trong một phân khúc nhất định.



Hình Cohort Explanation (CE)

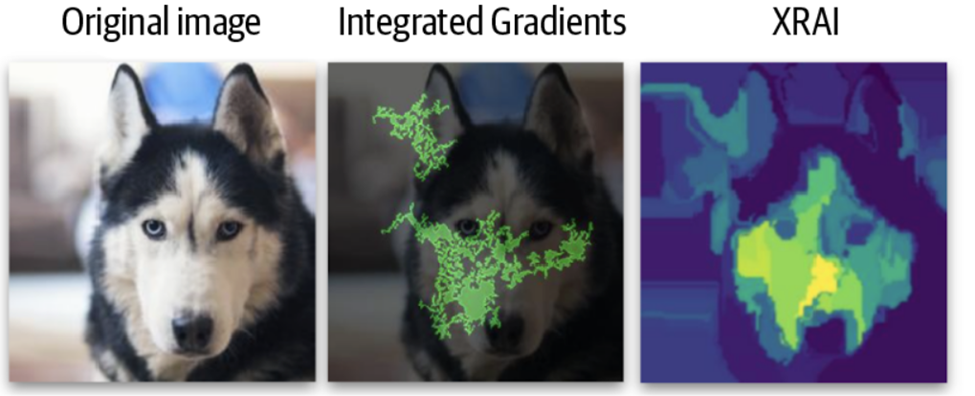
### Attribution Explanation (AE) – Giải thích dựa trên độ quan trọng của feature

AE là một trong những phương pháp phổ biến nhất, đặc biệt hữu ích cho các technical users.

**Cách hoạt động:**

* AE tập trung làm nổi bật các **feature quan trọng nhất** trong một dự đoán.
* Trong bài toán **object detection**, AE có thể hiển thị vùng ảnh quan trọng nhất để nhận diện đối tượng.
* Trong **regression/classification**, AE liệt kê danh sách feature có trọng số cao nhất từ trên xuống dưới.

**Ứng dụng:**

* Hỗ trợ Data Scientists và ML Engineers hiểu rõ hơn về cách mô hình hoạt động.
* Kiểm tra xem mô hình có đang sử dụng đúng feature quan trọng hay không.

Hình Attribution Explanation (AE)

### Example-Based Explanations (EBE) – Giải thích dựa trên ví dụ

Không giống AE, phương pháp EBE không tập trung vào độ quan trọng của từng feature mà thay vào đó, nó tìm các dữ liệu có đặc điểm tương tự để hỗ trợ giải thích.

**Ví dụ:**

* Nếu một mô hình AI chẩn đoán bệnh cho một bệnh nhân, EBE sẽ tìm những bệnh nhân có hồ sơ tương tự để hỗ trợ giải thích quyết định.
* EBE có thể sử dụng thuật toán K-Nearest-Neighbors (KNN) để tìm các điểm dữ liệu "hàng xóm" gần nhất.

**Ứng dụng:**

* Hữu ích trong bài toán nhận dạng hình ảnh, phân loại văn bản hoặc chẩn đoán y khoa.
* Giúp người dùng hiểu mô hình bằng cách so sánh với các trường hợp tương tự.

Hình A collage of dogs

AI-generated content may be incorrect. Example-Based Explanations (EBE)

### Counterfactual Explanations – Giải thích phản thực tế

Phương pháp này không chỉ tập trung vào "tại sao một dự đoán được đưa ra" mà còn xem xét **"điều gì cần thay đổi để có một kết quả khác"**.

**Ví dụ:**

* Hiện tại: Một bệnh nhân có mức Glucose = 195, BMI = 22, Exercise = 19 phút, Calories = 2000 bị chẩn đoán là Diabetic.
* Phản thực tế: Nếu Calories giảm từ 2000 xuống 1500, mô hình dự đoán người này không bị tiểu đường (Not Diabetic).

**Ứng dụng:**

* Dễ hiểu hơn đối với **non-technical users**, giúp họ biết cách cải thiện kết quả.
* Hữu ích trong các hệ thống ra quyết định như tài chính, y tế và tuyển dụng.

