Explainable AI: Basic and Recent Developments - Wojciech Samek

Thanh Nam - 19120301 Gia Thuan - 19120389

22/12/2022

Phần 0: Các vấn đề hiện nay [1]

Các mô hình học máy là những mô hình hộp đen. Chúng ta không thể xác thực, kiểm tra, kiểm chứng tính đúng đắn của các kết quả mà các mô hình học máy hộp đen này trả lời.

"Explaining more than classifiers": với sự phát triển ngày nay, bên cạnh các bài toán phân loại, nhu cầu về các mô hình phân cụm (clustering) trong học máy ngày càng lớn. Các mô hình phân cụm hiện nay đều gặp các vấn đề về phát hiện outlier nhưng con người không thể phân tích được những tính toán bên trong mô hình. Do đó cần phải phát triển các các phương pháp để có thể giải thích được những vấn đề đó.

Con người cần khả năng giải thích, diễn giải của các mô hình hộp đen để: - Xác thực độ tin cậy của mô hình - Đảm bảo khía cạnh pháp lý (đặc biệt là việc áp dụng các mô hình vào các lĩnh vực pháp lý hoặc các lĩnh vực có rủi ro cao) - Học hỏi từ các hệ thống, mô hình - Cải tiến các hệ thống, mô hình.

Đưa ra lời giải thích cho các mô hình hộp đen đồng nghĩa với việc con người có thể trả lời được câu hỏi **"Tại sao mô hình lại cho ra được kết quả như thế này?"**

Phần 1: Một số phương pháp giải thích (explanation method)

1. Perturbation-Based:

   Một số phương pháp tiêu biểu:

   - Occlusion-Based (Zeiler & Fergus 14)
   - Meaningful Perturbations (Fong & Vedaldi 17)

   Đánh giá mức độ quan trọng của các pixel bằng cách đo lường sự phản ứng (kết quả đầu ra) của mạng nơ-ron khi pixel đó thay đổi. Nói cách khác, che (masking) một phần nhất định của đầu vào và đánh giá sự ảnh hưởng của chúng tới quá trình phân loại của mô hình.

2. Function-Based:

   Xem mạng nơ-ron là một hàm số (function) và sử dụng các phép biến đổi để khảo sát (khai triển Taylor, tính toán đạo hàm, …)

   - Sensitivity Analysis (Simonya et al. 14)
   - Gradient x Input (Shrikumar et al. 16)

[1] Explainable AI - Methods, Applications & Recent Developments - Dr. Wojciech Samek | ODSC Europe 2019: [https://youtu.be/AFC8yWzypss]

3. Surrogate-Based / Sampling-Based

   Xấp xỉ một mô hình khác, hoặc mạng nơ-ron về những mô hình đơn giản có thể giải thích được bằng các ước tính dự đoán của mô hình đó dựa trên các tính toán phức tạp.

   - LIME (Ribeiro et al. 16)
   - SmoothGrad (Smilkov et al.16)

4. Structure-Based

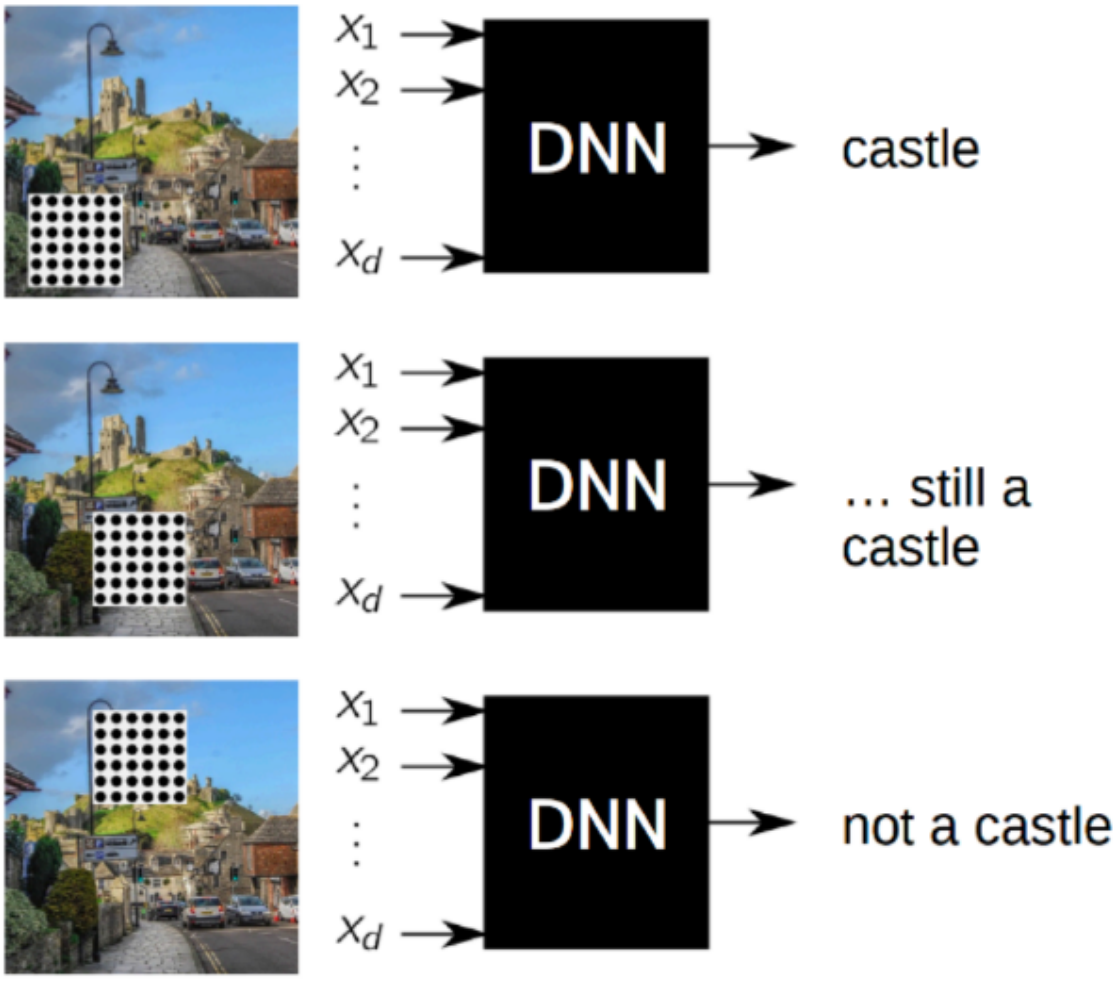
   Sử dụng cấu trúc bên trong của mạng nơ-ron để giải thích cho kết quả.

   - LRP (Bach et al.15)
   - Deep Taylor Decomposition (Montavon et al. 17)
   - Excitation Backprop (Zhang et al. 16)

Approach 1: Perturbation

Ý tưởng: Kiểm tra sự quan trọng của một đặc trưng bằng cách xóa hoặc làm xáo trộn hoặc che (masking) chúng.

Ví dụ: Bài toán phân lớp hình ảnh (hình ảnh một tòa lâu đài)



Tiến hành xóa hoặc che một phần của ảnh gốc ban đầu, khi đó sẽ có 2 trường hợp xảy ra:

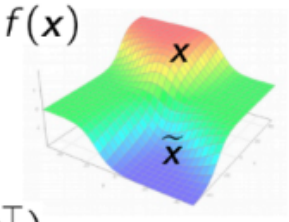
- TH1: mô hình vẫn nhận diện và phân lớp chính xác lâu đài → Xác định được các thành phần không quan trọng trong việc phân lớp (vì khi che những phần đấy đi, kết quả phân lớp vẫn chính xác).
- TH2: mô hình không phân lớp chính xác → xác định được thông tin phần vừa che rất quan trọng (ảnh hưởng đến kết quả phân lớp)

Nhược điểm:

- Chậm
- Giả định rằng các đặc trưng quan trọng nằm trong một vùng nhất định (locality). VD: các đặc trưng quan trọng có thể nằm ở hai hoặc nhiều vùng nhưng vẫn có quan hệ với nhau.
- Có thể vô tình tạo ra những đối tượng có đặc trưng hợp lệ. VD: Che một tòa lâu đài bằng một vùng xám, mô hình vẫn có thể xác định đó là tòa lâu đài vì một vùng xám nhìn từ xa khá giống tòa lâu đài.

Approach 2: (Simple) Taylor expansions

Ý tưởng: Ước lượng mức độ đóng góp/độ quan trọng của một đặc trưng bằng cách xấp xỉ Taylor đến bậc nhất.



Taylor Expansion

$$f(x) = f(\tilde{x}) + \sum_{i=1}^{d} \underbrace{[\nabla f(\tilde{x})]_i \cdot (x_i - \tilde{x}_i)}_{R_i} + \mathcal{O}(xx^\top)$$

Ưu điểm:

- Có thể áp dụng với mọi mô hình ML (có đạo hàm(differentiable) và gần phi tuyến(mildly non-linear)).

Nhược điểm:

- Khá khó tìm được điểm root tốt để thực hiện khai triển.

Approach 3: Gradient x Input

Motivation: Tính toán lời giải thích cho mỗi đầu vào bằng cách nhân đạo hàm với input mà không cần phải tối ưu hoặc tìm điểm root hợp lí.

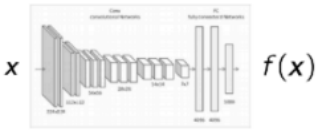
Observation: Lời giải thích thường bị nhiễu(noise). Nguyên nhân:

- Các hiệu ứng toàn cục không quan sát được khi khảo sát ở mức cục bộ.
- Các biến cục bộ của hàm số tăng theo cấp số nhân với độ sâu.

Gradient x Input

$$\forall_i : \; R_i = [\nabla f(x)]_i \cdot x_i$$

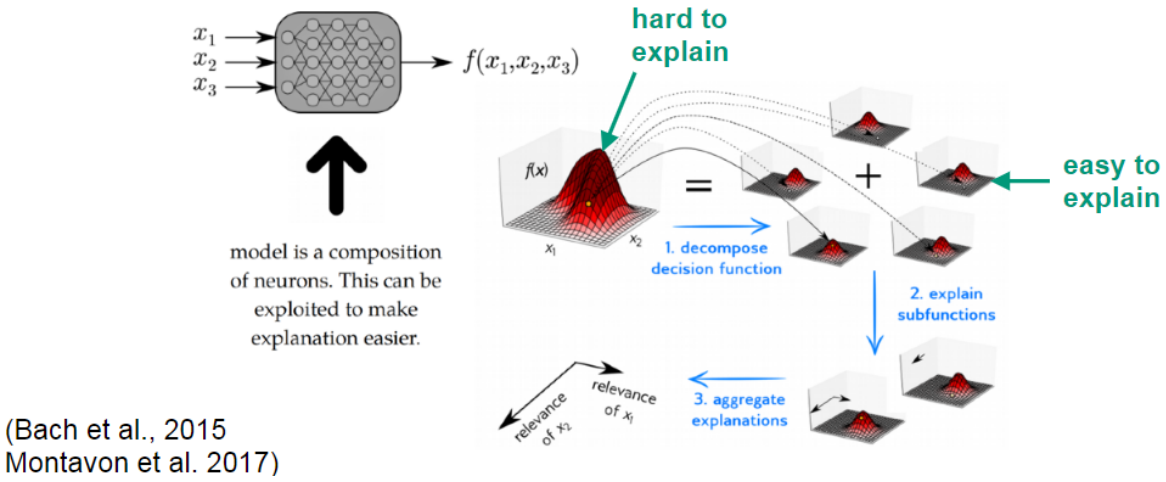
$$R = \nabla f(x) \odot x$$



Approach hiện tại: Layer-wise relevance propagation[2]

Ý tưởng: sử dụng chính kiến trúc để làm đơn giản hóa việc giải thích. Thay vì giải thích toàn bộ cùng lúc, chia nhỏ mạng nơ-ron thành từng phần nhỏ dễ giải thích hơn. Các bước: phân rã mô hình thành các decision function → giải thích các function → tổng hợp các lời giải thích.

[2]Layer-wise Relevance Propagation: An overview [https://iphome.hhi.de/samek/pdf/MonXAI19.pdf]



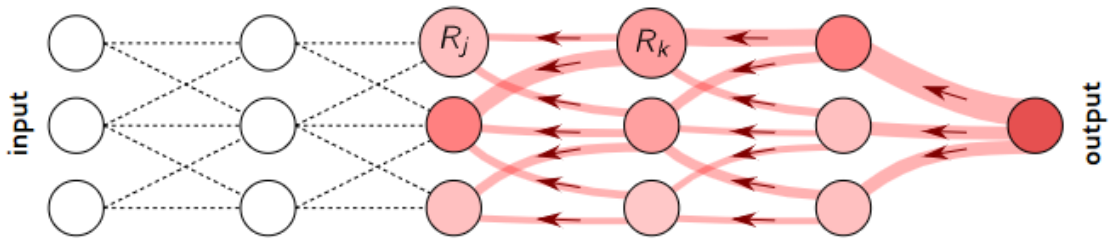
(Bach et al., 2015
Montavon et al. 2017)

Quy trình lan truyền của LRP phải đảm bảo tính bảo toàn, nghĩa là những gì nơ-ron nhận được phải phân phối lại cho lớp dưới với lượng như nhau, tương tự như định luật bảo toàn của Kirchoff trong các mạch điện. Với j và k là những nơ-ron thuộc hai tầng liên tiếp của mô hình, khi đó việc truyền điểm liên quan (relevance scores) $(R_k)_k$ của một tầng nhất định lên những nơ-ron của tầng thấp hơn đạt được bằng cách áp dụng quy tắc:

$$R_j = \sum_k \frac{z_{jk}}{\sum_j z_{jk}} R_k$$

Trong đó, đại lượng z_{jk} mô hình hóa mức độ mà nơ-ron j đã đóng góp để làm cho nơ-ron k có liên quan, mẫu số là tổng z_{jk} nhằm bảo toàn thông tin. Quá trình lan truyền kết thúc tại các đặc trưng đầu vào. Nếu sử dụng quy tắc trên cho tất cả nơ-ron trong mạng ta có thể dễ dàng xác minh thuộc tính bảo toàn theo lớp $\sum_j R_j = \sum_k R_k$, và mở rộng tính bảo toàn lên toàn cục $\sum_i R_i = f(x)$.

