1. **Introduction**

Con người có một hệ thống thị giác chuyên về các vật thể ở một khoảng cách và kích thước cố định. Ví dụ, khi chúng ta tập trung vào các vật thể gần đó, các vật thể ở xa sẽ bị mờ. Trong tự nhiên, nó rất khó để chúng ta tập trung vào các đối tượng có quy mô khác nhau cùng một lúc. Chúng tôi chỉ xử lý một trường nhìn nhỏ tại bất kỳ thời điểm nào và thích nghi bỏ qua nội dung hình ảnh còn lại trong ảnh. Tuy nhiên, thuật toán máy tính được thiết kế cho các tác vụ nhận dạng hình ảnh ở mức độ như phát hiện đối tượng xuất phát từ cách xử lý thông tin hình ảnh tự nhiên này . Để có được một đại diện mạnh mẽ theo tỷ lệ, các thuật toán phát hiện phổ biến như Faster-RCNN / Mask-RCNN được đào tạo trên một kim tự tháp hình ảnh đa tỷ lệ. Vì mỗi pixel được xử lý ở mỗi thang đo, cách tiếp cận này để xử lý thông tin hình ảnh làm tăng đáng kể thời gian đào tạo. Ví dụ: xây dựng một kim tự tháp hình ảnh 3 tỷ lệ (ví dụ: scale = 1x, 2x, 3x) yêu cầu xử lý gấp 14 lần số pixel có trong ảnh gốc. Vì lý do này, việc sử dụng đào tạo đa quy mô trong nhiều tình huống là không thực tế. Hiện tại, người ta thấy rằng bỏ qua độ dốc của các đối tượng có độ phân giải cực cao là có lợi trong khi sử dụng nhiều thang đo trong quá trình đào tạo. Ví dụ, khi xây dựng một kim tự tháp hình ảnh gồm 3 tỷ lệ, độ dốc của các đối tượng lớn và nhỏ nên được bỏ qua ở độ phân giải lớn và nhỏ tương ứng. Nếu đây là trường hợp, một câu hỏi trực quan phát sinh là, chúng ta có cần xử lý toàn bộ hình ảnh ở độ phân giải 3x không? Liệu nó có đủ để lấy mẫu RoI (chip) nhỏ hơn nhiều xung quanh các vật thể nhỏ ở độ phân giải này không? Mặt khác, nếu hình ảnh đã có độ phân giải cao và các đối tượng trong đó cũng có kích thước lớn, liệu có lợi ích gì trong việc lấy mẫu hình ảnh đó không?

Mặc dù bỏ qua các phần quan trọng của hình ảnh sẽ tiết kiệm tính toán, một con chip nhỏ hơn cũng sẽ thiếu bối cảnh cần thiết để nhận dạng. Một phần đáng kể của nền cũng sẽ bị bỏ qua ở độ phân giải cao hơn. Vì vậy, có một sự đánh đổi giữa tính toán, bối cảnh và khai thác tiêu cực trong khi đẩy nhanh đào tạo đa quy mô. Cuối cùng, chúng tôi trình bày một thuật toán đào tạo mới có tên là Scale Normalization cho Kim tự tháp hình ảnh với khả năng lấy mẫu hiệu quả (SNIPER), lấy mẫu một cách thích nghi các chip từ nhiều tỷ lệ của một kim tự tháp hình ảnh, dựa trên nội dung hình ảnh. Chúng tôi lấy mẫu chip tích cực dựa trên các trường hợp thực tế và chip âm tính dựa trên các đề xuất được tạo bởi mạng đề xuất khu vực. Trong cùng điều kiện (chuẩn hóa lô cố định),chúng tôi cho thấy SNIPER thực hiện cũng như chiến lược đa quy mô được đề xuất trong SNIP trong khi giảm số lượng pixel được xử lý theo hệ số 3 trong khi đào tạo trên bộ dữ liệu COCO. Vì SNIPER được đào tạo trên các chip kích thước 512x512, nó có thể gặt hái những lợi ích của kích thước lô lớn và đào tạo với chuẩn hóa hàng loạt trên một nút GPU. Đặc biệt, chúng tôi có thể sử dụng kích thước lô 20 trên mỗi GPU (dẫn đến tổng kích thước lô là 160), ngay cả với trình phát hiện Faster-RCNN dựa trên ResNet-101. Mặc dù hiệu quả, SNIPER có được hiệu suất cạnh tranh trên bộ dữ liệu phát hiện COCO ngay cả với các kiến ​​trúc phát hiện đơn giản như Faster-RCNN.

