

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

GVHD: GS.TS. Lê Hoài Bắc

Nhóm 3:

- Chung Kim Khánh (19127644)
- Nguyễn Anh Tuấn (21C11040)

MỤC LỤC

Chương 1:

GIỚI THIỆU

Chương 2:

CÁC CÔNG TRÌNH NGHIÊN
CỨU LIÊN QUAN

Chương 3:

PHƯƠNG PHÁP

Chương 1: Giới thiệu

1. Động lực nghiên cứu

a) Ý nghĩa khoa học

Mục tiêu của bất kỳ thuật toán học có giám sát: phải tìm ra được một hàm ánh xạ tốt nhất, một tập hợp các yếu tố đầu vào tới đầu ra chính xác của nó.

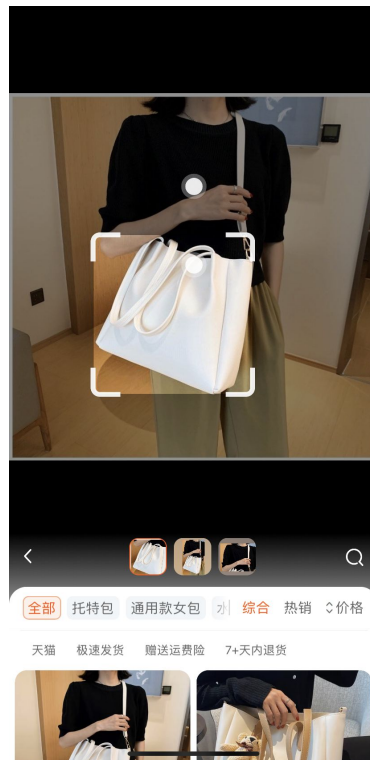
Ví dụ: Tác vụ nhận dạng đối tượng (Object recognition), trong đó, đầu vào là một hình ảnh của một con vật, và đầu ra chính xác sẽ là tên của con vật đó.



Ví dụ: Năm 2017, một hệ thống Image recognition đã “gian lận” bằng cách tìm kiếm các thẻ copyright liên kết với hình ảnh con ngựa thay vì học để nhận biết tại sao đó lại là con ngựa.

b) Ứng dụng thực tiễn

Tìm kiếm bằng hình ảnh: đưa ra gợi ý các hình ảnh/sản phẩm tương tự. Hiện ứng dụng này đang có trên Google, Shopee, Lazada, Taobao...



b) Ứng dụng thực tiễn

Hệ thống camera giám sát: Xác định các thông tin đối tượng và truy vết đối tượng → Phòng chống tội phạm.



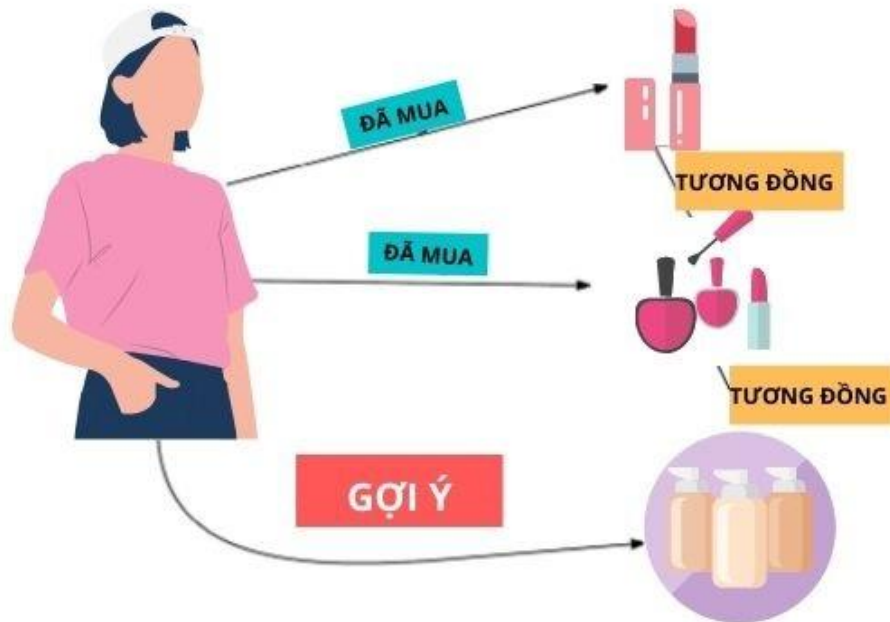
b) Ứng dụng thực tiễn

Hệ thống chấm công: Áp dụng nhận biết khuôn mặt, tuổi, vân tay trong các công ty → nhân viên không cần mang theo thẻ để quét/chấm công.



