

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN  
KHOA CÔNG NGHỆ THÔNG TIN

NHÓM 3

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

BÀI BÁO CÁO  
**PHƯƠNG PHÁP XAI – CHỈ  
DẪN LAN TRUYỀN NGƯỢC  
(GUIDED  
BACKPROPAGATION)**

Ngành: Khoa học máy tính



## MỤC LỤC

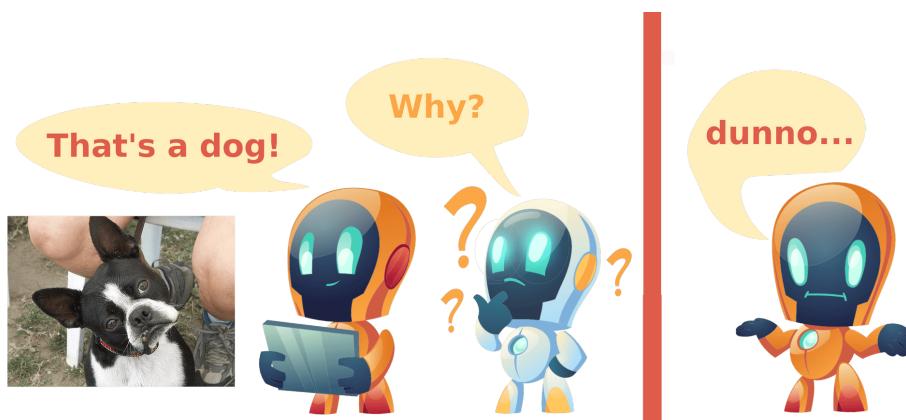
<i>Chương 1: Giới thiệu .....</i>	2
1. Động lực nghiên cứu .....	2
a) Ý nghĩa khoa học .....	2
b) Ứng dụng thực tiễn .....	2
2. Phát biểu bài toán.....	5
a) Yêu cầu mục đích .....	5
b) Thách thức .....	5
<i>Chương 2: Các công trình nghiên cứu liên quan .....</i>	7
Visualizing and Understanding Convolutional Networks (2013) [5] .....	7
Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps (2014) [6] .....	8
Striving For Simplicity: The All Convolutional Net (2014) [7] ....	8
<i>Chương 3: Phương pháp .....</i>	9
1. Đặt vấn đề .....	9
2. Phương thức hoạt động của Guided Backpropagation [9] ..	10
<i>Tài liệu tham khảo:.....</i>	13

## 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 nào cũng là 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. Một ví dụ đó là tác vụ nhận dạng đối tượng (Object recognition) đơn giản, trong đó, đầu vào là một hình ảnh của một con vật, và đầu ra chính xác sẽ là tên của con vật đó [4].

