Trường Đại học Khoa học Tự nhiên
Đại học Quốc gia TP.HCM

Phát hiện đối tượng trọng yếu dựa vào mạng nơ-ron tích chập

Sinh viên thực hiện: Cao Lê Minh Hiếu.
Giáo viên hướng dẫn: TS. Nguyễn Đức Hoàng Hạ

Mục lục

Giới thiệu đề tài
Phát biểu bài toán

Phát hiện đối tượng trọng yếu là bài toán phát hiện một hoặc nhiều vật thể quan trọng, nổi bật nhất trong bức ảnh, không phân biệt lớp vật thể.

Tất cả vật thể được xem là trọng yếu sẽ được đánh nhãn **vùng trắng** (**vùng trọng yếu**) như hình bên và phần còn lại của bức ảnh là **vùng đen** (vùng **không trọng yếu**)

Dòng 1 và 2: ảnh đầu vào và nhãn tương ứng.
Cột 1 là dữ liệu được lấy trong tập huấn luyện **DUTS-TR**. **Cột 2, 3** là dữ liệu được lấy từ tập kiểm thử **DUTS-TE.**

Giới thiệu đề tài
Phát biểu bài toán

Bài toán là một nhánh trong lĩnh vực **phát hiện đối tượng** và có mối tương quan với **phân đoạn ngữ nghĩa**. Tuy nhiên vẫn có sự khác biệt rõ trong các tác vụ trên (đầu ra của bài toán):

- **Phát hiện đối tượng** dự đoán tất cả các vật thể trong ảnh, vị trí của khung bao quanh và lớp của các vật thể đó. Trong khi **phát đối tượng trọng yếu** là bài toán phân lớp trên từng điểm ảnh và chỉ dự đoán các vật thể trọng yếu thay vì tất cả.
- **Phát hiện đối tượng trọng yếu** nhằm phân vùng ảnh chứa đối tượng trọng yếu(vùng trắng) và không trọng yếu (vùng đen), không kể số lương cũng như lớp vật thể. Trong khi **phân đoạn ngữ nghĩa** dự đoán các vùng trên bức ảnh ở mức độ lớp vật thể.

Đầu vào và đầu ra của bài toán
Phát hiện đối tượng (cột 1)
Phát hiện đối tượng trọng yếu (cột 2)
Phân đoạn ngữ nghĩa (cột 3)

Giới thiệu đề tài
Đối tượng trọng yếu là gì?

Ảnh mang hai thuộc tính quan trọng là đặc trưng thị giác và ngữ nghĩa. Đối tượng trọng yếu là một vùng trong ảnh nên cũng sẽ mang hai thuộc tính này với một số biểu hiện đặc thù như sau:

- Về đặc trưng thị giác: vùng có độ dị biệt về đặc trưng thị giác rõ nét so với phần còn lại của ảnh
- Về ngữ nghĩa: vùng có ngữ nghĩa quan trọng trong ảnh.
- Về vị trí, kích thước: thường ở trung tâm của bức ảnh.

Đối tượng trọng yếu chưa được định nghĩa và vẫn là đề tài nghiên cứu trong khoa học nhận thức, tâm lý học,…

Dòng 1 và 2: ảnh đầu vào và nhãn tương ứng.
Cột 1 là dữ liệu được lấy trong tập huấn luyện **DUTS-TR**. **Cột 2, 3** là dữ liệu được lấy từ tập kiểm thử **DUTS-TE.**