A screenshot of a computer

AI-generated content may be incorrect.

Hình Counterfactual Explanations

## Các Kỹ Thuật Model-Agnostic XAI

### Dữ Liệu Dạng Bảng

#### Premodeling Explainability (EDA) – Giải Thích Trước Khi Huấn Luyện Mô Hình

A diagram of stages of modeling

Description automatically generated

Hình 15 Premodeling Explainability (EDA)

Premodeling Explainability, hay còn gọi là phân tích dữ liệu khám phá (Exploratory Data Analysis - EDA), là bước đầu tiên và không thể thiếu trong quá trình phát triển mô hình máy học. Ở giai đoạn này, thay vì tập trung vào kết quả đầu ra của mô hình, chúng ta phân tích dữ liệu đầu vào để hiểu rõ hơn về tập dữ liệu, qua đó cải thiện chất lượng và hiệu suất mô hình.

**Mục tiêu của EDA:**

* Phát hiện vấn đề dữ liệu: Tìm ra giá trị thiếu, ngoại lệ, hoặc các đặc điểm không mong muốn trong tập dữ liệu.
* Phân tích quan hệ giữa các đặc trưng: Xem xét mối quan hệ giữa từng đặc trưng với nhau cũng như với biến mục tiêu bằng thống kê đơn biến và đa biến.
* Nghiên cứu phân phối dữ liệu: Xác định cách các đặc trưng tác động đến kết quả dự đoán, từ đó điều chỉnh mô hình phù hợp.

**Các phương pháp chính trong EDA:**

* Phân tích dữ liệu khám phá (Exploratory Data Analysis - EDA): Sử dụng các kỹ thuật thống kê và trực quan hóa để hiểu rõ tập dữ liệu.
* Chuẩn hóa mô tả tập dữ liệu (Dataset Description Standardization): Định nghĩa các tiêu chuẩn nhất quán giúp mô tả dữ liệu một cách dễ hiểu.
* Kỹ thuật tạo đặc trưng có thể giải thích (Explainable Feature Engineering): Tạo ra các đặc trưng mới giúp mô hình dễ hiểu và dễ diễn giải hơn.
* Phương pháp tóm tắt tập dữ liệu (Dataset Summarization Methods): Tạo ra các biểu diễn cô đọng, giúp người dùng nắm bắt nhanh các thông tin quan trọng.

**Lợi ích của Premodeling Explainability:**

* Giúp xây dựng nền tảng dữ liệu vững chắc trước khi huấn luyện mô hình.
* Hạn chế các vấn đề về bias, dữ liệu sai lệch, hoặc dữ liệu không đầy đủ.
* Giúp các nhà khoa học dữ liệu (Data Scientists) hiểu rõ hơn về tập dữ liệu, từ đó tối ưu hóa mô hình hiệu quả hơn.