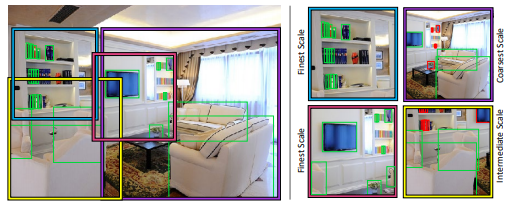
1. **Background**

Các thuật toán phát hiện đối tượng dựa trên học sâu đã phát triển chủ yếu từ máy dò R-CNN, bắt đầu với các đề xuất đối tượng được tạo bằng thuật toán không giám sát, đã thay đổi kích thước các đề xuất này thành hình ảnh kích thước chuẩn 224x224 và phân loại chúng bằng mạng nơ ron tích chập. Mô hình này là bất biến tỷ lệ, nhưng chi phí tính toán cho đào tạo và suy luận cho R-CNN quy mô tuyến tính với số lượng đề xuất. Để giảm bớt nút thắt tính toán này, Fast-RCNN đã đề xuất dự án đề xuất khu vực lên bản đồ tính năng tích chập mức cao và sử dụng các tính năng được gộp chung làm đại diện ngữ nghĩa cho các đề xuất khu vực. Trong quá trình này, tính toán được chia sẻ cho các lớp chập và chỉ các lớp kết nối hoàn toàn nhẹ mới được áp dụng cho mỗi đề xuất. Tuy nhiên, tích chập cho các đối tượng có kích thước khác nhau được thực hiện ở một tỷ lệ duy nhất, phá hủy các thuộc tính bất biến tỷ lệ của R-CNN. Do đó, suy luận ở nhiều thang đo được thực hiện và phát hiện từ nhiều thang đo được kết hợp bằng cách chọn các tính năng từ một cặp thang đo liền kề gần với độ phân giải của mạng được đào tạo trước. Mô hình Fast-RCNN đã trở thành phương pháp tiếp cận thực tế để phân loại các đề xuất khu vực vì nó nhanh và cũng nắm bắt được nhiều hơn trong các đặc tính của nó, Đó là cái thiếu của R-CNN.

Điều đáng chú ý là trong quá trình traning đa quy mô, các mẫu trên và mẫu dưới của Fast- RCNN từng đề xuất (dù nhỏ hay lớn) trong ảnh. Điều này không giống với R-CNN, trong đó mỗi đề xuất được thay đổi kích thước thành kích thước chuẩn là 224x224 pixel. Các đối tượng lớn không được ghép lại và các đối tượng nhỏ không được ghép xuống trong R-CNN. Về vấn đề này, R-CNN một cách thích hợp hơn không tăng / giảm mẫu mỗi pixel trong ảnh mà chỉ ở những vùng có khả năng chứa các đối tượng có độ phân giải phù hợp. Tuy nhiên, R-CNN không chia sẻ các tính năng tích chập cho các đề xuất gần đó như Fast-RCNN, khiến cho nó chậm. Để kết thúc này, chúng tôi đề xuất SNIPER, duy trì lợi ích của cả hai phương pháp này bằng cách tạo các vùng ngữ cảnh cụ thể (chip) bao gồm các đề xuất tối đa ở một quy mô cụ thể. SNIPER phân loại tất cả các đề xuất bên trong một con chip như Fast-RCNN cho phép chúng tôi thực hiện phân loại hiệu quả nhiều đề xuất trong một con chip. Vì SNIPER không đảo lộn hình ảnh khi có các vật thể lớn và cũng không xử lý các vùng nền dễ dàng, nên nó nhanh hơn đáng kể so với máy dò Fast-RCNN được đào tạo trên một kim tự tháp hình ảnh.

SNIP cũng được đào tạo về hầu hết tất cả các pixel của kim tự tháp hình ảnh (như Fast-RCNN), mặc dù độ dốc phát sinh từ các đối tượng có độ phân giải cực cao bị bỏ qua. Đặc biệt, 2 độ phân giải của kim tự tháp hình ảnh (480 và 800 pixel) luôn tham gia vào đào tạo và nhiều loại 1000 pixel được lấy mẫu từ độ phân giải 1400 pixel của hình ảnh ở tỷ lệ tốt nhất. SNIPER thực hiện quy trình cắt xén này đến một mức cực đoan bằng cách lấy mẫu các cây trồng 512 pixel từ 3 tỷ lệ của một kim tự tháp hình ảnh. Ở quy mô cực lớn (như 3x), SNIPER quan sát ít hơn một phần mười nội dung ban đầu có trong hình ảnh! Thật không may, vì các chip SNIPER chỉ được tạo bằng các trường hợp thực tế là rất nhỏ so với độ phân giải của ảnh gốc, một phần đáng kể của nền không tham gia vào đào tạo. Điều này làm cho tỷ lệ dương tính giả tăng lên. Do đó, điều quan trọng là

Hình 1: Lựa chọn chip tích cực SNIPER. SNIPER thích nghi lấy mẫu các vùng ngữ cảnh (còn gọi là chip) dựa trên sự hiện diện của các đối tượng bên trong hình ảnh. Mặt trái: Hình ảnh, ground-truth boxes (được miêu tả bởi các đường màu xanh lá cây) và các chip trong tỷ lệ hình ảnh gốc (được biểu thị bằng các hình chữ nhật màu xanh, vàng, hồng và tím). Phía bên phải: Lấy mẫu xuống / lên được thực hiện xem xét kích thước của các đối tượng. Các đối tượng được che phủ được hiển thị bằng màu xanh lá cây và các đối tượng không hợp lệ theo tỷ lệ tương ứng được hiển thị dưới dạng hình chữ nhật màu đỏ.



tạo chip cho các vùng nền là tốt. Trong SNIPER, điều này đạt được bằng cách lấy mẫu ngẫu nhiên một số chip cố định (tối đa là 2 trong bài viết này) từ các khu vực có khả năng bao gồm các kết quả chưa tích cực. Để tìm các khu vực như vậy, chúng tôi đào tạo một mạng RPN nhẹ với một lịch trình ngắn. Các đề xuất của mạng này được sử dụng để tạo chip cho các khu vực có khả năng chứa dương tính giả (điều này có khả năng có thể được thay thế bằng các đề xuất không được giám sát như EdgeBoxes). Sau khi thêm lấy mẫu chip âm, hiệu suất của SNIPER phù hợp với SNIP, nhưng nó nhanh hơn 3 lần! Vì chúng tôi có thể đạt được hiệu suất tương tự bằng cách quan sát ít hơn một phần mười hình ảnh, điều đó ngụ ý rằng bối cảnh rất lớn trong quá trình đào tạo không quan trọng đối với việc đào tạo các máy dò hiệu suất cao nhưng các khu vực lấy mẫu có chứa nhiều tiêu cực.