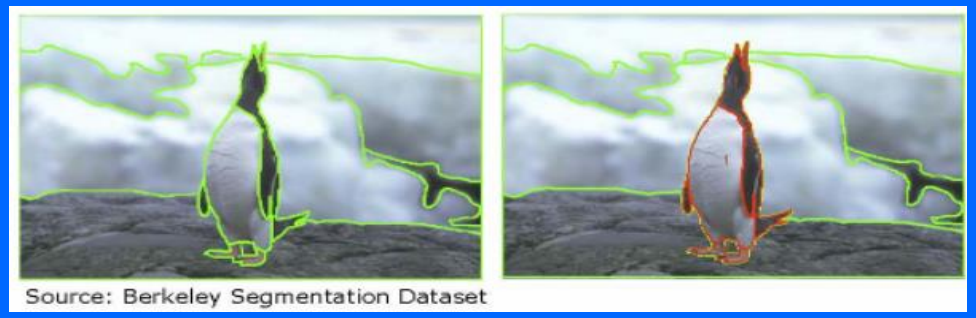
Giới thiệu đề tài
Những thách thức

- Chưa có định nghĩa đối tượng trọng yếu.
- Với những ảnh chỉ chứa một đối tượng, rõ ràng vật thể đó là trọng yếu của bức ảnh.
- Nhưng với ảnh gồm nhiều vật thể, việc nhận định đối tượng trọng yếu mang tính chủ quan và thiếu nhất quán vì góc nhìn của mỗi người là khác nhau (phụ thuộc vào tâm trạng, giới tính, sở thích và văn hóa của người nhìn).
- Dữ liệu đánh nhãn dựa trên khả năng nhận thức của con người.
- Phương pháp tích hợp các khối đặc trưng đa bậc hiệu quả.

Giới thiệu đề tài
Cách thức đánh nhãn dữ liệu

Có hai cơ chế đánh nhãn phổ biến nhất:

1. Các vật thể trong ảnh đã được đóng biên cạnh, người đánh nhãn sẽ nhập chuột để chọn các đối tượng mà theo họ là trọng yếu nhất. Trước đó người này sẽ được bảo hãy lựa chọn các đối tượng mà theo họ là nổi bật nhất trong bức ảnh.
2. Sử dụng hệ thống theo dõi điểm nhìn của mắt (eye-tracker) để thu thập các điểm nhìn của người đó trong bức ảnh từ 3-5s, vật thể nào chứa số điểm được phát hiện bởi hệ thống nhiều hơn một lượng lớn so với vật thể khác sẽ được xem là trọng yếu.


Source: Berkeley Segmentation Dataset

Giới thiệu đề tài
Học thuật

- Đề tài có nguồn gốc từ khoa học nhận thức (Cognitive Science), khoa học thần kinh (Neuro Science) và tâm lý học.
- Mô phỏng cơ chế tập trung về mặt thị giác (Visual Attention) của con người nhờ vào chính bộ dữ liệu thị giác được tạo ra bởi nhiều người đánh nhãn.

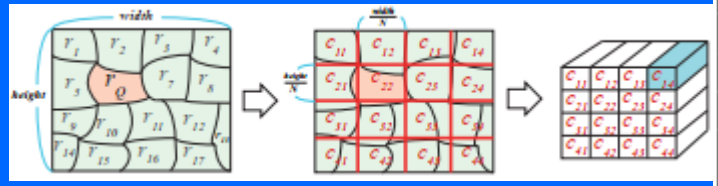
Giới thiệu đề tài
Ứng dụng

Được ứng dụng rộng rãi trong nhiều tác vụ cần rút trích thông tin tóm tắt,tổng quan trong ảnh:

- Phát hiện sự kiện trong bức ảnh (event detection).
- Tự động chú thích ảnh (image captioning).
- Nén ảnh (image compression)
- Trong đồ họa máy tính như cắt ảnh tự động (image cropping), làm nổi bật đối tượng.
- …

Các nghiên cứu liên quan
Chia bài toán thành nhiều giai đoạn

- Bài toán nhìn chung gồm bốn giai đoạn chính là đề xuất ứng viên, rút trích và chọn lọc đặc trưng, phân lớp, hậu xử lý.