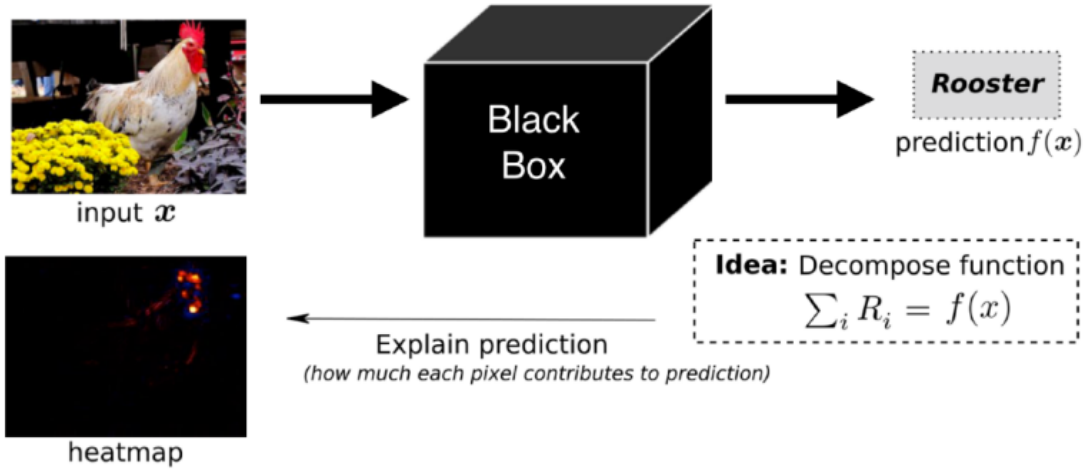
$$\sum_i R_i = \ldots = \sum_i R_i^{(l)} = \sum_j R_j^{(l+1)} = \ldots = f(x)$$



Minh họa về quy trình LRP. Nguồn: G. Montavon et al.[3]

Mặc dù LRP rõ ràng khác với phân rã Taylor, ở phần sau ta sẽ thấy rằng mỗi bước của quy trình lan truyền có thể được mô hình hóa như một phép phân rã Taylor riêng được thực hiện trên các đại lượng cục bộ trong đồ thị.

[3]Layer-wise Relevance Propagation: An overview [https://iphome.hhi.de/samek/pdf/MonXAI19.pdf]



Nguồn: Bach et al. PLOS ONE, 2015[4]

Ví dụ về bức ảnh nhận diện gà trống: giải thích các pixel đóng góp như thế nào vào kết quả dự đoán. Decompose function: $\sum R_i = f(x)$. Ý tưởng cho việc này là tái phân phối các bằng chứng cho lớp gà trống về không gian ảnh, khởi tạo $R_j^{(l+1)} = f(x)$.

Luật LRP cho Mạng chỉnh lưu sâu (Deep Rectifier Networks) Mạng nơ-ron sâu với tính phi tuyến của bộ chỉnh lưu (ReLU) là lựa chọn phổ biến nhất hiện nay trong các ứng dụng thực tế. Mạng nơ-ron chỉnh lưu bao gồm các nơ-ron thuộc loại $a_k = max(0, \sum_{0,j} a_j w_{jk})$. Tổng $\sum_{0,j}$ đi qua tất cả các kích hoạt của tầng dưới $(a_j)_j$, cùng với một nơ-ron thể hiện bias. Cụ thể, $a_0 = 1$ và w_{0k} là nơ-ron bias. Dưới đây là một số luật lan truyền LRP:

Simple LRP rule ($LRP-0$) (Bach et al. 2015):

$$R_i^{(l)} = \sum_j \frac{x_i \cdot w_{ij}}{\sum_{i'} x_{i'} \cdot w_{i'j}} R_j^{(l+1)}$$

Mỗi nơ-ron nhận phần chia sẻ của nó về mức độ liên quan được phân phối lại.

Epsilon LRP rule ($LRP-\epsilon$) (Bach et al. 2015):

$$R_j = \sum_k \frac{a_j w_{jk}}{\epsilon + \sum_0^j a_j w_{jk}} R_k$$

Bổ sung ϵ vào mẫu số nhằm mục đích hấp thụ một số liên quan khi những đóng góp vào kích hoạt của nơ-ron k là yếu hoặc mâu thuẫn. Khi ϵ càng lớn thì những yếu tố giải thích nổi bật nhất mới tồn tại trong quá trình hấp thụ. Điều này dẫn đến những lời giải thích thưa thớt hơn ở các đặc trưng đầu vào và ít bị nhiễu hơn.

Gamma LRP rule ($LRP-\gamma$)

[4]On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation: [https://doi.org/10.1371/journal.pone.0130140]

$$R_j = \sum_k \frac{a_j(w_{jk} + \gamma w_{jk}^+)}{\sum\limits_0^j a_j(w_{jk} + \gamma w_{jk}^+)} R_k$$

Một sự cải tiến khác nhằm tăng sự chú trọng với những đóng góp tích cực hơn là đóng góp tiêu cực. Tham số γ quyết định mức độ mà đóng góp tích cực được chú ý. Khi γ càng tăng, những đóng góp tiêu cực sẽ dần biến mất. Sự phổ biến của những đóng góp tích cực có tác động hạn chế đến độ lớn liên quan giữa tích cực và tiêu cực có thể phát triển trong giai đoạn lan truyền. Điều này giúp đưa ra những lời giải thích ổn định hơn. Ý tưởng xử lý các đóng góp tích cực và tiêu cực theo cách bất đối xứng với quy tắc $LRP - \alpha\beta$. Ngoài ra, việc chọn $\gamma \to \infty$ cho phép $LRP - \gamma$ trở nên tương đương với $LRP - \alpha_1\beta_0$, quy tắc z^+ và "excitation-backprop".

Alpha-beta LRP rule (Bach et al. 2015):

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

where $\alpha + \beta = 1$

Tương tự như Simple LRP nhưng chia ra 2 phần âm dương để đảm bảo rằng mẫu số sẽ không bằng 0.

Luật LRP dưới dạng Deep Taylor Decomposition Nhắc lại Simple Taylor Decomposition:

$$f(x) = f(\tilde{x}) + \sum_{i=1}^{d} [\nabla f(\tilde{x})]_i \cdot (x_i - \tilde{x_i}) + \mathcal{O}(xx^\top)$$

- Ý tưởng: sử dụng khai triển Taylor để tái phân phối sự liên quan từ output về input.
- Nhược điểm: khó tìm được điểm root tốt, gradient shattering.

Luật lan truyền có thể được giải thích trong khuôn khổ Deep Taylor Decomposition. DTD xem LRP như một chuỗi Taylor mở rộng được thực hiện cục bộ tại mỗi nơ-ron. Cụ thể hơn, điểm liên hệ R_k được biểu diễn dưới dạng hàm của các kích hoạt ở tầng thấp hơn $(a_j)_j$ được biểu thị bằng vector a, sau đó khai triển Taylor đến bậc nhất của $R_k(a)$ tại một số điểm tham chiếu trong không gian kích hoạt:

$$R_k(a) = R_k(\tilde{a}) + \sum_j [\nabla R_k(\tilde{a})]_j \cdot (a_i - \tilde{a_i}) + \mathcal{O}(aa^\top)$$

Phần hệ số của đạo hàm bậc nhất sẽ quyết định lượng R_k phân phối xuống các nơ-ron ở tầng thấp hơn. Vì quan hệ giữa a và R_k có thể phức tạp nên có thể sẽ khó tìm điểm tham chiếu thích hợp và tính gradient cục bộ.

Các công thức dùng cho Deep Taylor Decomposition:

1. Relevance model
   Để có được một biểu thức dạng đóng cho các số hạng của phương trình trên, người ta cần thay thế hàm liên quan thực sự $R_k(a)$ bằng một mô hình phù hợp $\hat{R}_k(a)$ dễ phân tích hơn. Một trong những mô hình như vậy là *modulated ReLU activation*:

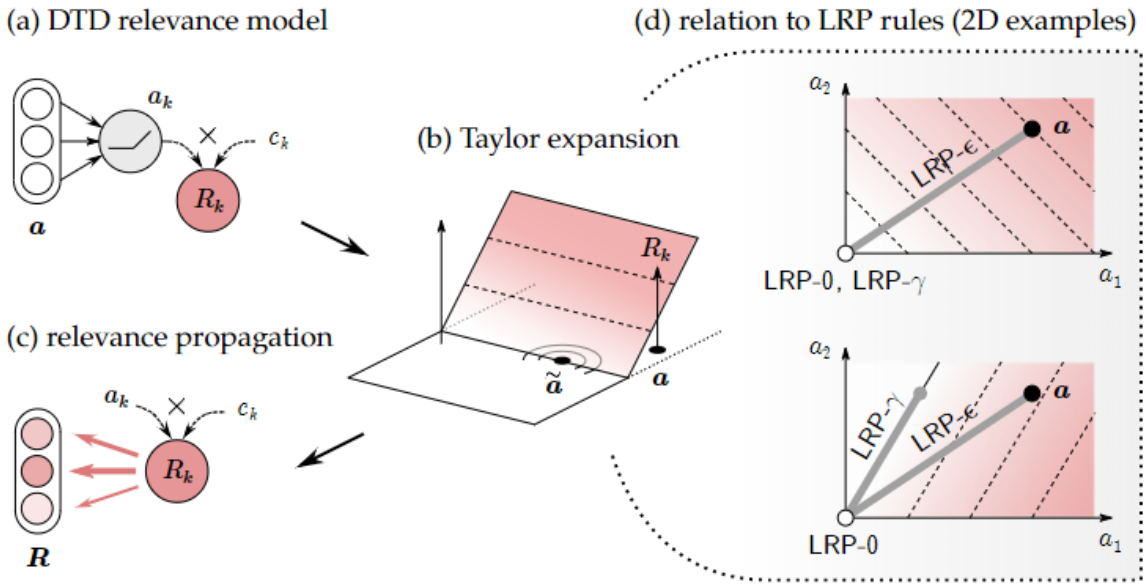
$$\hat{R}_k(a) = max(0, \sum_j a_j w_{jk}) c_k$$

Module c_k là hằng số và được đặt sao cho $\hat{R}_k(a) = R_k(a)$ tại điểm dữ liệu hiện tại. Có thể xem c_k là hằng số khi R_k là kết quả của việc áp dụng $LRP-0/\epsilon/\gamma$ ở các tầng cao hơn.

2. Khai triển Taylor

$$\hat{R}_k(a) = \hat{R}_k(\tilde{a}) + \sum_j \underbrace{(a_j - \tilde{a}_j) \cdot w_{jk} c_k}_{R_{j \leftarrow k}} + 0$$

Hệ số tại đạo hàm bậc hai hoặc cao hơn bằng 0 do tính tuyến tính trong miền kích hoạt của hàm ReLU. Hệ số tại đạo hàm bậc không có thể làm nhỏ tùy ý bằng cách chọn điểm tham chiếu gần bản lề ReLU. Khi một điểm tham chiếu được chọn, khai triển đến bậc nhất có thể tính toán một cách dễ dàng.



*Minh họa DTD: (a) mô hình liên quan dưới góc nhìn đồ thị, (b) mô hình liên quan và điểm tham chiếu khi thực hiện khai triển Taylor dưới góc nhìn hàm số, (c) lan truyền khai triển đến đạo hàm bậc nhất về các tầng thấp hơn.

3. Lựa chọn điểm tham chiếu

$$\tilde{a}^{(k)} = 0 \Leftrightarrow \rho = (\cdot), \epsilon = 0 (\text{LRP-0})$$
$$\tilde{a}^{(k)} = a - t \cdot a \Leftrightarrow \rho = (\cdot), \epsilon = (t^{-1} - 1) \cdot a_k (\text{LRP-}\epsilon)$$
$$\tilde{a}^{(k)} = a - t \cdot a \odot 1_{w_k \succ 0} \Leftrightarrow \rho = \max(0, \cdot) (\text{LRP-}\gamma)$$

Công thức Deep Taylor Decomposition cho các tầng trong mạng nơ-ron:

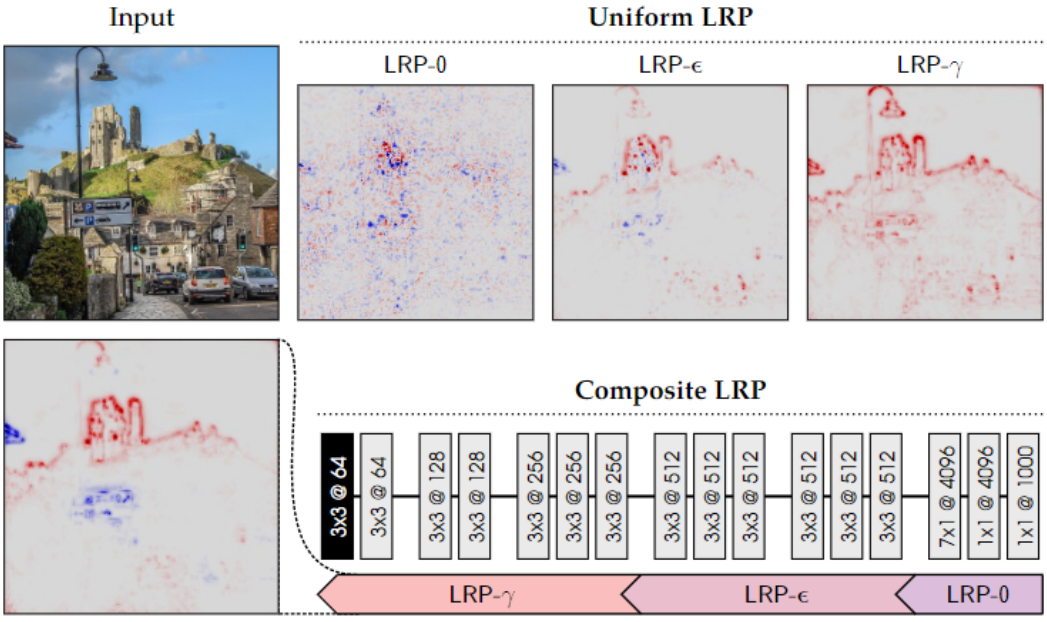
Name	Formula	Usage	DTD
$LRP-0$	$R_j = \sum_k \frac{a_j w_{jk}}{\sum_j^0 a_j w_{jk}} R_k$	upper layers	
$LRP-\epsilon$	$R_j = \sum_k \frac{a_j w_{jk}}{\epsilon + \sum_j^0 a_j w_{jk}} R_k$	middle layers	