b) Ứng dụng thực tiễn

Hệ thống recommender: Xác định tuổi tác, giới tính, các thông tin chi tiết → phục vụ cho marketing để gợi ý các sản phẩm cho người tiêu dùng.



b) Ứng dụng thực tiễn

- **Y tế:** Phân loại, nhận dạng thuốc
- **An ninh:** Kiểm tra hành lý ở sân bay



- **Pháp y:** Lấy chứng cứ

2. Động lực nghiên cứu



a) Yêu cầu mục đích

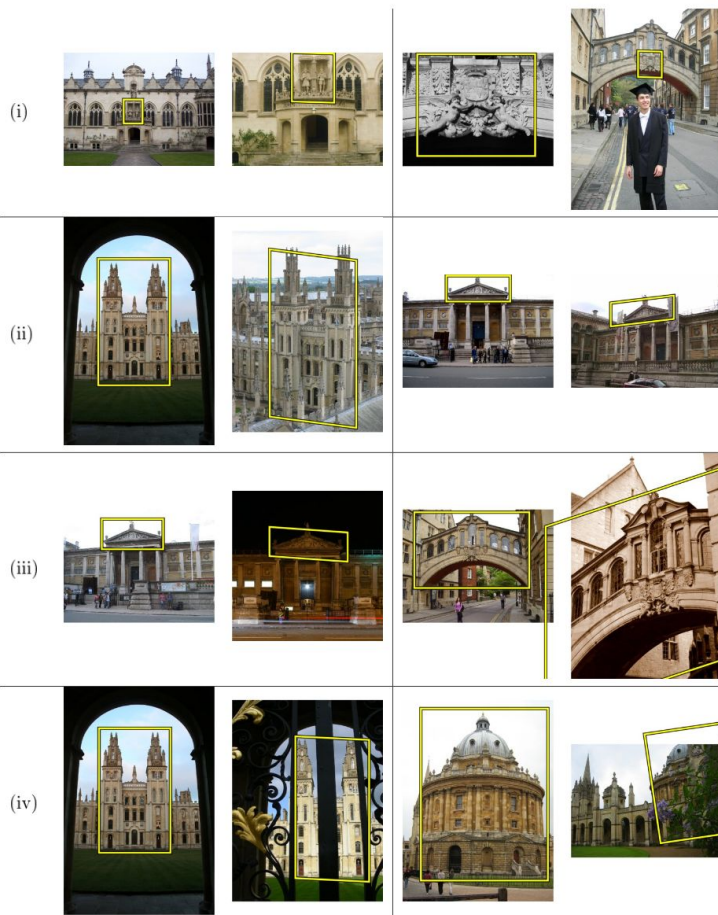
- **Input:** Ảnh
- **Output:**

(Thông tin đối tượng): Tên/Nhãn

(Heat map)

(Ảnh): Những hình ảnh tương tự

b) Thách thức



Thách thức về điều kiện hình ảnh trong thực tế (Điều kiện tự nhiên)

(i) Thay đổi về tỷ lệ của đối tượng đối với khung hình (Scale changes)

(ii) Thay đổi góc nhìn (Viewpoint changes)

(iii) Thay đổi về điều kiện ánh sáng (Lighting changes)

(iv) Che khuất một phần đối tượng (Partial occlusion)