Features of a superpixel ($f(r_c)$)	Feature Index
Average RGB value	1-3
Average LAB value	4-6
Average HSV value	7-9
Gabor filter response	10-33
Maximum Gabor response	34
Center location	35-36
RGB color histogram	37-61
LAB color histogram	62-86
HSV color histogram	87-110

Cơ chế chia ảnh thành super-pixel **(bên trái)** và rút trích đặc trưng thủ công **(bên phải)** trong "Deep Saliency with Encoded Low level Distance Map and High Level Features

Các nghiên cứu liên quan
Xu hướng phát triển
Mô hình đầu cuối

[1] Amulet Aggregating Multi-level Convolutional Features for Salient Object Detection
[2] BASNet Boundary-Aware Salient Object Detection
[3] A Bi-Directional Message Passing Model for Salient Object Detection
[4] Pyramidal Feature Shrinking for Salient Object Detection
[5] PiCANet Learning Pixel-Wise Contextual Attention for Saliency Detection
[6] TRACER Extreme Attention Guided Salient Object Tracing Network

- Lược bỏ đi hai giai đoạn đề xuất ứng viên và hậu xử lý. Sử dụng một mạng nơ-ron tích chập đầu cuối để rút trích, chọn lọc, tích đặc trưng đa bậc và phân lớp.
- Kiến trúc được sử dụng để rút trích đặc trưng phải hiệu quả, ít tham số và số phép toán thực hiện: dần thay thế VGGNet [1][3][5] bởi Resnet [2][4], EfficientNet[6].
- Phát hiện đối tượng đa kích thước, tăng vùng nhìn thấy của nơ-ron (receptive field): sử dụng tích chập giãn nở (dilated convolutions) [2][3][4][6].
- Kĩ thuật Attention được sử dụng để chọn lọc đặc trưng: spatial-attention[6], channel attention[4][6].
- Kết hợp các hàm mất mát: Binary Cross Entropy (BCE), Intersection over Union (IoU), L1.

Phương pháp đề xuất
Rút trích đặc trưng – Resnet50

- Giảm số lượng tham số so với VGG16 (23 triệu so với 138 triệu).
- Kiến trúc nhiều lớp, sâu hơn và vùng nhìn thấy của nơ-ron rộng hơn.
- Số phép toán giảm đi đáng kể khi ảnh đầu vào nhanh chóng bị giảm kích thước khi đi qua vài lớp tích chập đầu tiên.

Phương pháp đề xuất
Rút trích đặc trưng – Resnet50

Tên khối tích chập	Resnet-50	Kích thước đầu ra	Tên đầu ra
Conv1	7x7, 64 stride 2	$\frac{H}{2} x \frac{W}{2} x\ 64$	**L1**
Conv2_x	3x3 maxpool, stride 2		**L2**
	$\begin{bmatrix} 1x1,64 \\ 3x3,64 \\ 1x1,256 \end{bmatrix}$ x3	$\frac{H}{4} x \frac{W}{4} x256$	
Conv3_x	$\begin{bmatrix} 1x1,128 \\ 3x3,128 \\ 1x1,512 \end{bmatrix}$ x4	$\frac{H}{8} x \frac{W}{8} x512$	**L3**
Conv4_x	$\begin{bmatrix} 1x1,256 \\ 3x3,256 \\ 1x1,1024 \end{bmatrix}$ x6	$\frac{H}{16} x \frac{W}{16} x1024$	**L4**
Conv5_x	$\begin{bmatrix} 1x1,512 \\ 3x3,512 \\ 1x1,2048 \end{bmatrix}$ x3	$\frac{H}{32} x \frac{W}{32} x2048$	**L5**

Phương pháp đề xuất
Chọn lọc và tích hợp đặc trưng

- Các khối đặc trưng (feature map) đầu tiên L1, L2 rút trích được các biên cạnh (không phân biệt vật thể trọng yếu).
- Khối đặc trưng L5 chứa thông tin toàn cục, giàu ngữ nghĩa nhất. Việc phát hiện đối tượng trọng yếu chính xác phụ thuộc rất nhiều vào lượng thông tin này.
- Là bài toán phân lớp trên từng điểm ảnh, quá trình tích hợp đặc trưng đa bậc hiệu quả khôi phục độ phân giải là điều tất yếu.
- Do có sự cách biệt về độ phân giải cũng như mức độ ngữ nghĩa được rút trích, quá trình tích hợp chỉ nên thực hiện trên các khối đặc trưng liền kề, có sự cách biệt thấp nhất.