Name	Formula	Usage	DTD
$LRP-\gamma$	$R_j = \sum_k \frac{a_j(w_{jk}+\gamma w^+_{jk})}{\sum_j \sum_0 a_j(w_{jk}+\gamma w^+_{jk})} R_k$	lower layers	
$LRP-\alpha\beta$	$R_j = \sum_k \left(\alpha \frac{(a_j w_{jk})^+}{\sum_j \sum_0 (a_j w_{jk})^+} - \beta \frac{(a_j w_{jk})^-}{\sum_j \sum_0 (a_j w_{jk})^-} \right) R_k$	lower layers	*
flat	$R_j = \sum_k \frac{1}{\sum_j 1} R_k$	lower layers	
w^2-rule	$R_i = \sum_j \frac{w^2_{ij}}{\sum_i w^2_{ij}} R_j$	first layers ($\mathbb{R}^d$)	
$z^{\mathcal{B}}-rule$	$R_i = \sum_j \frac{x_i w_{ij} - l_i w^+_{ij} - h_i w^-_{ij}}{\sum_i x_i w_{ij} - l_i w^+_{ij} - h_i w^-_{ij}} R_j$	first layers (pixels)	

* Chỉ diễn giải DTD trong trường hợp $\alpha = 1, \beta = 0$
Các kí hiệu $(\cdot)^+ = max(0, \cdot)$ và $(\cdot)^- = min(0, \cdot)$. Với $LRP-\alpha\beta$, các tham số α, β tuân theo ràng buộc bảo toàn $\alpha = \beta + 1$. Với luật $z^{\mathcal{B}}$, các tham số l_i, h_i xác định ràng buộc của miền đầu vào $\forall i : l_i \leq x_i \leq h_i$

Best practice for LRP

Nguyên tắc: giải thích từng loại lớp (input, conv, dense, ...) với các quy tắc tối ưu theo Deep Taylor Decomposition.



Giải thích mô hình dự đoán tòa lâu đài với các luật khác nhau ở các tầng ($\gamma = 0.25$ và $\epsilon = 0.25$). Nguồn: G. Montavon et al.[5]

[5]Layer-wise Relevance Propagation: An overview [https://iphome.hhi.de/samek/pdf/MonXAI19.pdf]

Phần 2: Đánh giá các lời giải thích (Evaluating Explanations)

Desiderata for Explanations

1. Fidelity: lời giải thích phải phản ánh số lượng được giải thích chứ không phải thứ gì khác.
2. Understandability: lời giải thích phải dễ hiểu đối với người nhận.
3. Sufficiency: lời giải thích phải đưa ra những thông tin hợp lí về việc model đã đưa ra kết quả như thế nào.
4. Low overhead: lời giải thích không được làm cho model dự đoán kém chính xác hoặc kém hiệu quả.
5. Runtime efficiency: lời giải thích phải được tính toán trong khoảng thời gian hợp lí.

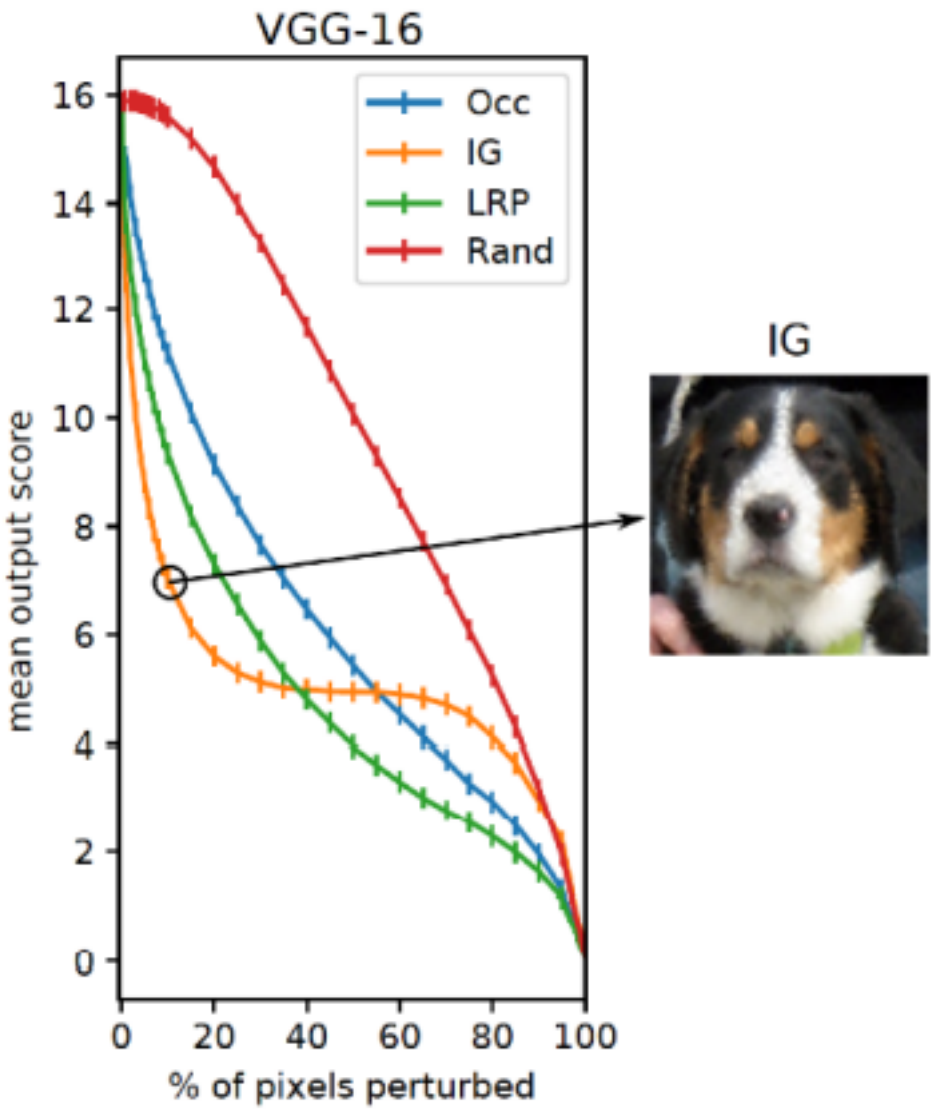
QUANTUS - Toolkit cho việc đánh giá lời giải thích của mạng nơ-ron

Evaluating Fidelity: Pixel-flipping

Ý tưởng: so sánh tính chọn lọc (Bach'15,Samek'17). "Nếu một đặc trưng được cho là có liên quan thì khi xóa chúng sẽ làm mất đi các bằng chứng ở kết quả đầu ra của mạng nơ-ron."

Thuật toán "Pixel Flipping":

```
Sort pixel/sort by relevance
Iterate
    destroy pixel/patch
    evaluate f(x)
Measure decrease of f(x)
```



- Tất cả các phương pháp giải thích phải tốt hơn một lời giải thích ngẫu nhiên.
- IG là phương pháp đúng nhất nếu chỉ xét ở vài pixel đầu tiên, sau đó kết quả gần như không thay đổi.
- Mặc dù không thể nhận diện bởi VGG-16, những pattern liên quan đến lớp vẫn nằm ở đó sau khi bị flip (chúng ta vẫn thấy con chó). Như vậy IG có thực sự giải thích lỗ hổng của VGG-16 thay vì hành vi điển hình của nó không?

Axiomatic approach

Ý tưởng: Sử dụng các tiên đề để đánh giá, giải thích cách hoạt động của mạng nơ-ron nhận tạo.

Axiomatic (tiên đề): là một thuộc tính hiển nhiên của một lời giải thích. Tiên đề được đặt ra đòi hỏi phải đúng (hiển nhiên) đối với các lớp đầu vào và cả mô hình.

Các yếu tố để xem xét, đánh giá:

- Conservation (tính bảo tồn)
- Continuity (tính liên tục)
- Implementation invariance (tính bất biến trong triển khai)

Tìm các tiên đề đến từng nơ-ron riêng lẻ của mạng nơ-ron nhân tạo, từ đó rút ra được những lời giải thích đang được hình thành tại nơ-ron đang xét đó. Điều này sẽ cho phép chúng ta xác định được các thành phần chính trong các kĩ thuật giải thích xem chúng có thỏa hoặc không thỏa tiên đề hay không.

Ground-truth Based evaluation

Ý tưởng: đánh giá dựa trên ground truth có sẵn, so sánh kết quả dự đoán với ground truth xem có chính xác hay không.

Tuy nhiên phương pháp đánh giá này khó có thể áp dụng được đối với các bài toán học không giám sát. Vì các bài toán này hầu như không có sẵn ground-truth, tuy nhiên vẫn có thể thực hiện thủ công để có được bộ ground-truth dùng để đánh giá hoặc dùng một số phương pháp đặc biệt sẽ được giới thiệu ở phần dưới (neuralization trick).

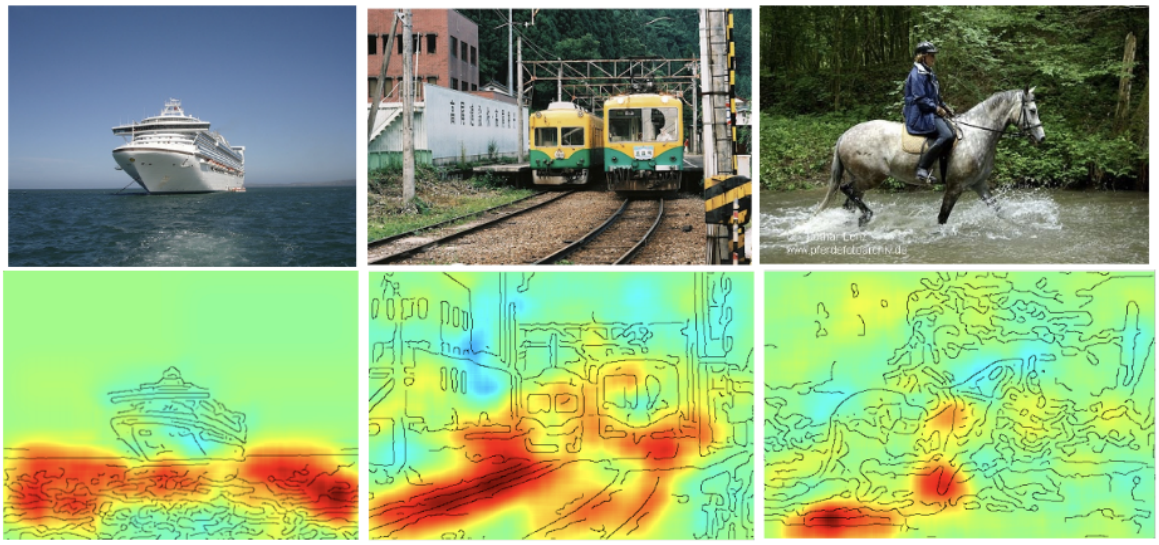
Phần 3: Ứng dụng của XAI (XAI applications)

Ứng dụng của LRP trong những vấn đề khác nhau:

Có thể áp dụng trong nhiều lĩnh vực để giải quyết các vấn đề khác nhau. Không những có thể áp dụng cho hình ảnh, mà còn áp dụng cho âm thanh, văn bản, video hoặc dữ liệu trong lĩnh vực y tế, lĩnh vực khoa học, lĩnh vực công nghiệp,...

1. PASCAL VOC challenge (2005-2012)

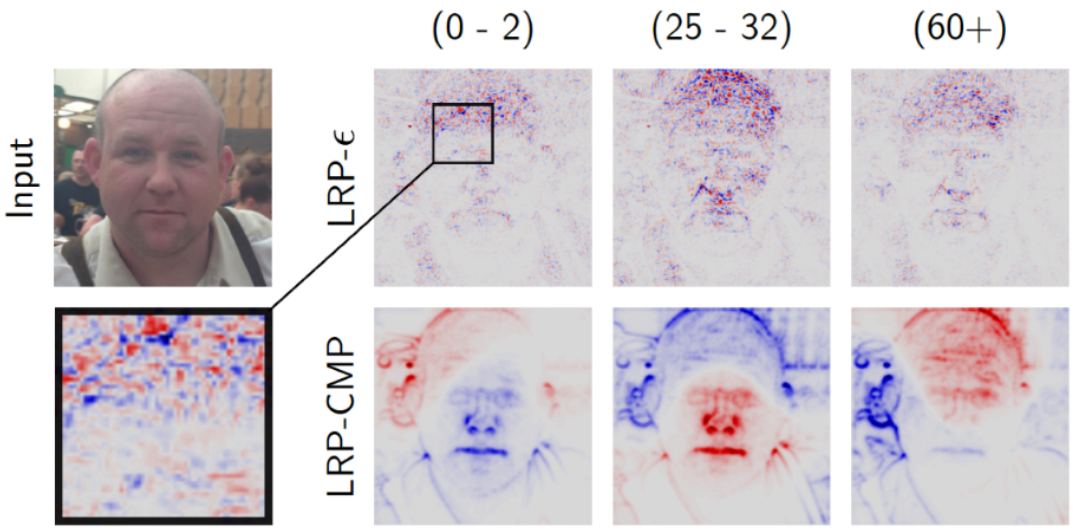
Mô hình dự đoán đúng nhưng các đặc trưng mà mô hình học được lại không hợp lí. VD: với bức ảnh con tàu thì mô hình tập trung vào vùng có nước, với bức ảnh đoàn tàu thì mô hình tập trung vào đường ray, với ảnh con ngựa thì mô hình tập trung vào phần copyright text (vì có nhiều ảnh con ngựa khác cũng có chung copyright text). Từ đây ta thấy XAI giúp các nhà nghiên cứu hiểu về cách học của mạng nơ-ron trong tác vụ phân loại ảnh nhằm sửa lỗi/cải tiến mô hình.