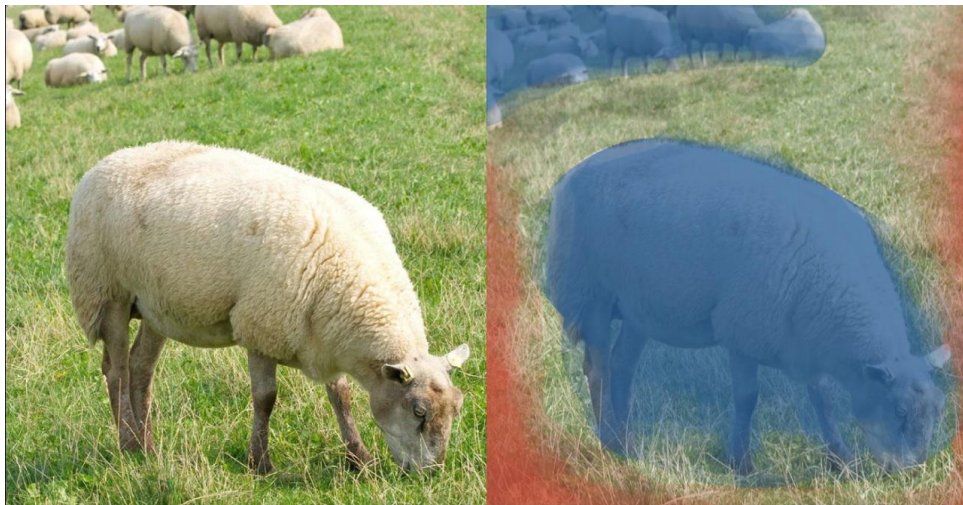
b) Thách thức



Thách thức về tốc độ truy xuất

(v) Dữ liệu quy mô lớn (Large-scale)

b) Thách thức



Thách thức về xử lý thông tin

(vi) Phân tách nền và đối tượng chính

Chương 2: Các công trình nghiên cứu liên quan



Visualizing and Understanding Convolutional Networks (2013)

- **Nguyên lý**

Mô hình Image Classification

- **Phương pháp**

Một kỹ thuật trực quan hóa mới giúp hiểu rõ hơn về chức năng của các lớp tính năng trung gian và hoạt động của bộ phân loại.

- **Tập dữ liệu**


Caltech-101 và Caltech-256

- **Hiệu suất**

Kết quả vượt trội so với các kết quả của phương pháp hiện đại nhất thời điểm đó

- **Ưu/khuyết điểm**

Độ nhiễu cao hơn so với GBP



Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps (2014)

- **Nguyên lý**

Mô hình Image Classification

- **Phương pháp**

Hai kỹ thuật trực quan hóa, dựa trên việc tính toán gradient của điểm lớp đối với hình ảnh đầu vào:

- Đầu tiên, tạo ra một hình ảnh, giúp tối đa hóa điểm số của lớp, qua đó trực quan hóa khái niệm về lớp, được ghi lại bởi ConvNet.
- Thứ hai, tính toán một lớp saliency map, cụ thể cho một hình ảnh và lớp nhất định.

- **Tập dữ liệu**

ILSVRC-2013 (1.2M ảnh đào tạo, dán nhãn thành 1000 lớp)

- **Hiệu suất**

Xử lý tập dữ liệu quy mô lớn

- **Ưu/khuyết điểm**

Độ nhiễu cao hơn so với GBP



Striving For Simplicity: The All Convolutional Net (2014)

- **Nguyên lý**

Nhận dạng đối tượng (Object recognition) từ các hình ảnh nhỏ với các mạng tích chập, đặt câu hỏi về sự cần thiết của các thành phần khác nhau trong quy trình.

- **Phương pháp**

Max-pooling có thể được thay thế đơn giản bằng một lớp tích chập với bước tiến tăng lên mà không làm giảm độ chính xác trên một số tiêu chuẩn nhận dạng hình ảnh.

- **Tập dữ liệu**

Bộ dữ liệu nhận dạng đối tượng (CIFAR-10, CIFAR-100, ImageNet)