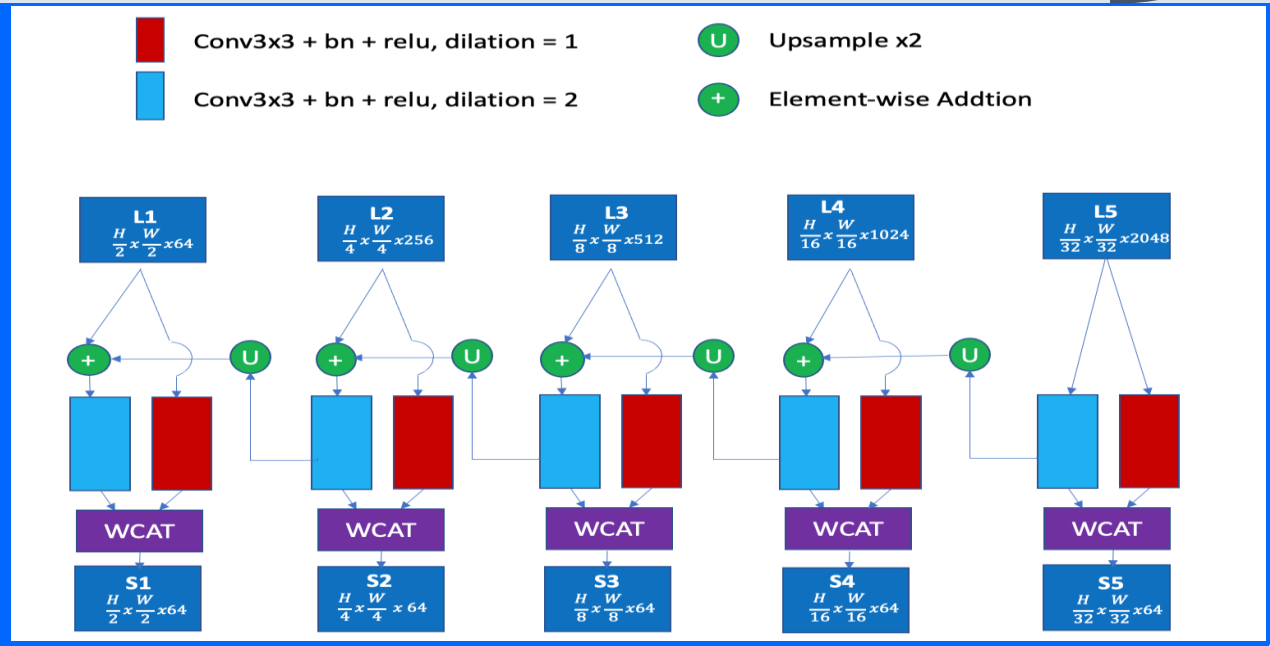
Phương pháp đề xuất

Chọn lọc và tích hợp đặc trưng

Mô-đun Scale-aware Enrichment Module (SEM)

Mô-đun nhận đầu vào là các khối đặc trưng **Li**, nhằm đạt được các mục đích sau:

- Truyền thông tin toàn cục của khối đặc trưng L5 đến các khối đặc trưng trước để lọc các điểm ảnh trọng yếu lần một.
- Sử dụng tích chập giãn nở để có thể phát hiện các đối tượng đa kích thước cũng như tăng vùng nhìn thấy của nơ-ron.
- Giảm số kênh (channels) của các khối đặc trưng nhằm giảm tham số mô hình thông qua các lớp tích chập ít số lượng bộ lọc hơn.