Nguồn: Lapuschkin et al. 2019[6]

2. Kiểm chứng mô hình phân loại ảnh mặt người:

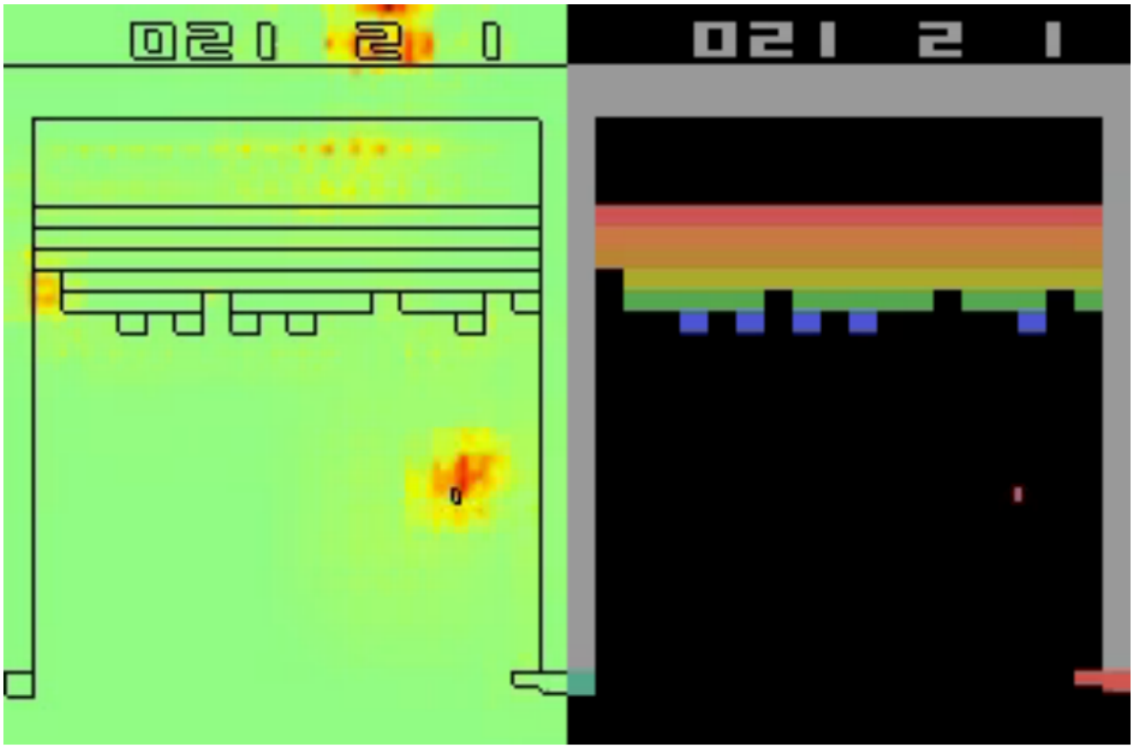
Mô hình dự đoán giới tính và độ tuổi của người trong ảnh (độ tuổi sẽ bao gồm các khoảng từ (0-2),(4-6),(8-13),(15-20),(25-32),(38-43),(48-53),(60+)). Bằng cách trực quan hóa người trong ảnh thành biểu đồ nhiệt (heatmap), mô hình giúp chúng ta dễ dàng xác định được các yếu tố ảnh hưởng đến kết quả dự đoán. Bao gồm xác định các chi tiết, yếu tố trên khuôn mặt người khiến cho mô hình có dự đoán như thế, nghĩa là ta có thể giải thích được vì sao mô hình lại dự đoán kết quả giới tính và độ tuổi như thế.



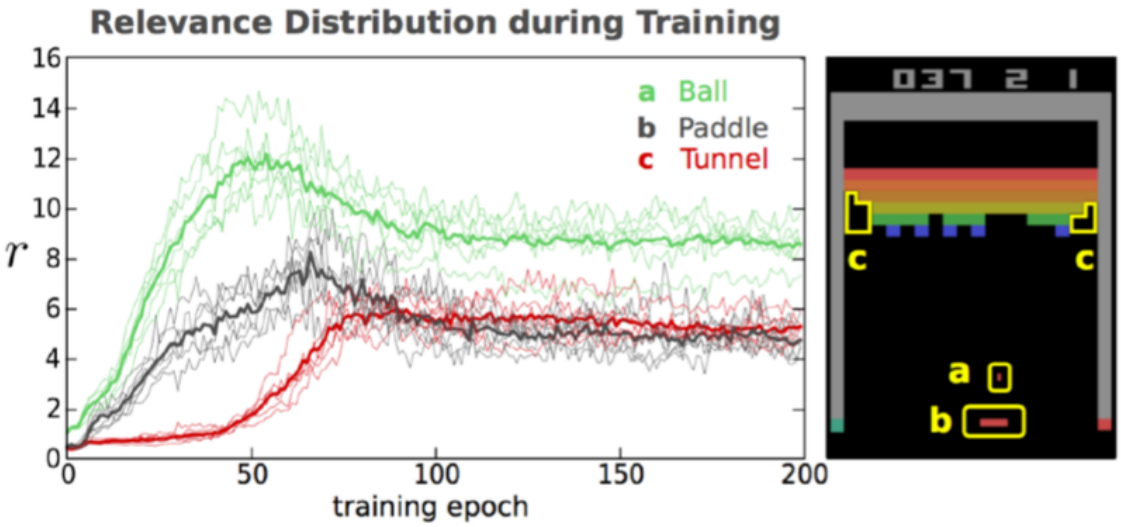
3. Hiểu về hành vi học:

Qua ví dụ trò chơi Atari Breakout

[6] Unmasking Clever Hans predictors and assessing what machine really learn:[https://www.nature.com/articles/s41467-019-08987-4]



Giao diện và cách thức chơi của trò Atari Breakout. Nguồn: Lapuschkin et al. 2019[7]



Biểu đồ phân bố sự tương quan trong suốt quá trình huấn luyện mô hình. Nguồn: Lapuschkin et al. 2019[8]

Quan sát biểu đồ trực quan mô hình được huấn luyện qua từng epoch.

[7] Unmasking Clever Hans predictors and assessing what machine really learn:[https://www.nature.com/articles/s41467-019-08987-4]

[8] Unmasking Clever Hans predictors and assessing what machine really learn:[https://www.nature.com/articles/s41467-019-08987-4]

- Đường màu xanh thể hiện quá bóng, là một yếu tố quan trọng trong trò chơi vì nếu để mất bóng (quả bóng rơi xuống dưới) thì trò chơi sẽ kết thúc.
- Đường màu đen thể hiện yếu tố bàn đạp (quan trọng không kém).
- Đường màu đỏ (tunnel) là đường hầm (vị trí nên bắn quả bóng) sao cho hiệu quả nhất mà mô hình tìm được thông qua các training epoch.

Từ đó ta có thể hiểu được hành vi học của mô hình. Đầu tiên là quan sát chuyển động của quả bóng, sau đó tập trung vào bàn đạp và cuối cùng là tập trung vào đường hầm để giúp ăn được nhiều điểm nhất đồng thời rủi ro thua thấp nhất.

4. XAI trong các ngành khoa học khác:

XAI được sử dụng như một công cụ để phân tích các dự đoán, chiến lược (hành vi) học của mô hình, đồng thời được sử dụng để phân tích dữ liệu. Ví dụ trong phân tích ảnh y tế, Thomas et al. đã sử dụng cách tiếp cận CNN+LSTM để phân tích các ảnh đầu vào, sau đó sử dụng LRP để trực quan hóa các kết quả.

Kết luận:

- Khả năng giải thích từng loại cung cấp nhiều thông tin giá trị về hành vi của mô hình.
- Luôn tồn tại các cách tiếp cận lý thuyết đối với XAI (Deep Taylor, Shapley). Điều đó cho phép tính toán, suy luận các lời giải thích thực sự có ý nghĩa, hơn nữa là đối với các mạng học sâu.
- Sự quan tâm dành cho XAI là rất lớn trong cộng đồng nghiên cứu khoa học nói chung.
- Một số tình huống con người không hiểu được các lời giải thích: các lĩnh vực không thể diễn giải được (chuỗi thời gian, trình tự, ngữ cảnh) hay "Pixel-wise" thay vì sử dụng các khái niệm của con người.

Phần 4: XAI trong các tập dữ liệu (Dataset-wide XAI)

Thông thường trong nghiên cứu, ta sẽ tìm lời giải thích đối với từng hình ảnh riêng lẻ, nhưng trên thực tế tập dữ liệu rất lớn (ví dụ như hàng triệu tấm ảnh). Việc này gây ra khó khăn rất lớn vì chúng ta không thể nào tìm lời giải thích cho hàng ngàn, hàng triệu dữ liệu trong tập data-wide.

→ Giải pháp cho vấn đề trên là thực hiện gom nhóm (clustering). Tìm các lời giải thích cho mô hình hộp đen, sau đó thực hiện tính toán để xem xét liệu có tồn tại các cấu trúc chung nào hay không để gom nhóm các lời giải thích lại với nhau.

Hiện tượng *Clever Hans*: Clever Hans là tên một con ngựa có khả năng thực hiện các phép tính toán đơn giản. Tuy nhiên, sau đó một nhà tâm lý học trong nghiên cứu của mình đã chứng minh được rằng con ngựa không thực sự biết thực hiện các tác vụ đó. Nó thực chất đã tuân theo các gợi ý mà người huấn luyện vô tình tạo ra. Trong AI, hiện tượng Clever Hans được mô tả là các mô hình học đã tiếp thu được những mối liên hệ giả trong dữ liệu huấn luyện, từ đó đưa ra dự đoán chính xác.

Trong bài báo [9], nhóm tác giả đã sử dụng LRP chỉ ra những vấn đề tồn tại kể cả trong những phương pháp hàng đầu.

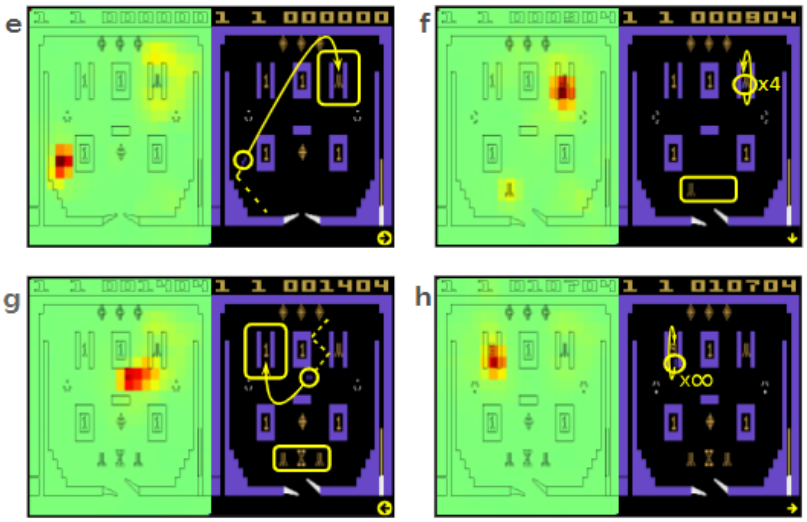
- Đầu tiên là mô hình dựa trên Fisher vectors được huấn luyện trên tập dữ liệu PASCAL VOC 2007. Mô hình này cùng với đối thủ của nó (một mạng nơ-ron học sâu được huấn luyện trước) được fine-tune trên PASCAL VOC đều cho ra kết quả ở mức SOTA. Tuy nhiên, khi áp dụng LRP để tạo ra bản đồ nhiệt, nhóm tác giả nhận thấy rằng những đặc trưng mà mô hình học được là không đúng. Với những bức ảnh con ngựa, đáng lẽ mô hình sẽ tập trung vào vận động viên cưỡi ngựa và con ngựa nhưng bản đồ nhiệt lại cho thấy các pixel ở góc phải dưới mới quan

[9] Unmasking Clever Hans predictors and assessing what machine really learn:[https://www.nature.com/articles/s41467-019-08987-4]

trọng. Đây là vị trí ghi nguồn của bức ảnh. Khi kiểm tra các bức ảnh thì các bức ảnh con ngựa đều có chung đặc điểm này, do đó, các mô hình bị overfitting mặc dù đặc trưng mà chúng học được là không chính xác.

Nguồn: Lapuschkin et al. 2019[10]

- Atari Pinball: mô hình học đã tập trung vào những khu vực chứa pixel là đối tượng ghi điểm và bỏ qua hai thanh chèo. Đầu tiên, mô hình sẽ đưa quả bóng vào Atari rollover mà không sử dụng thanh chèo, sau đó tìm cách "huých" vào bàn pinball sao cho quả bóng qua lại vĩnh viễn Left rollover (đây là khu vực ghi nhiều điểm nhất) mà không làm nghiêng bàn bi lắc. Mô hình học đã học được cách lạm dụng ngưỡng "nhúc nhích" được thực hiện thông qua cơ chế nghiêng trong Atari Pinball. Ở góc độ trò chơi, việc khai thác cơ chế trò chơi có thể xem là hợp lí, tuy nhiên trong thực tế, các phần mềm trò chơi sẽ được lập trình để thoát khỏi vòng lặp như thế.

