TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

NHÓM 3

CHUNG KIM KHÁNH (19127644), NGUYỄN ANH TUẤN (21C11040)

BẢN DỊCH ĐỀ TÀI

PHƯƠNG PHÁP XAI – CHỈ DẪN LAN TRUYỀN NGƯỢC (GUIDED BACKPROPAGATION)

Ngành: Khoa học máy tính

Thành phố Hồ Chí Minh - 2022

Phương pháp XAI – Chỉ Dẫn Lan Truyền Ngược (Guided Backpropagation)

Guided Backpropagation là gì?

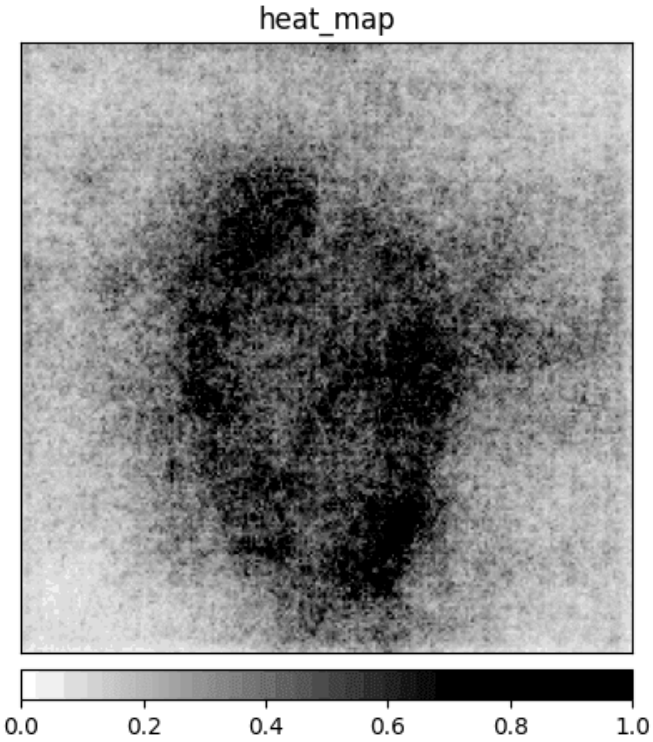
Guided Backpropagation (GBP) [3] là một phương pháp tiếp cận được thiết kế bởi Springenberg và cộng sự, dựa trên các ý tưởng của ***Deconvolution*** [1] và ***Saliency*** [2]. Các tác giả cho rằng phương pháp tiếp cận được thực hiện bởi Simonyan và cộng sự [2] có vấn đề với những luồng gradients âm (flow of negative gradients), làm giảm độ chính xác của các lớp cao hơn mà chúng ta đang cố gắng biểu diễn. Ý tưởng của họ là kết hợp hai phương pháp tiếp cận và thêm một "chỉ dẫn" cho *Saliency* với sự trợ giúp của tích chập ngược.

Để đạt được điều đó, chúng ta cần tập trung vào hàm kích hoạt ***ReLU*** trong CNN. Khi tính giá trị tại thành phần *Rectification* của *deconvnet*, chúng ta tạo mặt nạ toàn bộ giá trị non-positive với *ReLU*. Trong lớp đó, những giá trị đã tính toán được tính bằng cách chỉ dựa vào những tín hiệu đầu (sự tái cấu trúc từ các lớp cao hơn), và đầu vào đã bị bỏ qua. Mặt khác, trong phương pháp *Saliency*, chúng ta đang tập trung vào các giá trị gradient đã được tính dựa trên hình ảnh đầu vào. Nếu chúng ta sử dụng mặt nạ *deconvnet* của lớp *Rectification* và áp dụng nó lên các giá trị gradient của phương pháp *Saliency*, chúng ta có thể loại bỏ nhiễu gây ra bởi các giá trị gradient âm. Sự loại bỏ nhiễu này là lý do tại sao phương pháp có tiền tố "đã chỉ dẫn". Với sự trợ giúp của deconvolution (tích chập ngược) đã chỉ dẫn các giá trị backpropagation (lan truyền ngược) của phương pháp *Saliency* để tạo ra những hình ảnh sắc nét hơn (Hình. 1(c)).

(a) Ảnh gốc.

(b) Kết quả giải mã.

(c) Kết quả GBP



(d) Kết quả độ mặn

Hình 1: Trực quan hóa bản đồ độ mặn (bản đồ saliency) do phương pháp Saliency tạo ra [Hình. 1(d)], Giải mã (Deconvolution) [Hình. 1(b)] và GBP [Hình. 1(c)] của cùng một hình ảnh đầu vào [Hình. 1(a)] cho một lớp "weimaraner". Tất cả các bản đồ

được tạo bằng cùng một mô hình (ResNet18). Nguồn hình ảnh: Stanford Dogs Dataset | Kaggle

Như chúng ta có thể thấy ở Hình 1(c), việc sử dụng "chỉ dẫn" làm giảm đáng kể độ nhiễu gây ra bởi phương pháp Saliency (Hình. 1(d)). ***Ý tưởng GBP thường bị hiểu nhầm và được hiểu là "áp dụng kết quả deconvolution trên kết quả saliency"***. Điều này là không đúng bởi mặt nạ *ReLU* trích xuất từ deconvnet được sử dụng trên mọi cấp độ và do đó ảnh hưởng đến toàn bộ các giá trị gradient xuống đến các đầu vào của CNN, không chỉ ở cấp độ đầu tiên của CNN.

Đọc thêm

Tôi đã quyết định tạo ra một chuỗi các bài báo giải thích các phương pháp XAI quan trọng nhất hiện đang được sử dụng trong thực tế. Đây là bài báo chính: XAI Methods - The Introduction - Blog by Kemal Erdem

Tài liệu tham khảo:

[1] M. D. Zeiler, R. Fergus. Visualizing and Understanding Convolutional Networks, 2013.

[2] K. Simonyan, A. Vedaldi, A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

[3] J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller. Striving for Simplicity: The All Convolutional Net, 2014.

[4] A. Khosla, N. Jayadevaprakash, B. Yao, L. Fei-Fei. Stanford dogs dataset. https://www.kaggle.com/jessicali9530/stanford-dogs-dataset, 2019. Accessed: 2021-10-01.