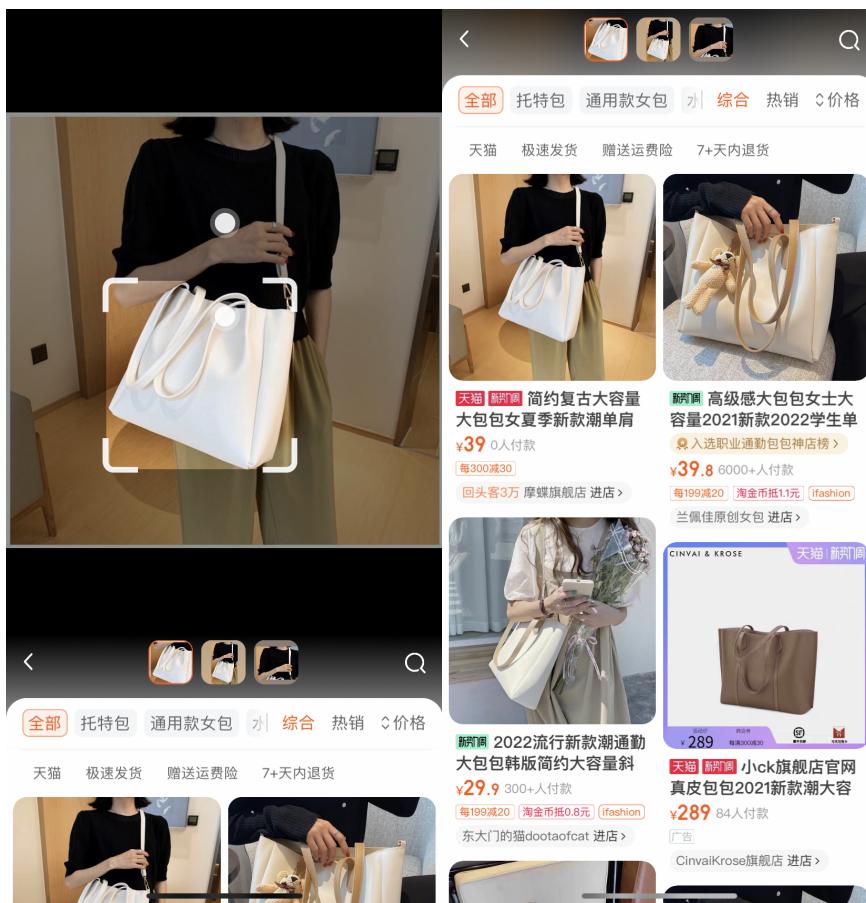


Trong một bức ảnh, việc hiểu bằng mắt và bằng hệ thống qua bức ảnh từ thô đến tinh. Thô là ID (tên của đối tượng) và tinh là các thuộc tính chi tiết hơn của đối tượng. Từ xưa, khi nhận dạng hình ảnh ta chỉ biết và quan tâm đến đến kết quả đúng hay sai nhưng lại không để ý đến việc lý giải được tại sao ta lại có kết quả đó. Để con người chấp nhận các thuật toán mà máy sử dụng để học, họ cần phải tin tưởng chúng để tránh các "gian lận" trong việc học.

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 [2].