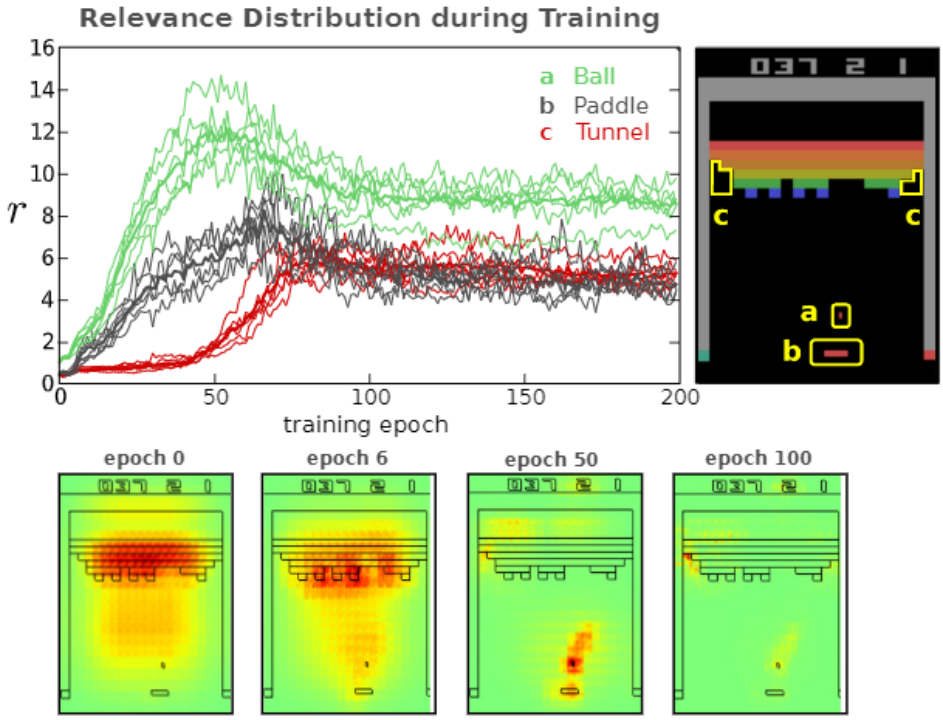


Nguồn: Lapuschkin et al. 2019[11]

[10] Unmasking Clever Hans predictors and assessing what machine really learn:[https://www.nature.com/articles/s41467-019-08987-4]

[11] Unmasking Clever Hans predictors and assessing what machine really learn:[https://www.nature.com/articles/s41467-

- Atari Breakout: Giai đoạn đầu, mô hình học cách kiểm soát quả bóng, do đó phần thanh đỡ sẽ sáng nhất, giai đoạn sau mô hình sẽ tìm cách đưa quả bóng vào khu vực có dạng phễu. Đây cũng là chiến lược mà con người sử dụng trong trò chơi này.



Nguồn: Lapuschkin et al. 2019[12]

Spectral Relevance Analysis - SpRAy (tạm dịch: phân tích tương quan phổ)

- Bước 1: Tính toán biểu đồ tương quan cho các mẫu cần quan tâm. Các biểu đồ phải được tính toán bằng LRP và chứa thông tin về vị trí mà mô hình phân loại tập trung vào khi phân loại bức ảnh.
- Bước 2: Thu nhỏ biểu đồ tương quan về cùng kích thước nhằm làm tăng tốc độ tính toán và dễ dàng phân tích.
- Bước 3: Phân tích cụm phổ (spectral cluster analysis) trên biểu đồ tương quan nhằm phân tích các cấu trúc của phân phối và gom lại thành hữu hạn các cụm (cluster).
- Bước 4: Xác định các cụm cần quan tâm bằng cách phân tích eigengap (khoảng cách giữa 2 trị riêng - eigenvalue). Eigengap càng lớn thì các cụm càng được phân tách rõ ràng.
- Bước 5 (Optional): Trực quan hóa bằng cách sử dụng t-SNE trên biểu đồ tương quan.

**Biểu đồ tương quan: heatmap*

Un'Hans the Dataset: (tạm dịch: loại bỏ những tương quan giả)

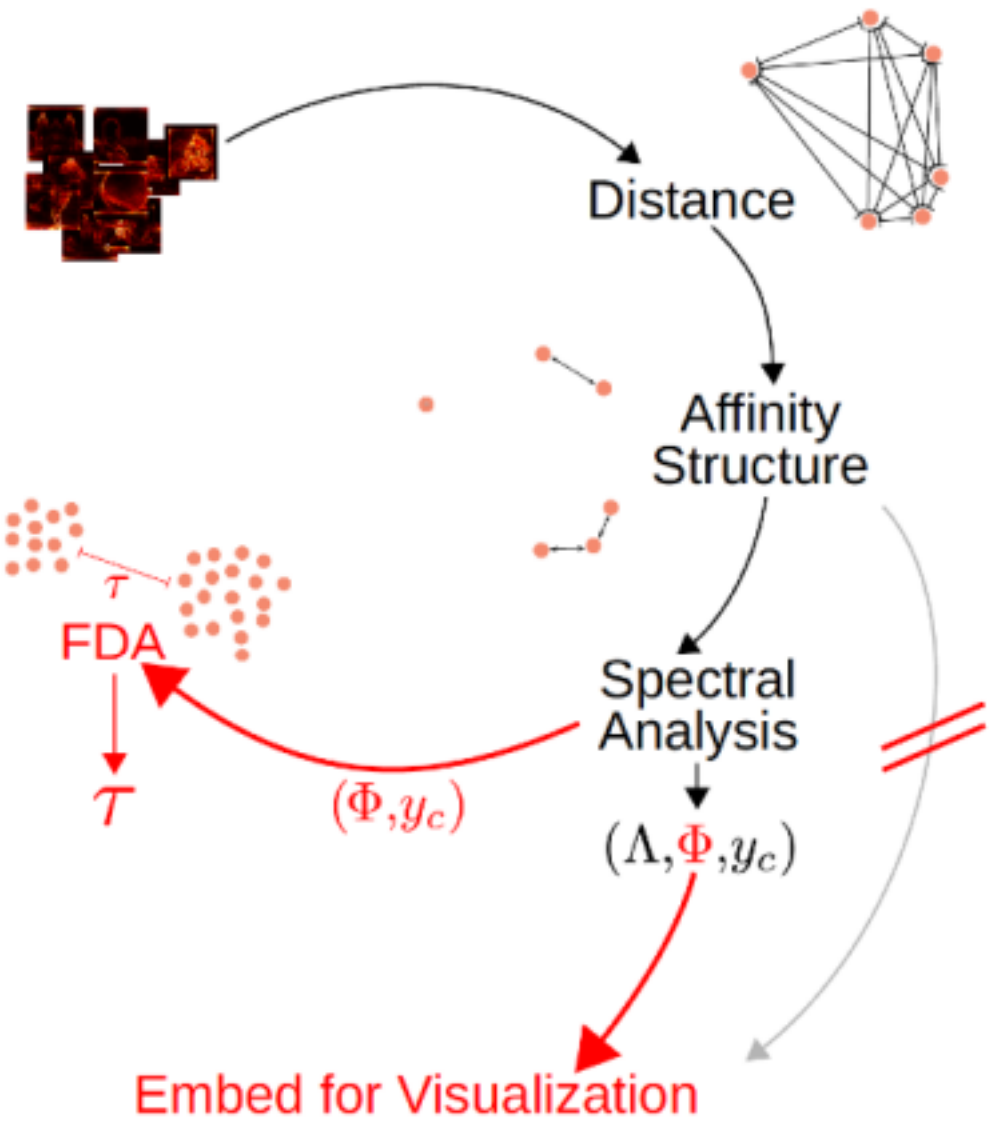
1. Tìm kiếm những Hans (liên hệ/tương quan giả)
2. Mô tả/mô hình hóa Hans.

019-08987-4]

[12]Unmasking Clever Hans predictors and assessing what machine really learn:[https://www.nature.com/articles/s41467-019-08987-4]

3. Loại bỏ Hans.

Automating Clever Hans Detection



Trong bài báo của *Anders et al. 2019* [13], tác giả đã đưa ra một số thay đổi trong phương pháp SpRAy:

- Sử dụng khoảng cách Gromov-Wasserstein thay vì khoảng cách Euclid trong đồ thị lân cận (neighborhood graph).
- Tự động hóa hơn nữa việc khám phá các tương quan giả bằng cách phân tích Φ sử dụng Fisher Discriminant Analysis.

[13] Analyzing ImageNet with Spectral Relevance Analysis: Towards ImageNet un-Hans'ed:[https://www.semanticscholar.org/paper/Analyzing-ImageNet-with-Spectral-Relevance-Towards-Anders-Marinc/5db1b78d1cc34ee388392c3214cd05ed7fe4317f]

- Trực quan hóa nhúng quang phổ (spectral embedding) Φ thay vì các cấu trúc tương đồng.

Algorithm:

Đầu vào: Mẫu đầu vào $X = x$, mô hình f tính toán trên X

Đầu ra: - Trị riêng $\Lambda = \{\lambda\}$

- Spectral embedding $\Phi \in R^{n \times q}$
- Nhãn các cluster $Y = \{y\}$
- Điểm phân lớp cluster τ
- Visualization embeddings $Z \in R^2$

Tính toán các thuộc tính cho $x \in X$

$R = \{\}$
    for $x \in X$ do
        $R_x = LRP(f, x)$;
        $R.append(R_x)$;
    end

/* tiền xử lí R */

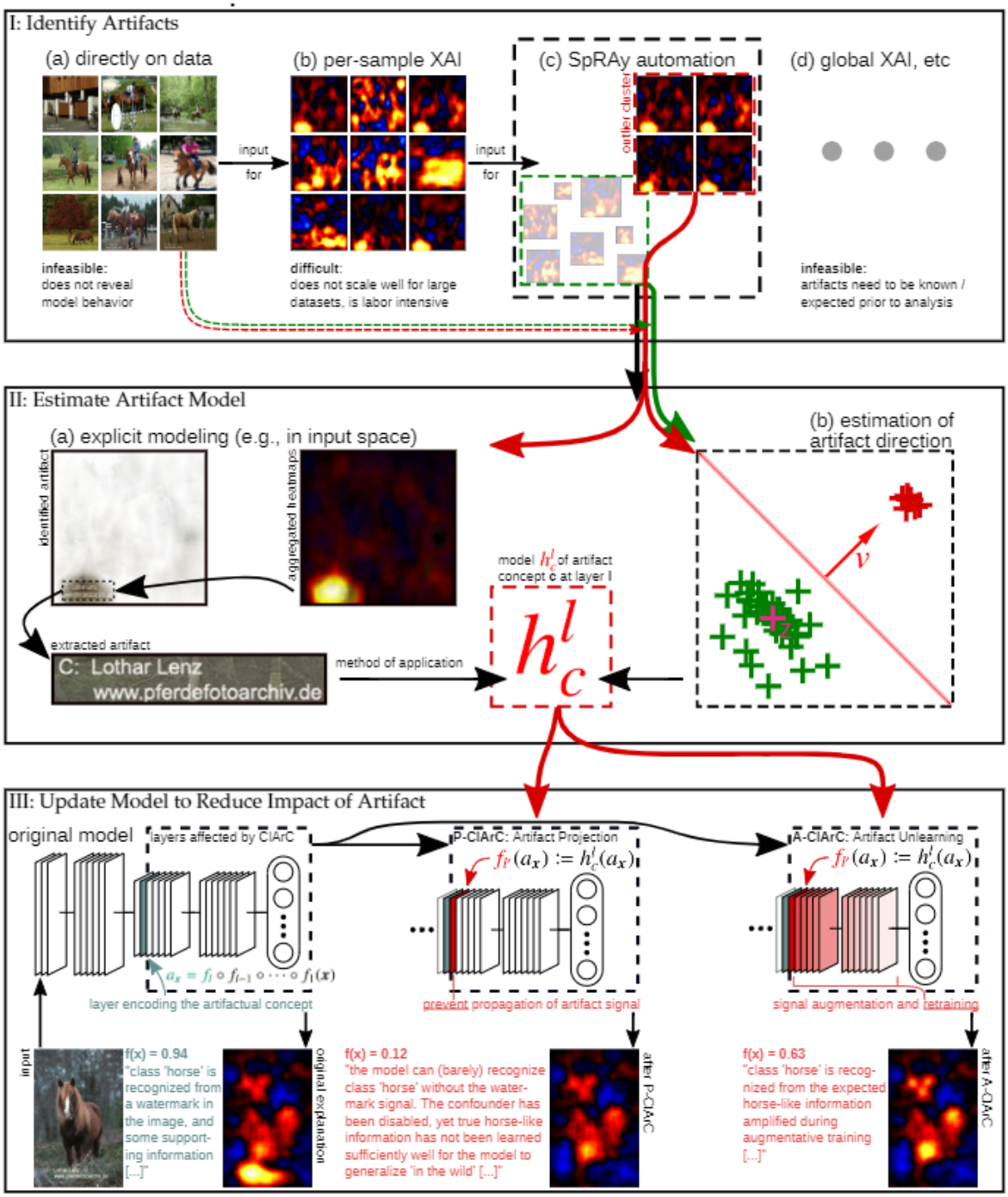
for $R_x \in R$ do
    $R_x \leftarrow maybe\_preprocess(R_x)$;
end

/* tính affinity và laplacian */

$A = affinity(R)$;
$L = laplacian(A)$;

Tính toán các đại lượng phân tích và trực quan

$\Lambda, \Phi, Y = spectral(L)$;
$\tau = FDA(\Phi, Y)$
$Z = visualization\_embedding(\Phi)$



Ba bước loại bỏ Hans. Nguồn: Anders et al.[14]

[14]Finding and Removing Clever Hans: Using Explanation Methods to Debug and Improve Deep Models: [https://arxiv.org/pdf/1912.11425.pdf]

Phần 5: Explanation-Guided Training

Mục tiêu: hướng dẫn mô hình dựa trên bằng chứng được học từ hình ảnh khi dự đoán những từ phổ biến.