#### Partial Dependence Plots (PDPs) – Biểu Đồ Phụ Thuộc Một Phần

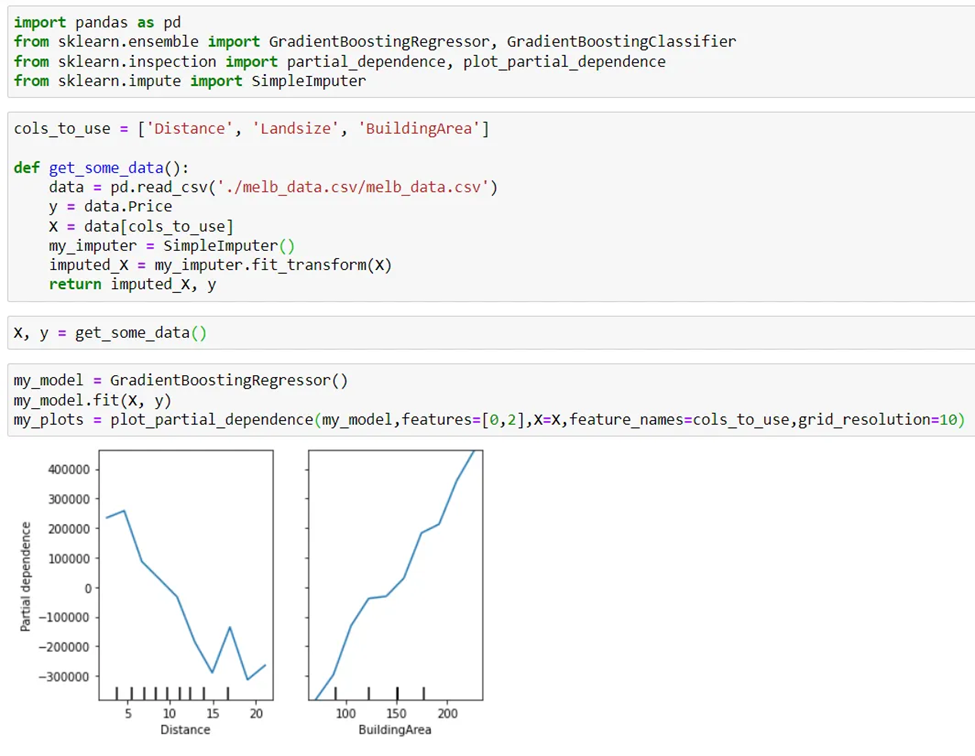
Partial Dependence Plots (PDPs) là một công cụ trong XAI giúp trực quan hóa mức độ ảnh hưởng của một đặc trưng đến kết quả dự đoán của mô hình. PDP thể hiện cách giá trị dự đoán trung bình thay đổi khi một đặc trưng cụ thể thay đổi, trong khi giữ nguyên các đặc trưng khác.

**Ứng dụng của PDPs:**

* Giúp hiểu ảnh hưởng của từng đặc trưng lên dự đoán của mô hình, đặc biệt với các mô hình phức tạp như mạng nơ-ron nhân tạo (Neural Network) hay rừng ngẫu nhiên (Random Forest).
* Phát hiện mối quan hệ tuyến tính hoặc phi tuyến tính giữa đặc trưng và biến mục tiêu.
* Xác định bias hoặc hạn chế của mô hình, giúp phát hiện các đặc trưng có tác động mạnh đến dự đoán.

**Ví dụ minh họa:**

* Giả sử ta có một mô hình dự đoán giá nhà. PDP có thể giúp chúng ta hiểu:
  + Khi diện tích nhà tăng, giá nhà có xu hướng tăng theo.
  + Khi khoảng cách từ nhà đến trung tâm thành phố tăng, giá nhà giảm đáng kể.
  + Ví dụ, nếu một căn nhà cách trung tâm 15km thì giá trung bình có thể giảm khoảng 300.000 USD.

Hình 16 Ví dụ dự đoán giá nhà của PDPs

**Hạn chế của PDPs:**

* PDP giả định rằng các đặc trưng là độc lập (tức là không có hiện tượng đa cộng tuyến), nhưng trong thực tế, nhiều đặc trưng có mối liên hệ với nhau.
  + Ví dụ: tuổi thường có mối tương quan dương với mức lương.
* Khi có sự phụ thuộc giữa các đặc trưng, PDP có thể dẫn đến diễn giải sai lệch hoặc gây hiểu lầm.

**Lưu ý quan trọng khi sử dụng PDPs:**

* Kiểm tra xem các đặc trưng có mối quan hệ với nhau hay không trước khi sử dụng PDP.
* Nếu có đa cộng tuyến, nên kết hợp với các kỹ thuật khác như SHAP hoặc ICE (Individual Conditional Expectation) để có cái nhìn chính xác hơn.