1. **SNIPER**

Chúng tôi mô tả các thành phần chính của SNIPER trong phần này. Một là khai thác chip dương / âm và thứ hai là gán nhãn sau khi chip được tạo. Cuối cùng, chúng tôi sẽ thảo luận về lợi ích của việc đào tạo với SNIPER.

**3.1 Thế hệ chip**

SNIPER tạo ra chip Ci ở nhiều tỷ lệ {s1, s2, .., si ..sn} trong hình ảnh. Đối với mỗi tỷ lệ, hình ảnh đầu tiên được định cỡ lại theo chiều rộng (Wi) và chiều cao (Hi). Trên khung vẽ này, chip pixel K × K được đặt ở các khoảng d pixel bằng nhau (chúng tôi đặt d thành 32 trong bài viết này). Điều này dẫn đến một mảng chip hai chiều ở mỗi thang đo.

**3.2 Lựa chọn chip tích cực**

Đối với mỗi tỷ lệ, có một phạm vi khu vực mong muốn Ri = [rimin; rimax], i [1; n] xác định hộp / đề xuất thực tế nào tham gia đào tạo cho từng thang đo i. Danh sách hợp lệ các hộp giới hạn sự thật nằm trong Ri được gọi là Gi. Sau đó, các chip được chọn một cách tham lam sao cho số hộp tối đa hợp lệ (G i) được bảo hiểm tối đa. Một hộp sự thật được cho là được bảo hiểm nếu nó được bao bọc hoàn toàn bên trong một con chip. Tất cả các chip tích cực từ một tỷ lệ được kết hợp trên mỗi hình ảnh và được gọi là Cipos. Đối với mỗi hộp giới hạn mặt đất, luôn tồn tại một con chip bao phủ nó. Do Ri liên tiếp chứa các khoảng chồng lấp, hộp giới hạn mặt đất có thể được gán cho nhiều chip ở các tỷ lệ khác nhau. Cũng có thể là hộp giới hạn mặt đất giống nhau có thể ở nhiều chip từ cùng một tỷ lệ. Các trường hợp thực tế có sự chồng chéo một phần (IoU > 0) với một con chip bị cắt. Tất cả các hộp sự thật được cắt (hợp lệ hoặc không hợp lệ) được giữ lại trong chip và được sử dụng trong gán nhãn.

Hình 2: Lựa chọn chip âm SNIPER. Hàng đầu tiên: hình ảnh và các hộp sự thật. Hàng dưới cùng: các đề xuất tiêu cực không được đề cập trong các chip dương (được biểu thị bằng các vòng tròn màu đỏ nằm ở trung tâm của mỗi đề xuất cho rõ ràng) và các chip âm được tạo dựa trên các đề xuất (được biểu thị bằng hình chữ nhật màu cam).

Theo cách này, mọi hộp sự thật được bao phủ ở quy mô thích hợp. Vì kích thước crop nhỏ hơn nhiều so với độ phân giải của hình ảnh (tức là nhỏ hơn 10 lần đối với hình ảnh có độ phân giải cao), SNIPER không xử lý hầu hết nền ở độ phân giải cao. Điều này dẫn đến tiết kiệm đáng kể trong tính toán và yêu cầu bộ nhớ trong khi xử lý hình ảnh độ phân giải cao. Chúng tôi minh họa điều này bằng một ví dụ được hiển thị trong Hình 1. Phía bên trái của hình hiển thị hình ảnh với các hộp sự thật được thể hiện bằng các hộp giới hạn màu xanh lá cây. Các hình chữ nhật màu khác ở phía bên trái của hình cho thấy các chip được tạo bởi SNIPER ở độ phân giải hình ảnh gốc bao gồm tất cả các đối tượng. Các chip này được minh họa ở phía bên phải của hình với cùng màu viền. Các hộp giới hạn màu xanh lá cây và màu đỏ đại diện cho các đối tượng sự thật mặt đất hợp lệ và không hợp lệ tương ứng với quy mô của chip. Có thể thấy, trong ví dụ này, SNIPER xử lý hiệu quả tất cả các đối tượng chân thực ở quy mô phù hợp bằng cách hình thành 4 chip có độ phân giải thấp.