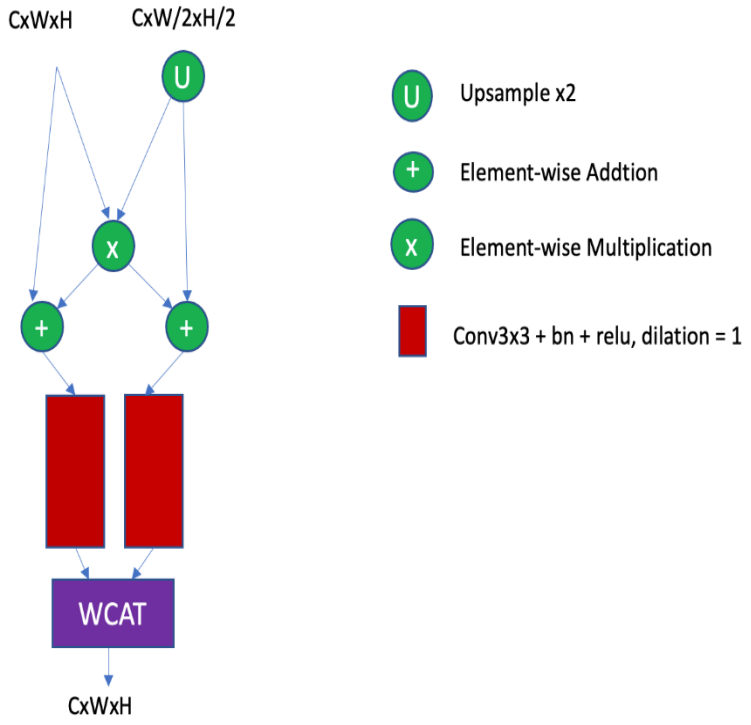
Phương pháp đề xuất

Chọn lọc và tích hợp đặc trưng

Mô-đun Adjacent Fusion Module (AFM)

Tiếp tục tích hợp các khối đặc trưng liền kề **Si** do lọc lần một nhiễu xuất hiện trong quá trình truyền luồng thông tin trọng yếu.

Mô-đun được thiết kế nhằm loại bỏ đi nhiễu cũng như tăng cường thông tin trọng yếu.