|  |  |  |  |
| --- | --- | --- | --- |
| **Kỹ thuật** | **Mục đích** | **Ứng dụng** | **Hạn chế** |
| Premodeling Explainability (EDA) | Hiểu dữ liệu trước khi huấn luyện mô hình | Phát hiện lỗi dữ liệu, hiểu quan hệ giữa các đặc trưng | Không giúp hiểu cách mô hình ra quyết định |
| Partial Dependence Plots (PDPs) | Xác định ảnh hưởng của từng đặc trưng lên kết quả dự đoán | Giúp phát hiện mối quan hệ tuyến tính và phi tuyến tính | Giả định đặc trưng độc lập, có thể gây hiểu lầm khi có đa cộng tuyến |

Bảng So sánh 2 kỹ thuật (EDA) và (PDPs)

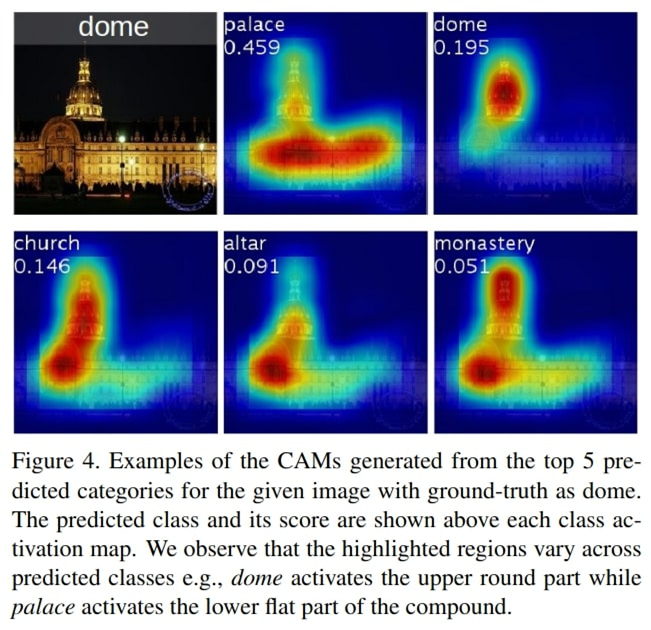
### Dữ Liệu Dạng Ảnh

#### CAM (Class Activation Map) – Bản Đồ Kích Hoạt Lớp

Trước khi tìm hiểu về Grad-CAM, ta cần hiểu về CAM (Class Activation Map) một trong những phương pháp đơn giản nhất giúp trực quan hóa cách mà một mô hình học sâu nhận diện các đối tượng trong ảnh.

**Nguyên lý hoạt động của CAM:**

* CAM sử dụng Global Average Pooling (GAP) để lấy giá trị trung bình của các feature maps từ lớp Convolution cuối cùng.
* Mỗi feature map có thể được coi như một đặc trưng quan trọng đối với một class cụ thể trong bài toán phân loại ảnh.
* Khi dự đoán đầu ra, mỗi class sẽ có một tập trọng số riêng (weight), và CAM sẽ sử dụng các trọng số này để highlight lên những vùng quan trọng trong ảnh – tức là các vùng giúp mô hình nhận diện class đó.



A diagram of a system

AI-generated content may be incorrect.

Theo bức ảnh bên trên, giả sử bài toán phân loại ảnh của ta có 10 class. Khi tới lớp Conv cuối cùng, ta thực hiện GAP để lấy average cho 10 class đó. Mỗi class tương ứng với 1 weight khi thực hiện prediction, ở đây ta có 10 class tương ứng với 10 weight. Cuối cùng ta dùng CAM để highlight lên feature map dựa vào weight tương ứng với từng class.

**Hạn chế của CAM:**

* Chỉ áp dụng được với những mô hình sử dụng GAP trước khi đưa vào lớp dự đoán đầu ra (**feature maps → GAP → output prediction**).
* Không thể dùng cho các mô hình CNN phức tạp hơn (không có GAP).

A drawing of a diagram

AI-generated content may be incorrect.

Hình 17 Sơ đồ hoạt động của CAM

#### Grad-CAM – Phiên Bản Cải Tiến Của CAM

Grad-CAM (Gradient-weighted Class Activation Mapping) là một phương pháp tổng quát hóa của CAM, giúp khắc phục những hạn chế của CAM và có thể áp dụng với hầu hết các mô hình CNN mà không cần thay đổi kiến trúc.