**3.3 Lựa chọn chip âm**

Mặc dù các chip tích cực bao gồm tất cả các trường hợp tích cực, một phần đáng kể của nền không được bao phủ bởi chúng. Nền phân loại không chính xác làm tăng tỷ lệ dương tính giả. Trong các thuật toán phát hiện đối tượng hiện tại, khi thực hiện đào tạo đa tỷ lệ, mọi pixel trong ảnh được xử lý ở tất cả các tỷ lệ. Mặc dù đào tạo trên tất cả các thang đo làm giảm tỷ lệ dương tính giả, nhưng nó cũng làm tăng tính toán. Chúng tôi khẳng định rằng một lượng đáng kể của nền dễ dàng phân loại và do đó, chúng tôi có thể tránh thực hiện bất kỳ tính toán nào trong các khu vực đó. Vậy, làm thế nào để chúng ta loại bỏ các khu vực dễ phân loại? Một cách tiếp cận đơn giản là sử dụng các đề xuất đối tượng để xác định các khu vực nơi các đối tượng có khả năng có mặt. Xét cho cùng, trình phân loại của chúng tôi hoạt động dựa trên các đề xuất khu vực và nếu không có đề xuất khu vực nào trong một phần của hình ảnh, điều đó ngụ ý rằng rất dễ dàng để phân loại làm nền. Do đó, chúng ta có thể bỏ qua những phần của hình ảnh trong quá trình đào tạo.

Để kết thúc này, để khai thác chip âm, trước tiên chúng tôi đào tạo RPN cho một vài kỷ nguyên. Không có chip tiêu cực được sử dụng để đào tạo mạng này. Nhiệm vụ của mạng này là hướng dẫn đại khái cho chúng tôi trong việc lựa chọn các khu vực có khả năng chứa thông tin sai, do đó không cần thiết phải chính xác. RPN này được sử dụng để tạo các đề xuất trên toàn bộ tập huấn luyện. Chúng tôi giả định rằng nếu không có đề xuất nào được tạo ra trong một phần chính của hình ảnh bằng RPN, thì không có khả năng chứa một thể hiện đối tượng. Đối với lựa chọn chip âm, đối với mỗi thang đo i, trước tiên, chúng tôi sẽ xóa tất cả các đề xuất đã được đề cập trong Cipos.

Sau đó, với mỗi thang đo i, chúng tôi tham lam chọn tất cả các chip bao gồm ít nhất M đề xuất trong Ri. Điều này tạo ra một tập hợp các chip âm cho mỗi tỷ lệ trên mỗi hình ảnh, Cineg. Trong quá trình đào tạo, chúng tôi lấy mẫu ngẫu nhiên một số lượng chip âm cố định trên mỗi epoch (trên mỗi hình ảnh) từ nhóm chip âm tính được tạo ra từ tất cả các tỷ lệ, tức là Uin = 1 Cineg. Hình 2 cho thấy các ví dụ về các chip âm được tạo bởi SNIPER. Hàng đầu tiên hiển thị hình ảnh và các hộp sự thật. Ở hàng dưới cùng, chúng tôi hiển thị các đề xuất không được đề cập bởi Cipos và các chip âm tương ứng được tạo ra (các hộp màu cam). Tuy nhiên, để rõ ràng, chúng tôi đại diện cho mỗi đề xuất bằng một vòng tròn màu đỏ ở trung tâm của nó. Như được minh họa, SNIPER chỉ xử lý các khu vực có khả năng chứa dương tính giả, dẫn đến thời gian xử lý nhanh hơn.

**3.4 Phân công nhãn**

Mạng của chúng tôi được đào tạo từ đầu đến cuối trên các chip này như Faster-RCNN, tức là nó học cách tạo ra các đề xuất cũng như phân loại chúng với một mạng duy nhất. Trong khi đào tạo, các đề xuất được tạo bởi RPN được gán nhãn và các mục tiêu hộp giới hạn (cho hồi quy) dựa trên tất cả các hộp thực tế có trong chip. Chúng tôi không lọc các hộp sự thật dựa trên Ri. Thay vào đó, các đề xuất không thuộc Ri được bỏ qua trong quá trình đào tạo. Vì vậy, một hộp sự thật lớn được cắt, có thể tạo ra một đề xuất hợp lệ nhỏ. Giống như Fast-RCNN, chúng tôi đánh dấu bất kỳ đề xuất nào có tỷ lệ trùng lặp lớn hơn 0,5 với hộp thực tế là tích cực và chỉ định các mục tiêu hộp giới hạn cho đề xuất. Mạng của chúng tôi được đào tạo từ đầu đến cuối và chúng tôi tạo ra 300 đề xuất cho mỗi chip. Chúng tôi không áp dụng bất kỳ ràng buộc nào rằng một phần của các đề xuất này nên được lấy mẫu lại dưới dạng tích cực [29], như trong Fast-RCNN. Chúng tôi đã không sử dụng OHEM [32] để phân loại và sử dụng một mất mát entropy chéo đơn giản để phân loại. Để gán nhãn RPN, chúng tôi sử dụng các hộp sự thật mặt đất hợp lệ để gán nhãn và hộp sự thật mặt đất không hợp lệ để vô hiệu hóa các neo, như được thực hiện trong SNIP [33].