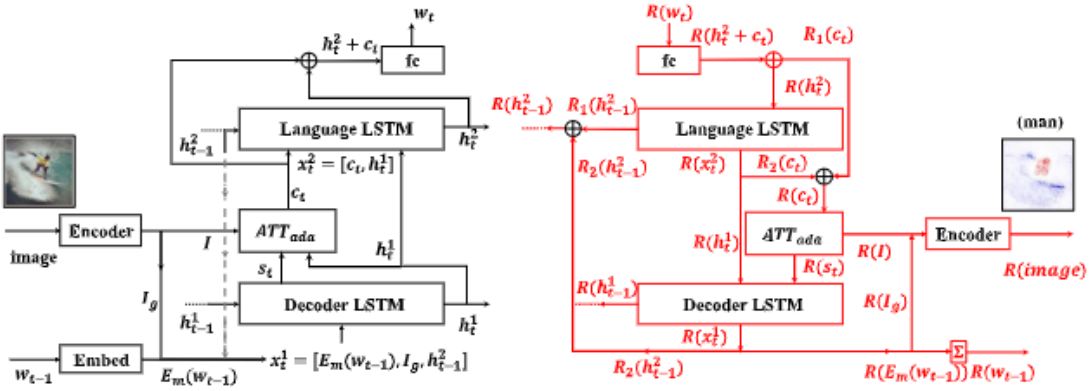
$$\mathcal{L} = \lambda \mathcal{L}_{ce}(p, y) + (1 - \lambda)\mathcal{L}_{ce}(\hat{p}, y)$$

Trong bài báo của Sun et al. 2021[15], tác giả sử dụng adaptive attention và multi-head attention cho các image captioning models. Mô hình Ada-LSTM sử dụng một module adaptive attention và 1 LSTM nối với 1 lớp fully-connected để đoán từ. Mô hình MH-FC sử dụng một module multi-head attention với 1 lớp fully-connected với chức năng tương tự Ada-LSTM. Hàm loss được sử dụng ở cả 2 mô hình là $\mathcal{L} = \mathcal{L}_{ce}(p, y)$ với $p = (p_t)_{t=0}^{l}$ là điểm dự đoán trên tập từ điển, l là độ dài câu và y là nhãn thực tế. Mô hình sau đó được tối ưu bằng thuật toán Self-critical Sequence Training (SCST). SCST tối ưu trên các hàm đánh giá bất khả vi (VD: CIDEr) bằng học tăng cường:

$$R = \mathbb{E}_{St, S^{greedy} \sim p}[\text{metric}(S^s, S^{gt}) - \text{metric}(S^{greedy}, S^{gt})]$$

với R là phần thưởng, S^s là câu được chọn từ phân phối được dự đoán $p = (p_t)_{t=0}^{l}$, S^{greedy} là câu dự đoán bằng greedy search, S^gt là câu chú thích mẫu.

- Với Grad-CAM và Guided Grad-CAM, các phương thức này sẽ lan truyền ngược gradient của một dự đoán về visual feature (tạm dịch: đặc trưng thị giác) I, được biểu diễn bởi $g(I) \in \mathbb{R}^{n_v \times d_v}$. Sau đó, ta có được ma trận trọng số theo thứ tự channel từ $g(I)$ cho đặc trưng thị giác I, với $w_I = \Sigma_{k=1}^{n_v} g(I)_{(k)} \in \mathbb{R}^{d_v}$. I sau đó được tính tổng theo chiều của đặc trưng, trọng số $w(I)$ để tạo ra activation map phản ánh độ quan trọng của các pixel trên bản đồ đặc trưng, $\text{CAM} = ReLU(\Sigma_{k=1}^{d_v} w_{I_{(k)}} I_{(k)}) \in \mathbb{R}^{n_v}$. Grad-CAM tái định hình (reshape) và tăng cường mẫu (up-samples) lớp activation map để tạo ra lời giải thích cho hình ảnh. Để tạo ra hình ảnh giải thích mịn và có độ phân giải cao, Grad-CAM nhân theo thứ tự phần tử (element-wise) với GuidedBackpropagation, do đó nó có tên là Guided Grad-CAM.
- Với LRP, nhóm tác giả xem cơ chế attention là một sự kết hợp tuyến tính trên tập các đặc trưng với trọng số theo thứ tự thành phần (component-wise) sao cho LRP relevance scores không lan truyền ngược thông qua trọng số mà chỉ thông qua các đặc trưng. Bằng cách này, nhóm tác giả có thể áp dụng trực tiếp LRP để phân phối relevance score của biểu diễn ngữ cảnh (context representation) về các đặc trưng thị giác dựa theo trọng số attention và bỏ qua các tính toán trong cơ chế attention.



[15]Explain and improve: LRP-inference fine-tuning for image captioning models:[https://www.sciencedirect.com/science/article/pii/S1566253521001494]

Mô hình Ada-LSTM gốc và khi áp dụng LRP. Nguồn: Sun et al.[16]

Ví dụ với mô hình Ada-LSTM, LRP tuân theo cấu trúc gốc của mô hình và di chuyển cùng hướng với lan truyền ngược của gradient (ngoại trừ cơ chế attention) dọc theo các cạnh của đồ thị có hướng không tuần hoàn. Điểm khác biệt là thay thế đạo hàm từng phần trên cạnh bằng các luật tái phân phối LRP dựa trên deep Taylor framework. Đầu tiên khởi tạo một relevance score cho từ mục tiêu, $R(w_T)$, từ kết quả của tầng fully-connected. Các luật LRP được tính toán trên tất cả các tầng $\text{fc}, \oplus, \text{Language LSTM}, ATT_{ada}, \text{Decoder LSTM}$, và Encoder. Với mỗi từ được giải thích, LRP gán một giá trị relevance score cho tất cả pixel của ảnh đầu vào ($R(image)$) và mỗi từ trong chuỗi đầu vào ($R(w_{T-1}), ..., R(w_1)$). Ta có thể trực quan hóa lời giải thích bằng heatmap khi tính trung bình $R(image)$ qua chiều các channel. Relevance score của một từ là tổng relevance score của các từ trước trên word embedding.

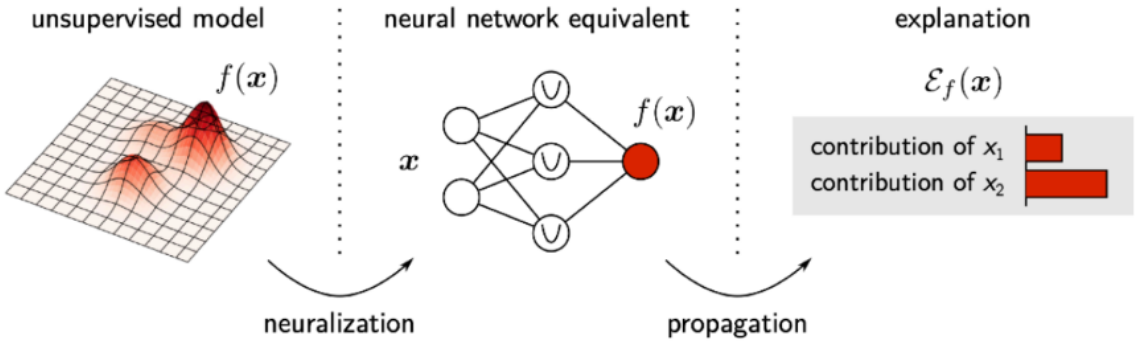
Phần 6: Giải thích các mô hình phi nơ-ron (Explaining Non-Neural Models) [17]

Về nguyên tắc, chúng ta có thể sử dụng các kỹ thuật trong XAI để có thể đưa ra lời giải thích như Predict difference analysis, hoặc LIME, đây đều là những kỹ thuật vượt trội để áp dụng cho bất kì mô hình để đưa ra lời giải thích. Tuy nhiên, những cách tiếp cận trên sẽ đòi hỏi phải đánh giá nhiều lần để kiểm tra tác động của các input đầu vào (do đầu vào đã được xáo trộn perturbation), dẫn đến việc chậm chạp khi dữ liệu có nhiều chiều. Ngoài ra, việc perturbation cục bộ có thể mô tả không trung thực về tổng thể đóng góp của một đặc trưng trong việc ảnh hưởng đến sự quyết định gom cụm của mô hình.

Ý tưởng: Chuyển những mô hình học máy không giám sát thông thường thành mạng nơ-ron và sau đó tìm lời giải thích cho chúng

Đối với các thuật toán học máy (như SVM, K-means,..), tiến hành giải thích bằng 2 bước sau:

- Chuyển đổi các mô hình đó thành một mạng nơ-ron (còn gọi là nơ-ron hóa mô hình)
- Giải thích mạng nơ-ron với các phương pháp lan truyền (LRP).



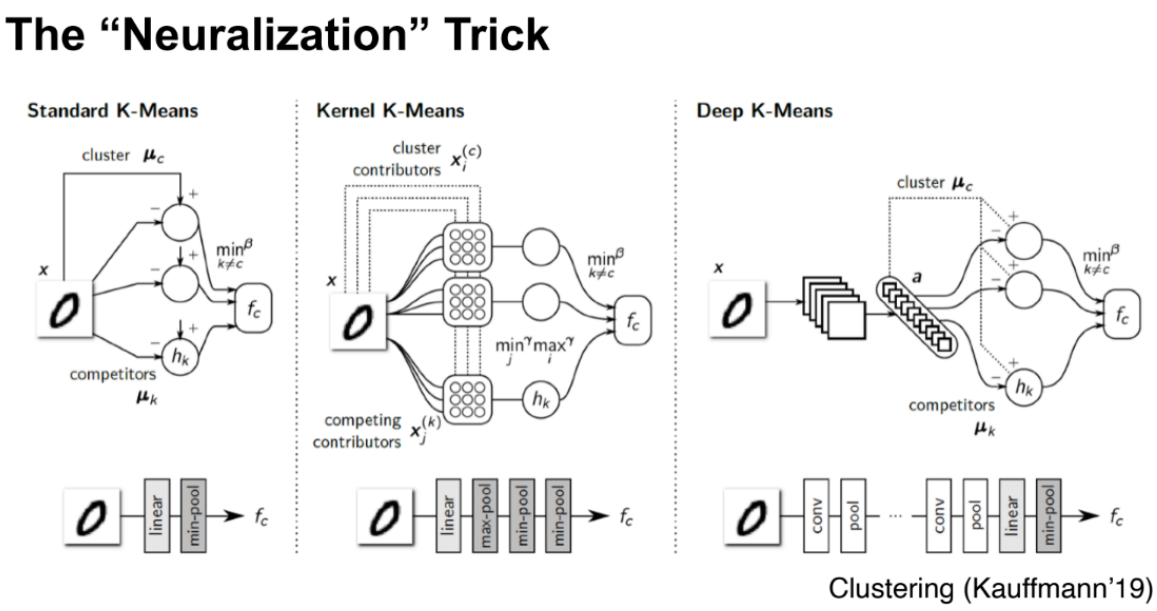
Các bước thực hiện kĩ thuật neuralization trick. Nguồn: Wojciech Samek et al.[18]

Nói tóm lại, từ một mô hình học máy không giám sát, ta tiến hành nơ-ron hóa mô hình đấy thu được một mạng nơ-ron tương ứng. Từ mạng nơ-ron thu được, tiến hành phương pháp lan truyền ngược để biến đổi mạng nơ-ron thành một hàm số (phương trình). Dựa vào hàm số thu được ta có thể đánh giá được tính thiết yếu của từng đặc trưng cũng như tìm được lời giải thích cho mô hình.

[16] Explain and improve: LRP-inference fine-tuning for image captioning models:[https://www.sciencedirect.com/science/article/pii/S1566253521001494]

[17] From Clustering to Cluster Explanations via Neural Networks:[https://arxiv.org/pdf/1906.07633.pdf]

[18] Explainable AI - Methods, Applications & Recent Developments - Dr. Wojciech Samek | ODSC Europe 2019: [https://youtu.be/AFC8yWzypss]



Các kĩ thuật Neuralization trick. Nguồn: Jacob Kauffmann et al.[19]

1. Standard K-Means

A. Neuralization of the Cluster Assignment

Hàm quyết định có thể được tái tạo bởi một mạng nơ-ron hai lớp bao gồm một lớp tuyến tính (standard linear) và một lớp min-pooling:

Neuralized K-means:

$$h_k = w_k^T x + b_k \qquad \text{(layer 1)} f_c = min_{k \neq c}\{h_k\} \qquad \text{(layer 2)}$$

Lớp đầu tiên (standard linear) tương ứng với một tập hợp các hàm tuyến tính được tinh chỉnh với các trọng tâm của các cụm khác nhau. Lớp thứ hai (min-pooling) sẽ quyết định chọn hàm tuyến tính nào để kích hoạt tại một vị trí nhất định. Hai lớp này cùng nhau tạo nên một hàm tuyến tính từng phần (nghĩa là ở mỗi vị trí khác nhau, sẽ sử dụng một hàm tuyến tính khác nhau sao cho phù hợp).

B. Propagation of the Cluster Assignment

Sau khi đi qua mạng nơ-ron 2 lớp (standard layer và min-pooling layer), đầu ra của mạng là hàm f_c - đóng vai trò như bước đầu tiên cho các nơ-ron trong lớp trung gian $(h_k)_k$ bằng cách lan truyền qua các hàm tối tiểu. Đảm bảo tuân theo chiến lược đầu vào nhỏ nhất sao cho hàm số nhận được phần lớn nhất để phân phối lại:

$$R_k = \frac{exp(-\beta h_k)}{\Sigma_{k \neq c} exp(-\beta h_k)} f_c$$

Trong đó: - R_k (điểm số trung gian): là mức độ phù hợp của nơ-ron h_k với cluster assignment f_c

- β: là một siêu tham số độ cứng. Tham số độ cứng nội suy giữa một chiến lược phân phối đều ($\beta = 0$) và chiến lược tận dụng tối thiểu ($\beta \to \infty$).