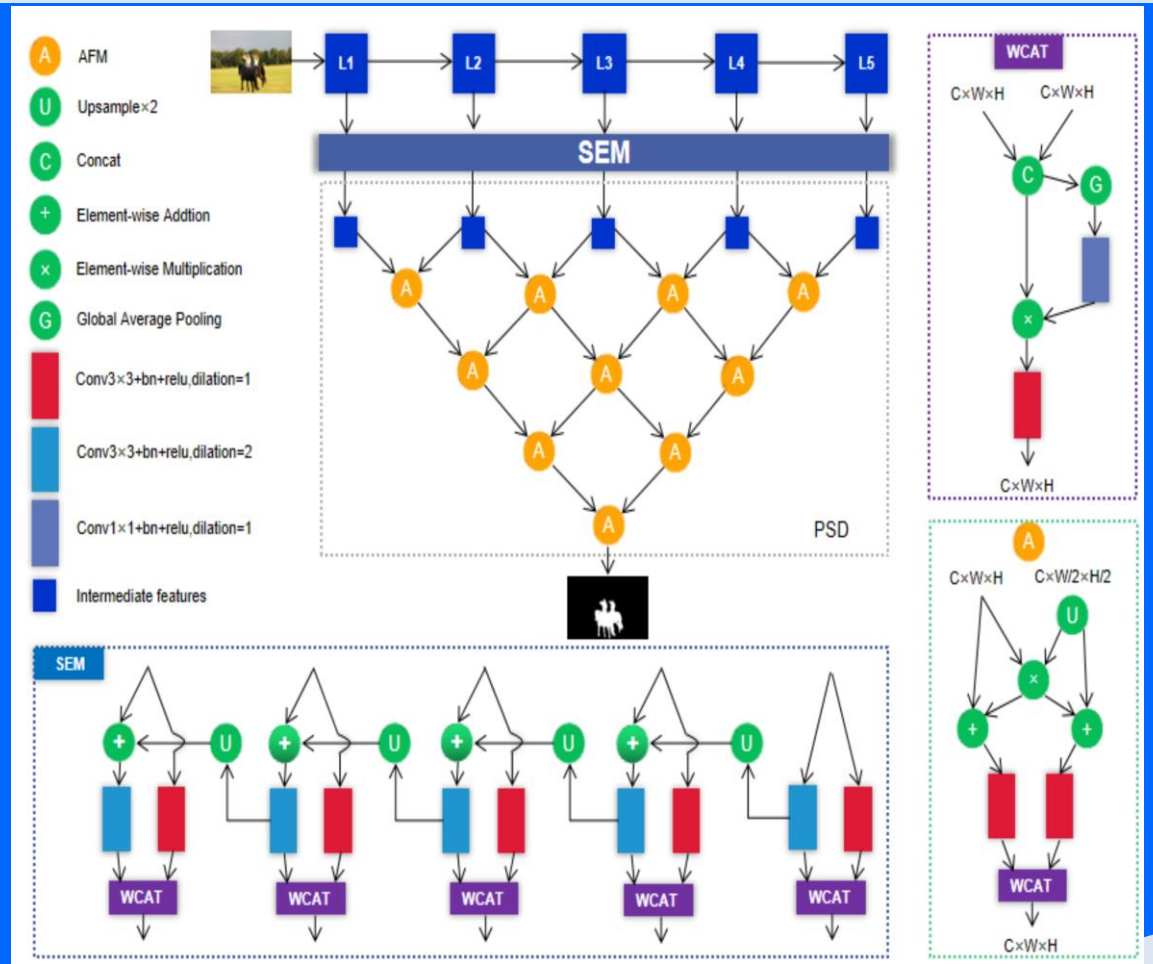
Phương pháp đề xuất

Cả quá trình rút trích và tích hợp – Mô hình gốc Pyramidal Feature Shrinking Network (PFSNet)

[4] Pyramidal Feature Shrinking for Salient Object Detection

Đầu vào: ảnh có kích thước HxWx3.

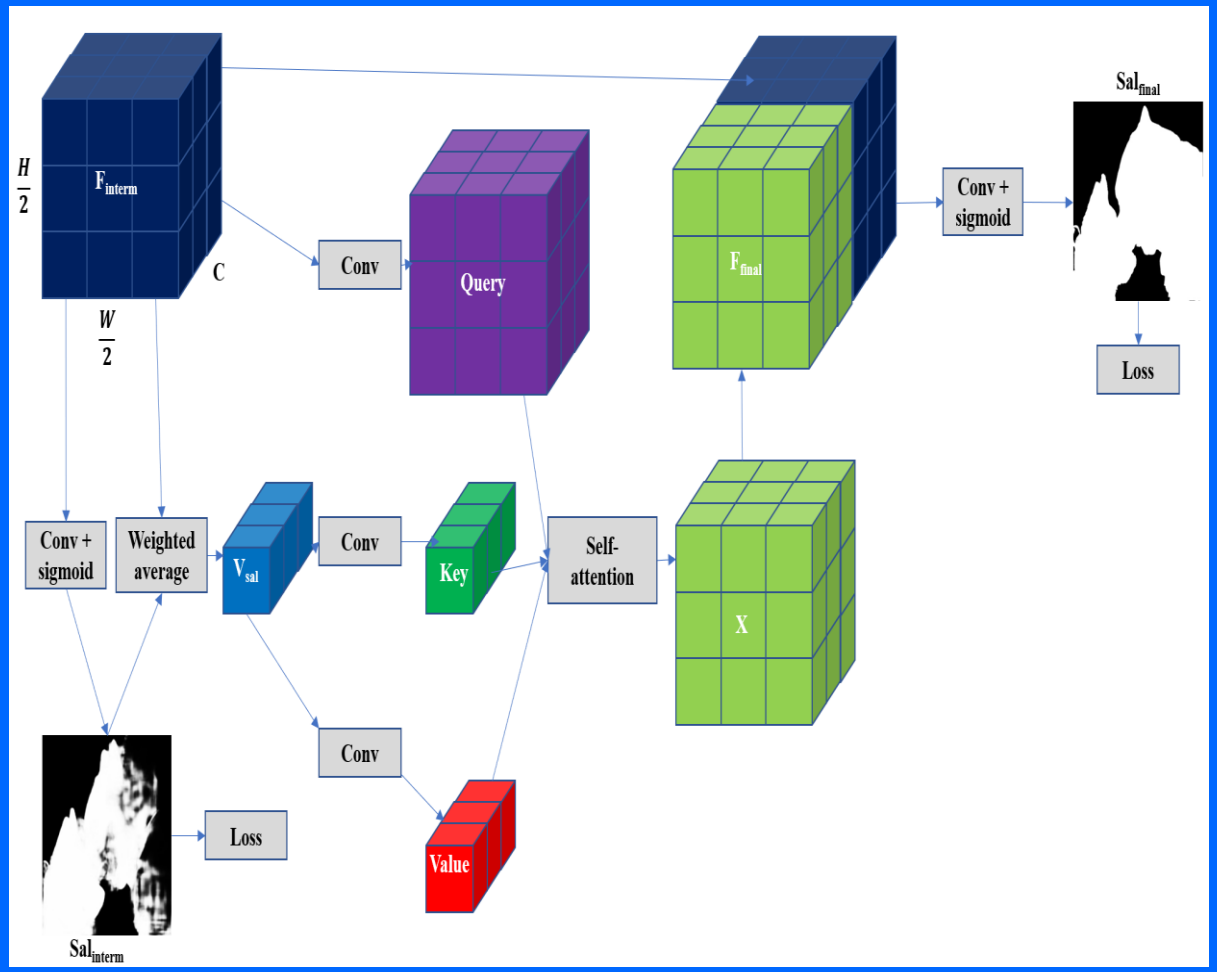
Đầu ra: khối đặc trưng $\mathbf{F_{interm}}$ có kích thước $\frac{H}{2} x \frac{W}{2} x\ 64$ và bản đồ điểm quan trọng (saliency map) $\mathbf{Sal_{interm}}$ được phân lớp trên khối đặc trưng trên.