**3.5 Lợi ích**

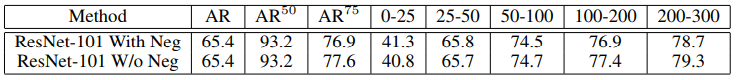
Để đào tạo, chúng tôi lấy mẫu ngẫu nhiên các chip từ toàn bộ tập dữ liệu để tạo một lô. Trung bình, chúng tôi tạo ra 5 chip có kích thước 512x512 mỗi hình ảnh trên bộ dữ liệu COCO (bao gồm cả chip âm) khi đào tạo theo ba tỷ lệ (512 / ms 2, 1: 667, 3). Con số này chỉ nhiều hơn 30% so với số pixel được xử lý trên mỗi hình ảnh khi thực hiện đào tạo tỷ lệ đơn lẻ với độ phân giải hình ảnh là 800x1333. Vì tất cả các hình ảnh của chúng tôi có cùng kích thước, dữ liệu được đóng gói tốt hơn nhiều dẫn đến việc sử dụng GPU tốt hơn, dễ dàng vượt qua 30% chi phí. Nhưng quan trọng hơn, chúng tôi gặt hái những lợi ích của đào tạo đa quy mô trên 3 quy mô, quy mô lô lớn và đào tạo với chuẩn hóa hàng loạt mà không làm chậm hiệu suất trên một nút GPU 8!.

Người ta thường tin rằng hình ảnh có độ phân giải cao (ví dụ: 800x1333) là cần thiết cho các tác vụ nhận dạng mức cá thể. Do đó, đối với các tác vụ nhận dạng mức, chẳng hạn, không thể đào tạo với số liệu thống kê chuẩn hóa hàng loạt được tính trên một GPU. Các phương pháp như chuẩn hóa hàng loạt đồng bộ [22, 41] hoặc đào tạo trên 128 GPU [28] đã được đề xuất để giảm bớt vấn đề này. Chuẩn hóa hàng loạt được đồng bộ hóa làm chậm đáng kể việc đào tạo và đào tạo trên 128 GPU cũng không thực tế đối với hầu hết mọi người. Do đó, chuẩn hóa nhóm [35] gần đây đã được đề xuất để các tác vụ nhận dạng mức cá thể có thể được hưởng lợi từ một hình thức chuẩn hóa khác trong một cài đặt lô thấp trong quá trình đào tạo. Với SNIPER, chúng tôi cho thấy nút thắt độ phân giải hình ảnh có thể được giảm bớt cho các tác vụ nhận dạng mức cá thể. Miễn là chúng tôi có thể bao phủ các tiêu cực và sử dụng các phương pháp bình thường hóa quy mô phù hợp, chúng tôi có thể đào tạo với một lô lớn các chip có độ phân giải thấp được ghép lại, ngay cả trên các bộ dữ liệu đầy thách thức như COCO. Kết quả của chúng tôi cho thấy bối cảnh vượt ra ngoài một lĩnh vực nhất định có thể không có lợi trong quá trình đào tạo. Cũng có thể trường tiếp nhận hiệu quả của các mạng nơ ron sâu không đủ lớn để tận dụng các pixel ở xa trong ảnh, như được đề xuất trong [24].

Trong các bộ dữ liệu rất lớn như OpenImagesV4 [18] chứa 1: 7 triệu hình ảnh, hầu hết các đối tượng đều lớn và hình ảnh được cung cấp có độ phân giải cao (1024x768), do đó, việc lấy mẫu hình ảnh lên 3 sẽ ít quan trọng hơn. Trong trường hợp này, với SNIPER, chúng tôi tạo ra 3: 5 triệu chip có kích thước 512x512 bằng tỷ lệ (512 / ms, 1). Lưu ý rằng SNIPER cũng thực hiện lấy mẫu xuống thích ứng. Vì các thang đo nhỏ hơn, các chip sẽ bao phủ nhiều nền hơn, do đó tác động của việc lấy mẫu âm tính bị giảm đi. Trong trường hợp này (lựa chọn chip dương), SNIPER chỉ xử lý một nửa số pixel so với đào tạo đa quy mô ngây thơ trên các thang đo đã đề cập ở trên trong OpenImagesV4. Do đó, chúng tôi đã có thể đào tạo Faster-RCNN với xương sống ResNet-101 trên 1: 7 triệu hình ảnh chỉ trong 3 ngày trên một nút 8 GPU!

1. **Chi tiết thí nghiệm**

Chúng tôi đánh giá SNIPER trên bộ dữ liệu COCO để phát hiện đối tượng. COCO chứa 123.000 hình ảnh trong tập huấn luyện và xác nhận và 20.288 hình ảnh trong bộ thử nghiệm. Chúng tôi đào tạo về tập huấn luyện và xác nhận kết hợp và báo cáo kết quả trên tập kiểm thử. Do việc thu hồi các đề xuất không được cung cấp bởi máy chủ đánh giá, chúng tôi đào tạo trên 118.000 hình ảnh và báo cáo thu hồi trên 5.000 hình ảnh còn lại (thường được gọi là bộ minival).



Bảng 1: Chúng tôi vẽ sơ đồ thu hồi cho SNIPER có và không có phủ định. Đáng ngạc nhiên, việc thu hồi không bị ảnh hưởng bởi việc lấy mẫu chip âm.