**Cách hoạt động của Grad-CAM:**

* Grad-CAM sử dụng gradient của lớp đầu ra theo feature maps để xác định các vùng quan trọng trong ảnh.
* Dựa vào thông tin gradient này, Grad-CAM có thể tạo ra một heatmap trực quan, giúp highlight lên những vùng có đóng góp cao nhất vào dự đoán của mô hình.
* Không giống CAM, Grad-CAM không yêu cầu sử dụng GAP, nên có thể áp dụng với bất kỳ mô hình CNN nào, miễn là có các lớp khả vi (differentiable layers).

**Tại sao Grad-CAM thường sử dụng lớp Convolution cuối?**

* Grad-CAM có thể áp dụng lên bất kỳ feature maps của bất kỳ layer nào trong mạng. Tuy nhiên, paper gốc của Grad-CAM đã chứng minh rằng:
  + Lớp Convolution cuối thường chứa thông tin quan trọng nhất, giúp mô hình tập trung vào các vùng đặc trưng của ảnh.
  + Vì vậy, sử dụng Grad-CAM tại lớp Conv cuối cùng sẽ giúp mô hình có tính giải thích tốt nhất.

**Đặc Điểm Chính Của Grad-CAM**

* Grad-CAM sử dụng heatmap để highlight các vùng quan trọng trong ảnh, giúp hiểu rõ hơn về cách mô hình đưa ra dự đoán.
* Không yêu cầu thay đổi kiến trúc mạng, có thể áp dụng cho bất kỳ mô hình CNN nào có các lớp khả vi.
* Là một bản cải tiến của CAM, giúp giải thích trực quan cách CNN hoạt động.

**Hạn Chế Của Grad-CAM**

* Có thể highlight sai các vùng không thực sự quan trọng do hiện tượng gradient vanishing (gradient bị mất dần).
* Chưa chính xác 100% trong việc xác định vùng đóng góp chính xác của mô hình vào dự đoán.

**Các phương pháp cải tiến Grad-CAM:**

* HiRes-CAM: Cải thiện độ phân giải của heatmap, giúp highlight chính xác hơn.
* Ablation-CAM: Loại bỏ dần các phần không quan trọng để tăng tính chính xác của heatmap.

A drawing of a group of squares

AI-generated content may be incorrect.

Hình 18 Sơ đồ hoạt động của Grab-CAM

#### LIME (Local Interpretable Model-Agnostic Explanations)

LIME là một kỹ thuật giải thích mô hình dựa trên nhiễu động, không phụ thuộc vào mô hình cụ thể. Nó được sử dụng để giải thích dự đoán của các mô hình học máy, chủ yếu là các mô hình phân loại, nhưng cũng có thể được áp dụng cho các mô hình hồi quy.

LIME hoạt động bằng cách tạo nhiễu các đầu vào xung quanh dự đoán cần giải thích (ví dụ, bật hoặc tắt các vùng của hình ảnh hoặc thay đổi các giá trị của các đặc trưng trong dữ liệu dạng bảng) và đánh giá ảnh hưởng của các thay đổi này đến đầu ra của mô hình. Nó sử dụng một mô hình đơn giản và có thể diễn giải (ví dụ: hồi quy tuyến tính) để đánh giá tầm quan trọng của các tính năng đầu vào đối với kết quả dự đoán cục bộ.