- **Hiệu suất**

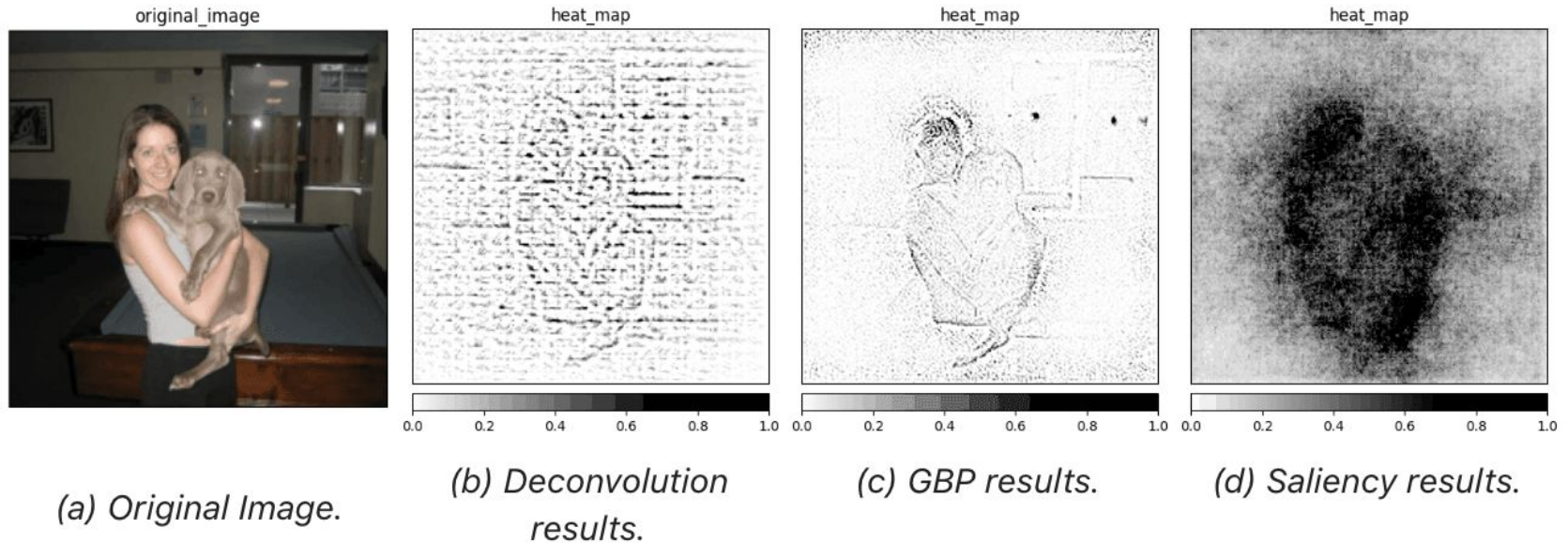
Có thể được áp dụng cho phạm vi cấu trúc mạng rộng hơn so với các phương pháp hiện có.