Trên COCO, chúng tôi đào tạo SNIPER với quy mô lô là 128 và với tỷ lệ học tập là 0,015. Chúng tôi sử dụng kích thước chip 512×512 pixel. Thang đo đào tạo được đặt thành (512/ms, 1.667, 3) trong đó ms là chiều rộng và chiều cao giá trị tối đa của hình ảnh. Phạm vi khu vực mong muốn (ví dụ Ri) được đặt thành (0,802), (322, 1502) và (1202, inf) cho mỗi thang đo tương ứng. Đào tạo được thực hiện trong tổng số 6 epoch với bước xuống ở cuối epoch 5. Lật hình ảnh được sử dụng như một kỹ thuật tăng cường dữ liệu. Mỗi kỷ nguyên cần 11.000 lần lặp. Để đào tạo RPN mà không có tiêu cực, mỗi kỷ nguyên cần 7000 lần lặp. Chúng tôi sử dụng RPN để tạo chip âm và huấn luyện nó trong 2 epoch với tỷ lệ học tập cố định là 0,015 mà không có bất kỳ bước xuống nào. Do đó, đào tạo RPN trong 2 epoch đòi hỏi ít hơn 20% tổng thời gian đào tạo. Đề xuất RPN được trích xuất từ ​​tất cả các quy mô. Lưu ý rằng suy luận mất 1/3 thời gian cho một lần vượt qua hoàn toàn về phía trước và chúng tôi không thực hiện bất kỳ lật để trích xuất các đề xuất. Do đó, quá trình này cũng hiệu quả. Chúng tôi sử dụng đào tạo chính xác hỗn hợp như được mô tả trong **[27]**. Để kết thúc này, chúng tôi tái phân chia tỷ lệ phân rã theo trọng số 100, giảm tỷ lệ học tập xuống 100 và giảm độ dốc xuống 100. Điều này đảm bảo rằng chúng tôi có thể đào tạo với kích hoạt độ chính xác một nửa (và do đó kích thước lô lớn hơn gấp 2 lần) mà không bị mất độ chính xác. Chúng tôi sử dụng trọng số fp32 cho lớp chập đầu tiên, lớp chập cuối cùng trong RPN (phân loại và hồi quy) và các lớp được kết nối đầy đủ trong Faster-RCNN.

Chúng tôi đánh giá SNIPER bằng trình phát hiện phổ biến, Faster-RCNN với ResNets [**14, 15**] và MobileNetV2 [**31**]. Các đề xuất được tạo bằng cách sử dụng RPN trên các tính năng của conv4 và việc phân loại được thực hiện sau khi kết hợp các tính năng của conv4 và conv5. Trong nhánh conv5, chúng tôi sử dụng các cấu trúc có thể biến dạng và stride là 1. Chúng tôi sử dụng bản đồ tính năng 512 chiều trong RPN. Đối với nhánh phân loại, trước tiên, chúng tôi chiếu sơ đồ tính năng được nối thành 256 kích thước và sau đó thêm 2 fully connected với 1024 hidden units. Đối với các mạng nhẹ như MobileNetv2 [**31**], để bảo toàn khả năng xử lý tính toán của mạng, chúng tôi không thực hiện bất kỳ thay đổi kiến ​​trúc nào đối với mạng như thay đổi bước tiến của mạng hoặc thêm các kết cấu có thể biến dạng. Chúng tôi đã giảm kích thước RPN xuống 256 và kích thước của các lớp fc xuống 512 từ 1024. RPN và nhánh phân loại đều được áp dụng trên lớp với sải chân 32 cho MobileNetv2.

SNIPER tạo ra 1,2 triệu chip cho bộ dữ liệu COCO sau khi hình ảnh được lật. Điều này dẫn đến khoảng 5 chip cho mỗi hình ảnh. Trong một số hình ảnh chứa nhiều phiên bản đối tượng, SNIPER có thể tạo ra tối đa 10 chip và các chip khác trong đó có một đối tượng nổi bật lớn, nó sẽ chỉ tạo ra một chip duy nhất. Theo một nghĩa nào đó, nó làm giảm sự mất cân bằng về độ dốc được truyền đến một thể hiện mức độ có mặt trong các máy dò được đào tạo về hình ảnh độ phân giải đầy đủ. Ít nhất là về lý thuyết, đào tạo về hình ảnh độ phân giải đầy đủ được thiên về các trường hợp đối tượng lớn.

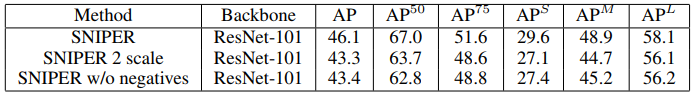
**4.1 Phân tích thu hồi**

Chúng tôi quan sát rằng việc thu hồi (trung bình trên nhiều ngưỡng chồng chéo 0,5: 0,05: 0,95) cho RPN không giảm nếu chúng tôi không thực hiện lấy mẫu âm tính. Điều này là do thu hồi không tính đến dương tính giả. Như được chỉ ra trong Mục **4.2**, điều này trái ngược với mAP để phát hiện trong đó lấy mẫu âm tính đóng vai trò quan trọng. Hơn nữa, trong lấy mẫu chip dương tính, chúng tôi bao gồm mọi mẫu sự thật mặt đất. Do đó, để tạo các đề xuất, chỉ cần đào tạo trên các mẫu dương tính là đủ. Kết quả này tiếp tục thúc đẩy chiến lược tìm kiếm tiêu cực dựa trên RPN trong đó việc đào tạo được thực hiện chỉ trên các mẫu tích cực.