[19] From Clustering to Cluster Explanations via Neural Networks:[https://arxiv.org/pdf/1906.07633.pdf]

Lưu ý rằng đối với hai trường hợp của siêu tham số β, cách tiếp cận ở đây cho phép ngữ cảnh hóa lời giải thích (nghĩa là không phân phối lại trên các cụm đối thủ khác nhau quá xa và không liên quan), đồng thời đảm bảo tính liên tục của lời giải thích khi chúng ta chuyển từ một cụm đối thủ này sang một cụm đối thủ khác. Có thể dụng heuristic để chọn siêu tham số β:

$$\beta = \mathbb{E}[fc]^{-1}$$

Trong đó, kỳ vọng được tính toán trên toàn bộ tập dữ liệu. Nói cách khác, xét f_c như là một điểm số "điển hình" trong pool, và tham số độ cứng β sẽ tỉ lệ nghịch với f_c

Tiếp theo, xem xét làm thế nào để tiếp tục phân phối lại điểm số trung gian R_k cho lớp đầu vào, trong đó kích thước sẽ tương ứng với các đại lượng quan sát (được giả định rằng người dùng có thể hiểu được). Để đạt được điều này, ta có thể áp dụng luật lan truyền LRP:

$$R_k = \Sigma_{k \neq c} \frac{(x_i - m_{ik}) w_{ik}}{\Sigma_i (x_i - m_{ik}) w_{ik}} R_k$$

Trong đó, $m_k = (\mu_c + \mu_k)/2$ là trung điểm giữa trọng tâm của cụm lợi ích và cụm đối thủ cạnh tranh. Nói cách khác, chính là gán các kích thước nơi các điểm đầu vào kích hoạt so với trung điểm $(x - m_k)$ phù hợp với phản hồi của mô hình w_k.

2. Kernel K-Means

Thuật toán phân cụm standard K-Means có một số giới hạn nhất định về khả năng biểu diễn, vì nó chỉ cho phép biểu diễn các cụm có thể phân tách tuyến tính theo cặp. Mô hình kernel K-Means là một phần mở rộng đơn giản của K-Means, khi đó dữ liệu được ánh xạ lần đầu tiên tới một không gian đặc trưng (feature space) thông qua một số map $x \to \Phi(x)$ được tạo ra bởi một số hàm kernel $K(x, u)$. Hàm quyết định được tính bằng kernel K-means:

$$\forall_{k \neq c} : \|\Phi(x) - \mu_c\|^2 < \|\Phi(x) - \mu_k\|^2 (1)$$

Trong đó, trọng tâm được xác định trong không gian đặc trưng (featrure space)

Nếu ta áp dụng cùng một cách giải thích như trong phần trước, ta sẽ có được một lời giải thích về mặt kích thước của feature space, và sau đó chúng ta sẽ cần phải lan truyền ngược mở rộng thông qua feature map Φ. Ta có thể thay thế một công thức trực quan hơn (cụ thể cho trường hợp Gaussian kernel), khi mà khoảng cách đến một cụm cụ thể được mô hình hóa bằng khoảng cách tối thiểu mềm (soft minimum distance) đến các điểm dữ liệu thành viên khác trong cùng một cụm. Cụ thể, ta sẽ có công thức cho hàm quyết định thay thế cho công thức bên trên:

$$\forall k \neq c : LME^{-\gamma}_{i \in C_c} \|x - u_i\|^2 < LME^{-\gamma}_{j \in C_k} \|x - u_j\|^2 (2)$$

Trong đó,

- $(u_i)_i$ và $(u_j)_j$ là một tập hợp các điểm dữ liệu đại diện cho hai cụm
- $C_c, C_k \subset \mathbb{N}$ là các tập chỉ số không trùng lặp của các vector support đại diện cho các cụm này và khi đó $LME^{-\gamma}$ có giá trị tổng quát F-mean với $F(t) = e^{-\gamma t}$

$$LME^{-\gamma}_{i \in C} s_i = -\frac{1}{\gamma} \Big(\frac{1}{|C|} \Sigma_{i \in C} exp(-\gamma s_i) \Big) (3)$$

Phần phía sau có thể hiểu là một soft min-pooling và chúng sẽ hội tụ về hard min-pooling khi $\gamma \to \infty$.

Hàm quyết định theo công thức (2) có thể được tái biểu diễn bằng một mạng nơ-ron 4 lớp (bao gồm 1 lớp tuyến tính và 3 lớp pooling phia sau)

Neuralized kernel K-means:

$$h_{ijk} = w_{ij}^T x + b_{ij} \qquad \text{(layer 1)}$$
$$h_{jk} = LME^{\gamma}_{i \in C_c}\{h_{ijk}\} \qquad \text{(layer 2)}$$
$$h_k = LME^{-\gamma}_{j \in C_k}\{h_{jk}\} \qquad \text{(layer 3)}$$
$$f_c = min_{k \neq c}\{h_k\} \qquad \text{(layer 4)}$$

Khi $w_{ij} = 2(u_i - u_j)$ và $b_{ij} = \|u_j\|^2 - \|u_i\|^2$ thì LME^{γ} và $LME^{-\gamma}$ có thể được xem như là soft max-pooling và soft min-pooling, và được gán cho cụm c nếu $f_c(x) > 0$.

Mạng nơ-ron vừa đề xuất có thể được dùng để hỗ trợ quy trình giải thích. Bởi vì mạng nơ-ron này bao gồm lớp tuyến tính và lớp pooling, luật lan truyền ngược được đề xuất cho trường hợp K-means vẫn được áp dụng.

3. Deep K-means

Không giống như kernel K-means, deep K-mean sử dụng feature map được cho trước dưới dạng một chuỗi các ánh xạ theo lớp $\Psi(x) = \Psi_L \circ .... \circ \Psi_1(x)$ và feature map thường được học thông qua việc lan truyền ngược để tạo ra các cụm có kết cấu như mong muốn.

Deep K-means sẽ sử dụng hàm quyết định giống mô hình với K-means standard nhưng thay bằng không gian đặc trưng (feature space):

$$\forall_{k \neq c} : \|\Psi(x) - \mu_c\|^2 < \|\Psi(x) - \mu_k\|^2$$

Neuralized Deep K-means:

$$a = \Psi_L \circ ... \circ \Psi_1(x) \qquad \text{(layer 1...L)}$$
$$h_k = w_k^T a + b_k \qquad \text{(layer L+1)}$$
$$f_c = min_{k \neq c}\{h_k\} \qquad \text{(layer L+2)}$$

Trong đó, $w_k = 2(\mu_c - \mu_k)$ và $b_k = \|\mu_k\|^2 - \|\mu_c\|^2$

Phần 7: Kỹ thuật Pruning và Quantization dựa vào XAI (XAI-Based Pruning and Quantization)

Ngày nay, sự phát triển của mạng nơ-ron nhân tạo (DNN) đã tạo được nhiều thành công trong các lĩnh vực khác nhau. Tuy nhiên, đi đôi với việc mạng học sâu ngày càng trở nên phức tạp là sự gia tăng các tham số cũng như là các phép tính toán cần thực hiện. Sự gia tăng về bộ nhớ và nhu cầu tính toán khiến việc học sâu trở nên khó khăn đối với các thiết bị có nên tảng phần cứng hạn chế (như là điện thoại di động, các thiết bị ngoại vi, IoT,...). Do đó nhu cầu giảm thiểu chi phí chung đồng thời duy trì hiệu suất của mô hình rất được quan tâm (bao gồm kỹ thuật rút gọn tham số, quantization, pruning, kỹ thuật nén không mất mát dữ liệu,...).

XAI-Based Pruning [20]

Để có thể giảm bộ nhớ lưu trữ cũng như chi phí tính toán đối với các mạng DNN, kỹ thuật cắt tỉa mạng nơ-ron được đánh giá rất cao.

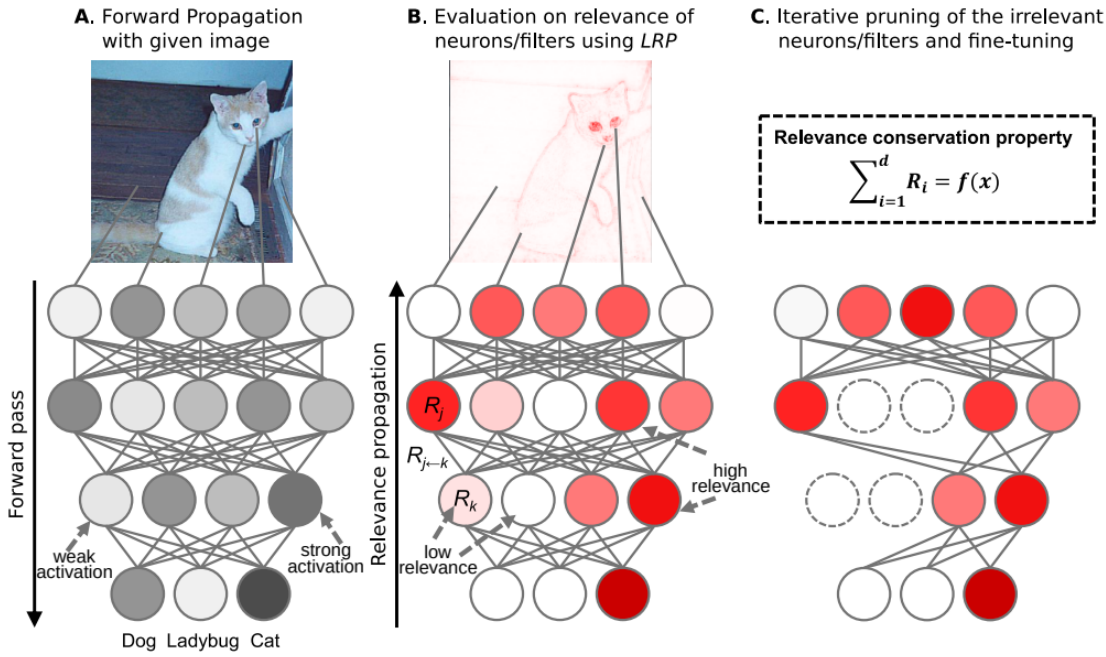
Ý tưởng phương pháp pruning: Cắt tỉa (loại bỏ) để loại bỏ các tập con của các đơn vị mạng nơ-ron (trọng số hoặc filter) ít quan trọng. Xác định tập con không liên quan (không có ảnh hưởng cao trong mạng) của các tham số để thực hiện xóa chúng.

XAI với Pruning

Kỹ thuật cắt tỉa này kết hợp với việc sử dụng XAI (cụ thể là LRP). LRP được phát triển như một phương pháp để giải thích việc gán các điểm quan trọng. Với các kích thước đầu vào khác nhau của mạng nơ-ron phản ánh sự đóng góp của kích thước đầu vào đấy đến quyết định của mô hình. Mức độ liên quan (relevance) được lan truyền ngược từ đầu ra (output) đến đầu vào (input) và được gán cho từng đơn vị trong mạng. Bởi vì điểm số liên quan (relevance scores) được tính cho mọi lớp từ đầu ra đến đầu vào nên những điểm số này về cơ bản sẽ phản ánh tầm quan trọng của từng đơn vị riêng lẻ trong một mô hình và sự đóng góp của nó bên trong mạng. Do đó đây có thể được xem là một phương pháp tiềm năng được dùng làm tiêu chí để cắt tỉa.

LRP-Based Network Pruning

Ý tưởng chính trong việc cắt tỉa các đơn vị trong mạng nơ-ron là sử dụng "the relevance quantity computed with LRP". LRP sẽ phân tách một quyết định phân loại thành những đóng góp của từng đơn vị trong mạng đến sự phân loại tổng thể, gọi là "Mức độ liên quan" (Relevances).



Vận dụng LRP vào mạng học sâu trong bài toán phân lớp. Đơn vị nào có mạng càng đậm thì càng đóng góp nhiều vào quyết định phân lớp của mạng. Nguồn: Seul-Ki Yeom et al. [21]

Tính toán kích thước đầu vào của mạng CNN và thể hiện chúng dưới dạng biểu đồ nhiệt. Các mức độ liên quan (Relevances) sẽ làm nổi bật những đầu vào quan trọng (có ảnh hưởng lớn đến quyết định

[20] Pruning by Explaining: A Novel Criterion for Deep Neural Network Pruning: [https://arxiv.org/pdf/1912.08881.pdf]
[21] Pruning by Explaining: A Novel Criterion for Deep Neural Network Pruning: [https://arxiv.org/pdf/1912.08881.pdf]

phân loại).

Đặc điểm chính của LRP là lan truyền ngược qua mạng mà trong đó đầu ra của mạng được phân phối lại cho tất cả các đơn vị của mạng theo kiểu từng lớp. Lan truyền ngược như thế này có cấu trúc tương tự như lan truyền ngược gradient (gradient backpropagation) nên thời gian chạy là tương tự như nhau. Sự phân phối lại này sẽ dựa trên nguyên tắc bảo tồn sao cho các mối liên quan có thể ngay lập tức được hiểu là sự đóng góp mà một đơn vị trong mạng đóng góp vào đầu ra của mạng.

Explainability-Driven Quantization [22]

Đối với việc các phép tính toán trong các mạng học sâu, hầu hết tham số được tính toán trong các phần cứng như CPU, GPU đều có dạng số chấm động 32 bit. Điều này dẫn đến chi phí tính toán cao, tiêu thụ nhiều điện năng, độ trễ lớn và nhu cầu bộ nhớ cấp phát rất cao. Bắt nguồn từ vấn đề đó, người ta đã đề xuất ra phương pháp quantization.

Ý tưởng phương pháp quantization: Kỹ thuật quantization được áp dụng để giảm thiểu số bit cần dùng để biểu diễn các tham số, trọng số bằng cách ánh xạ các giá trị tương ứng thành một tập hữu hạn gồm các mức quantization rời rạc (các cụm). Nếu có n cụm, ta chỉ cần $log_2 n$ bit để biểu diễn một điểm dữ liệu, nghĩa là càng N càng nhỏ (càng ít cụm) thì càng cần ít bit để biểu diễn cho điểm dữ liệu đó. Tuy nhiên, việc liên tục giảm số cụm (giảm n) có thể dẫn đến việc gây ra lỗi và hiệu suất của mô hình ngày càng giảm.