- **Ưu/khuyết điểm**

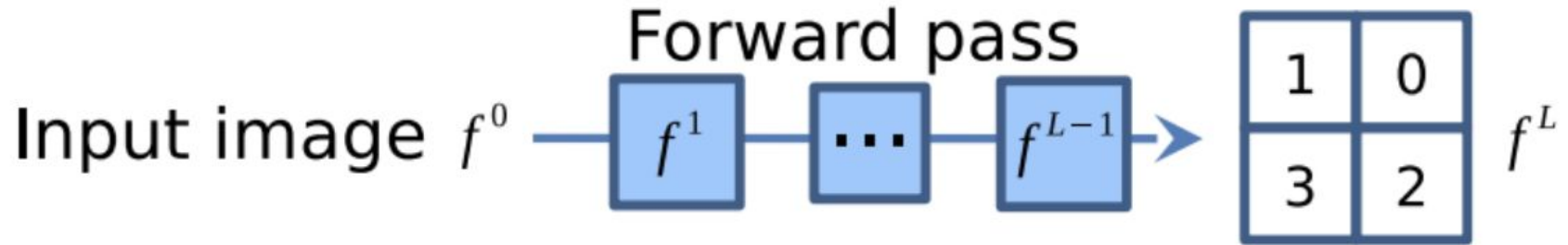
Giảm độ nhiễu

Chương 3: Phương pháp

Guided backpropagation: Chỉ dẫn lan truyền ngược



Guided backpropagation: Forward pass



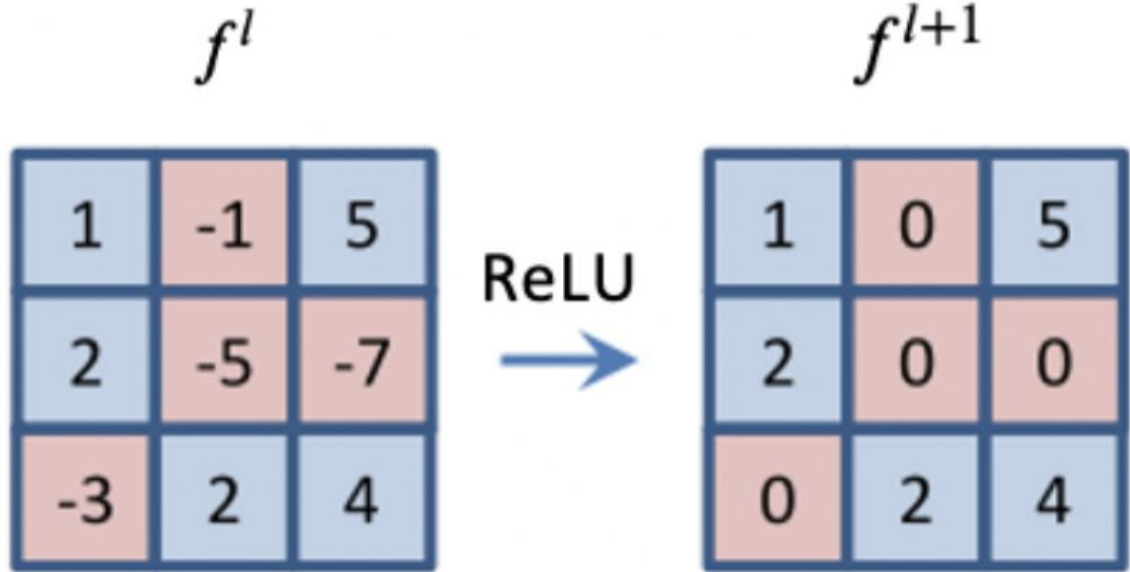


Guided backpropagation: Forward pass

$$f_i^{l+1} = \text{ReLU}(f_i^l) = \max(f_i^l, 0)$$

Guided backpropagation: Forward pass

Forward pass



Guided backpropagation: traditional backpropagation

Backward pass:
backpropagation

gradients for layer l

-2	0	-1
6	0	0
0	-1	3



gradients for layer $l+1$

-2	3	-1
6	-3	1
2	-1	3

The neurons marked **red** are dead, and the gradients can not be backpropagated.

Guided backpropagation

Backward pass:
*guided
backpropagation*

gradients for layer l

0	0	0
6	0	0
0	0	3

gradients for layer $l+1$

-2	3	-1
6	-3	1
2	-1	3



The gradients for both the red and yellow neurons are not backpropagated.

Guided backpropagation: summary

c) activation: $f_i^{l+1} = \text{relu}(f_i^l) = \max(f_i^l, 0)$

backpropagation: $R_i^l = (f_i^l > 0) \cdot R_i^{l+1}$, where $R_i^{l+1} = \frac{\partial f^{out}}{\partial f_i^{l+1}}$

backward
'deconvnet': $R_i^l = (R_i^{l+1} > 0) \cdot R_i^{l+1}$

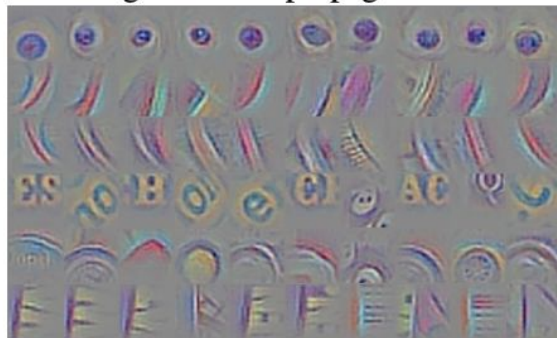
guided
backpropagation: $R_i^l = (f_i^l > 0) \cdot (R_i^{l+1} > 0) \cdot R_i^{l+1}$