#### 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... Phía dưới là tìm kiếm sản phẩm trên Taobao bằng hình ảnh.



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



**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. Tăng độ chính xác, tránh tình trạng gian lận



**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.



**Y tế:** Phân loại, nhận dạng thuốc

**An ninh:** Kiểm tra hành lý ở sân bay



**Pháp y:** Lấy chứng cứ

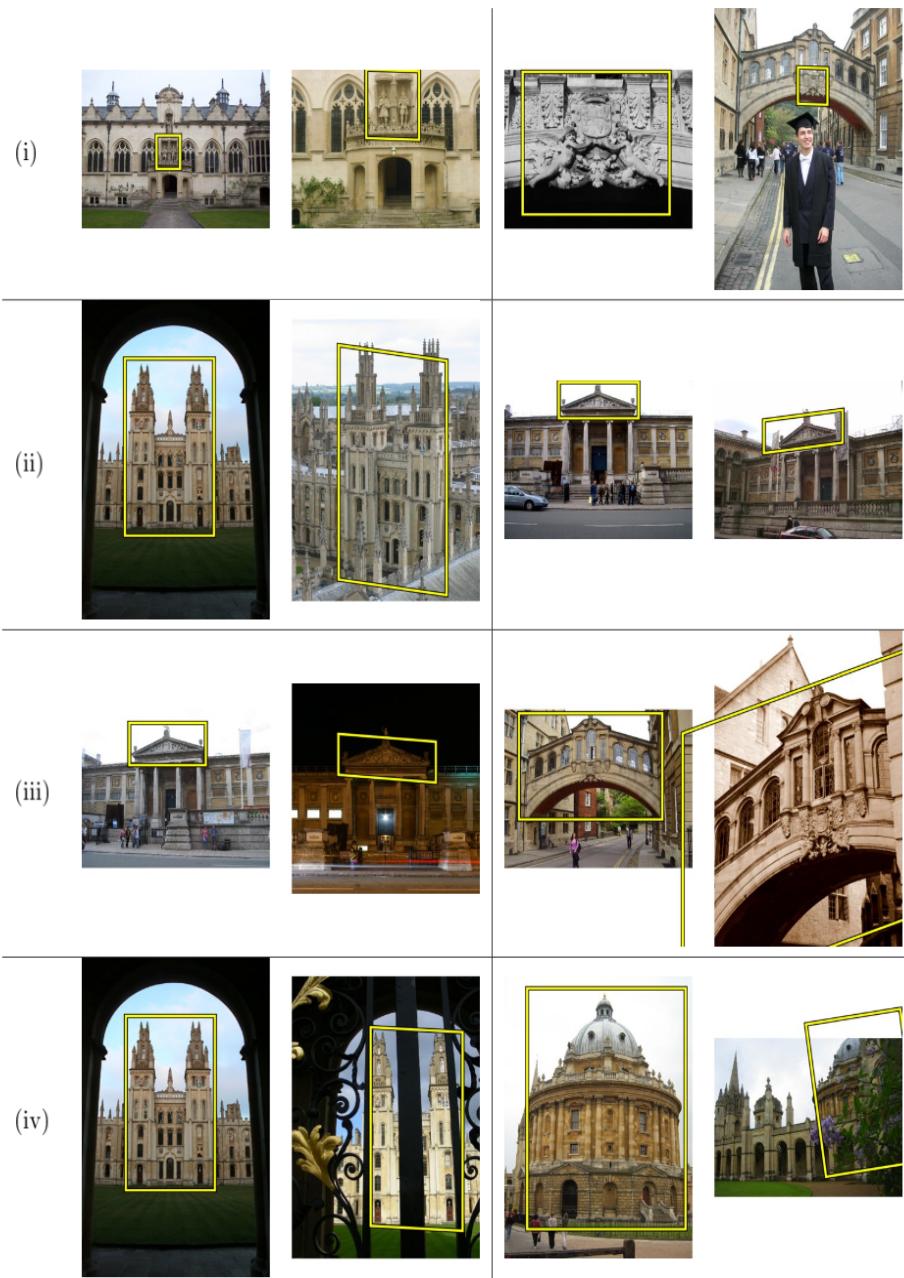
## 2. Phát biểu bài toán

### 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)

### Thách thức về tốc độ truy xuất



Dữ liệu quy mô lớn (Large-scale)

Thách thức về xử lý thông tin



Phân tách nền và đối tượng chính

## Chương 2: Các công trình nghiên cứu liên quan

Visualizing and Understanding Convolutional Networks  
(2013) [5]

- 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) [6]

- **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) [7]

- **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

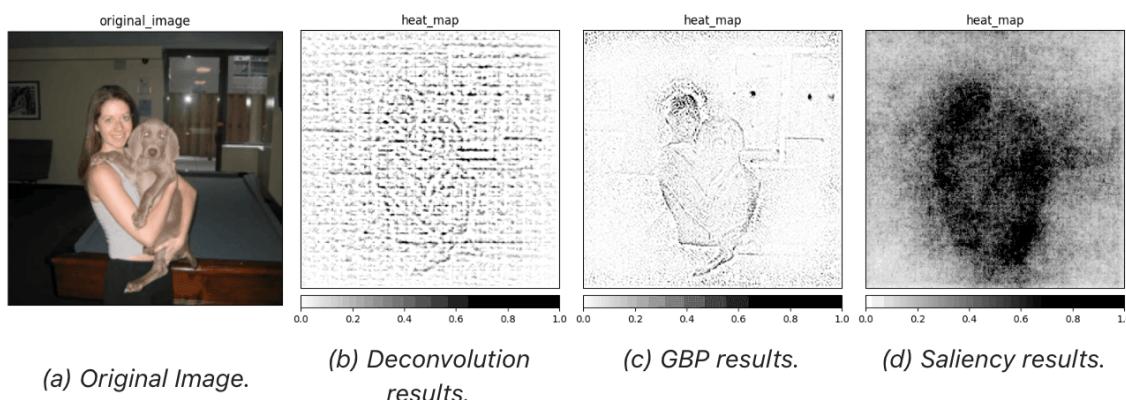
Có thể được áp dụng cho phạm vi cấu trúc mạng rộng hơn so với các phương pháp hiện có.