Phương pháp đề xuất
Cơ chế Object Context Representation – cải tiến

- Mô-đun này cho kết quả tốt trong bài toán phân đoạn ngữ nghĩa "**Object-Contextual Representations for Semantic Segmentation**".
- Áp dụng mô-đun này cho bài toán phát hiện đối tượng trọng yếu lần đầu tiên.
- Mô-đun dựa trên cơ chế self-attention, sẽ điều chỉnh mức độ trọng yếu của điểm ảnh dựa vào sự tương đồng giữa véc-tơ đặc trưng của nó và véc-tơ đại diện cho mức độ trọng yếu.

Phương pháp đề xuất
Cơ chế OCR dựa



	Phân đoạn ngữ nghĩa	Phát hiện đối tượng trọng yếu
Phân lớp	Dự đoán **N** lớp vật thể	Phân lớp các đối tượng trọng yếu, không trọng yếu
Số lượng véc-tơ đại diện	**N** véc-tơ đại diện cho **N** lớp vật thể	**1** véc-tơ đại diện cho mức độ trọng yếu (xem véc-tơ đại diện cho không trọng yếu là phần bù)

Phương pháp đề xuất
Hàm lỗi Adaptive Pixel Intensity (API)[6]

Kết hợp ba hàm lỗi **BCE**, **IoU** và **L1** với trọng số đánh mạnh vào biên cạnh đối tượng:

$$\mathbf{BCE_W} = -\frac{\sum_{i=1}^{H}\sum_{j=1}^{W}(1+w_{ij})\sum_{c=0}^{1}(y_c \log(\hat{y}_c)+(1-y_c)\log(1-\hat{y}_c))}{\sum_{i=1}^{H}\sum_{j=1}^{W}(1+w_{ij})}$$

$$\mathbf{IoULoss_W} = 1 - \frac{\sum_{i=1}^{H}\sum_{j=1}^{W}(y_{ij}\,\hat{y}_{ij})(1+w_{ij})}{\sum_{i=1}^{H}\sum_{j=1}^{W}(y_{ij}+\hat{y}_{ij}-y_{ij}\,\hat{y}_{ij})(1+w_{ij})}$$

$$\mathbf{L1Loss_W} = \sum_{i=1}^{H}\sum_{j=1}^{W}|y_{ij}-\hat{y}_{ij}|w_{ij}$$

$$W_{i,j} = 0.5\sum_{k\in\{3,15,31\}}\left|\frac{\sum_{h=i-\frac{k}{2},w=j-\frac{k}{2}}^{h=i+\frac{k}{2},w=j+\frac{k}{2}} y_{hw}}{k^2} - y_{i,j}\right| y_{i,j}$$