Guided backpropagation: kết quả

deconv



guided backpropagation



corresponding image crops



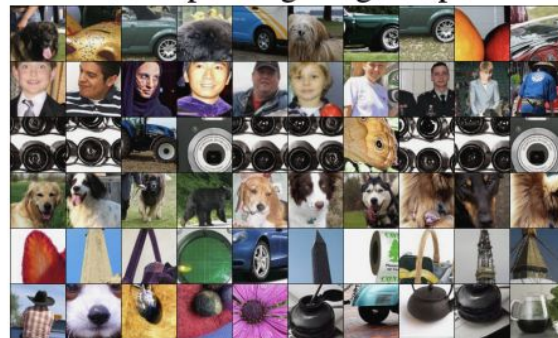
deconv



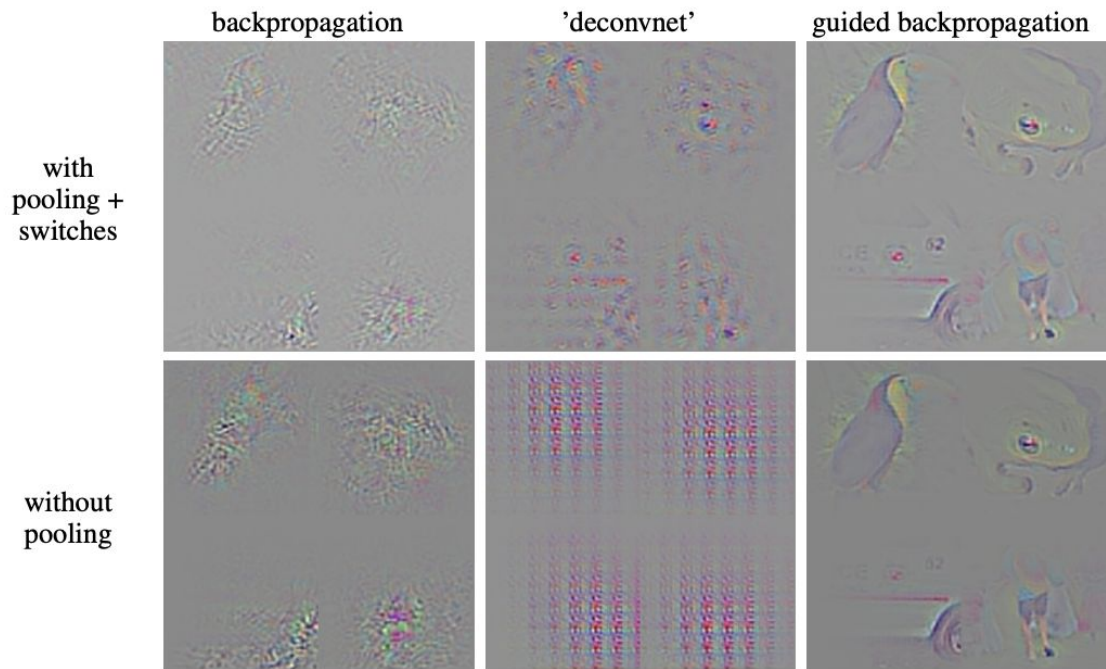
guided backpropagation



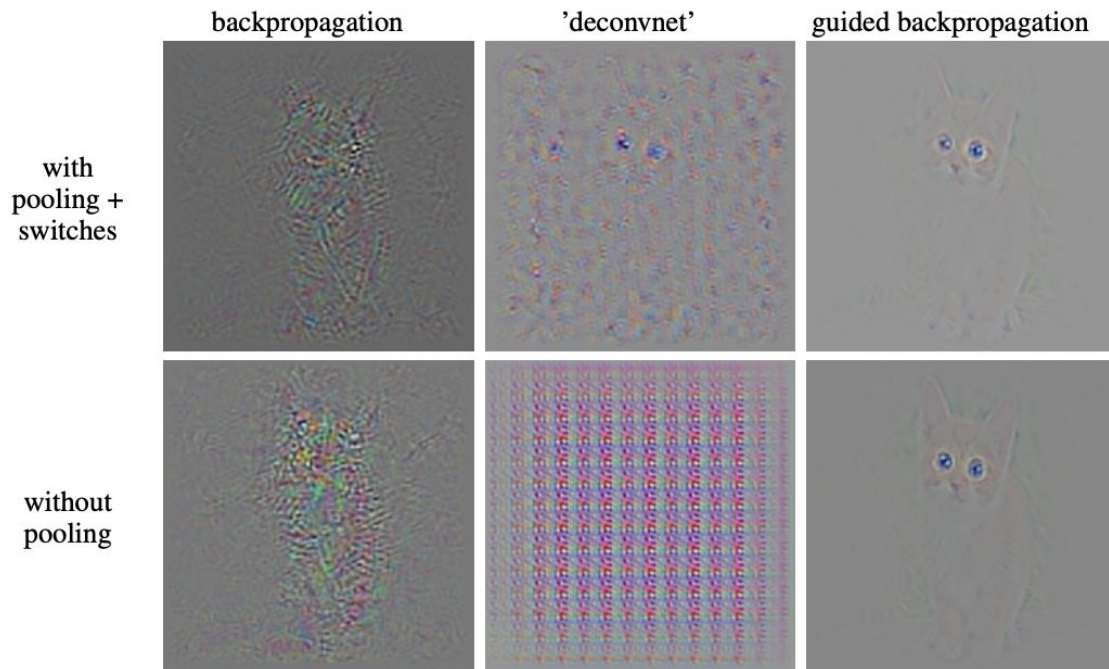
corresponding image crops