- Ưu khuyết điểm

Giảm độ nhiễu

## Chương 3: Phương pháp

### 1. Đặt vấn đề



Heat map của Deconvolution và Saliency (hình (b) và hình (d)) dùng để chứng minh với con người thông qua quan sát bằng mắt thường để thấy mô hình CNN thật sự không “gian lận”. Nó chỉ ra được vị trí của đối tượng tuy nhiên vẫn còn nhiễu nhiều và rất khó thấy. Mục đích của việc giảm bớt nhiễu để giải quyết thách thức về phân tách nền với đối tượng, tăng khả năng nhận dạng đối tượng. Để giảm nhiễu, ta sử dụng phương pháp Guided Backpropagation.

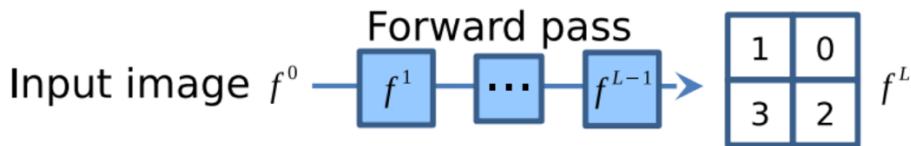
Guided Backpropagation (GBP) [7] là một phương pháp tiếp cận được thiết kế bởi Springenberg và cộng sự, dựa trên các ý tưởng của Deconvolution [5] và Saliency [6].

GBP là một kỹ thuật trực quan hóa dựa trên gradient giúp trực quan hóa gradient đối với hình ảnh khi lan truyền ngược thông qua hàm kích hoạt ReLU.

## 2. Phương thức hoạt động của Guided Backpropagation [9]

Cho một ảnh đầu vào và một mạng đã được đào tạo trước, phương pháp Guided Backpropagation có thể cho chúng ta biết các pixels nào trong ảnh đầu vào quan trọng trong việc cung cấp thông tin cho mô hình dự đoán. Phương pháp này gọi là Guided Backpropagation (chỉ dẫn lan truyền ngược) bởi vì chúng ta sẽ chọn những nơ-ron nào sẽ được kích hoạt khi Backpropagation.

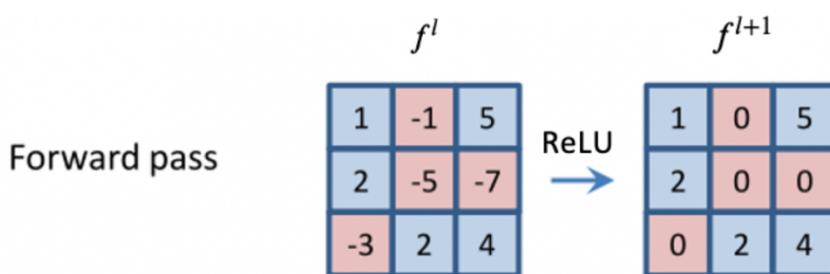
Để thực hiện Guided Backpropagation, đầu tiên chúng ta sẽ thực hiện một bước forward pass tới layer mà chúng ta đang quan tâm:



Ở đây chúng ta chỉ chú ý đến các bản đồ đặc trưng ( $f^1, f^2, \dots, f^L$ ) và hàm kích hoạt ReLU giữa hai bản đồ đặc trưng ( $f^l, f^{l+1}$ ). Do đó, ở bước forward pass, chúng ta có:

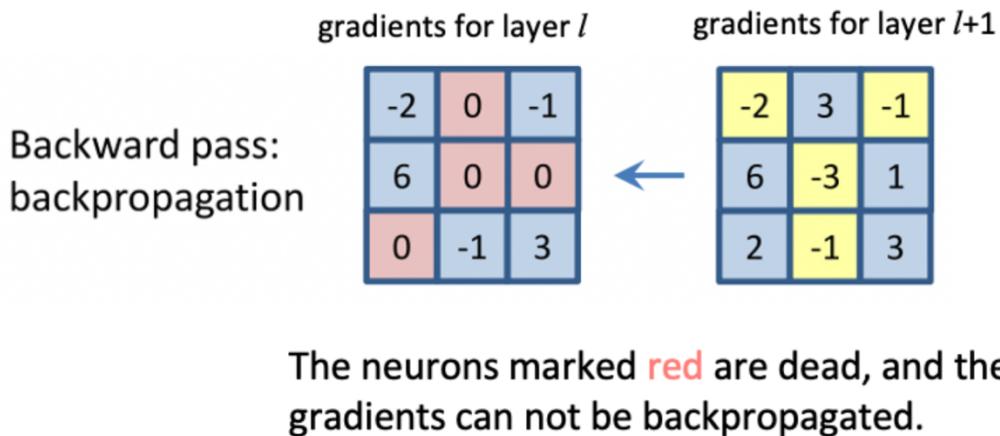
$$f_i^{l+1} = \text{ReLU}(f_i^l) = \max(f_i^l, 0)$$

Điều này có nghĩa rằng nếu giá trị đầu vào của lớp liền trước bị âm, thì đầu ra sau khi thực hiện phép ReLU sẽ đặt giá trị đó về 0, ví dụ:



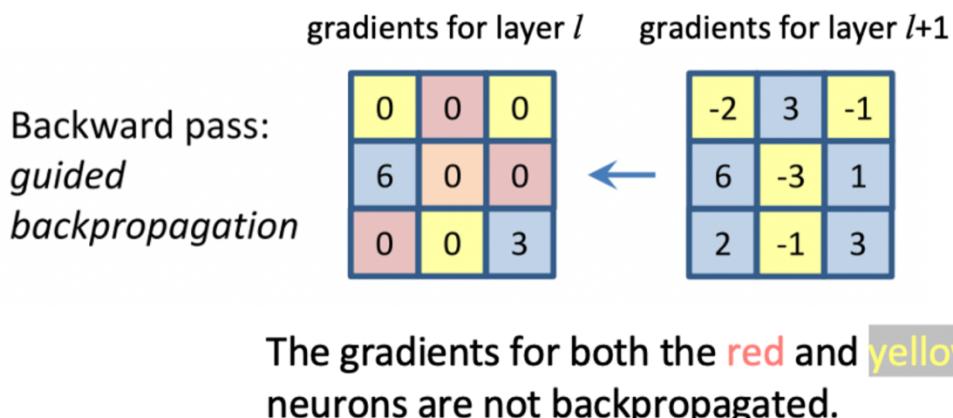
Trong trường hợp này, chúng ta có thể nói rằng các nơ-ron có giá trị âm đã “chết” (nghĩa là sẽ không còn được sử dụng ở bước backpropagation nữa).

Đối với ReLU, khi chúng ta thực hiện backpropagation truyền thống, nó chỉ truyền các giá trị gradient tới lớp trước đó nếu thông tin đầu vào ban đầu có giá trị dương (nghĩa là các nơ-ron chưa “chết”), ví dụ:

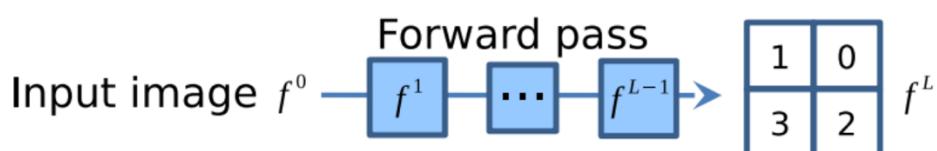


Như hình bên trên, giá trị gradient cho các nơ-ron màu đỏ (nơ-ron đã “chết”) tại lớp thứ 1 được set giá trị 0 mặc dù gradient tại lớp  $l + 1$  nó vẫn có giá trị.