**Chi tiết về cách hoạt động của LIME:**

LIME là một kỹ thuật giải thích hậu kỳ, không phụ thuộc vào mô hình, dựa trên nhiễu động. Điều đó có nghĩa là nó có thể được áp dụng cho bất kỳ mô hình học máy nào (ví dụ: neural networks, SVM, random forest, v.v.) và được áp dụng sau khi mô hình đã được đào tạo. Về bản chất, LIME xử lý mô hình đã được đào tạo như một API, lấy một ví dụ và tạo ra giá trị dự đoán. Để giải thích lý do tại sao mô hình đưa ra một dự đoán nhất định cho một ví dụ đầu vào nhất định, thuật toán LIME hoạt động bằng cách truyền rất nhiều ví dụ nhiễu động nhẹ của đầu vào ban đầu vào mô hình và sau đó đo lường cách các dự đoán của mô hình thay đổi với những thay đổi đầu vào nhỏ này. Các nhiễu động xảy ra ở cấp độ tính năng của ví dụ đầu vào; tức là, hình ảnh, pixel và vùng pixel được sửa đổi để tạo ra một đầu vào nhiễu động mới. Theo cách này, các pixel hoặc vùng pixel ảnh hưởng nhiều nhất đến dự đoán của mô hình được đánh dấu là có ảnh hưởng nhiều hơn hoặc ít hơn đến đầu ra dự đoán của mô hình đối với ví dụ đầu vào nhất định.

Để đi sâu hơn vào chi tiết, chúng ta hãy giải thích thêm hai thành phần chính của việc triển khai LIME: thứ nhất, cách tạo nhiễu loạn cho hình ảnh và thứ hai, ý nghĩa của việc đo lường sự thay đổi trong dự đoán của mô hình dựa trên những nhiễu loạn này.

Đối với một dự đoán nhất định, hình ảnh đầu vào được chia thành các thành phần có thể diễn giải được hoặc các vùng của hình ảnh. LIME phân đoạn hình ảnh thành các vùng được gọi là siêu điểm ảnh. Một siêu điểm ảnhlà một nhóm dựa trên sự tương đồng của các pixel riêng lẻ của một hình ảnh thành các thành phần tương tự.

Ví dụ, Hình 19 cho thấy cách hình ảnh của loài vẹt mào lưu huỳnh có thể được phân đoạn thành các siêu pixel. Các vùng siêu pixel biểu diễn các thành phần có thể diễn giải của hình ảnh.

A white bird with yellow feathers

AI-generated content may be incorrect.

Hình 19 Phân đoạn thành các siêu pixel

Thuật toán Quickshift được sử dụng trong phân đoạn hình ảnh để tạo ra các vùng siêu điểm ảnh, như trong Hình 19, trong quá trình triển khai LIME (Local Interpretable Model-Agnostic Explanations). Thuật toán này áp dụng phép xấp xỉ dịch chuyển trung bình hạt nhân và tính toán phân đoạn phân cấp trên nhiều thang đo. Quickshift có ba tham số quan trọng:

* Sigma: Xác định thang đo của phép xấp xỉ mật độ cục bộ.
* Max\_dist: Xác định mức phân đoạn phân cấp.
* Ratio: Kiểm soát tỷ lệ giữa khoảng cách trong không gian màu và không gian hình ảnh khi so sánh độ tương đồng của hai điểm ảnh.

Thuật toán này được tích hợp trong thư viện scikit-image và là phương pháp phân đoạn mặc định trong LIME.

A white background with black text

AI-generated content may be incorrect.

LIME (Local Interpretable Model-Agnostic Explanations) tạo ra các ví dụ nhiễu loạn của hình ảnh đầu vào bằng cách thay đổi giá trị của các điểm ảnh trong mỗi vùng siêu điểm ảnh thành màu xám. Điều này tạo ra các biến thể của hình ảnh gốc, mỗi biến thể sẽ được đưa vào mô hình đã được đào tạo (ví dụ như mạng nơ-ron sâu) để dự đoán lại kết quả. Các dự đoán này sẽ có xác suất cho từng lớp, như trong ví dụ với con vẹt mào lưu huỳnh trong Hình 19.

Các dự đoán từ những hình ảnh đã bị nhiễu loạn sẽ tạo thành một tập dữ liệu dùng để huấn luyện mô hình tuyến tính của LIME. Mô hình tuyến tính này sẽ phân tích mức độ đóng góp của từng thành phần có thể diễn giải (các vùng siêu điểm ảnh) vào dự đoán ban đầu của mô hình phức tạp. Cách này giúp làm rõ cách thức mà mô hình đã học và đưa ra các quyết định dựa trên các yếu tố cụ thể trong hình ảnh, từ đó cung cấp sự giải thích cho dự đoán của mô hình.