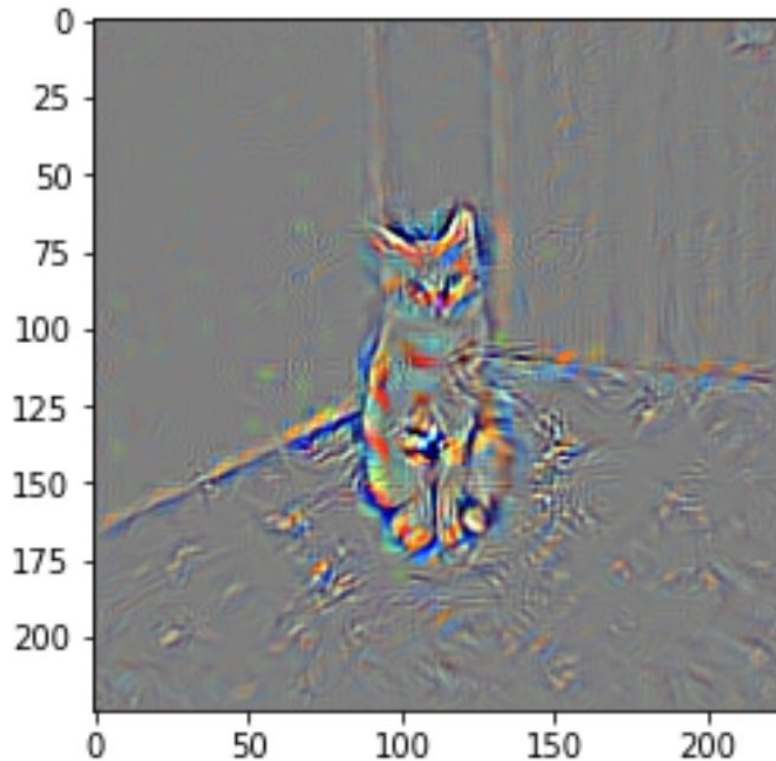
Guided backpropagation: kết quả



Guided backpropagation: kết quả

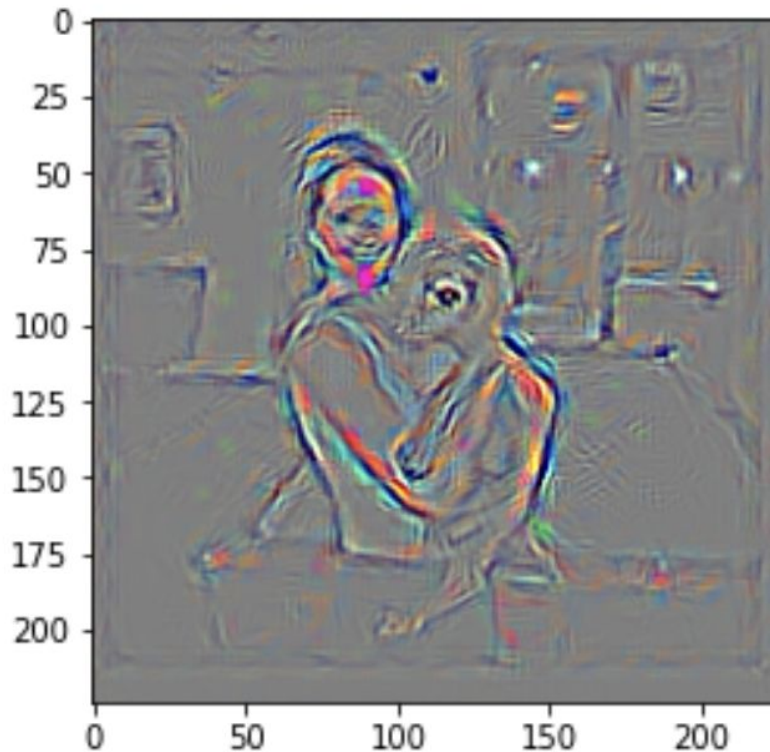


Guided backpropagation: thực nghiệm



Guided backpropagation: thực nghiệm

original_image



Cảm ơn thầy đã lắng nghe!



Cuối cùng
cũng xong
:>