XAI với Quantization

Các mô hình XAI có thể được áp dụng để tìm các đặc trưng liên quan ở đầu vào. Bằng cách sử dụng thông tin về mức độ phù hợp của trọng số từ LRP (được nhắc đến ở phần đầu). Do đó, phương pháp này sẽ kết hợp hai khái niệm XAI và lý thuyết thông tin, nghĩa là thay vì chỉ gán các giá trị trọng số dựa trên khoảng cách của chúng tới các cụm quantization tương ứng, giờ đây mô hình sẽ xét thêm mức độ liên quan của trọng số dựa trên LRP rồi mới quyết định gán một giá trị trọng số vào cụm quantization nào.

Phần 8: Kết luận - Tương lai của XAI (Conclusion - Future of XAI) [23]

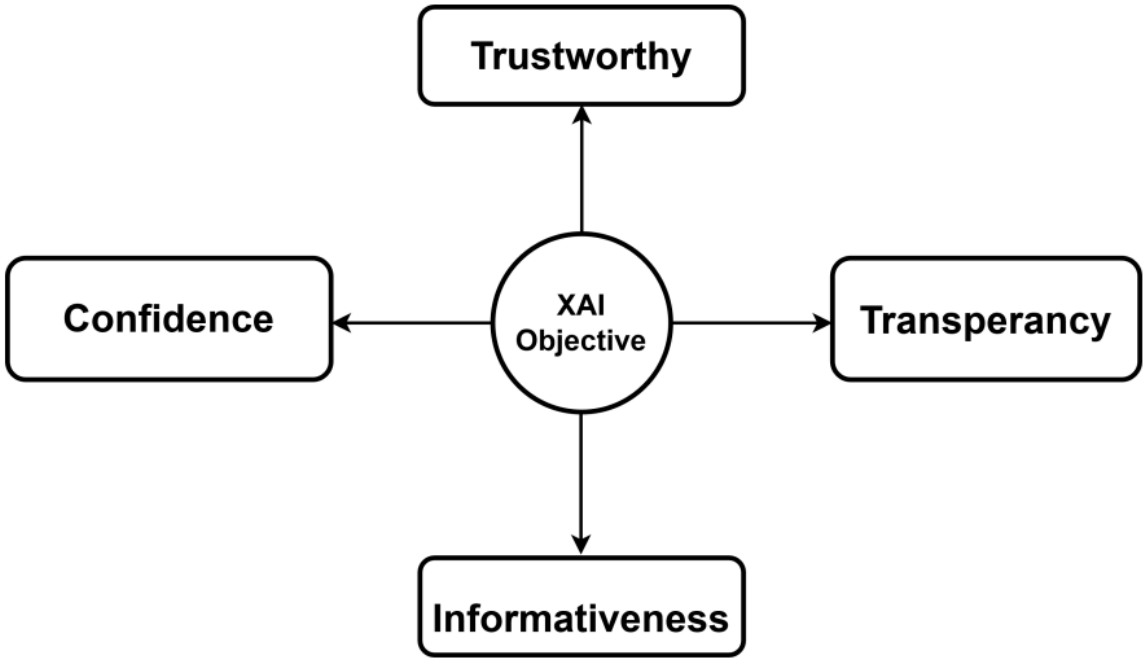
Trong những năm gần đây, các ứng dụng của trí tuệ nhân tạo AI đã được sử dụng trong rất nhiều lĩnh vực như khoa học, tài chính, kinh doanh, mạng xã hội, và đạt được nhiều thành công rực rỡ. Các thuật toán dựa trên AI đã được áp dụng thành công cho tất cả các loại hình dữ liệu (như văn bản, hình ảnh, âm thanh, video) trong các lĩnh vực khác nhau, chẳng hạn như chăm sóc sức khỏe, quốc phòng, luật pháp và trật tự, quản trị, công nghiệp tự trị,... Thuật toán AI giờ đây có thể hiệu quả giải quyết một số phân loại, hồi quy, phân cụm, học chuyển giao hoặc vấn đề tối ưu hóa.

Trong đó, việc tìm các lời giải thích cho các mô hình đóng một vai trò rất quan trọng để đảm bảo được sự tin cậy cũng như tính minh bạch của các mô hình. Một ví dụ trong lĩnh vực y tế, khi bác sĩ muốn chắc chắn về kết quả dự đoán bệnh của một mô hình AI, họ phải kiểm tra lại bằng cách phân tích ảnh chụp CT (một phương pháp đòi hỏi tốn nhiều chi phí và thời gian) bởi vì mô hình AI thường không chính xác 100%. Do đó, nếu có một cái nhìn chi tiết về cách đưa ra kết quả của mô hình AI, sẽ giúp cho bác sĩ có thể đánh giá được độ tin cậy của kết quả dự đoán, đồng thời giúp tránh được nguy hiểm đến tính mạng của bệnh nhân nếu xảy ra lỗi.

Các mô hình XAI sẽ có thể trả lời được các câu hỏi "Wh" (what, why, when - cái gì, tại sao, khi nào) mà các mô hình truyền thống khó có thể trả lời được. Từ đó, các mô hình XAI có thể được ứng dụng trong các lĩnh vực như chăm sóc sức khỏe, quốc phòng, luật,...- đều là những lĩnh vực đòi hỏi sự tin cậy, minh bạch và con người có thể lý giải được.

[22] ECQx: Explainability-Driven Quantization for Low-Bit and Sparse DNNs:[https://arxiv.org/pdf/2109.04236.pdf]
[23] Explainable AI: current status and future directions: [https://arxiv.org/pdf/2107.07045.pdf]

XAI AS A TOOL TO OPEN BLACK BOX



Bốn yếu tố chính của XAI. Nguồn: Prashant Gohel et al.[24]

Mục tiêu của XAI

a. Tính minh bạch và thông tin

Các mô hình XAI có thể nâng cao sự công bằng cũng như tính minh bạch bằng cách đưa ra lời giải thích mà người bình thường cũng có thể hiểu được. Tính minh bạch là yếu tố quan trọng để đánh giá hiệu quả hoạt động của một mô hình XAI và những lý giải của nó.

b. Tính tin cậy

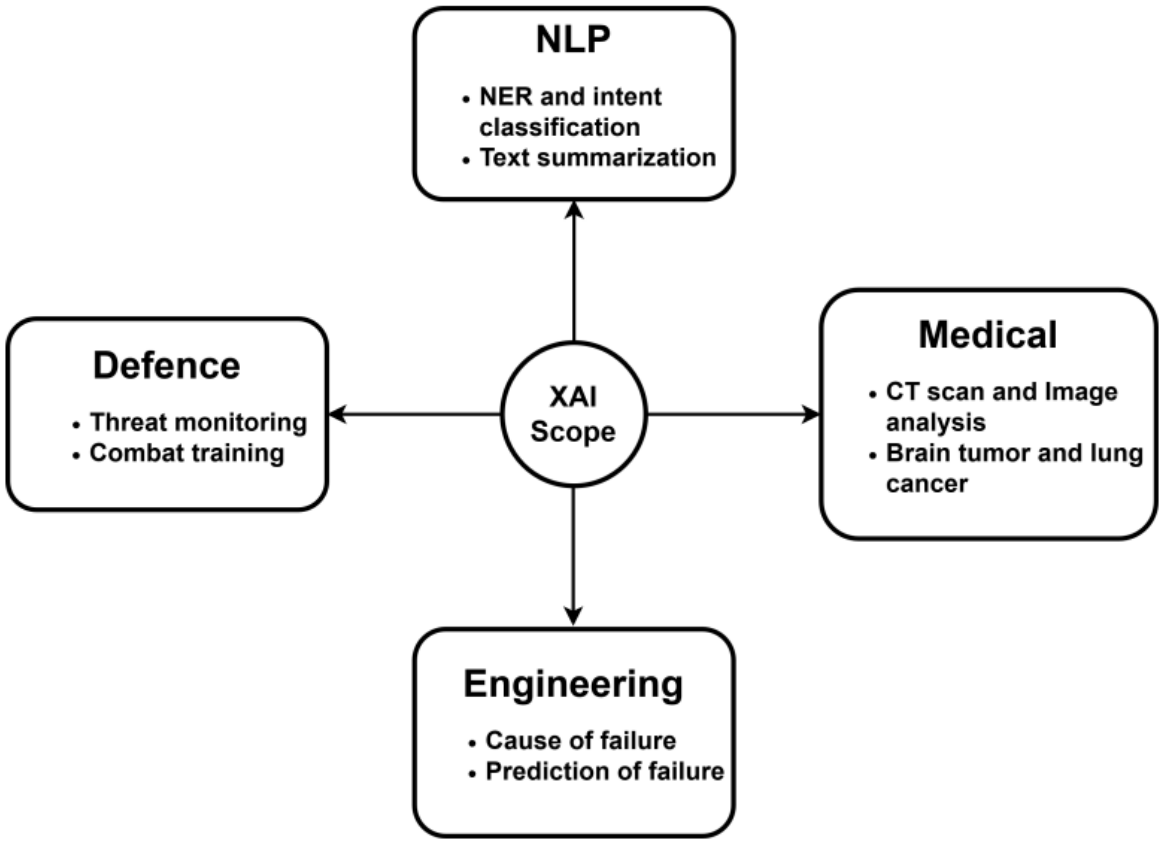
Sự tin cậy là một trong những yếu tố quan trọng khiến con người quyết định sử dụng, phụ thuộc vào một công nghệ nào đó. Con người sẽ càng tin cậy các thuật toán AI/ML nếu chúng đưa ra được những lời giải thích xác thực, hợp lý và khoa học.

c. Tính công bằng

Bias-variance trade off trong các mô hình AL/ML giúp XAI gia tăng sự công bằng và giúp giảm thiểu bias (bias-variance trade off) của dự đoán tại thời điểm giải thích, diễn giải.

Tầm vực của XAI

[24]Explainable AI: current status and future directions: [https://arxiv.org/pdf/2107.07045.pdf]



Bốn lĩnh vực mà XAI đóng vai trò quan trọng. Nguồn: Prashant Gohel et al.[25]

XAI có thể được ứng dụng trong nhiều lĩnh vực khác nhau, nhưng hầu hết đều là những lĩnh vực đòi hỏi sự tin cậy, minh bạch và con người có thể hiểu được:

a. Xử lý ngôn ngữ tự nhiên (NLP): Các mô hình XAI có thể đem đến sự tin cậy trong các bài toán phân loại từ ngữ, văn bản, phát hiện tin giả mạo, phát hiện lừa đảo. Góp phần tăng tính tự động hóa trong thời đại công nghiệp 4.0.

b. Y tế (Medical): XAI có thể phân tích tình trạng, bệnh tình của bệnh nhân bằng cách quan sát lịch sử bệnh tật của họ. Bằng cách sử dụng các thuật toán AI/ML trong xử lý hình ảnh y học, các chuyên gia y tế có thể dễ dàng phát hiện các bệnh nguy hiểm ở giai đoạn sớm nhất (khối u ác tính, bệnh về phổi, bệnh da liễu, xương tủy,...).

c. Quốc phòng (Defence): Ngày nay trong chiến tranh, vũ khí tự động hóa và tên lửa dẫn đường tự động, máy bay không người lái ngày càng thể hiện uy lực mạnh mẽ, do đó việc phát triển XAI trong lĩnh vực quốc phòng giúp các thiết bị, vũ khí chiến đấu có tính hiệu quả cao hơn, chính xác hơn và ổn định hơn.

d. Ngân hàng (Banking): Các hệ thống ngân hàng là một trong những cơ sở quan trọng nhất của một quốc gia nói riêng và toàn thế giới nói chung. Trong thời đại số hóa hiện nay, nhiều kẻ xấu đã lợi dụng lỗ hổng để lừa đảo, gian lận các giao dịch nhằm trục lợi cho bản thân. Việc phát triển một mô hình XAI tốt có thể giúp để phân tích, điều tra được những giao dịch gian lận và giảm các thông tin sai lệch nhằm lừa đảo khách hàng.

[25] Explainable AI: current status and future directions: [https://arxiv.org/pdf/2107.07045.pdf]

Nói tóm lại, XAI là một lĩnh vực quan trọng và tiềm năng về lợi ích trong cả nghiên cứu lẫn thực tế. XAI sẽ ngày càng chứng minh được khả năng vượt trội các mô hình ML/AI truyền thống vốn không thể giải thích được kết quả.

Reference

1. Explainable AI - Methods, Azpplications & Recent Developments - Dr. Wojciech Samek | ODSC Europe 2019: [https://youtu.be/AFC8yWzypss]
2. Layer-wise Relevance Propagation: An overview [https://iphome.hhi.de/samek/pdf/MonXAI19.pdf]
3. Unmasking Clever Hans predictors and assessing what machine really learn:[https://www.nature.com/articles/s41467-019-08987-4]
4. Analyzing ImageNet with Spectral Relevance Analysis: Towards ImageNet un-Hans'ed:[https://www.semanticscholar.org/paper/ImageNet-with-Spectral-Relevance-Towards-Anders-Marinc/5db1b78d1cc34ee388392c3214cd05ed7fe4317f]
5. Finding and Removing Clever Hans: Using Explanation Methods to Debug and Improve Deep Models: [https://arxiv.org/pdf/1912.11425.pdf]
6. Explain and improve: LRP-inference fine-tuning for image captioning models:[https://www.sciencedirect.com/science
7. From Clustering to Cluster Explanations via Neural Networks:[https://arxiv.org/pdf/1906.07633.pdf]
8. Pruning by Explaining: A Novel Criterion for Deep Neural Network Pruning: [https://arxiv.org/pdf/1912.08881.pdf]
9. ECQx: Explainability-Driven Quantization for Low-Bit and Sparse DNNs: [https://arxiv.org/pdf/2109.04236.pdf]
10. Explainable AI: current status and future directions: [https://arxiv.org/pdf/2107.07045.pdf]
11. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation: [https://doi.org/10.1371/journal.pone.0130140]