Đề xuất paper về XAI

- Các công trình về XAI cho tới hiện tại mình đọc có xu hướng đi tính mức độ ảnh hưởng của input (hoặc feature của input) đối với output. Từ việc tính toán này, có thể suy ra mức quan trọng của input đối với output.
- Nói cách khác, mô hình đang nhìn vào những input quan trọng để dự đoán output.

TCAV - Giải thích dựa trên khái niệm

- Source: https://arxiv.org/pdf/1711.11279.pdf
- Nội dung:
  - Khi nhìn vào con ngựa vằn, một lời giải thích do con người sinh ra, trả lời cho câu hỏi: "Tại sao đây lại là con ngựa vằn?" có thể là: "Vì nó có các vằn".
  - TCAV giả định rằng không gian đặc trưng của mô hình máy học E_m có thể được diễn giải bằng cách ánh xạ về không gian đặc trưng của con người (bao gồm các khái niệm cấp cao mà con người có thể hiểu được) E_h. Từ đó, bài toán đặt ra là tìm hàm số g:

$$g : E_m \rightarrow E_h$$

  - Sau khi học được các khái niệm từ hàm g bên trên, TCAV sẽ đánh giá mức độ quan trọng của khái niệm đó trong việc đưa ra dự đoán. Ví dụ: Khái niệm "vằn" là quan trọng khi đưa ra dự đoán "ngựa vằn".

SHAP - Giải thích dựa trên hướng loại bỏ đặc trưng

- Source: https://papers.nips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
- Nội dung:
  - Giả sử một tấm ảnh được model phân lớp "dog". Ta sẽ thử cắt ngẫu nhiên các phần của tấm ảnh đi và input ngược lại vào model. Nếu model vẫn nhận ra ảnh thuộc về phân lớp "dog" nghĩa là phận bị cắt không thực sự đóng góp cho quá trình phân lớp. Ngược lại, nếu model thực hiện phân lớp sai thì nghĩa là phần bị cắt rất quan trọng với model.

Layer-wise Relevance Propagation - Giải thích dựa trên lan truyền

- Source: https://arxiv.org/pdf/1604.00825.pdf
- Nội dung:
  - Cố gắng tính toán mức độ liên quan của input (ví dụ: từng pixel) với output bằng việc lan truyền ngược chiều forward nhằm mục đích xác định được input quan trọng đối với model trong việc dự đoán output.

Kế thừa sự khả diễn giải của mô hình Attention

- Source: https://openaccess.thecvf.com/content/CVPR2021/papers/Chefer_Transformer_Interpretability_Beyond_Attention_Visualization_CVPR_2021_paper.pdf
- Nội dung:
  - Giả định rằng phần dữ liệu quan trọng với mô hình trong việc dự đoán output là phần được chú ý nhiều nhất, có thể dựa trên cơ chế attention để xác định các phần đó.