A bird with blue arrows

AI-generated content may be incorrect.

Hình 20 Giá trị trung bình của các giá trị điểm ảnh riêng lẻ

Trong triển khai LIME, các siêu điểm ảnh được bật "bật" hoặc "tắt" bằng cách thay đổi giá trị điểm ảnh của một phân đoạn thành màu xám. Điều này tạo ra một tập hợp các hình ảnh nhiễu loạn được truyền đến mô hình để dự đoán. Thay vào đó, cũng có thể thay đổi các vùng siêu điểm ảnh thành giá trị trung bình của các giá trị điểm ảnh riêng lẻ trong vùng siêu điểm ảnh, như thể hiện trong Hình 20 .

A bird with a yellow beak

AI-generated content may be incorrect.

Hình 21 Kết quả của thuật toán LIME

LIME đo lường định lượng các đóng góp của các đầu vào siêu điểm ảnh bằng cách tạo ra một mô hình nhỏ hơn, có thể diễn giải, nhằm cung cấp giải thích cho mô hình ban đầu. Mô hình nhỏ này thường là mô hình tuyến tính vì tính dễ hiểu và khả năng diễn giải cao của nó: trọng số của mỗi tính năng trong mô hình tuyến tính trực tiếp chỉ ra tầm quan trọng của tính năng đó đối với dự đoán

## Framework

### PyTorch

PyTorch là một thư viện học sâu mã nguồn mở được phát triển bởi Facebook's AI Research Lab (FAIR), nổi tiếng với tính linh hoạt, dễ sử dụng và hiệu suất cao. Được xây dựng dựa trên ngôn ngữ lập trình Python, PyTorch giúp người dùng dễ dàng tạo ra và huấn luyện các mô hình học sâu cho nhiều ứng dụng khác nhau, bao gồm cả hệ thống khuyến nghị.

Trong lĩnh vực hệ thống khuyến nghị, PyTorch cung cấp các công cụ mạnh mẽ để xây dựng và tinh chỉnh các mô hình học sâu như mạng nơ-ron hồi quy (RNNs), mạng nơ-ron tái đạo (LSTMs), mạng nơ-ron tích chập (CNNs), và các mô hình kết hợp khác. Điều này cho phép người phát triển tạo ra các mô hình phức tạp có khả năng học được biểu diễn phức tạp từ dữ liệu người dùng và mặt hàng, từ đó dự đoán sở thích và hành vi cá nhân của họ.

PyTorch cung cấp một loạt các công cụ để huấn luyện mô hình, bao gồm cả việc tối ưu hóa hàm mất mát và điều chỉnh siêu tham số. Điều này giúp tạo ra các mô hình chính xác và hiệu quả cho hệ thống khuyến nghị.

Bên cạnh đó, PyTorch cũng hỗ trợ xử lý dữ liệu lớn thông qua việc tích hợp với các thư viện khác như TorchText và TorchVision. Điều này giúp người dùng có thể xử lý và phân tích một lượng lớn dữ liệu người dùng và mặt hàng một cách hiệu quả.

Cuối cùng, PyTorch cung cấp các công cụ để triển khai mô hình đã huấn luyện vào môi trường sản phẩm. Nhờ tích hợp dễ dàng vào ứng dụng hoặc trang web thực tế, hệ thống khuyến nghị có thể được tích hợp và triển khai một cách linh hoạt và hiệu quả.

# TBA

# TÀI LIỆU THAM KHẢO

Paper

[[2011.08891] Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks](https://arxiv.org/abs/2011.08891)

[[1312.4400v3] Network In Network](https://arxiv.org/abs/1312.4400v3)

[Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization | IEEE Conference Publication | IEEE Xplore](https://ieeexplore.ieee.org/document/8237336)

[[2310.08537] XAI Benchmark for Visual Explanation](https://arxiv.org/abs/2310.08537)