**4.2 Khai thác và quy mô chip âm**

SNIPER sử dụng khai thác chip âm để giảm tỷ lệ dương tính giả trong khi tăng tốc độ đào tạo bằng cách bỏ qua các **vùng dễ dàng** bên trong hình ảnh. Như đã đề xuất trong mục **3.3**, chúng tôi sử dụng mạng đề xuất khu vực được đào tạo với lịch trình học tập ngắn để tìm các khu vực đó. Để đánh giá hiệu quả của phương pháp khai thác tiêu cực, chúng tôi so sánh độ chính xác trung bình trung bình của SNIPER với một biến thể nhỏ chỉ sử dụng chip dương trong quá trình đào tạo (ký hiệu là SNIPER w/o neg). Tất cả các thông số khác vẫn giữ nguyên. **Bảng 2** so sánh hiệu suất của các mô hình này. Phương pháp khai thác chip âm được đề xuất cải thiện đáng kể AP cho tất cả các ngưỡng nội địa hóa và kích thước đối tượng. Đáng chú ý, khai thác chip âm cải thiện độ chính xác trung bình từ **43,4** đến **46,1**. Điều này trái ngược với phần cuối cùng nơi chúng tôi đang đánh giá các đề xuất. Điều này là do mAP bị ảnh hưởng bởi dương tính giả. Nếu chúng tôi không bao gồm các vùng trong hình ảnh có chứa các âm bản tương tự như các trường hợp tích cực, nó sẽ làm tăng tỷ lệ dương tính giả của chúng tôi và ảnh hưởng xấu đến hiệu suất phát hiện. SNIPER là một thuật toán đào tạo đa quy mô hiệu quả. Trong tất cả các thí nghiệm trong bài báo này, chúng tôi sử dụng ba thang đo đã nói ở trên (**Xem Phần 4** để biết chi tiết). Để chỉ ra rằng SNIPER được hưởng lợi một cách hiệu quả từ đào tạo đa quy mô, chúng tôi giảm số lượng thang đo **từ 3 xuống 2** bằng cách giảm thang độ phân giải cao. **Bảng 2** cho thấy độ chính xác trung bình trung bình của SNIPER trong hai cài đặt này. Có thể thấy, bằng cách giảm số lượng thang đo, hiệu suất luôn giảm một mức lớn trên tất cả các số liệu đánh giá.

Đối với tỷ lệ đầu tiên, zero-padding được sử dụng nếu phần nhỏ hơn của hình ảnh trở nên nhỏ hơn 512 pixel.



**Bảng 2**: Đào tạo hiệu ứng trên 2 thang đo (1.667 và kích thước tối đa 512). Chúng tôi cũng cho thấy tác động trong hiệu suất khi không có khai thác tiêu cực được thực hiện.

**4.3 Thời gian**

Phải mất 14 giờ để đào tạo SNIPER từ đầu đến cuối trên nút 8 GPU V100 với bộ phát hiện Faster-RCNN có xương sống ResNet-101. Điều đáng chú ý là chúng tôi đào tạo trên 3 tỷ lệ của một kim tự tháp hình ảnh (kích thước tối đa 512, 1.667 và 3). Đào tạo RPN hiệu quả hơn nhiều và chỉ mất 2 giờ cho đào tạo trước. SNIPER không chỉ hiệu quả trong đào tạo, nó còn có thể xử lý khoảng 5 hình ảnh mỗi giây trên một GPU V100. Để sử dụng tốt hơn các tài nguyên, chúng tôi chạy song song nhiều quy trình trong quá trình suy luận và tính thời gian trung bình cần để xử lý một lô 100 hình ảnh.

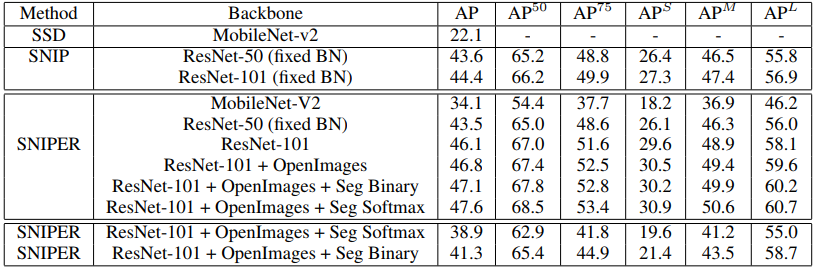
**4.4 Suy luận**

Chúng tôi thực hiện suy luận trên một kim tự tháp hình ảnh và chia tỷ lệ hình ảnh gốc theo các độ phân giải sau (480, 512), (800, 1280) và (1400, 2000). Phần tử đầu tiên là kích thước tối thiểu với điều kiện kích thước tối đa không vượt quá yếu tố thứ hai. Phạm vi hợp lệ cho đào tạo và suy luận tương tự như SNIP [**33**]. Để kết hợp các phát hiện, chúng tôi sử dụng Soft-NMS [**4**]. Chúng tôi không perform flipping [**39**], iterative bounding box regression [**9**] or mask tightening [**22**].

**4.5 So sánh với công nghệ tiên tiến**

Thật khó để so sánh một cách công bằng các máy dò khác nhau vì chúng khác nhau về kiến trúc xương sống (like ResNet [14], ResNext [36], Xception [6]), pre-training data (e.g. ImageNet-5k, JFT [16], OpenImages [18]), different structures in the underlying network (e.g multi-scale features [20, 26], deformable convolutions [8], heavier heads [28], anchor sizes, path aggregation [22]), test time augmentations like flipping, mask tightening, iterative bounding box regression etc.