Kết quả thử nghiệm
Công tác chuẩn bị

Quá trình huấn luyện

- Bộ dữ liệu **DUTS-TR** gồm 10553 bức ảnh và nhãn được sử dụng để huấn luyện (5% dữ liệu cho tập validation).
- **Tăng cường dữ liệu**: lật ảnh, xoay 90 độ, làm mờ và nhiễu Gauss, điều chỉnh độ tương phản, độ sáng.

Quá trình đánh giá

- Đánh giá mô hình trên các bộ dữ liệu **ECSSD** (1000 ảnh), **DUTS-TE** (5019 ảnh).
- **Độ đo**: F-measure, S-measure, MAE

Kết quả thử nghiệm

Quá trình huấn luyện

	PFSNet + OCR (key channels = 64)	PFSNet + OCR (key channels = 128)
Số lượng tham số	31.33 triệu	31.63 triệu
Quá trình huấn luyện	Cả mô hình được huấn luyện đầu cuối với learning rate cho backbone là 0.005 và các mô-đun khác là 0.05. Weight decay là 0.0005. Kích thước lô là 8. Kích thước ảnh đầu vào là 352x352 Chạy hơn 100 epochs với thời gian chạy ~10 phút/epoch. Tổng thời gian của cả quá trình là 18-19 giờ.	Huấn luyện với các siêu tham số như cấu hình 1. Chạy hơn 100 epochs với thời gian chạy ~13 phút/epoch. Tổng thời gian của cả quá trình là 22-23 giờ.

Kết quả thử nghiệm
Số liệu

Bộ dữ liệu	DUTS-TE			ECSSD		
Độ đo / **Mô hình**	**Avg F**	**S**	**MAE**	**Avg F**	**S**	**MAE**
PFSNet (số liệu theo paper)			.0360			.0310
PFSNet (chạy kiểm chứng)	.8533	**.8924**	.0359	.9260	**.9292**	.0314
PFSNet OCR 64	.8390	.8764	.0376	.9228	.9230	.0323
PFSNet OCR 128	**.8538**	.8864	**.0358**	**.9280**	.9264	**.0302**

Kết quả thực nghiệm

Kết luận

- **OCR** có thể điều chỉnh mức độ trọng yếu của từng điểm ảnh dựa trên độ mức độ trọng yếu của tất cả các điểm ảnh
- Ứng dụng mô-đun **OCR** từ bài toán phân đoạn ngữ nghĩa cho thấy sự cải thiện so với mô hình gốc.
- OCR hoạt động tốt trong các bài toán dự đoán chính xác(như trong phân đoạn ngữ nghĩa và phát hiện đối tượng trọng yếu) trên từng điểm ảnh khi các điểm ảnh đã được rút trích đặc trưng và có thể phân lớp độc lập.
- Các bài toán khác nhau yêu cầu việc rút trích véc-tơ đặc trưng cho điểm ảnh là khác nhau.

Hướng phát triển

- Tăng số kênh khóa của **OCR**, trước tiên là 256 (cũng là lựa chọn của bài toán phân đoạn ngữ nghĩa).
- Tăng số lượng kênh đầu vào của các lớp tích chập trong mô-đun **AFMs, SEM** nếu độ chính xác của mô hình bị gây ra bởi **bottleneck.**
- Sử dụng **backbone** hiệu quả hơn như EfficientNet, DenseNet, …
- Bất kể các kiến trúc nào có thể rút trích đặc trưng có thông tin toàn cục và tích hợp một cách có chọn lọc cho ra véc-tơ đặc trưng cho từng điểm ảnh sẽ cho kết quả tốt trong phát hiện đối tượng trọng yếu. Ví dụ như SegFormer dựa trên Transformer ở phân đoạn ngữ nghĩa.

Cảm ơn