Trong trường hợp của phương pháp guided backpropagation, chúng ta cũng không truyền ngược các giá trị gradient âm (các nơ-ron màu vàng trong hình bên trên) đến các lớp trước đó, ví dụ như hình dưới:

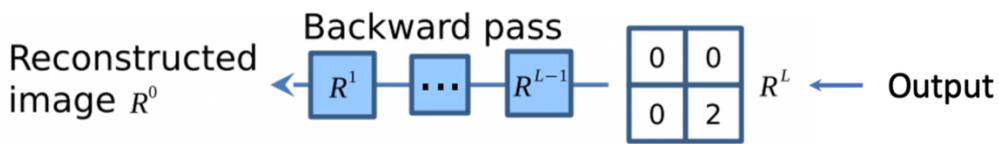


Tổng kết lại, ở bước forward pass như sau:



Cho một bản đồ đặc trưng  $f^L$ , cho rằng chúng ta chỉ quan tâm đến các nơ-ron tại góc dưới cùng bên phải có giá trị gradient bằng 2. Chúng ta sẽ giữ lại nơ-ron này và đặt

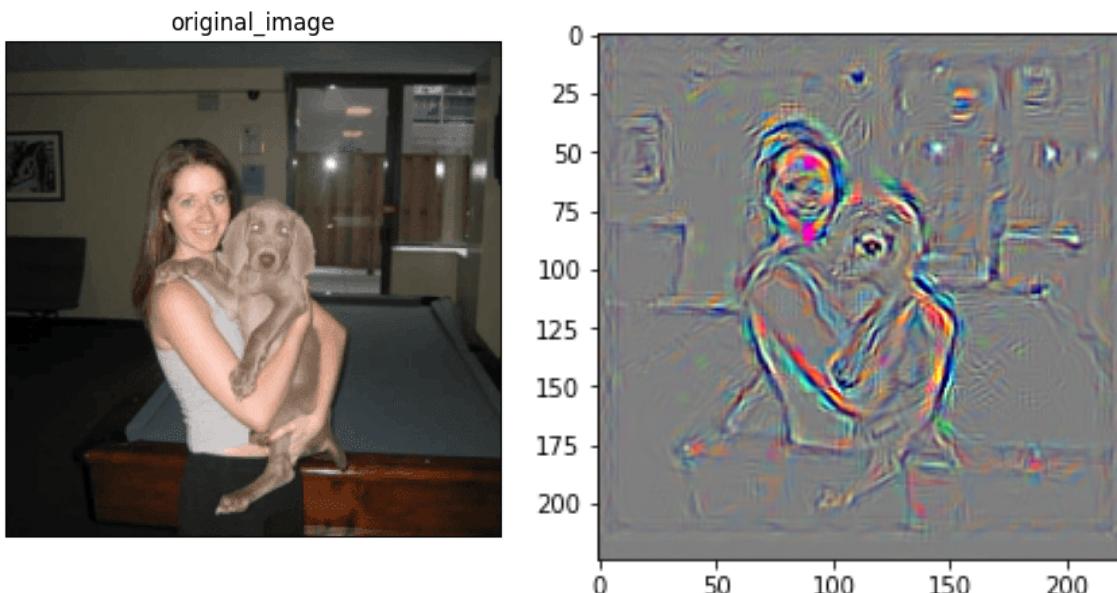
tất cả các nơ-ron khác về 0. Sau đó giá trị gradient của nơ-ron được chọn sẽ lan truyền ngược về tất cả các đầu vào để thực hiện việc tái tạo lại thông tin:

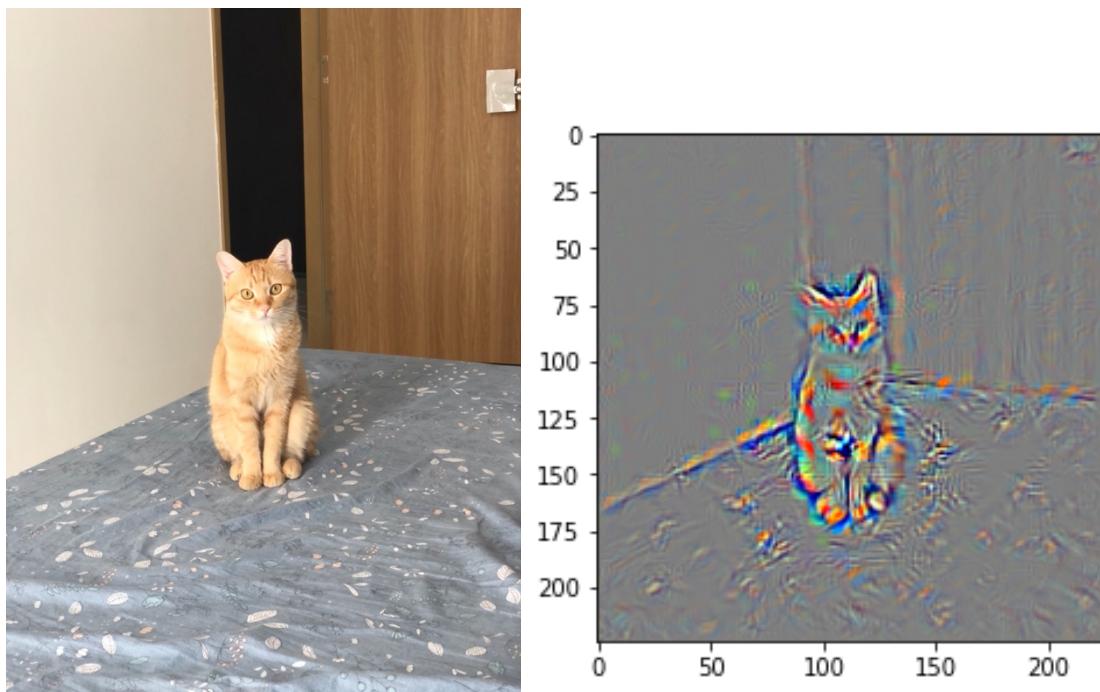


Chúng ta sẽ chọn các nơ-ron không bị đánh dấu màu vàng hoặc đỏ ở hình bên trên (các nơ-ron có giá trị gradient không âm) để thực hiện guided backpropagation. Do đó, phương trình backpropagation (đã được “chỉ dẫn - guided”) như sau:

$$R_i^l = (f_i^l > 0) \cdot (R_i^l + 1 > 0) \cdot R_i^{l+1} \text{ trong đó } R_i^{l+1} = \frac{\partial_{out}}{\partial f_i^{l+1}}$$

Đầu ra  $R^0$  sẽ có kết quả như hình dưới:





## Tài liệu tham khảo:

- [1] [Trí tuệ nhân tạo – Wikipedia tiếng Việt](#)
- [2] [Explainable artificial intelligence - Wikipedia](#)
- [3] [XAI Methods - The Introduction - Blog by Kemal Erdem](#)
- [4] [Truyền ngược – Wikipedia tiếng Việt](#)
- [5] M. D. Zeiler, R. Fergus. [Visualizing and Understanding Convolutional Networks](#), 2013.
- [6] K. Simonyan, A. Vedaldi, A. Zisserman. [Deep inside convolutional networks: Visualising image classification models and saliency maps](#), 2014.
- [7] J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller. [Striving for Simplicity: The All Convolutional Net](#), 2014.
- [8] 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.
- [9] [Deep Learning: Guided BackPropagation, LESLIE'S BLOG 2020-07-22](#)