Do đó, chúng tôi so sánh kết quả của chúng tôi với SNIP [**33**], đây là một phương pháp gần đây để đào tạo máy phát hiện đối tượng trên một kim tự tháp hình ảnh. Các kết quả được trình bày trong **Bảng 3**. Không sử dụng chuẩn hóa hàng loạt [**17**], SNIPER đạt được kết quả tương đương. Trong khi SNIP [**33**] xử lý gần như tất cả các kim tự tháp hình ảnh, mặt khác, SNIPER lại giảm chi phí tính toán bằng cách bỏ qua các vùng dễ dàng. Hơn nữa, vì SNIPER hoạt động với đầu vào có độ phân giải thấp hơn, nó làm giảm dung lượng bộ nhớ. Điều này cho phép chúng tôi tăng kích thước lô và không giống như SNIP [33], chúng tôi có thể hưởng lợi từ việc chuẩn hóa lô trong quá trình đào tạo. Với chuẩn hóa hàng loạt, SNIPER vượt trội đáng kể so với SNIP trong tất cả các số liệu. Cần lưu ý rằng không chỉ phương pháp được đề xuất là chính xác hơn, nó còn **nhanh hơn 3 lần** trong quá trình đào tạo. Theo hiểu biết tốt nhất của chúng tôi, đối với kiến ​​trúc Faster-RCNN với xương sống ResNet-101 (với các kết cấu có thể biến dạng), kết quả được báo cáo của chúng tôi là 46,1% là hiện đại. Kết quả này cải thiện đến 46,8% nếu chúng tôi đào tạo trước máy dò trên bộ dữ liệu OpenImagesV4. Thêm một đầu phân đoạn cá thể và đào tạo mạng phát hiện cùng với nó cải thiện hiệu suất lên 47,6%.



**Bảng 3**: Phân tích và so sánh với đào tạo độ phân giải đầy đủ. Hai hàng cuối hiển thị kết quả phân tích khi đầu mặt nạ được huấn luyện với mất softmax cách N + 1 và mất softmax nhị phân cho các lớp N.

Với đường dẫn suy luận hàng loạt hiệu quả của chúng tôi, chúng tôi có thể xử lý 5 hình ảnh mỗi giây trên một GPU V100 duy nhất và vẫn thu được mAP là 47,6%. Điều này ngụ ý rằng trên các GPU hiện đại, việc thực hiện suy luận trên một kim tự tháp hình ảnh bao gồm độ phân giải cao như 1400x2000 là thực tế. Chúng tôi cũng hiển thị kết quả cho Faster-RCNN được đào tạo với MobileNetV2. Nó đạt được mAP là 34,1% so với phiên bản SSDLite [**31**], thu được 22,1%. Điều này một lần nữa nhấn mạnh tầm quan trọng của kim tự tháp hình ảnh (và đào tạo SNIPER) khi chúng ta có thể cải thiện hiệu suất của máy dò 12%.

Chúng tôi cũng hiển thị kết quả cho phân khúc ví dụ. Kiến trúc mạng giống như Mask-RCNN [**12**], chỉ là chúng ta không sử dụng FPN [**20**] và sử dụng cùng kiến trúc phát hiện được mô tả để phát hiện đối tượng. Để đa tác vụ, chúng tôi đã thử hai biến thể của hàm mất để đào tạo nhánh mặt nạ. Một là chức năng softmax nền trước cho các lớp N và một là chức năng softmax N + 1 cách. Đối với phân đoạn ví dụ, mạng được đào tạo với mất Softmax 2 chiều cho mỗi lớp rõ ràng hoạt động tốt hơn. Nhưng, để phát hiện đối tượng, mất Softmax theo cách N + 1 dẫn đến kết quả tốt hơn một chút. Chúng tôi chỉ sử dụng 3 thang đo trong quá trình suy luận và không thực hiện lật, siết mặt nạ, hồi quy giới hạn hộp lặp, mặt nạ đệm trước khi thay đổi kích thước, vv Kết quả phân đoạn cá thể của chúng tôi là sơ bộ và cho đến nay chúng tôi mới chỉ đào tạo được 2 mô hình.

1. **Kết luận**

Chúng tôi đã trình bày một thuật toán để đào tạo đa quy mô hiệu quả, lấy mẫu các chip có độ phân giải thấp từ một kim tự tháp hình ảnh đa quy mô để tăng tốc đào tạo đa quy mô lên **gấp 3 lần**. Khi làm như vậy, SNIPER đã không thỏa hiệp về hiệu suất của bộ dò do các kỹ thuật lấy mẫu hiệu quả cho các chip dương và âm. Vì SNIPER hoạt động trên các chip có độ phân giải thấp được lấy mẫu lại, nên có thể được đào tạo với kích thước lô lớn trên một GPU duy nhất mang nó đến gần hơn với giao thức để phân loại hình ảnh. Điều này trái ngược với thực tiễn đào tạo phổ biến về hình ảnh độ phân giải cao cho các nhiệm vụ nhận dạng mức ví dụ. Trong tương lai, chúng tôi muốn tăng tốc suy luận đa quy mô, bởi vì một phần đáng kể của nền có thể được loại bỏ mà không cần thực hiện tốn kém tính toán. Sẽ rất thú vị khi đánh giá độ phân giải chip làm bối cảnh bắt đầu làm tổn hại đến hiệu suất của các máy dò đối tượng.